# PanelNet: Understanding 360 Indoor Environment via Panel Representation Supplementary Material

Haozheng Yu    Lu He    Bing Jian    Weiwei Feng    Shan Liu
Tencent America
{haozhengyu, lhluhe, bingjian, wfeng, shanl}@tencent.com

## A. Detail of panel geometry embedding

Given the stride $S$ and the interval $I$, we generate $N$ panels in resolution $H_e \times I$ from the input RGB ERP in resolution $H_e \times W_e$, where $N = \frac{W_e}{S}$. To make the network aware of ERP distortion, we generate the geometric features in the same size as the output of layer1 and add them together as the input of layer2. Note that the output feature map of layer1 is in shape$f_c \in \mathbb{R}^{C_c \times \frac{H_e}{4} \times \frac{I}{4} \times N}$. For a ResNet-34 encoder, $C_c = 64$. Thus, we generate the corresponding local and global 3D Cartesian coordinates of each point from an image in resolution $\frac{H_e}{4} \times \frac{W_e}{4}$, the size of each panel is $\frac{H_e}{4} \times \frac{I}{4}$. For a pixel $P_e(x_e, y_e)$ located on this low-resolution ERP, its corresponding azimuth angle $\varphi$ and the polar angle $\theta$ on a sphere are computed as:

$$\begin{cases} \varphi = \frac{8\pi x_e}{W_e} \\ \theta = \frac{4\pi y_e}{H_e} \end{cases} \tag{1}$$

Given the azimuth angle $\varphi$ and the polar angle $\theta$ of a point on a unit sphere, the absolute 3D Cartesian coordinates and relative 3D Cartesian coordinates are computed as described in Section 3.3. The input shape to the panel geometry embedding network is $f_g \in \mathbb{R}^{5 \times \frac{H_e}{4} \times \frac{I}{4} \times N}$ and the output is $f'_g \in \mathbb{R}^{C_c \times \frac{H_e}{4} \times \frac{I}{4} \times N}$.

## B. Ablation study of the backbone selection

| Method | MRE | MAE | RMSE | $\delta^1$ |
|---|---|---|---|---|
| ResNet-18 | 0.1206 | 0.2162 | 0.3537 | 0.8605 |
| ResNet-34 | **0.1033** | **0.1859** | **0.3212** | **0.8976** |

Table 1. Ablation study of the backbone used in PanelNet.

We test the impact of different backbones in our model. The experiment is performed on Stanford2D3D [1] for depth estimation. The models are constructed without the panel geometry embedding network and Local2Global Transformer. Since partitioning the entire panorama into panels with overlaps greatly increase the computational complexity, we test ResNet-18 and ResNet-34 rather than vision Transformers as backbones. As shown in Table 4, ResNet-34 surpasses its counterpart by a large margin with its deeper structure. We use ResNet-34 as our backbone for the experiments in our paper.

## C. Ablation study of the Local2Global Transformer structure

We further conduct an ablation study to show the effect of stacking the Window Blocks and Panel Blocks successively in a loop and using relative position embedding for Panel Blocks. All the networks are conducted using the best PanelNet structure described in Section 4.6 with different Transformer architectures. We train and evaluate these networks on Stanford2D3D [1] dataset for depth estimation. The stride is set to 32 and the interval is 128. As we observed in Table 2, changing the default Transformer blocks order and using relative position embedding for Panel Blocks reduces the power of our proposed Local2Global Transformer. For the best performance, we stack Window Blocks from low patch resolution to high and add Panel Blocks after all the Window Blocks. We use absolute position embedding for Panel Blocks.

| Method | MRE | MAE | RMSE | $\delta^1$ |
|---|---|---|---|---|
| Successively | 0.0903 | 0.1649 | 0.3016 | 0.9151 |
| Relative | 0.0856 | 0.1588 | 0.3016 | 0.9195 |
| Local2Global | **0.0829** | **0.1520** | **0.2933** | **0.9242** |

Table 2. Ablation study of different Transformer structures. ”Successively” stands for stacking Window Blocks and Panel Blocks successively in a loop. ”Relative” stands for using relative position embedding for Panel Blocks. ”Local2Global” stands for the Transformer structure used for our best result in the paper.

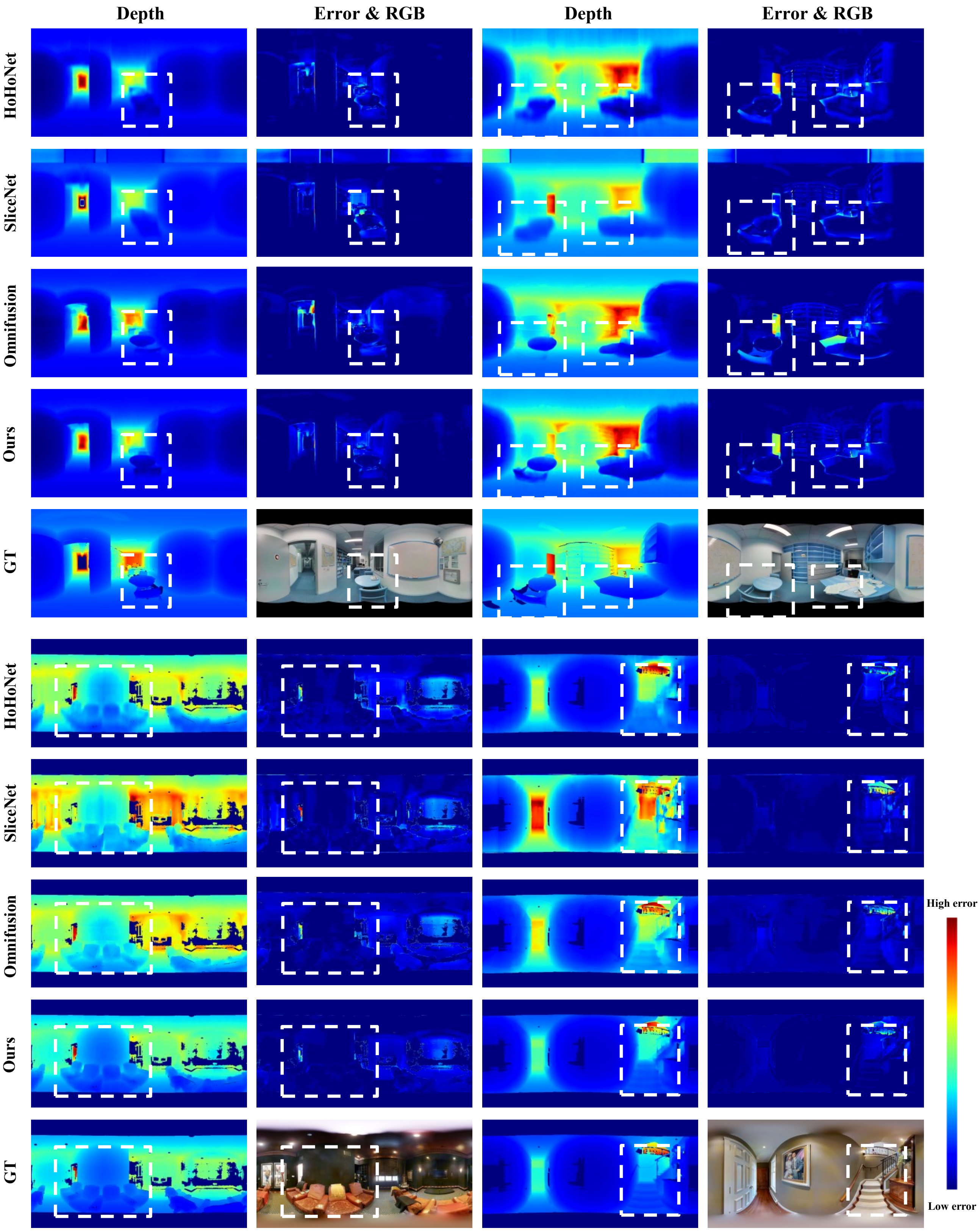


Figure 1. More qualitative results of depth estimation on Stanford2D3D [1] (top 5 rows) and Matterport3D [2] (down 5 rows). We also show the corresponding error maps. The darker in color the smaller in error.

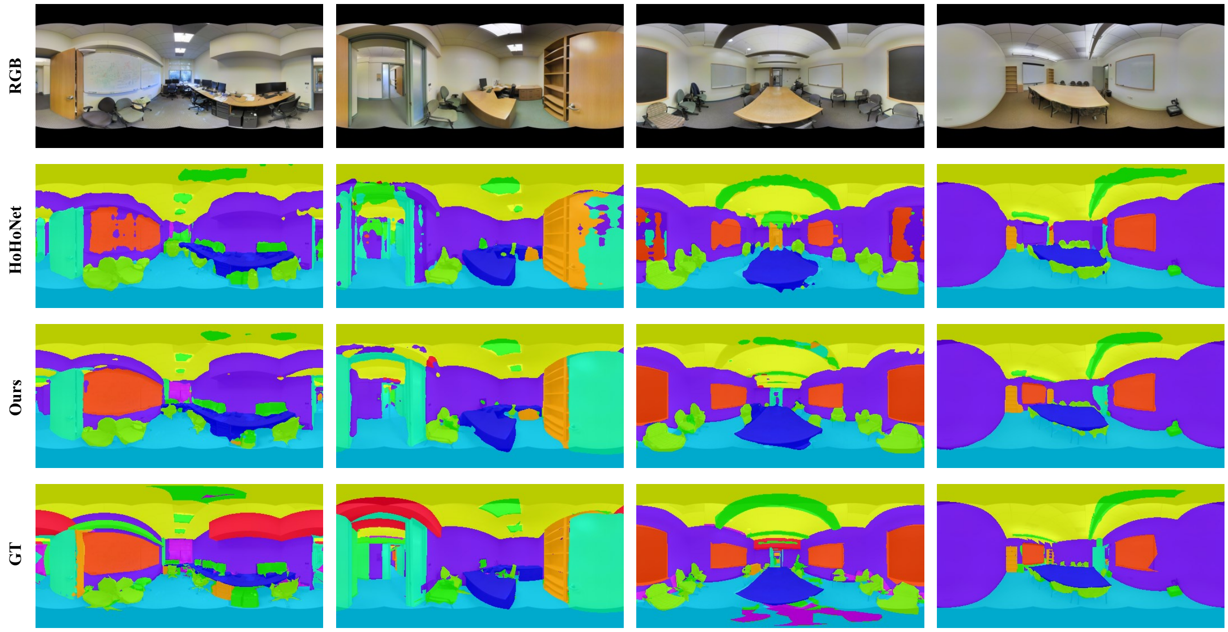
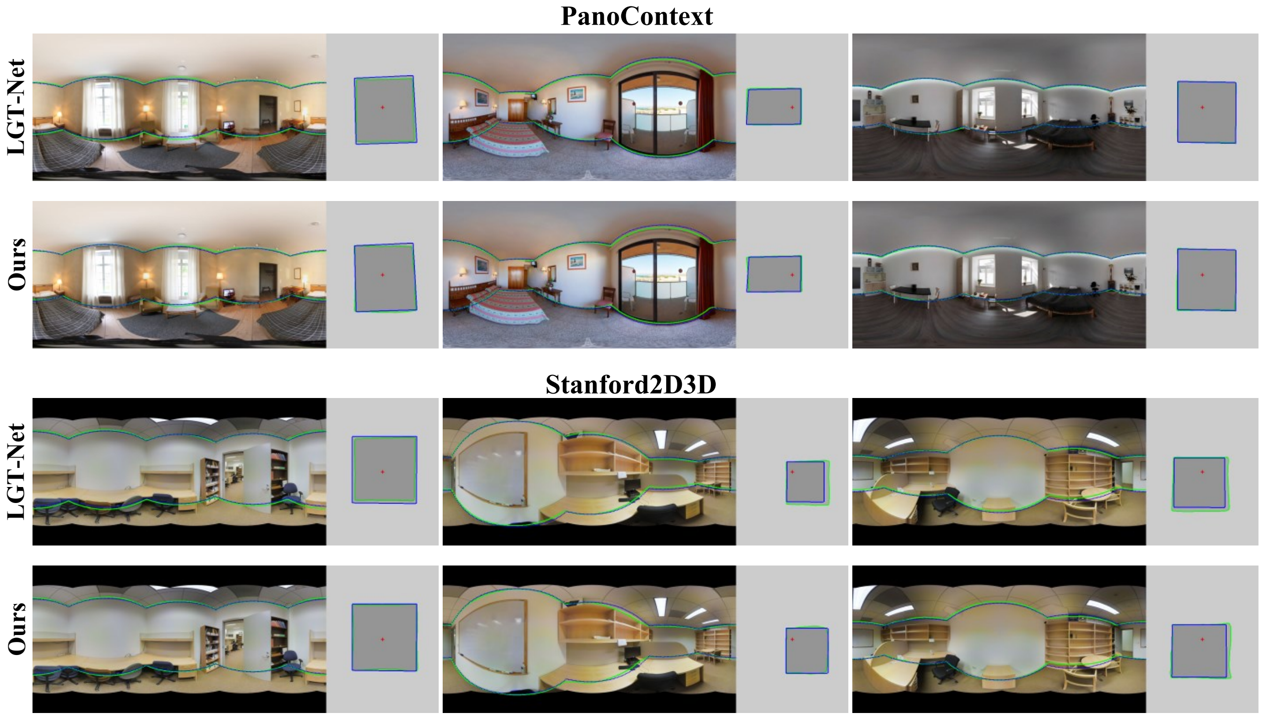


Figure 2. More qualitative results of indoor semantic segmentation on Stanford2D3D [1] dataset.


PanoContext

Stanford2D3D

Figure 3. Qualitative results of room layout estimation of our model against LGT-Net [3] on PanoContext [7] and Stanford2D3D [8]. We show the room layout predictions without pre-processing and post-processing. We show room layouts (left) together with the floor plan (right). The blue lines are the ground truth and the green lines are the predictions.

| Method | overall | beam | board | bookcase | ceiling | chair | clutter | column | door | floor | sofa | table | wall | window |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Per class mIoU% | | | | | | | | | | | | | | |
| HoHoNet [6] | 43.3 | 2.5 | 44.7 | 41.4 | **83.8** | **49.6** | **33.6** | 5.0 | **56.5** | **94.3** | 10.0 | 47.3 | 64.6 | 29.0 |
| Ours | **46.3** | **5.2** | **59.5** | **46.5** | 79.8 | 45.1 | 28.6 | **14.2** | 53.8 | 91.8 | **16.8** | **49.1** | **65.7** | **45.4** |
| Per class mAcc% | | | | | | | | | | | | | | |
| HoHoNet [6] | 53.9 | 10.0 | 54.0 | 56.4 | **95.4** | **66.7** | **47.5** | 7.4 | **68.6** | **97.8** | 11.2 | 70.2 | **82.5** | 33.0 |
| Ours | **58.7** | **20.9** | **71.8** | **64.5** | 93.2 | 58.7 | 42.4 | **22.5** | 63.6 | 95.8 | **22.3** | **74.6** | 81.9 | **51.5** |

Table 3. Perclass quantitative results of semantic segmentation on Stanford2D3D [1] dataset.

## D. Additional depth estimation results

We show more depth estimation results against HoHoNet [6], SliceNet [5] and Omnifusion [4], shown in Figure 1. We train a 2-iteration Omnifusion [4] model according to their official code on Matterport3D [2] dataset and show their results. We present the depth maps together with the error maps for evaluation. The results show that our model predicts smooth and continuous depth for indoor structures and captures more detail of the small objects such as chairs and tables. The overall depth error of our method is much lower than the previous methods.

## E. Additional semantic segmentation results

We show per-class IoU and per-class Acc for semantic segmentation on Stanford2D3D [1] in the resolution of 256 × 512. We compare our model against HoHoNet [6], results shown in Table 3. We achieve higher IoU in 8 classes among 13 classes and higher Acc in 7 classes among 13 classes using only RGB image as input. We show more qualitative results of indoor 360 semantic segmentation in Figure 2. Our model predicts more accurate and continuous object edges (e.g. whiteboards, doors) compared to HoHoNet [6].

## F. Qualitative results of layout estimation

We present the qualitative results of our model on two panorama layout estimation datasets, PanoContext [7] and the extended Stanford2D3D [8] dataset. Following the previous methods, we train our network using the mixture of PanoContext [7] and Stanford2D3D [8] dataset. Since we use the same indoor layout representation and similar training setup of LGT-Net [3], we compare our results against theirs, shown in Figure 3. The results of LGT-Net [3] are obtained from their official code and pre-trained weights. Our model shows comparable performance against LGT-Net [3] on both datasets.

## References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv*, 2017. 1, 2, 3, 4

[2] 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. In *3DV*, 2017. 2, 4

[3] Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *CVPR*, 2022. 3, 4

[4] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *CVPR*, 2022. 4

[5] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *CVPR*, 2021. 4

[6] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *CVPR*, 2021. 4

[7] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *ECCV*, 2014. 3, 4

[8] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. Layoutnet: Reconstructing the 3d room layout from a single rgb image. In *CVPR*, 2018. 3, 4