Supplementary Material
HoHoNet: 360 Indoor Holistic Understanding with Latent Horizontal Features

A. Network architecture diagram

We show the detailed architecture diagram in Fig. A. The shape of each feature tensor is denoted as "*# of channels, height, width*" within the box. The height and width of the input panorama are assumed to be 512 and 1024 respectively. D and E are hyperparameters. The $\mathrm{ConvSqueezeH}$ layer is a depthwise convolution layer with kernel size set to the prior known input feature height without padding, which produces output feature height 1.

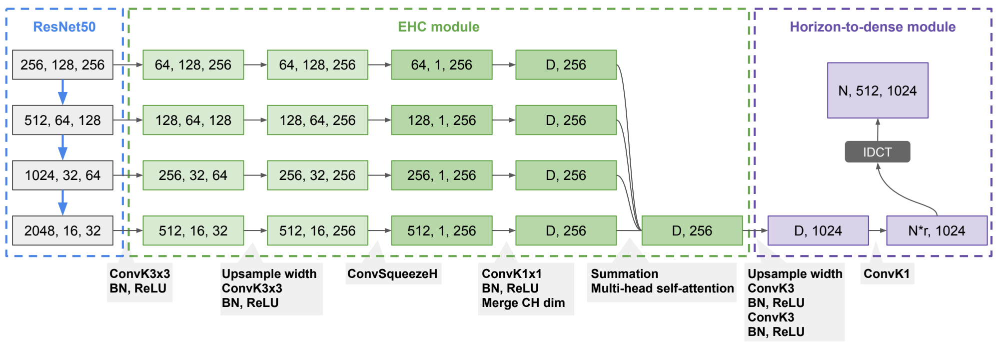
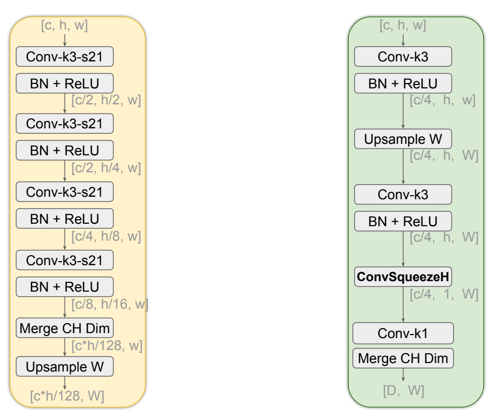


Figure A: The detailed network architecture with ResNet50 [4] backbone.

B. Comparing EHC block and HC block [8]

The height compression block aims to squeeze a 2D feature from the backbone to produce a 1D horizontal feature. Fig. B shows the architecture of our *Efficient Height Compression block* (EHC block) and the one of HC block [8] for comparison. The HC block [8] employs a sequence of convolution layers to gradually reduce the number of channels and heights, while we first use a convolution layer for channel reduction and then use bilinear upsampling and $\mathrm{ConvSqueezeH}$ layer to produce the features in horizontal shape. We show in our ablation experiments that replacing the HC block [8] with the proposed ECH block leads to better speed and accuracy.



(a) The HC block in [8]. (b) The proposed EHC block.

Figure B: Comparison of the proposed EHC block and the HC block in [8].

C. Detailed layout estimation results

We show detailed quantitative results for room layout under different numbers of ground truth 2D corners in Table A. Our training protocol and layout formalization are identical to HorizonNet [8], while we observe improvements (except rooms with six corners) by using our network architecture. In comparison with the most recent state-of-the-art—AtlantaNet [7], we show better results on scenes with fewer corners and similar accuracy on overall scenes; meanwhile, our model is 22× faster than AtlantaNet [7]. For depth-based evaluation proposed by LayoutNet v2 [13], we use an in-house implementation to synthesize layout depth as the ground truth layout depth is not available from the MatterportLayout dataset and the provided implementation for synthesizing layout depth produces invalid values (i.e., zero) for some cases. Note that our implementation is very different the [13]'s implementation so the results are not directly comparable. LayoutNet v2 [13], AtlantaNet [7] and our method achieve 0.28, 0.20 and 0.22 RMSE respectively; 0.90, 0.94 and 0.95 δ^1 respectively.

Method	# of corners				
	overall	4	6	8	10+
3D IoU (%)					
LayoutNet v2 [13]	75.82	81.35	72.33	67.45	63.00
DuLa-Net v2 [10]	75.07	77.02	78.79	71.03	63.27
HorizonNet [8]	79.11	81.88	**82.26**	71.78	68.32
AtlantaNet [7]	**80.02**	82.09	82.08	**75.19**	**71.61**
Ours	79.88	**82.64**	82.16	73.65	69.26

Method	# of corners				
	overall	4	6	8	10+
2D IoU (%)					
LayoutNet v2 [13]	78.73	84.61	75.02	69.79	65.14
DuLa-Net v2 [10]	78.82	81.12	82.69	74.00	66.12
HorizonNet [8]	81.71	84.67	**84.82**	73.91	70.58
AtlantaNet [7]	82.09	84.42	83.85	**76.97**	**73.18**
Ours	**82.32**	**85.26**	84.81	75.59	70.98

Table A: Detailed quantitative comparison for room layout estimation on MatterportLayout [13] under different numbers of ground-truth corners.

D. Detailed semantic segmentation results

We show detailed per-class IoU and per-class Acc for semantic segmentation in Table B. We achieve the best IoU on 10 out of 13 classes and superior overall mIoU; we achieve best Acc on 7 out of 13 classes and comparable overall mAcc.

Method	overall	beam	board	bookcase	ceiling	chair	clutter	column	door	floor	sofa	table	wall	window
Low-resolution RGB-D														
UGSCNN [5]	38.3	8.7	32.7	33.4	82.2	42.0	25.6	10.1	41.6	87.0	7.6	41.7	61.7	23.5
HexRUNet [12]	43.3	**10.9**	39.7	37.2	84.8	50.5	29.2	11.5	45.3	92.9	19.1	49.1	63.8	29.4
TangentImg [3]	37.5	**10.9**	26.6	31.9	82.0	38.5	29.3	5.9	36.2	89.4	12.6	40.4	56.5	26.7
Ours	40.8	3.6	43.5	40.6	81.8	41.3	27.7	9.2	52.0	92.2	9.4	44.6	61.6	23.4
High-resolution RGB-D														
TangentImg [3]	51.9	4.5	49.9	50.3	85.5	**71.5**	42.4	11.7	50.0	94.3	32.1	**61.4**	70.5	50.0
Ours	**56.3**	7.4	**62.3**	**55.5**	**87.0**	66.4	**44.3**	**19.2**	**66.5**	**96.1**	**43.3**	60.1	**72.9**	**51.4**

(a) Per-class IoU (%).

Method	overall	beam	board	bookcase	ceiling	chair	clutter	column	door	floor	sofa	table	wall	window
Low-resolution RGB-D														
UGSCNN [5]	54.7	19.6	48.6	49.6	93.6	63.8	43.1	28.0	63.2	96.4	21.0	70.0	74.6	39.0
HexRUNet [12]	58.6	23.2	56.5	62.1	94.6	66.7	41.5	18.3	64.5	96.2	41.1	**79.7**	77.2	41.1
TangentImg [3]	50.2	**25.6**	33.6	44.3	87.6	51.5	44.6	12.1	64.6	93.6	26.2	47.2	78.7	42.7
Ours	52.1	9.5	56.5	56.6	95.1	57.9	40.7	12.5	64.5	96.8	10.6	69.1	79.3	28.4
High-resolution RGB-D														
TangentImg [3]	**69.1**	22.6	62.0	70.0	90.3	**84.7**	55.5	**41.4**	76.7	96.9	**70.3**	73.9	80.1	**74.3**
Ours	68.9	16.7	**79.0**	**71.8**	**96.4**	79.2	**59.7**	26.9	**77.7**	**98.2**	58.0	79.6	**85.9**	66.3

(b) Per-class Acc (%).

Table B: Detailed quantitative per-class results on Stanford2D3D [1] with RGB-D as input.

E. More qualitative comparisons for depth estimation

We show more qualitative comparisons with the prior art—BiFuse [9]—in Fig. C. BiFuse's results are obtained from their official released model trained on the real-world Matterport3D [2] dataset.

Figure C: More qualitative comparisons of the estimated dense depth with the prior art—BiFuse [9]. The 'Advantage' column shows the MAE difference between ours and BiFuse's where the blue color indicates ours is better and the red color for vice versa.

F. Qualitative results for semantic segmentation

Qualitative results for semantic segmentation on Stanford2D3D [1] dataset are shown in Fig. D. We fail to build the prior art [3] from their public release for semantic segmentation on high-resolution panorama, so we only show our results.

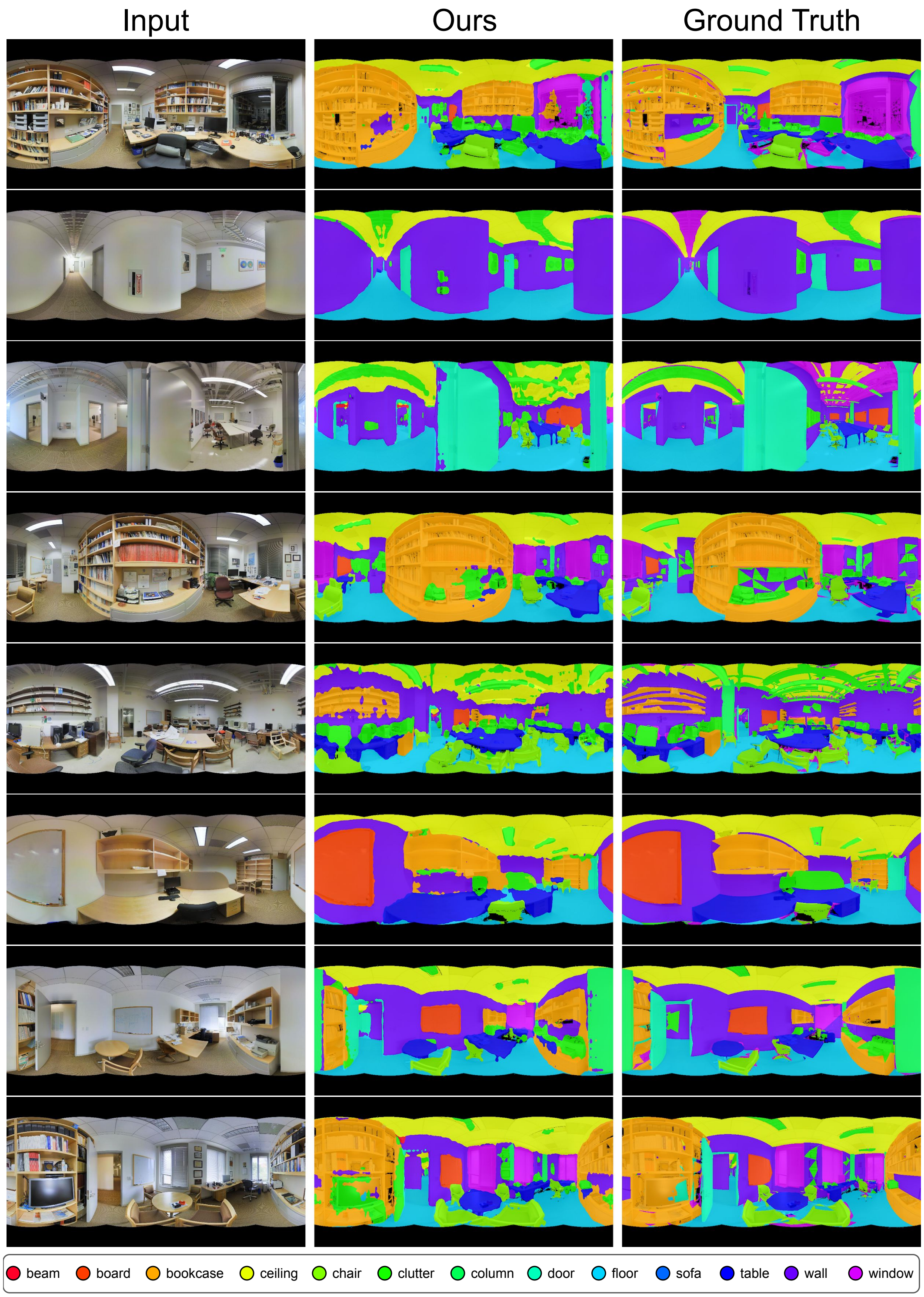


Figure D: Qualitative results for semantic segmentation on Stanford2D3D [1] dataset.

G. Qualitative comparisons for layout estimation

We show qualitative comparisons for room layout estimation with the prior art—AtlantaNet [7]—in Fig. E. The results of AtlantaNet are obtained from their official code and pre-trained weights. We use [8] post-processing algorithm to produce Manhattan layouts; AtlantaNet [7]'s algorithm generates less restrictive Atlanta layouts. Our model achieves promising results comparable to the most recent AtlantaNet [7], while our model runs $22\times$ faster.

Figure E: Qualitative comparisons for room layout estimated with the competitive AtlantaNet [7]. The green, magenta, and blue are the ground truth layout, AtlantaNet's results, and our results respectively.

H. 3D visualization for the estimated depth

Figure F: 3D visualization for the estimated depth by HoHoNet.

I. 3D visualization for the estimated layout

Figure G: 3D visualization for the estimated layout by HoHoNet.

J. Future work

There are plenty of potential future directions upon the proposed framework. *(i)* In this work, we only present two generic operations—linear interpolation and the inverse discrete cosine transform—to predict dense from the LHFeat. With the unified basis view between the two specific operations, we hope future work can find an even better basis or task-specialized basis. *(ii)* Developing the ERP distortion-aware technique upon our framework could also be more uncomplicated as our "decoder" is horizontal, enabling future work to focus only on the backbone layers. *(iii)* The effectiveness of the proposed HoHoNet is only showcased on each modality separately in this work. Extend the recent works on 360° multi-modalities regularization [6] or 360° cascade multi-stages modeling [11] upon the LHFeat is a potential direction. *(iv)* Finally, despite the generally good performance of HoHoNet, our visualization reveals the weakness of HoHoNet on the boundary region and high-frequency signal in a column. Future work on proposing a remedy for the observed issue is desirable.

References

[1] Iro Armeni, Sasha Sax, Amir Roshan Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *CoRR*, abs/1702.01105, 2017. 2, 4

[2] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pages 667–676. IEEE Computer Society, 2017. 3

[3] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12423–12431. IEEE, 2020. 2, 4

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 1

[5] Chiyu Max Jiang, Jingwei Huang, Karthik Kashinath, Prabhat, Philip Marcus, and Matthias Nießner. Spherical cnns on unstructured grids. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 2

[6] Lei Jin, Yanyu Xu, Jia Zheng, Junfei Zhang, Rui Tang, Shugong Xu, Jingyi Yu, and Shenghua Gao. Geometric structure based and regularized depth estimation from 360 indoor imagery. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 886–895. IEEE, 2020. 6

[7] Giovanni Pintore, Marco Agus, and Enrico Gobbetti. Atlantanet: Inferring the 3D indoor layout from a single 360 image beyond the manhattan world assumption. In *Proceedings of The European Conference on Computer Vision (ECCV)*, 2020. 2, 5

[8] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. HorizonNet: learning room layout with 1d representation and pano stretch data augmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1047–1056, 2019. 1, 2, 5

[9] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 459–468. IEEE, 2020. 3

[10] Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung-Kuo Chu. Dula-net: A dual-projection network for estimating room layouts from a single RGB panorama. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3363–3372. Computer Vision Foundation / IEEE, 2019. 2

[11] Wei Zeng, Sezer Karaoglu, and Theo Gevers. Joint 3d layout and depth prediction from a single indoor panorama image. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVI*, volume 12361 of *Lecture Notes in Computer Science*, pages 666–682. Springer, 2020. 6

[12] Chao Zhang, Stephan Liwicki, William Smith, and Roberto Cipolla. Orientation-aware semantic segmentation on icosahedron spheres. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3532–3540. IEEE, 2019. 2

[13] Chuhang Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung-Kuo Chu, and Derek Hoiem. 3d manhattan room layout reconstruction from a single 360 image. *CoRR*, abs/1910.04099, 2019. 2