Complementary Bi-directional Feature Compression for Indoor 360° Semantic Segmentation with Self-distillation: SUPPLEMENTARY MATERIAL

Zishuo Zheng[1,2], Chunyu Lin[1,2]*, Lang Nie[1,2], Kang Liao[1,2], Zhijie Shen[1,2], Yao Zhao[1,2]
[1]Institute of Information Science, Beijing Jiaotong University, Beijing, China
[2]Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China
{zszheng, cylin, nielang, kang_liao, zhjshen, yzhao}@bjtu.edu.cn

1. Overview

In this material, we first show a detailed illustration of our network architecture in Sec.2. This illustration offers details on all the network modules and is intended to complement the general description provided in the paper. Then we show class-level qualitative results for the panoramic semantic segmentation experiments at different input resolutions in Sec3. Furthermore, we illustrate more experimental results of our solution in Sec.4. Finally, We validate the extensibility of our model by the panoramic depth estimation in Sec. 5.

2. Detailed Network Architecture

We show the detailed network architecture in Fig.1. The proposed solution takes as input an equirectangular RGB image (256×512) and outputs a segmentation image at the same resolution of the input. To be more specific, we use the residual U-Net-style architecture [9] [3] as backbone to generate 4 levels feature maps. Then these features are fed into the feature pyramid network and Mix-MLP layer to yield powerful representations without changing sizes. Subsequently, these features are compressed in parallel, keeping the horizontal dimension unchanged and compressing the vertical one, and keeping the vertical dimension unchanged and compressing the horizontal one. To align different resolution 1D representations, we interpolate these tenors so that they have the same horizontal dimension (64) and vertical dimension (128). Finally, we concatenate them to obtain the bi-directional representations with different channels (1024 for horizontal and 2048 for vertical which are hyperparameters.) with the primary consideration is that the width of the panoramic image is twice the height. In the decoding process, we upsample two flattened representations to output two 2D feature maps having the same sizes (64×128). Particularly, for most padding operations, we use circular padding for the left-right boundaries of the feature maps. We exploit zero padding (**ZP**) in the Mix-MLP layer and decode part (*A-Conv* [5]).

*Corresponding author

3. Detailed Semantic Segmentation Results

Detailed per-class IoU and per-class Acc results are given in Table.1, and Table.2. For low-resolution RGB-D input (64×128), we achieve the highest IoU on 9 out of 13 classes and the best Acc on 8 out of 13 classes. For high-resolution RGB-D input (1024×2048), we achieve the highest IoU on 9 out of 13 classes and the best Acc on 7 out of 13 classes. More importantly, the rest classes without achieving the highest quantitative results can also be the second-highest. Furthermore, we can also observe that nearly every class benefits from our complementary bi-directional representation. This is especially noticeable for classes with horizontal distribution and large distortion shape, like *chair*, *ceiling*, *floor*, and *wall*.

4. More Qualitative Results

We exhibit more qualitative comparisons with the previous work—HoHoNet [6] in Fig.2 and Fig. 3, where our solution can deal with different indoor scenes and yield the best performance on visual appearance.

5. Algorithm extensibility

Theoretically, our network also can handle other pixel2pixel tasks. So we further validate our extensibility in other pixel2pixel tasks, such as panoramic depth estimation task in the same dataset. We removed the self distillation (because the loss function needs to be redesigned) and did not change any other structures. We strictly followed the experimental protocol in other solutions [8] [10]. As shown in the Table. 3, the results show that our model outperforms the current SOTA approaches in most metrics (especially in the most important metric, RMSE). It also demonstrates that the proposed model has the potential to solve other tasks. Fig. 4 shows the qualitative comparison results which indicate that the complementary features help our network build a better panoramic perception capability.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

[2] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12426–12434, 2020.

[3] Chiyu Max Jiang, Jingwei Huang, Karthik Kashinath, Philip Marcus, Matthias Niessner, et al. Spherical cnns on unstructured grids. In *International Conference on Learning Representations*, 2018.

[4] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11536–11545, 2021.

[6] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1047–1056, 2019.

[7] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2573–2582, 2021.

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 462–471, 2020.

[9] Chao Zhang, Stephan Liwicki, William Smith, and Roberto Cipolla. Orientation-aware semantic segmentation on icosahedron spheres. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3533–3541, 2019.

[10] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 448–465, 2018.

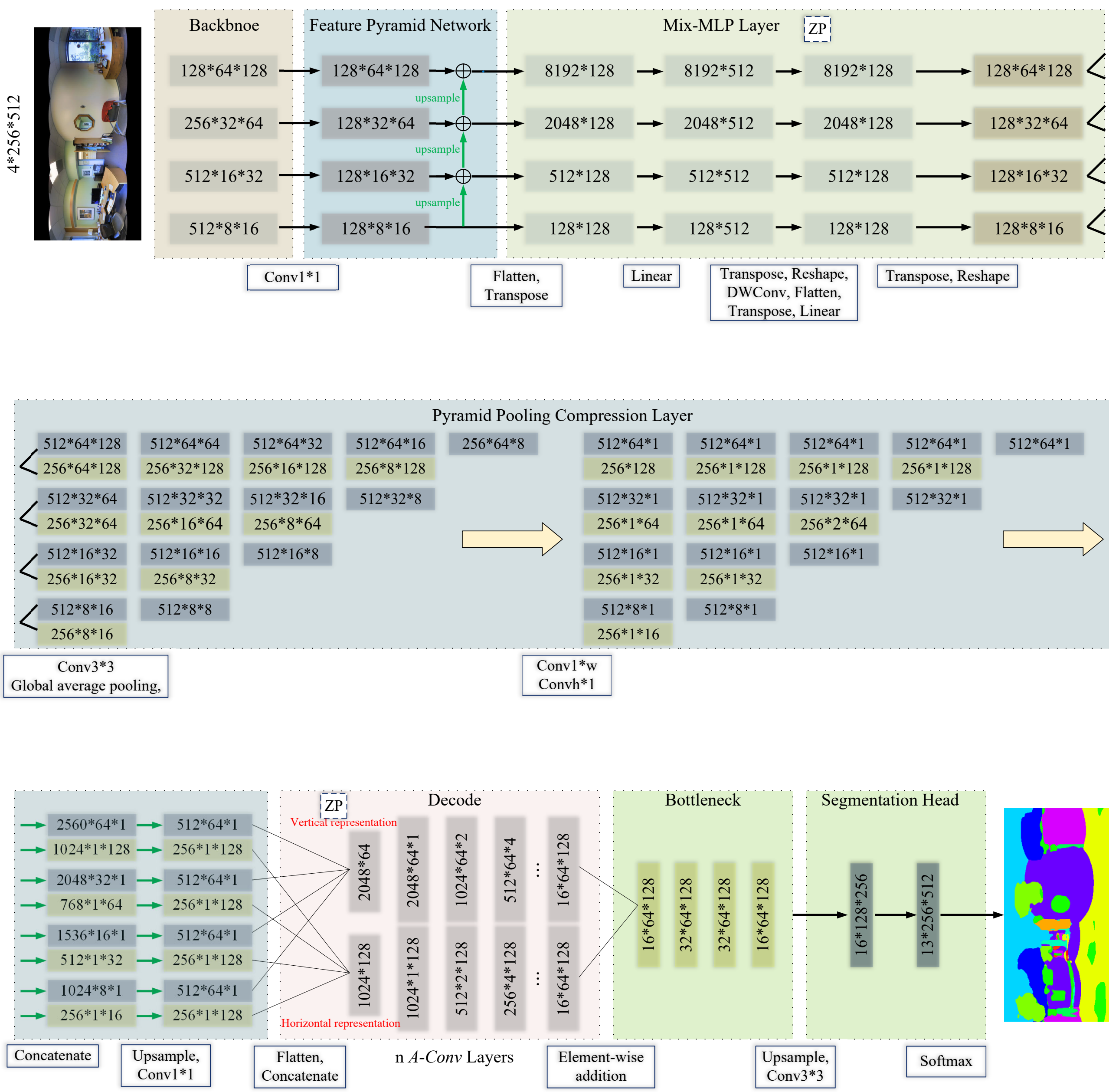

Figure 1. The detailed display of our network architecture.

Table 1. Detailed quantitative per-class IoU (%) results on Stanford2D3D [1]. The top two result are shown in red and blue.

Method	Overall	beam	board	bookcase	ceiling	chair	clutter	column	door	floor	sofa	table	wall	window
Low-resolution RGB-D														
UGSCNN [3]	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 [9]	**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 [2]	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
HoHoNet [7]	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
Ours	**47.2**	**8.9**	**50.1**	**44.0**	**85.1**	**47.7**	**35.2**	**11.5**	**54.6**	**93.9**	**18.6**	**48.5**	**66.2**	**35.0**
High-resolution RGB-D														
TangentImg [2]	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**
HoHoNet [7]	**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**
Ours	**57.1**	**10.0**	**59.9**	**55.0**	**88.6**	**72.9**	**46.8**	**19.2**	**63.9**	**96.6**	**44.3**	**63.7**	**73.4**	47.8

Table 2. Detailed quantitative per-class Acc (%) results on Stanford2D3D [1]. The top two result are shown in red and blue.

Method	Overall	beam	board	bookcase	ceiling	chair	clutter	column	door	floor	sofa	table	wall	window
Low-resolution RGB-D														
UGSCNN [3]	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 [9]	**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 [2]	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**
HoHoNet [7]	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
Ours	**61.2**	**26.3**	**68.6**	**58.9**	**95.3**	**65.0**	**48.5**	16.7	**70.0**	**97.3**	**32.4**	**74.0**	**81.5**	**44.4**
High-resolution RGB-D														
TangentImg [2]	**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**
HoHoNet [7]	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**
Ours	**69.9**	**22.8**	**77.9**	**71.0**	**96.9**	**84.9**	**61.1**	26.4	76.0	**98.3**	**60.8**	**79.9**	**86.8**	61.5

Table 3. Quantitative comparison for depth estimation on Stanford2D3D[1].

Method	MRE ↓	MAE ↓	RMSE ↓	RMSE(log) ↓	δ^1 ↑	δ^2 ↑	δ^3 ↑
FCRN [4]	0.1837	0.3428	0.5774	0.1100	0.7230	0.9207	0.9731
OmniDepth [10]	0.1996	0.3743	0.6152	0.1212	0.6877	0.8891	0.9578
Equi [8]	0.1428	0.2711	0.4637	0.0911	0.8261	0.9458	0.9800
Cube [8]	0.1332	0.2588	0.4407	0.0844	0.8347	0.9523	0.9838
BiFuse [8]	0.1209	0.2343	0.4142	0.0787	0.8660	0.9580	0.9860
HoHoNet [7]	**0.1014**	0.2027	0.3834	**0.0668**	**0.9054**	0.9693	0.9886
Ours	0.1039	**0.1957**	**0.3678**	0.0679	0.8933	**0.9747**	**0.9901**

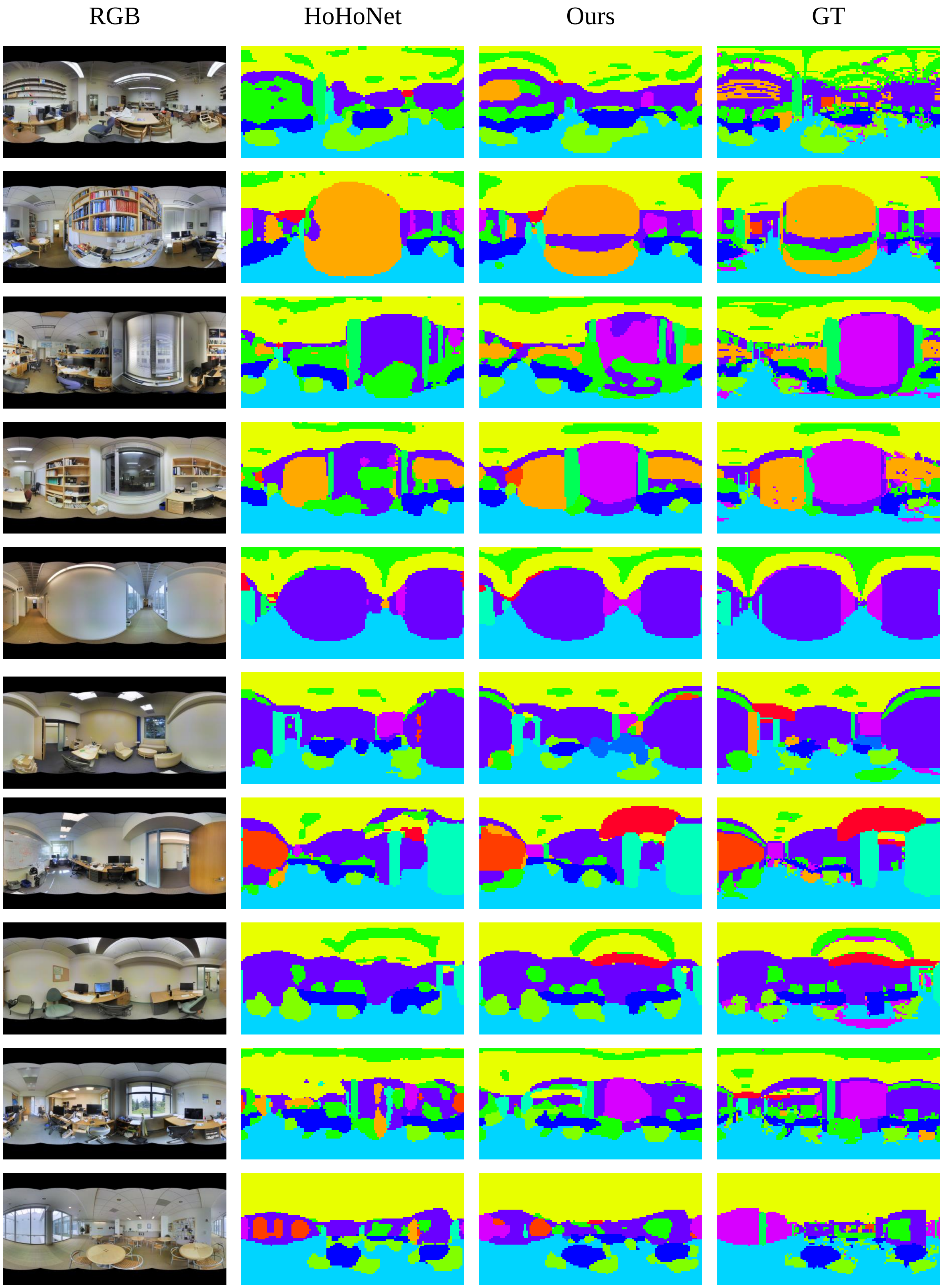

Figure 2. More results of our method on 64×128 resolution RGB-D input.

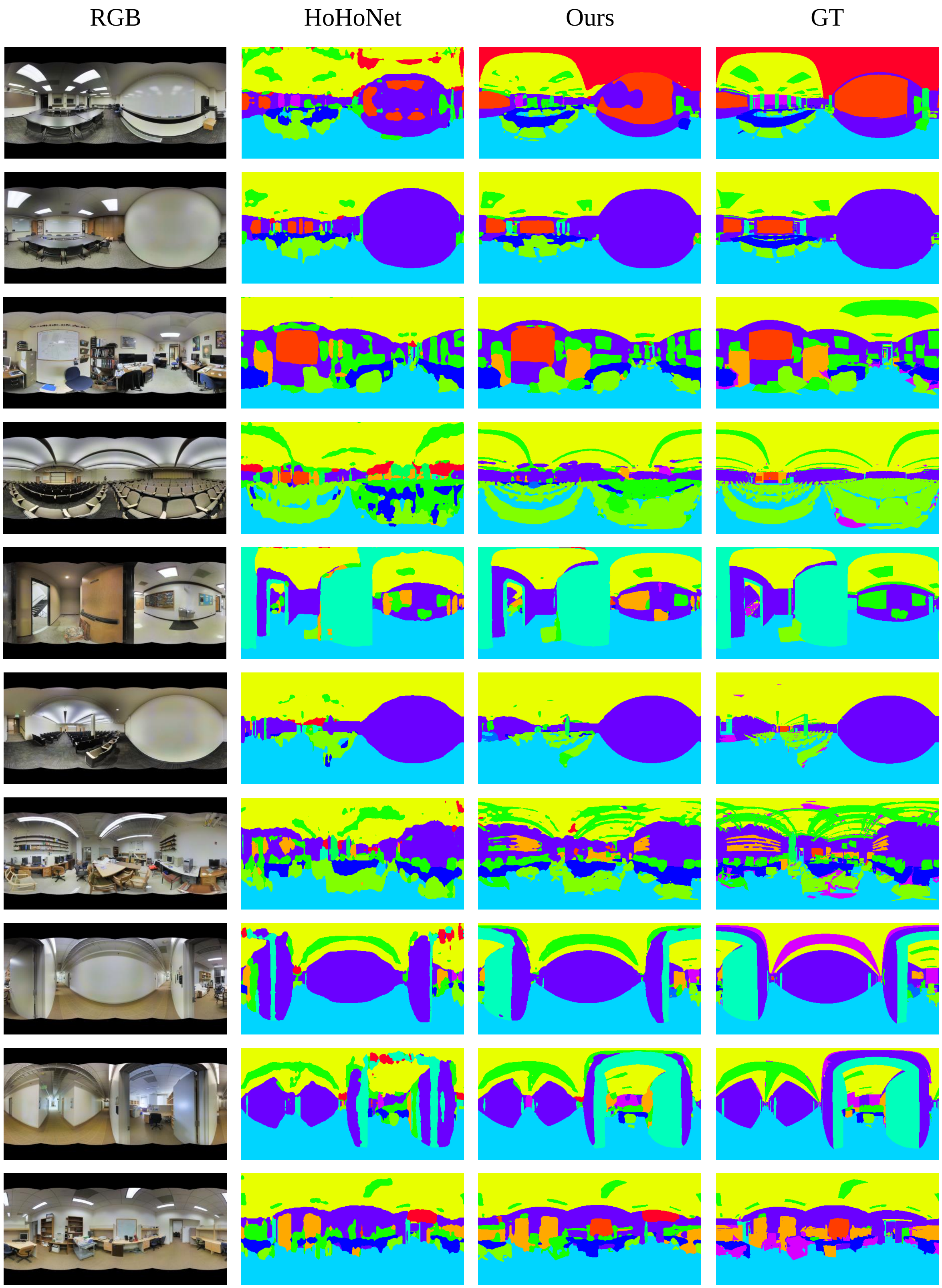

Figure 3. More results of our method on 256 $\times$ 512 resolution RGB-D input

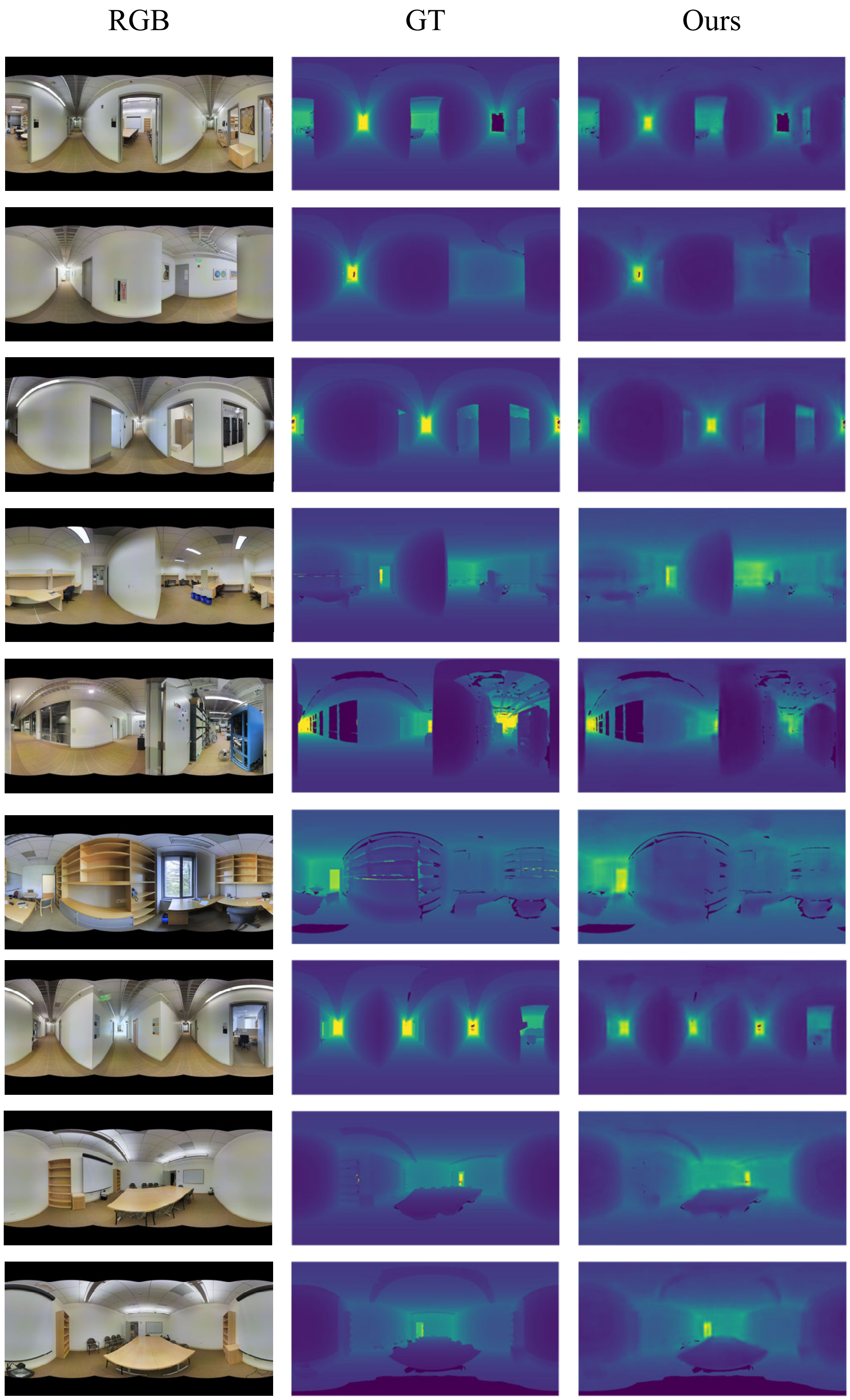

Figure 4. Qualitative comparison on the panoramic depth estimation.