HiMODE: A Hybrid Monocular Omnidirectional Depth Estimation Model (Supplementary Materials)

In this supplementary material, we provide more ablation studies and results of the proposed *HiMODE*.

A. Ablation Studies on Backbone

As introduced in the main paper, backbone module is an important part of our system. This section provides more ablation studies on the backbone module to demonstrate its superiority, quantitatively and qualitatively, to the other pre-trained backbones.

A.1. Detailed Architecture of the Backbone.

Our CNN-based backbone is referred to as depth-wise due to using depth-wise Conv layers in HNet blocks which are concatenated with the Conv layers. Depth-wise separable CNNs have less parameters and possibility for overfitting, such as MobileNet. HNet (as shown in Figure 1) is extracted from HardNet [4]. Comparing the number of layers, our backbone has only 40 layers (i.e. HNet=4×8, Conv=4, Concat=4 layers) which is significantly less than that of HardNet (i.e. 68 layers).

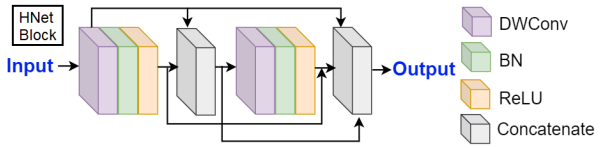

Figure 1. The overall architecture of the proposed HNet block extracted from the HardNet [4] structure.

A.2. The Effects of Input Resolution

The visual information is affected by the image resolution. High image resolution results in higher visual information and so better image quality. Generally, when the image resolution is reduced, the performance of the CNN-based networks degrades significantly [7]. On the other hand, lower input image resolution is desirable as it leads to a reduced number of features and the optimized number of parameters. Consequently, the risk of model overfitting is minimized [2]. Nevertheless, extensive lowering of the image resolution eliminates the information that is useful for classification. The effects of the input image resolution on the overall performance of the proposed system based on our novel CNN-based backbone is investigated and compared with four pre-trained models of ResNet34 [5], ResNet50 [5], DenseNet [6], and HardNet [4]. The evaluation results are presented in Table 1 in terms of four error-based evaluation metrics and three accuracy-based evaluation metrics. The terms "low" and "high" for image resolution refer to the image size of 256×512 and 512×1024, respectively. Comparing the results, our proposed backbone ranks first in all evaluation metrics on all three datasets, except for *Abs-Rel* and δ on Stanford3D, *RMSElog* on Matterport3D, and *RMSE* on PanoSunCG, at which our proposed backbone ranks second with a slight difference. The superiority of our proposed backbone is proven as the other models cannot surpass its performance even with high-resolution inputs. It is worth mentioning that the overall performance of our proposed system maintains almost the same when the resolution of the input images varies, demonstrating its independence and robustness to the input image size. Consequently, our *HiMODE* system is proposed based on the low-resolution input images so that the number of parameters is reduced without sacrificing the performance accuracy, as opposed to the other state-of-the-art approaches [10, 12] which were mostly based on 512×1024 input images.

A.3. Computation Cost of Different Backbones

In addition to the performance, the superiority of our proposed CNN-based backbone is further investigated by comparing its computation cost with that of four pre-trained models of ResNet34 [5], ResNet50 [5], DenseNet [6], and HardNet [4]. The results in terms of the number of parameters as computation cost with three accuracy-based evaluation metrics on Stanford3D [1] dataset are presented in Table 2 (for both low and high resolution). We can observe that the proposed *HiMODE* based on our novel CNN-based backbone has the least number of parameters for low resolution input images with the values of 79.67M as well as the best performance accuracy of 0.9711 and 0.9965 in terms of δ, δ^2, respectively. Its performance in terms of δ^3 is almost the same as that of HardNet. Replacing the other pre-trained models of ResNet34, ResNet50, DenseNet, and HardNet with our proposed backbone brings additional computation burden (i.e. parameters) of 7.29M, 10M, 6.48M, and

2.57M, respectively. Besides, accuracy also significantly decreases. The highest degradation in δ, and δ^2 occurs in DenseNet with the values of 0.9076 and 0.9839, respectively, while the poorest performance of 0.9880 in terms of δ^3 belongs to ResNet34. For high resolution input images, *HiMODE* based on our proposed CNN-based backbone still has the least number of parameters (98.89M) comparing with the others. Achieving the least computation cost with the highest performance accuracy proves the capabilities of our proposed backbone over the other pre-trained feature extractors.

A.4. Qualitative Results for Different Backbones

The performance of *HiMODE* based on our proposed CNN-based backbone is compared with the other pre-trained models qualitatively in Figures 2-4. As it is mentioned in the main paper, our depth-wise proposed backbone can extract high-resolution features near the edges to overcome distortion and artifact issues. On the depth maps estimated based on our proposed backbone, sharper edges and more details are recovered.

B. More Results on 3D Structure

B.1. Quantitative Results

The detailed quantitative results for 3D structure estimation under different number of ground-truth corners are presented in Table 3 as a supplement to the main paper to extend the quantitative studies. In comparison with the recent state-of-the-art approaches, our proposed *HiMODE* achieves the best results for 6 corners (82.23%) on the 2D IOU (intersection over union) metric, and both 4 (85.48%) and 6 (85.05%) corners in terms of 3D IOU. Overall, our proposed method can achieve state-of-the-art performance in 3D structure estimation with fewer corners. For higher number of corners, our method obtained comparable results although AtlantaNet [8] is the best performer.

B.2. Qualitative Results

Additional qualitative results for estimating 3D structures from monocular omnidirectional images on three datsets of Stanford3D [1], Matterport3D [3], and PanoSunCG [11] are demonstrated in Figures 5-7, respectively[1]. Our method was evaluated on different input images with various numbers of corners. Qualitatively, our *HiMODE* can successfully reconstruct the 3D structure by finding the corners and boundary between walls, floor, and ceiling, which is a vital task in VR/AR and robotics applications. The proposed *HiMODE* successfully reconstructs the 3D structure with different numbers of corners by finding the corners and boundary between walls, floor, and ceiling.

[1]Some samples of 3D structures are available at https://bit.ly/3HLh1Z3 in video format.

C. More Omnidirectional Depth Results

We show more qualitative results for depth map estimation by our *HiMODE* in Figures 8-10 on three datasets; Stanford3D, Matterport3D, and PanoSunCG. The results of our proposed *HiMODE* are compared with two other recent state-of-the-art approaches of Bifuse [12] and HoHoNet [10] on three datasets in Figures 11-13. These visual results further demonstrate the superior performance of the proposed *HiMODE* over the other two methods in recovering the details of the surfaces, even for the deep regions and small objects.

In addition, the effectiveness of combining the *HiMODE* output with the output of two recent state-of-the-art approaches; Bifuse [12] and HoHoNet [10], on three datasets is investigated. The qualitative results are illustrated in Figures 14-16. Very interestingly, we observe significant improvement in the depth map estimation when *HiMODE* is combined with Bifuse, HohoNet, or both methods via a simple concatenation of the respective outputs. The best qualitative results are achieved with the combination of three methods, whereby the resulting depth map mimics the groundtruth depth map very closely (the last columns of Figures 14-16).

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. 1, 2

[2] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on neural networks*, 5(4):537–550, 1994. 1

[3] 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. *arXiv preprint arXiv:1709.06158*, 2017. 2

[4] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. Hardnet: A low memory traffic network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3552–3561, 2019. 1, 3, 4, 5, 6

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3, 4, 5, 6

[6] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1, 3, 4, 5, 6

[7] Suresh Prasad Kannojia and Gaurav Jaiswal. Effects of varying resolution on performance of cnn based image classification: An experimental study. *Int. J. Comput. Sci. Eng*, 6(9):451–456, 2018. 1

Table 1. A quantitative comparison between the proposed CNN-based backbone with four pre-trained models on three datasets based on two input image resolutions of 256×512 (low) and 512×1024 (high).

Datasets	Backbones	Resolution	Errors				Accuracy		
			Abs-Rel	Sq-Rel	RMSE	RMSElog	δ	δ^2	δ^3
Stanford3D	ResNet34 [5]	High	0.0956	0.0824	0.3875	0.1577	0.9398	0.9817	0.9906
		Low	0.1128	0.0635	0.3665	0.1873	0.9149	0.9884	0.9880
	ResNet50 [5]	High	0.0666	0.0489	0.2897	0.1217	0.9512	0.9940	0.9968
		Low	**0.0509**	0.0682	0.3177	0.1185	0.9349	0.9906	0.9923
	DenseNet [6]	High	0.0823	0.0702	0.3346	0.1246	0.9451	0.9901	0.9944
		Low	0.1045	0.0624	0.3358	0.1621	0.9076	0.9839	0.9889
	HardNet [4]	High	0.0755	0.0461	0.2984	0.1038	0.9578	0.9947	0.9972
		Low	0.0789	0.0352	0.3041	0.1215	0.9234	0.9947	**0.9992**
	Proposed	High	0.0679	0.0223	0.2711	0.0963	0.9693	0.9959	0.9987
		Low	0.0532	**0.0207**	**0.2619**	**0.0821**	**0.9711**	**0.9965**	0.9989
Matterport3D	ResNet34 [5]	High	0.1026	0.0861	0.3956	0.1434	0.9487	0.9820	0.9777
		Low	0.1078	0.1139	0.4587	0.1786	0.8976	0.9792	0.9800
	ResNet50 [5]	High	0.0699	0.0586	0.3610	0.1003	0.9523	0.9928	0.9859
		Low	0.1014	0.0856	0.4189	0.1251	0.9257	0.9755	0.9945
	DenseNet [6]	High	0.0782	0.0545	0.3678	0.1165	0.9501	0.9893	0.9908
		Low	0.0935	0.0472	0.3548	0.1547	0.9138	0.9668	0.9829
	HardNet [4]	High	0.0630	0.0471	0.3355	**0.0873**	0.9562	0.9918	0.9938
		Low	0.0769	0.0244	0.3648	0.1174	0.9415	0.9831	0.9902
	Proposed	High	**0.0597**	**0.0213**	0.3146	0.0894	0.9601	0.9921	0.9981
		Low	0.0658	0.0245	**0.3067**	0.0959	**0.9608**	**0.9940**	**0.9985**
PanoSunCG	ResNet34 [5]	High	0.1006	0.0653	0.3989	0.1595	0.9466	0.9783	0.9849
		Low	0.1353	0.1471	0.4823	0.2379	0.9183	0.9947	0.9926
	ResNet50 [5]	High	0.0832	0.0474	**0.3259**	0.1339	0.9524	0.9864	0.9936
		Low	0.1094	0.1043	0.3847	0.2149	0.9524	0.9918	0.9989
	DenseNet [6]	High	0.0852	0.0427	0.3561	0.1226	0.9538	0.9889	0.9951
		Low	0.0949	0.0987	0.4283	0.1958	0.9245	0.9909	0.9895
	HardNet [4]	High	0.0715	0.0398	0.3303	0.1178	0.9615	0.9910	0.9978
		Low	0.0726	0.0557	0.3985	0.1305	0.9693	0.9897	0.9877
	Proposed	High	**0.0667**	**0.0347**	0.3265	**0.1013**	**0.9691**	0.9945	0.9990
		Low	0.0682	0.0356	0.3378	0.1048	0.9688	**0.9951**	**0.9992**

Table 2. Comparison between the proposed CNN-based backbone with four pre-trained models as backbone in terms of computation cost and accuracy (on Stanford3D dataset). The bold and underlined numbers indicate the best results for low and high resolution input images, respectively.

Backbones	Input	Computation Cost	Accuracy		
		Parameters	δ	δ^2	δ^3
ResNet34 [5]	High	103.55M	0.9398	0.9817	0.9906
	Low	86.96M	0.9149	0.9884	0.9880
ResNet50 [5]	High	107.28M	0.9512	0.9940	0.9968
	Low	89.67M	0.9349	0.9906	0.9923
DenseNet [6]	High	104.81M	0.9451	0.9901	0.9944
	Low	86.15M	0.9076	0.9839	0.9889
HardNet [4]	High	100.37M	0.9578	0.9947	0.9972
	Low	82.24M	0.9234	0.9947	**0.9992**
Proposed	High	<u>98.89M</u>	<u>0.9693</u>	<u>0.9959</u>	<u>0.9987</u>
	Low	**79.67M**	**0.9711**	**0.9965**	0.9989

Table 3. Quantitative comparison between our *HiMODE* and five state-of-the-art methods for 3D structure estimation on Stanford3D dataset in terms of 2D and 3D IOU. The best results are indicated with bold numbers.

IOU (%)	Approaches	# Corners				
		All	4	6	8	10+
2D	LayoutNet v2 [14]	75.82	81.35	72.33	67.45	63.00
	DuLa-Net v2 [13]	75.07	77.02	78.79	71.03	63.27
	HorizonNet [9]	79.11	81.88	82.26	71.78	68.32
	AtlantaNet [8]	**80.02**	82.09	82.08	**75.19**	**71.61**
	HoHoNet [10]	79.88	**82.64**	82.16	73.65	69.26
	HiMODE	79.74	82.40	**82.23**	72.87	69.03
3D	LayoutNet v2 [14]	78.73	84.61	75.02	69.79	65.14
	DuLa-Net v2 [13]	78.82	81.12	82.69	74.00	66.12
	HorizonNet [9]	81.71	84.67	84.82	73.91	70.58
	AtlantaNet [8]	82.09	84.42	83.85	**76.97**	**73.18**
	HoHoNet [10]	**82.32**	85.26	84.81	75.59	70.98
	HiMODE	81.41	**85.48**	**85.05**	74.38	70.10

[8] Giovanni Pintore, Marco Agus, and Enrico Gobbetti. Atlantanet: Inferring the 3d indoor layout from a single 360° image beyond the manhattan world assumption. In *European Conference on Computer Vision*, pages 432–448. Springer, 2020. 2, 3

[9] 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. 3

[10] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet:

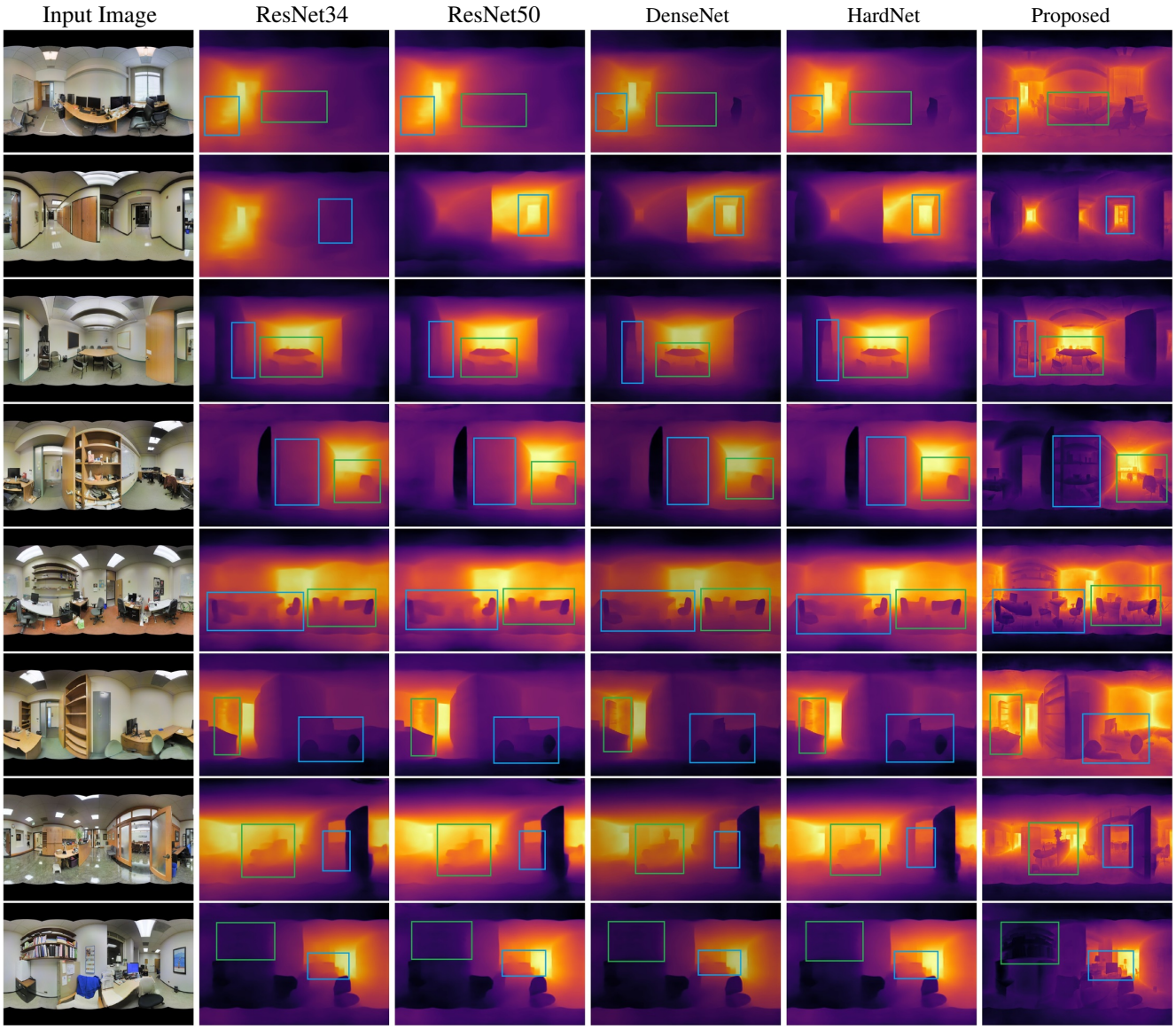

Figure 2. Qualitative comparisons for our *HiMODE* based on our proposed CNN-based backbone and four pre-trained models of ResNet34 [5], ResNet50 [5], DenseNet [6], and HardNet [4] on Stanford3D dataset. As demonstrated by rectangles, our *HiMODE* can accurately recover the details of the surface especially sharp edges even for the deep regions and small objects.

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. 1, 2, 3, 13, 14, 15, 16, 17, 18

[11] Fu-En Wang, Hou-Ning Hu, Hsien-Tzu Cheng, Juan-Ting Lin, Shang-Ta Yang, Meng-Li Shih, Hung-Kuo Chu, and Min Sun. Self-supervised learning of depth and camera motion from 360° videos. In *Asian Conference on Computer Vision*, pages 53–68. Springer, 2018. 2

[12] 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. 1, 2, 13, 14, 15, 16, 17, 18

[13] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3363–3372, 2019. 3

[14] 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. 2019. 3

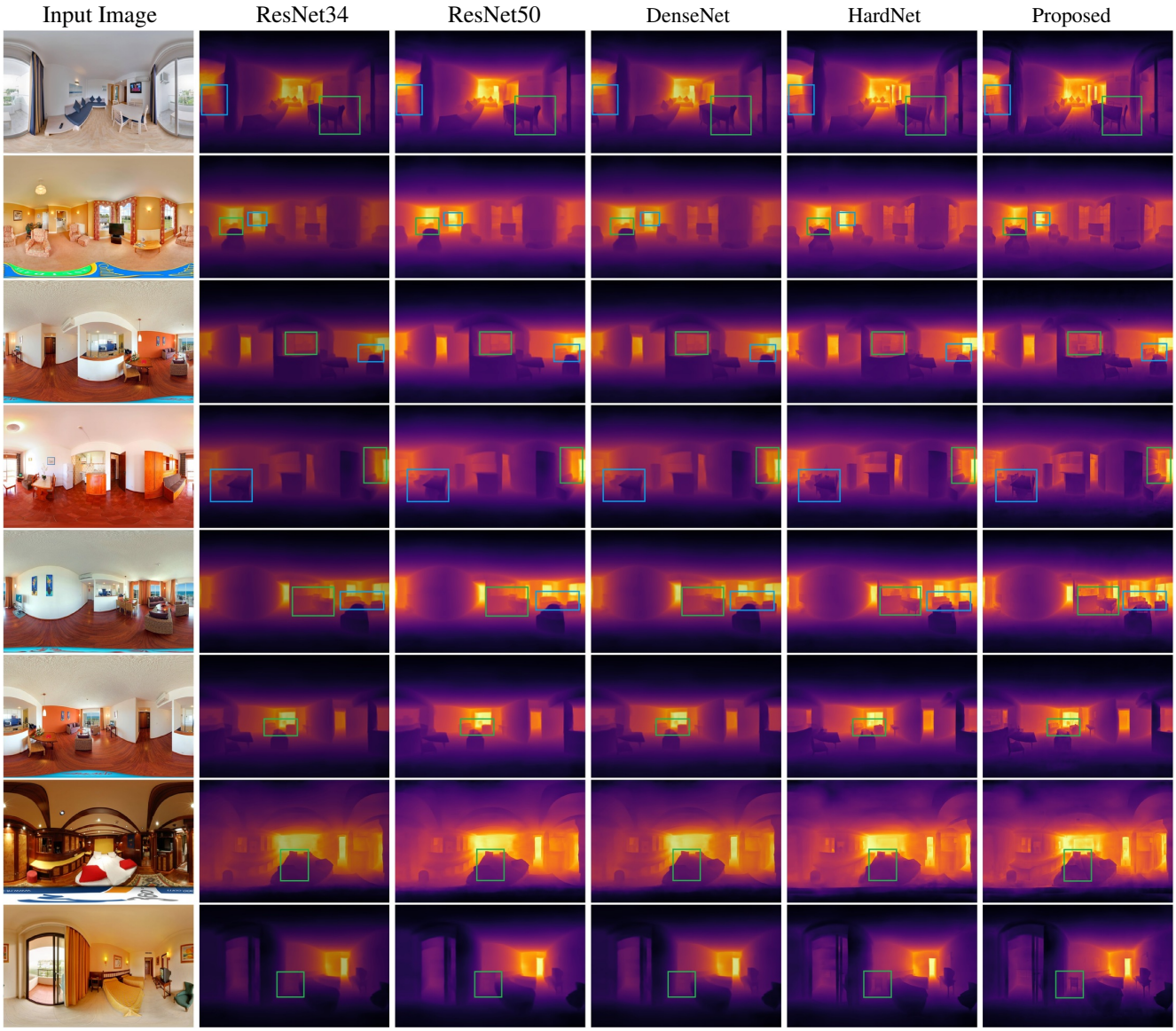

Figure 3. Qualitative comparisons for our *HiMODE* based on our proposed CNN-based backbone and four pre-trained models of ResNet34 [5], ResNet50 [5], DenseNet [6], and HardNet [4] on Matterport3D dataset. As demonstrated by rectangles, our *HiMODE* can accurately recover the details of the surface especially sharp edges even for the deep regions and small objects.

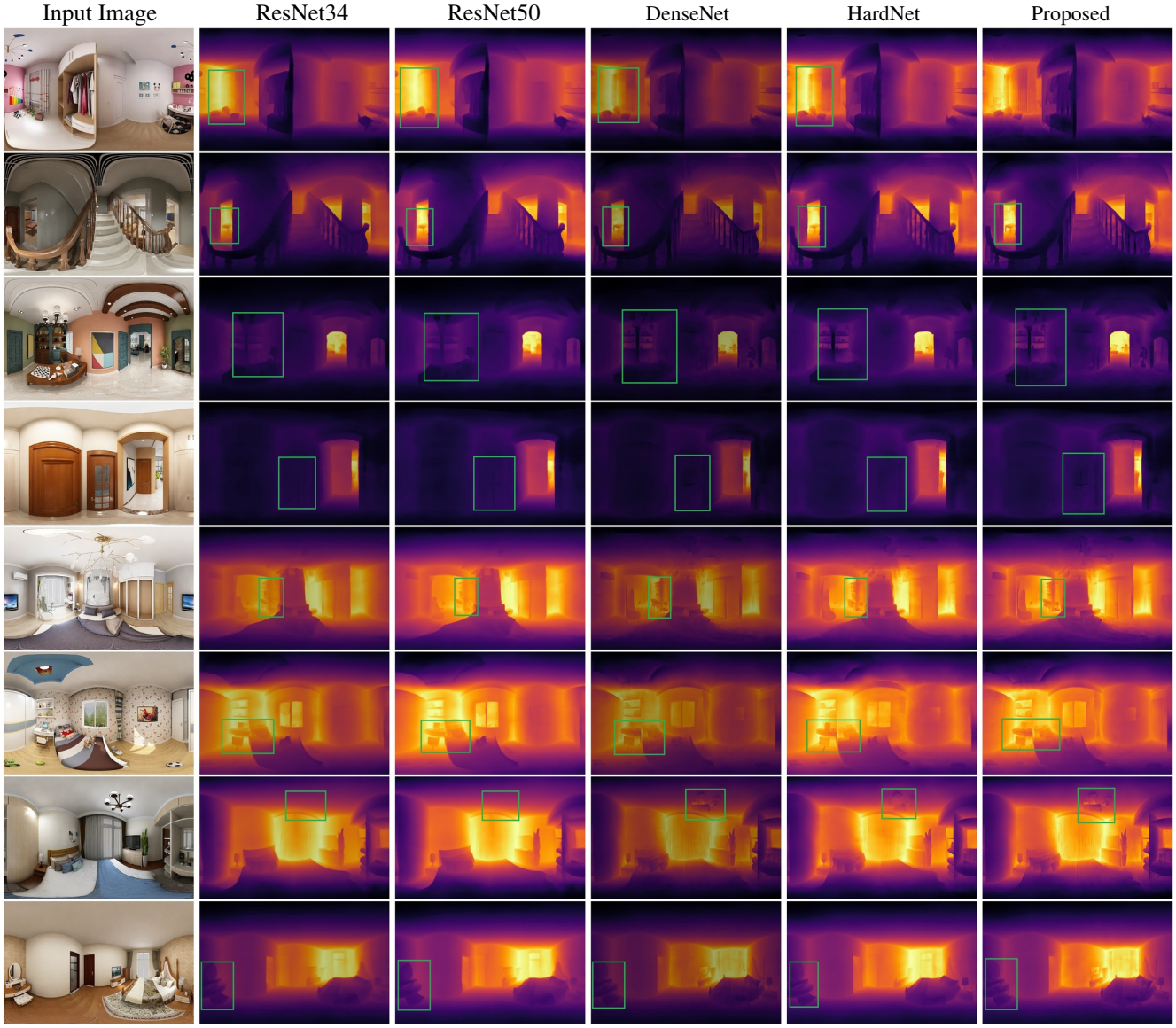

Figure 4. Qualitative comparisons for our *HiMODE* based on our proposed CNN-based backbone and four pre-trained models of ResNet34 [5], ResNet50 [5], DenseNet [6], and HardNet [4] on PanoSunCG dataset. As demonstrated by rectangles, our *HiMODE* can accurately recover the details of the surface especially sharp edges even for the deep regions and small objects.

Figure 5. 3D structures estimation on Stanford3D dataset using our *HiMODE*.

Figure 6. 3D structures estimation on Matterport3D dataset using our *HiMODE*.

Figure 7. 3D structures estimation on PanoSunCG dataset using our *HiMODE*.

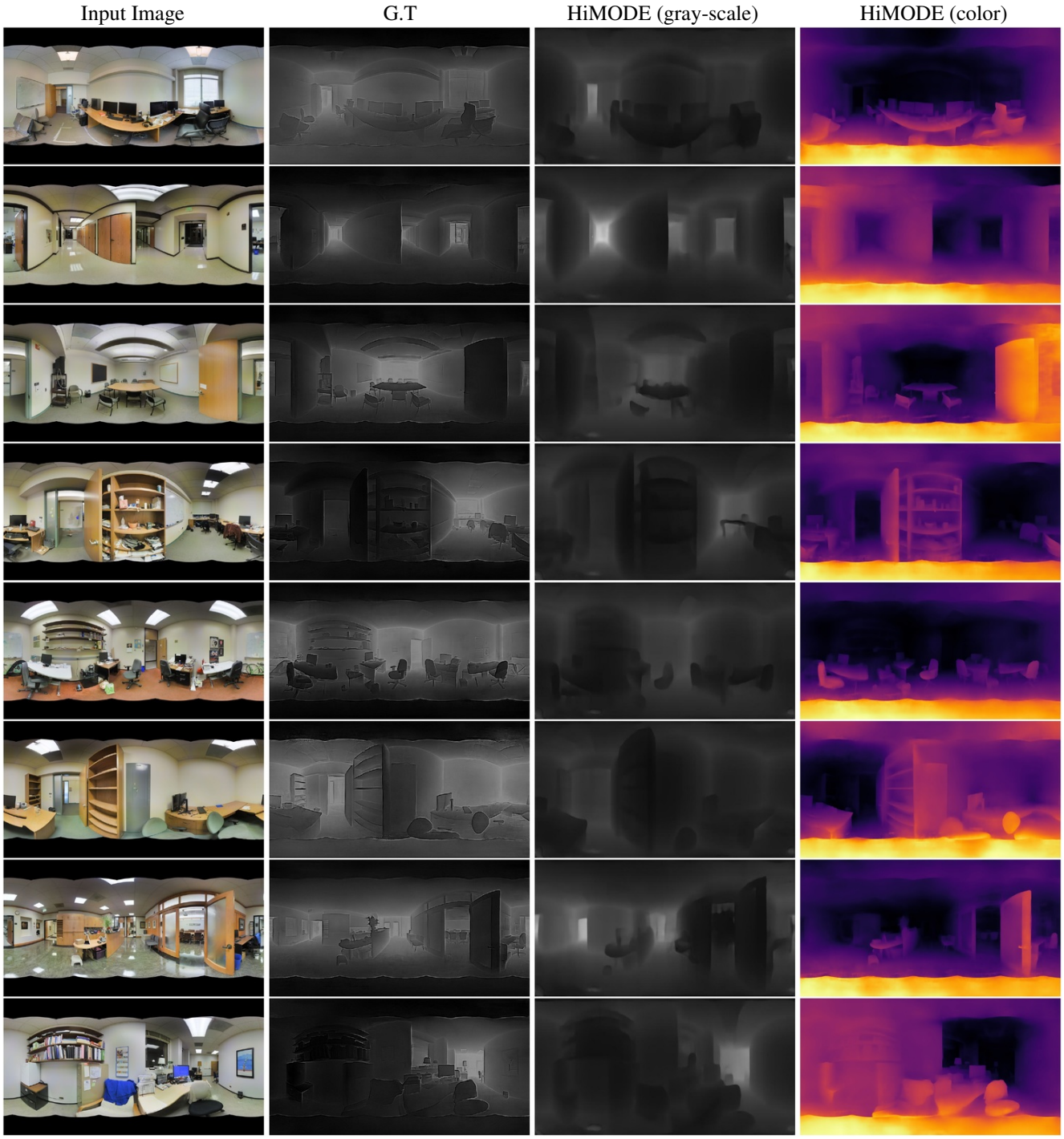

Figure 8. More qualitative results for omnidirectional depth map estimation based on our *HiMODE* on Stanford3D dataset.

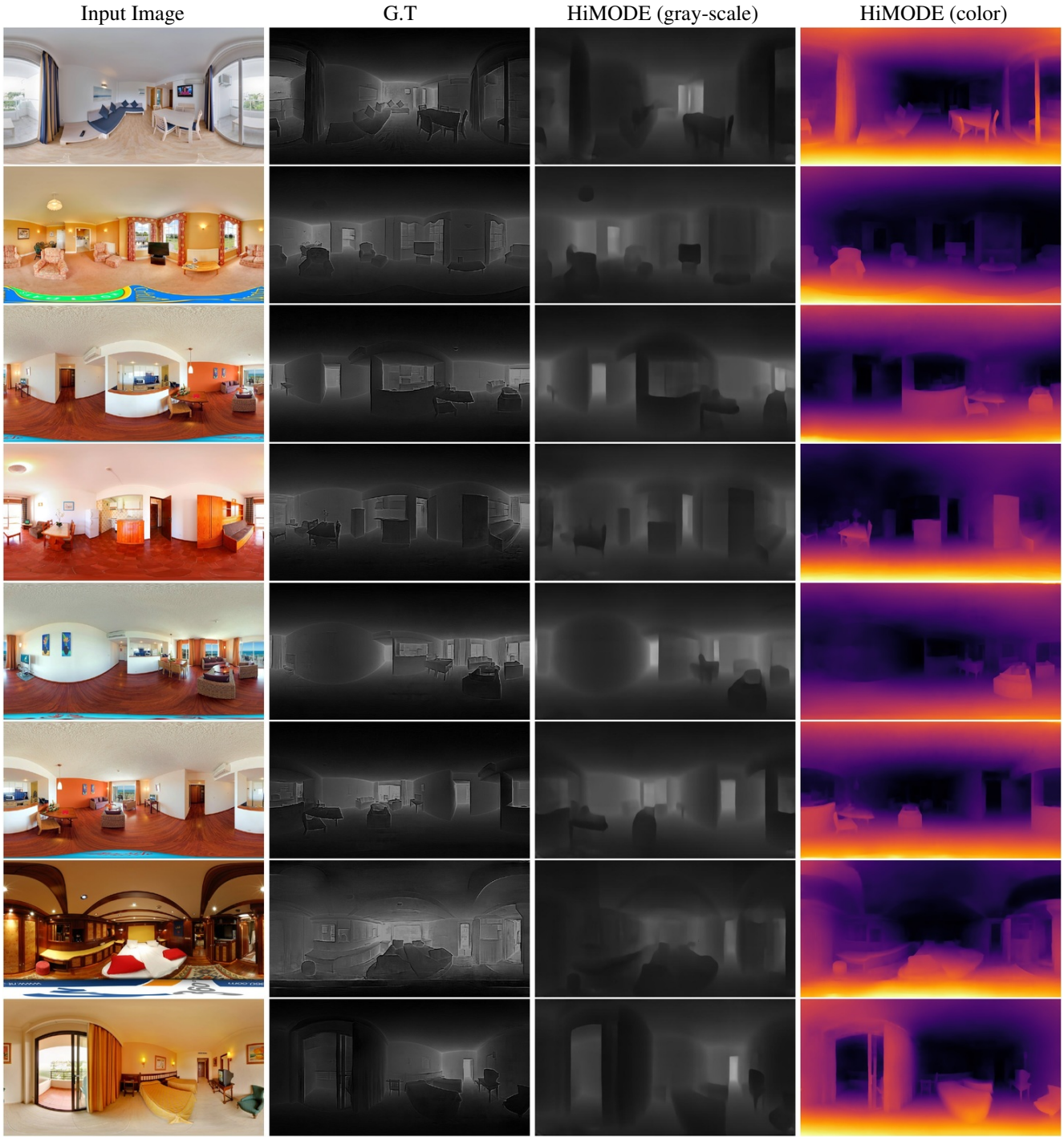

Figure 9. More qualitative results for omnidirectional depth map estimation based on our *HiMODE* on Matterport3D dataset.

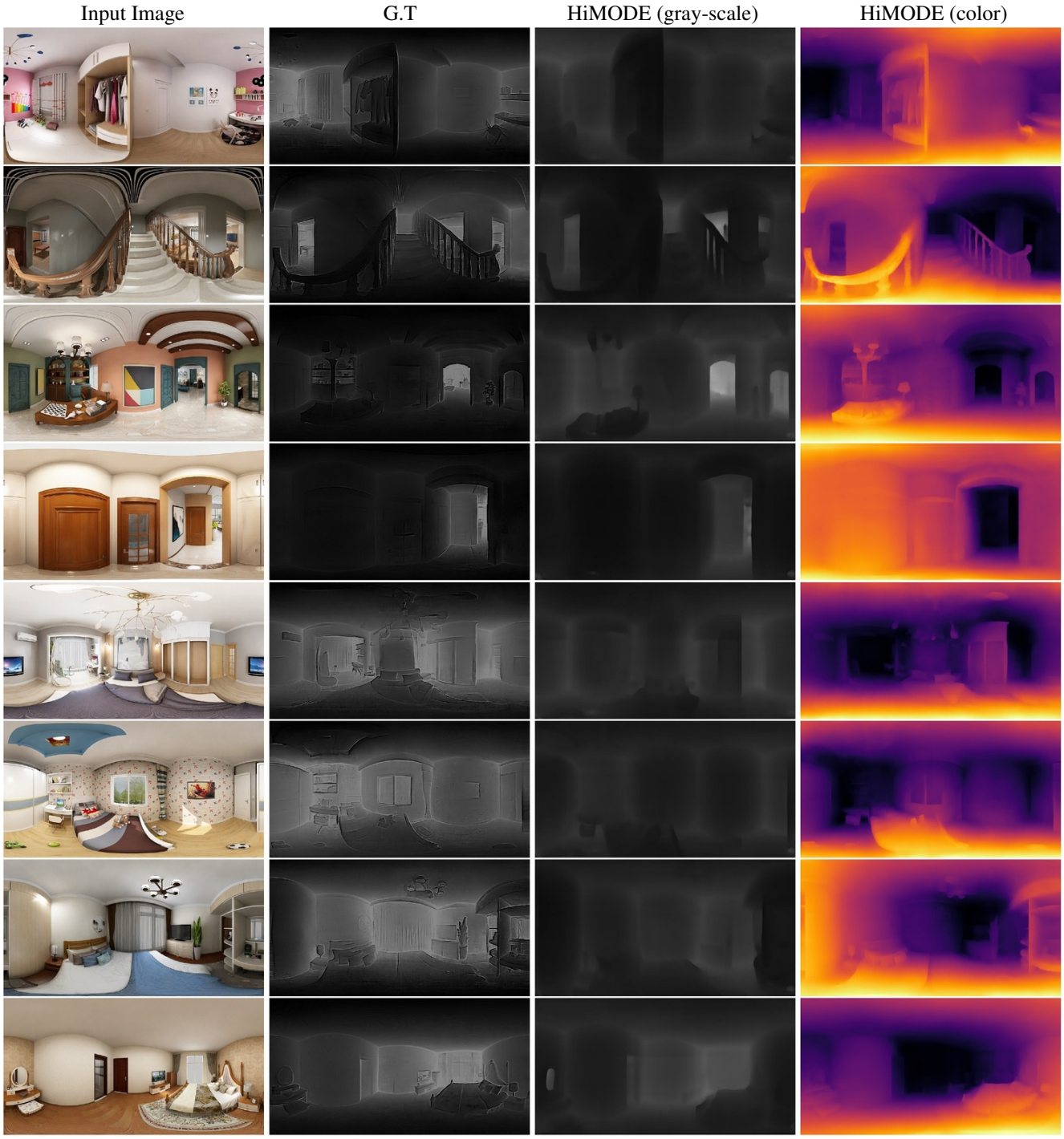

Figure 10. More qualitative results for omnidirectional depth map estimation based on our *HiMODE* on PanoSunCG dataset.

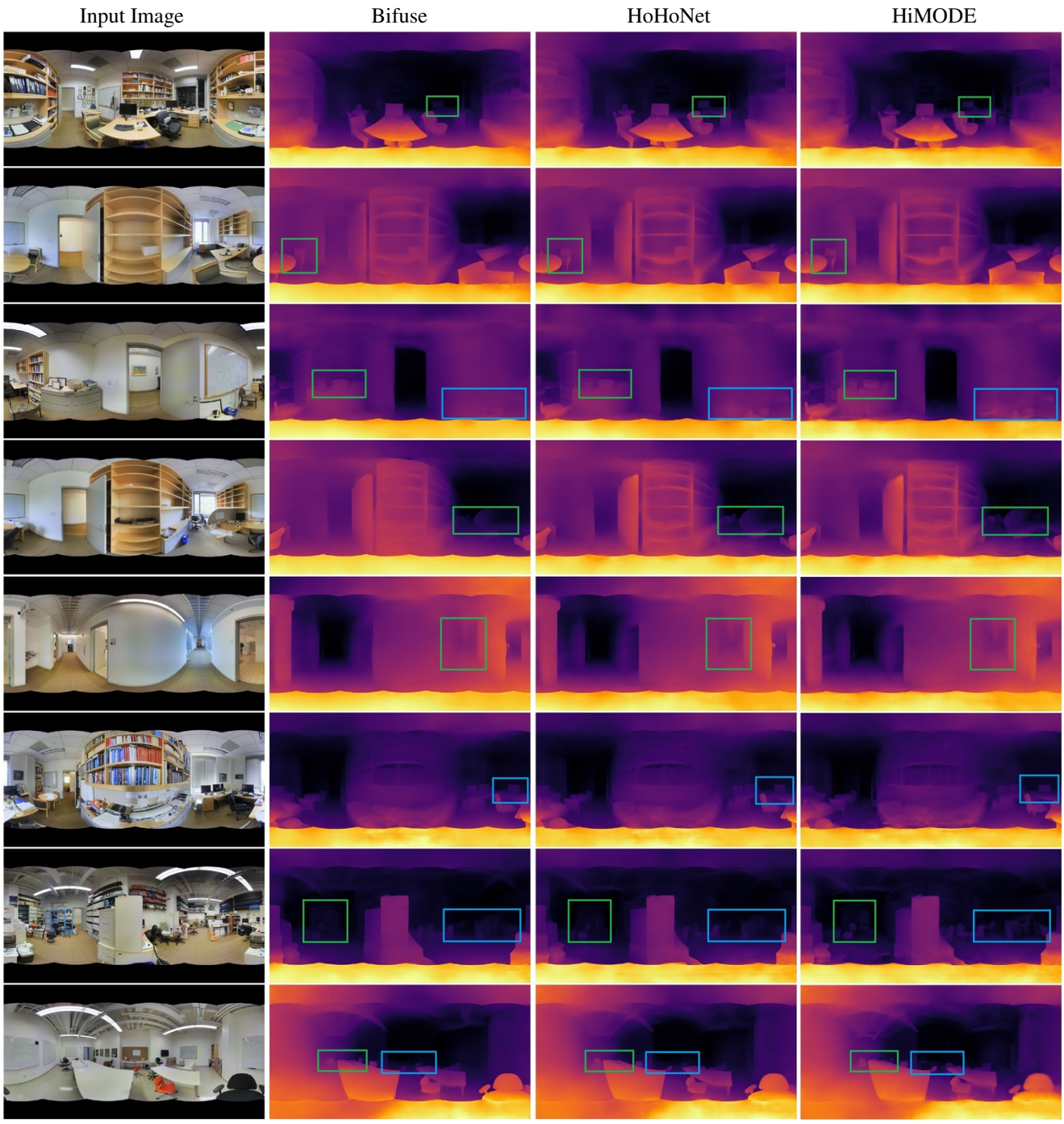

Figure 11. More qualitative comparisons between our *HiMODE* and two recent state-of-the-art approaches, Bifuse [12] and HoHoNet [10] on Stanford3D dataset. As demonstrated by rectangles, our *HiMODE* can accurately recover the details of the surface even for the deep regions with small objects.

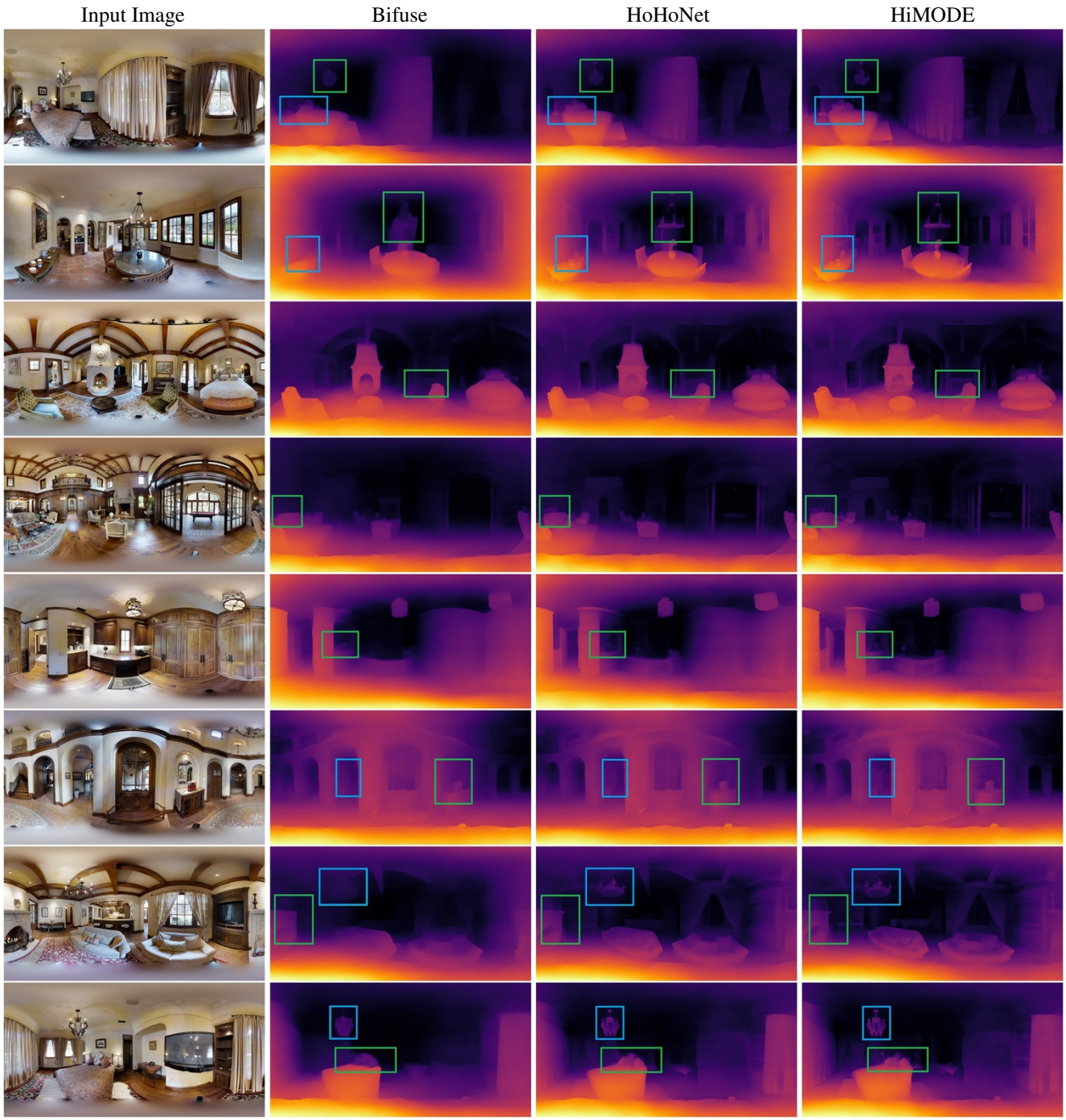

Figure 12. More qualitative comparisons between our *HiMODE* and two recent state-of-the-art approaches, Bifuse [12] and HoHoNet [10] on Matterport3D dataset. As demonstrated by rectangles, our *HiMODE* can accurately recover the details of the surface with sharp edges even for the deep regions and for small objects.

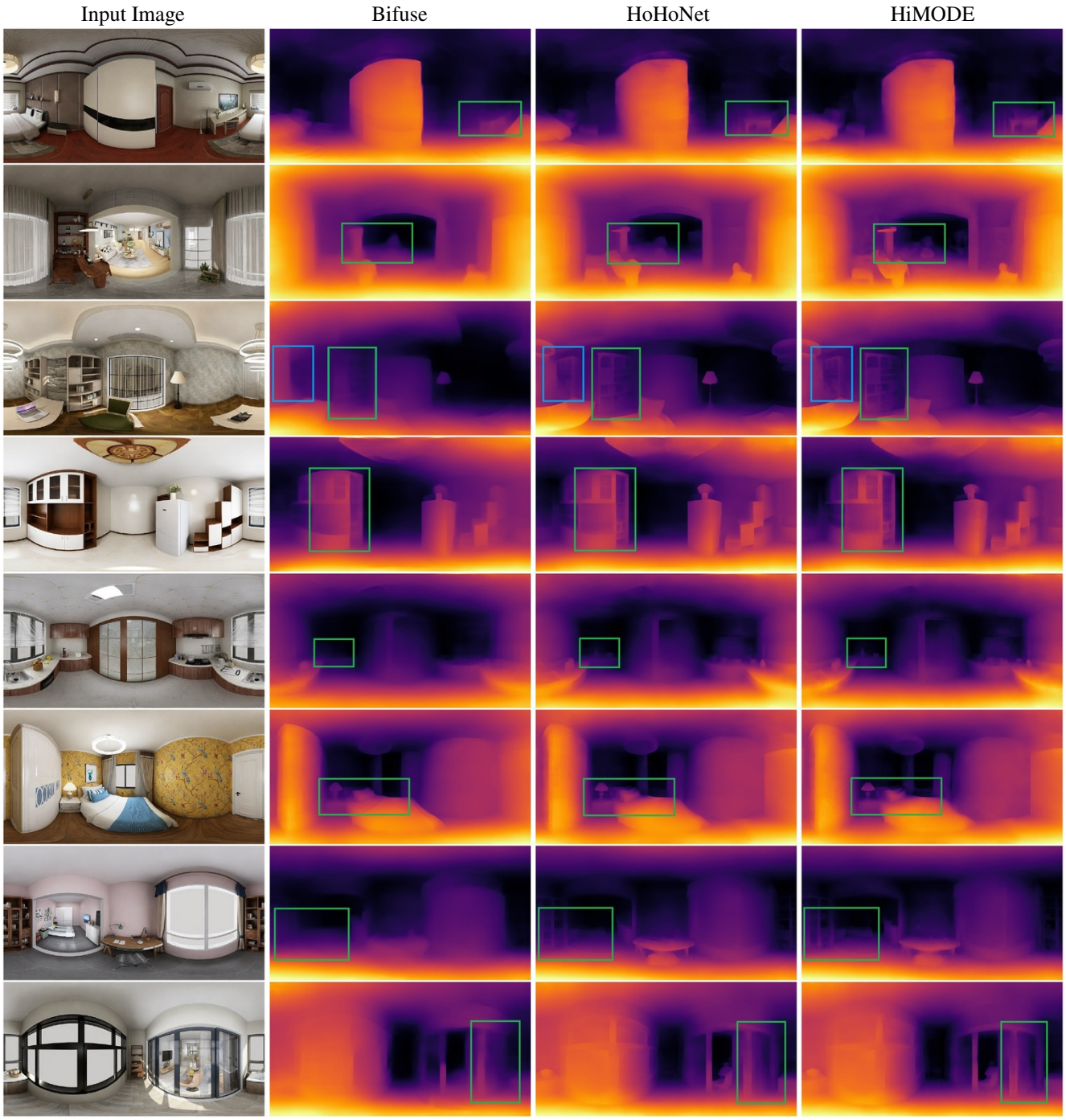

Figure 13. More qualitative comparisons between our *HiMODE* and two recent state-of-the-art approaches, Bifuse [12] and HoHoNet [10] on PanoSunCG dataset. As demonstrated by rectangles, our *HiMODE* can accurately recover the details of the surface with sharp edges even for the deep regions and for small objects.

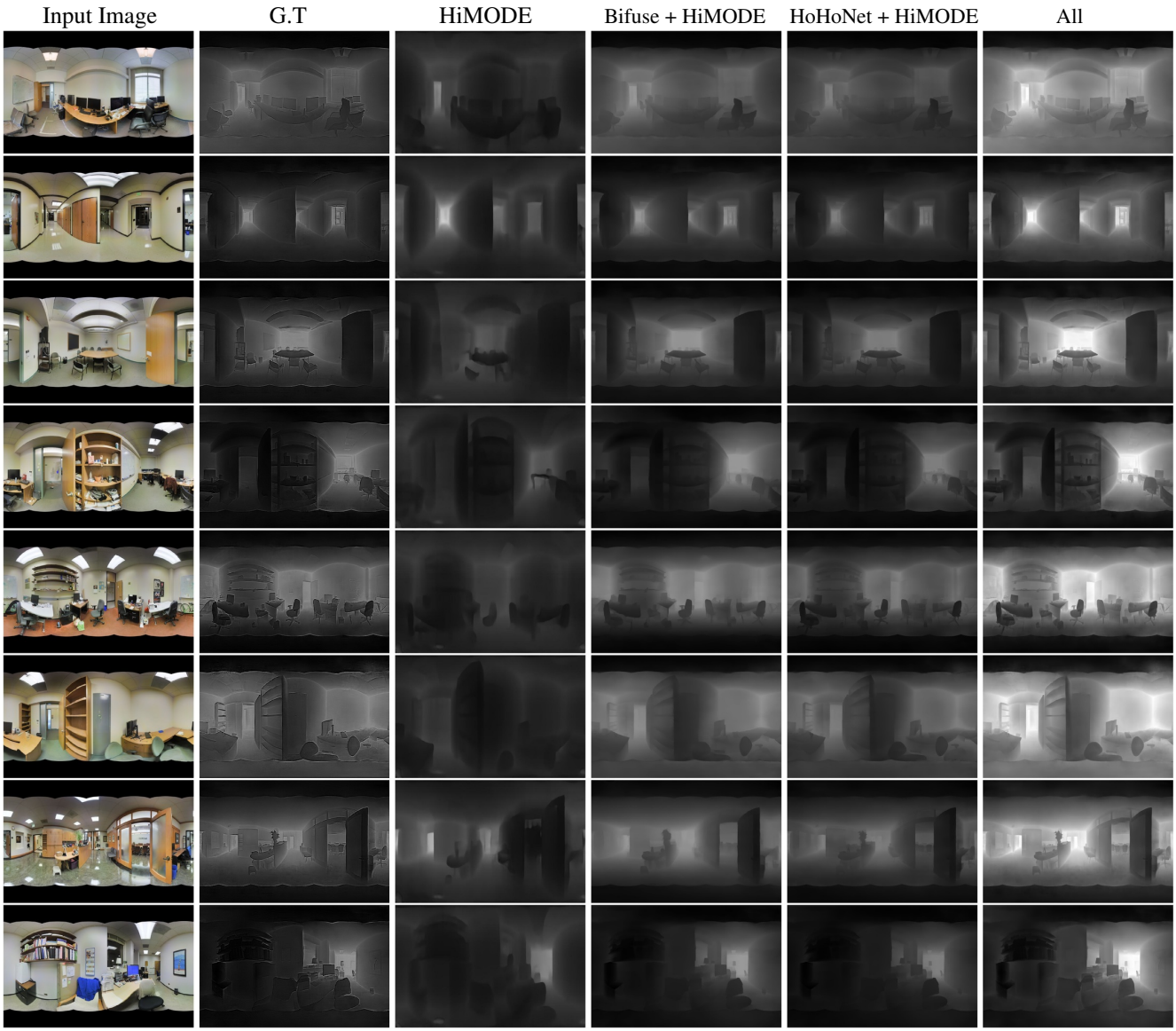

Figure 14. More qualitative results for omnidirectional depth map estimation based on our *HiMODE* along with its combination with two recent state-of-the-art approaches, Bifuse [12] and HoHoNet [10] on Stanford3D dataset.

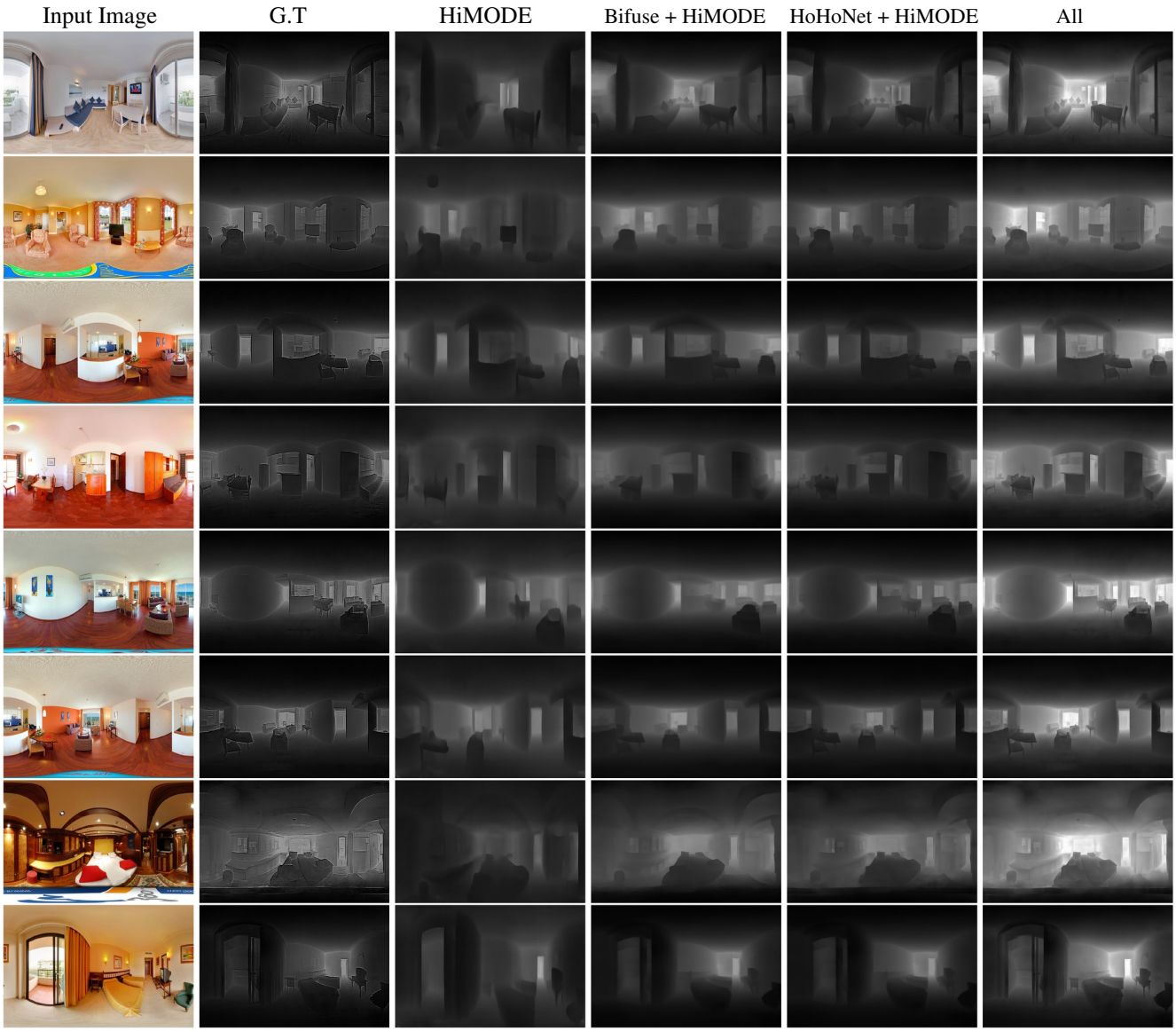

Figure 15. More qualitative results for omnidirectional depth map estimation based on our *HiMODE* along with its combination with two recent state-of-the-art approaches, Bifuse [12] and HoHoNet [10] on Matterport3D dataset.

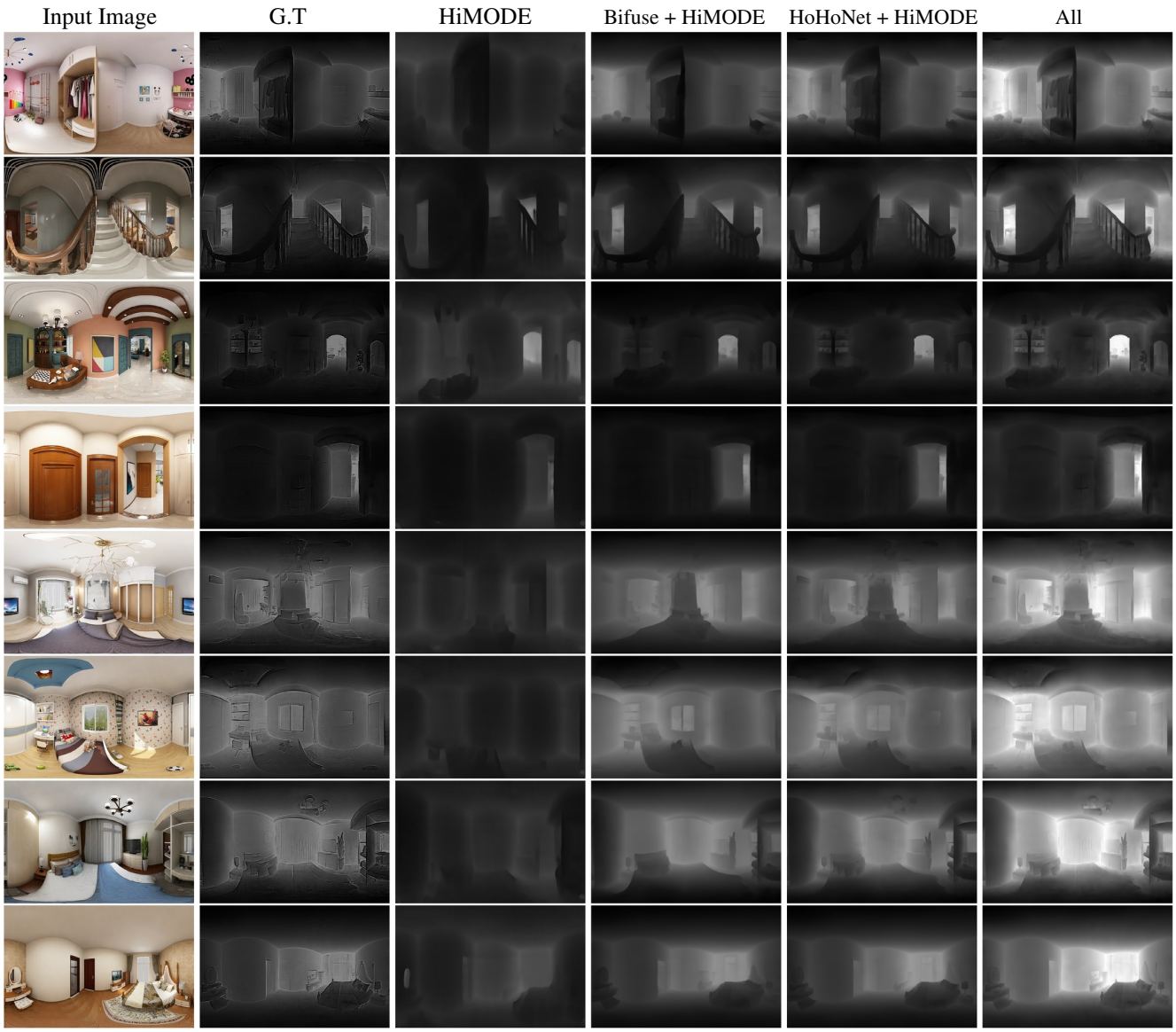

Figure 16. More qualitative results for omnidirectional depth map estimation based on our *HiMODE* along with its combination with two recent state-of-the-art approaches, Bifuse [12] and HoHoNet [10] on PanoSunCG dataset.