HUSH: Holistic Panoramic 3D Scene Understanding using Spherical Harmonics

Jongsung Lee[1], Harin Park[1], Byeong-Uk Lee[2], Kyungdon Joo[1*]
[1]Artificial Intelligence Graduate School, UNIST, [2]KRAFTON
{syniez, harinp33, kyungdon}@unist.ac.kr
byeonguk.lee@krafton.com

Overview

In Sec. 1, we introduce the preliminary about equirectangular projection (ERP). In Sec. 2, we show additional experiment results: qualitative comparison of depth and surface normal estimation and quantitative comparison of surface normal estimation. Additionally, we show ablation studies, including a comparison of performance with respect to the number of SH basis and the task-relevant index maps when using SH bases and learnable parameters as queries in the hierarchical attention module, as discussed in Sec. 3. Finally, we provide an application of the HUSH framework in Sec. 4.

1. Preliminaries

In this section, we introduce the relationship between ERP and spherical coordinates that are equivalent to each other.

Equirectangular projection. ERP is a popular way to represent panorama images that project 3D information from a sphere into a 2D image domain with severe distortions. This kind of projection method makes it easy to interpret the relationship between the panorama image and the sphere. As shown in Fig. 1, we can map 2D image pixel point p to 3D point P from 2D panorama image domain to the 3D space as follows:

$$p = \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \frac{W}{2\pi}\theta \\ \frac{H}{\pi}\phi \end{bmatrix}, \tag{1}$$

$$P = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \rho \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \rho \begin{bmatrix} sin(\phi) \cdot cos(\theta) \\ sin(\phi) \cdot sin(\theta) \\ cos(\phi) \end{bmatrix}, \tag{2}$$

where W and H denote the resolution of the panorama image, ρ is the depth value of 3D point P, ϕ and θ indicate elevation and azimuth angle, respectively. Using Eq. (1), we can represent pixel coordinate $[u, v]^T$ to the unit sphere coordinate $[\theta, \phi]^T$ ($\rho = 1$) according to the image resolution $[H, W]$. After representing the 2D pixel on the unit sphere, we can find the corresponding points p and P' each relies on the panorama image and the unit sphere. Then, we can also match the 3D point P with the points P' and p, because P' is the unit vector of the P (see Eq. (2)). Here, the transformation from the 3D point P to the 2D pixel point p is also available vice versa.

2. Additional Experiments

Additional qualitative results. Here, we provide additional qualitative comparisons with other methods in terms of depth estimation, and surface normal estimation. As shown in Fig. 6, HUSH outperforms previous methods [1, 2], especially on the object boundaries (red boxes). Furthermore, we visualize the predicted depths in 3D space, as illustrated in Fig. 7. From these results, we can observe that HUSH performs more accurately in planar regions than other methods. Figure 8 illustrates the qualitative comparison of normal estimation results between HUSH and Elite360M*. Overall, we can observe that HUSH provides clearer structural details and more distinct object boundaries. Additionally, HUSH maintains better surface normal consistency in plane regions, particularly in the Matterport3D [3] and Structured3D [8] datasets.

[*] Corresponding author.

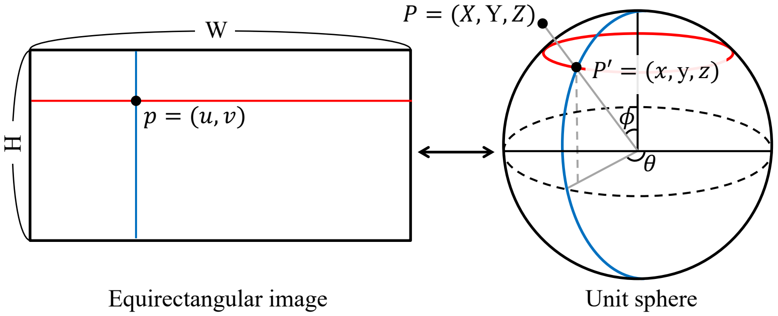

Figure 1. **Transformation between ERP and sphere**.

Quantitative results of surface normal estimation. We compare the surface normal estimation performance of HUSH with Elite360M*. Table 1 shows that HUSH achieves competitive results on three benchmark datasets. Because HUSH is a transformer-based framework, it takes a large number of training datasets to optimize the network and understand the SH basis in terms of the 3D scene understanding. Hence, HUSH shows better performance on the two large-scale datasets: Matterport3D and Structured3D, while Elite360M* shows better performance on the small dataset: Stanford2D3D dataset. Although HUSH performs worse than Elite360M* on some metrics, we observe from Fig. 8 that using the SH query enables a more effective understanding of the geometric information of the scene.

Dataset	Method	RMSE°↓	Mean°↓	Med°↓	11.5°↑	22.5°↑	30°↑
Stanford2D3D	Elite360M*	**20.625**	**10.685**	**3.573**	**0.778**	**0.852**	**0.884**
	Ours (D+N)	20.961	11.191	3.832	0.746	0.838	0.878
Matterport3D	Elite360M*	25.123	15.847	8.132	0.642	0.786	0.837
	Ours (D+N)	**24.863**	**15.035**	**7.065**	**0.663**	**0.797**	**0.847**
Structured3D	Elite360M*	20.621	8.798	**1.079**	**0.850**	0.889	0.907
	Ours (D+N)	**14.959**	**7.014**	1.638	0.834	**0.899**	**0.926**

Table 1. **Quantitative comparison of surface normal estimation on three benchmark datasets**.

3. Ablation Studies

Comparison of SH basis numbers. We conduct an ablation study to validate the robustness of HUSH according to the number of SH bases (*i.e.*, the level of SH basis function). Table 2 shows that the performance of depth estimation remains consistent regardless of the level l. However, the surface normal estimation shows improved performance as the level increases. Furthermore, we analyze which SH bases are primarily utilized for each task, as shown in Fig. 2. From this figure, we can see that each task predominantly relies on different SH bases (*e.g.*, in the case of the Stanford2D3D dataset, depth estimation employs overall SH bases, but surface normal estimation mostly uses SH bases around 20). Based on these observations, we ultimately set the level to 10 to achieve robust performance in both depth and surface normal estimation with the flexibility to expand to other 3D scene understanding tasks.

SH basis level (l)	Depth metric				Normal metric			
	Abs Rel↓	Sq Rel↓	RMSE↓	δ_1↑	RMSE°↓	Mean°↓	Med°↓	11.5°↑
2 (3 SH bases)	0.080	**0.054**	0.338	0.938	20.850	11.506	4.390	0.738
4 (10 SH bases)	0.082	0.055	0.337	0.936	20.972	11.734	4.840	0.734
6 (21 SH bases)	0.079	0.056	0.336	**0.940**	**20.843**	11.423	4.307	0.742
8 (36 SH bases)	**0.076**	0.055	0.337	**0.940**	21.021	11.643	4.462	0.733
10 (55 SH bases)	0.078	0.055	**0.333**	0.938	20.961	**11.191**	**3.832**	**0.746**

Table 2. **Ablation study for the number of SH bases on the Stanford2D3D dataset**.

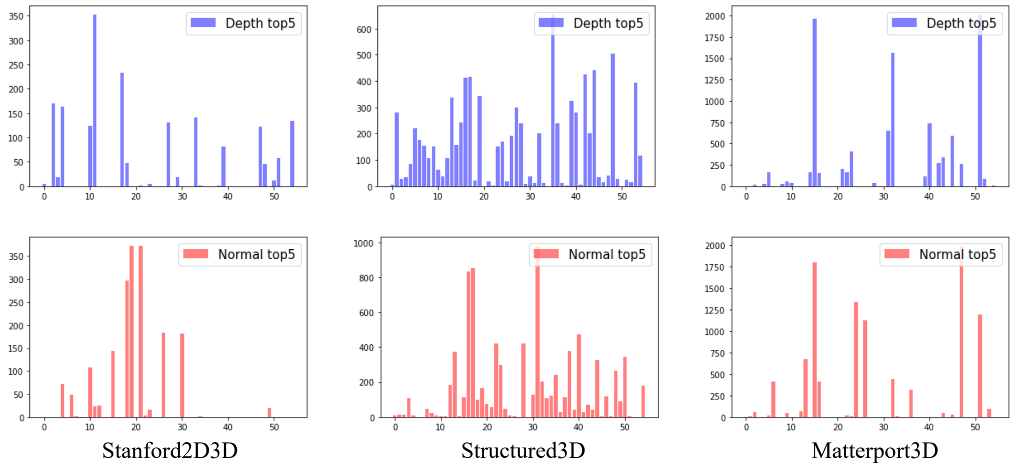

Figure 2. **Dominant basis histogram**. The first and second rows denote top-5 dominant SH basis index histograms on depth and surface normal estimation tasks, respectively. In each graph, the x and y axes indicate the basis index and the frequency of each basis.

SH queries *vs*. learnable queries. We provide additional visual results of the SH index map to demonstrate the impact of SH bases for the panoramic 3D indoor scene understanding task. We compare the SH query-based HUSH framework and learnable (LR) query-based HUSH framework by checking the top-5 dominant SH bases on each task (depth and surface normal estimation), as shown in Fig. 3. We further describe the effectiveness of the SH query on the 3D space in Fig. 4. The SH query-based HUSH framework shows a geometrically better index map than LR query-based HUSH. In particular, SH queries for surface normal estimation can behave like a plane segment module because the points with the same surface normal usually lie in the same plane. In terms of depth estimation, SH queries discretize the scene into several concentric spheres according to the indoor scene scale. *Note that the 2D and 3D index map visualization shares the same SH basis index colormap, which is described in Fig. 3.*

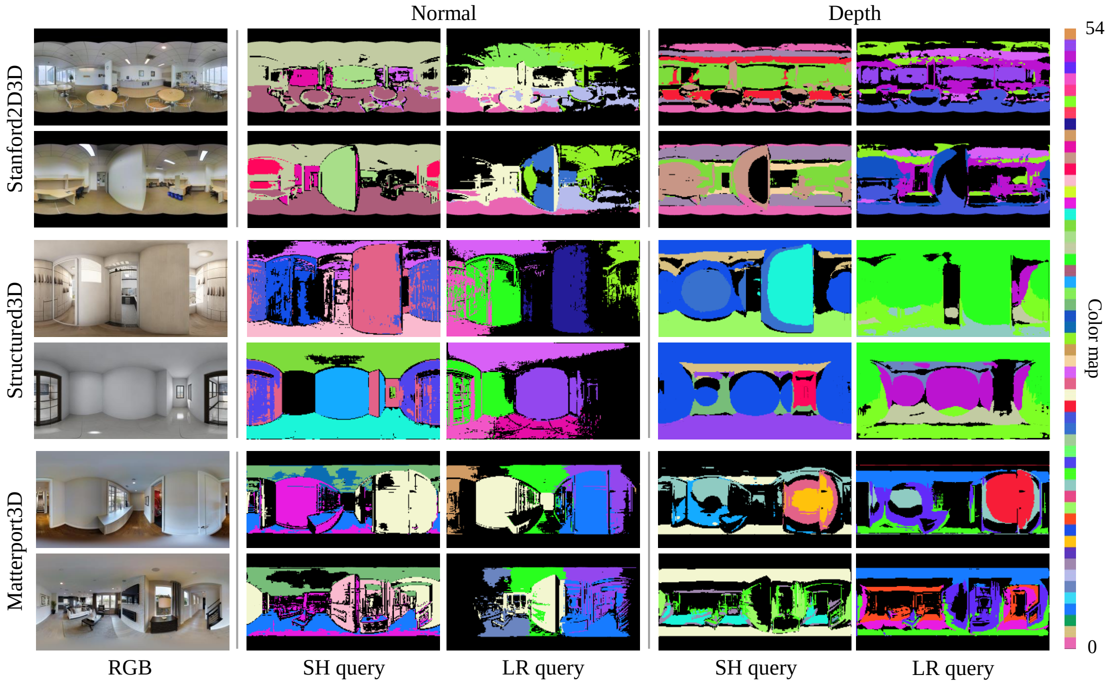

Figure 3. **Task-relevant index map visualization**. We select the top-5 dominant queries from the SH index map for each task. SH query and learnable (LR) query indicate our HUSH framework trained with SH bases and learnable parameters as a query for hierarchical attention module, respectively. We use a different colormap compared to the main paper (right part of the image). The black parts are unselected areas.

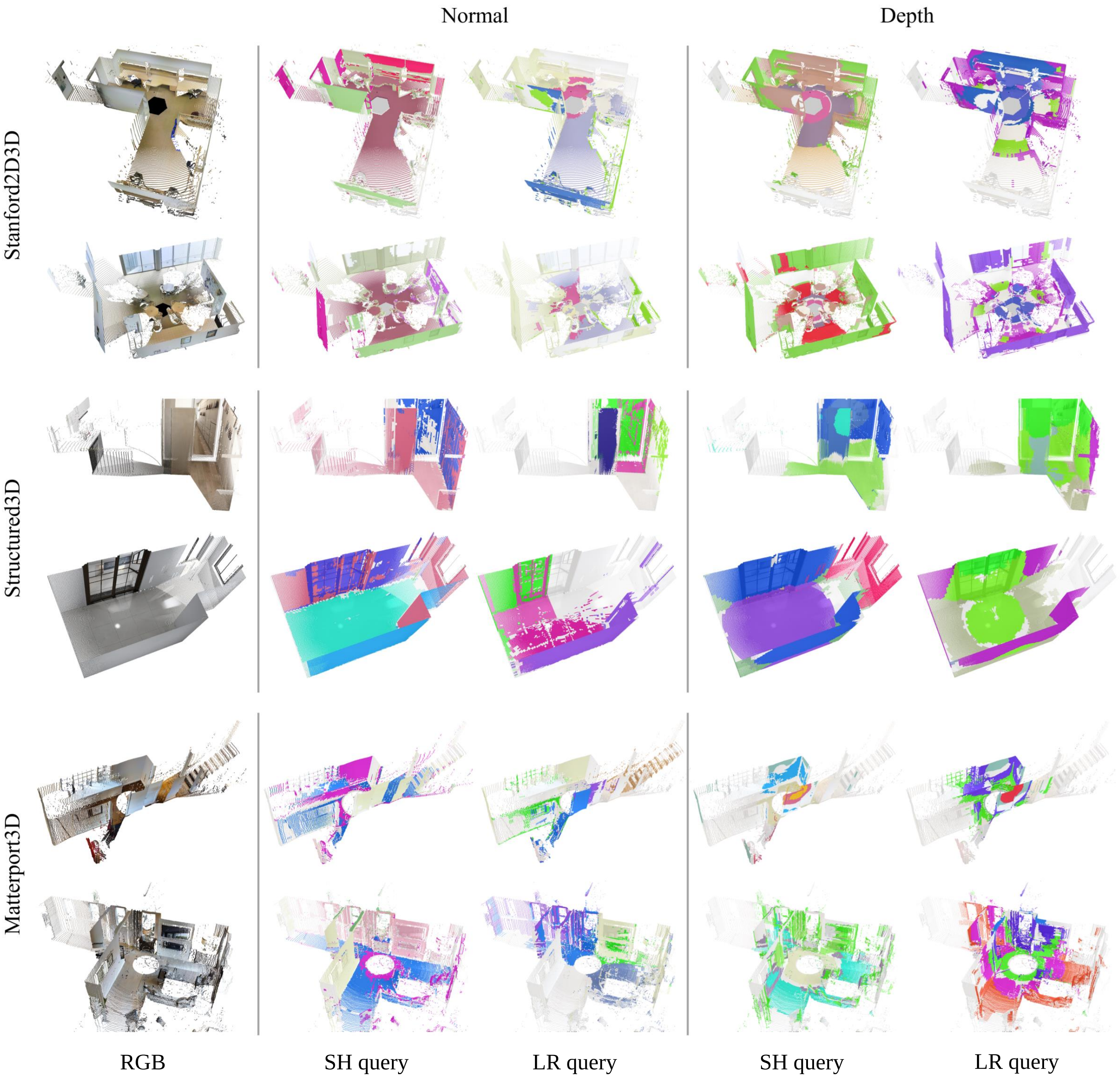

Figure 4. **Task-relevant index map visualization on 3D**. We show the task-relevant index maps on 3D space in the same scene as Fig. 3. For visualization purposes only, we use the ground truth depth map and remove the ceiling area.

4. Application: Layout Estimation

Layout estimation training details. Here, we describe how we train layout estimation tasks to validate our HUSH framework in case of its extensibility. To show the effect of our SH-integrated framework, we simply design a layout head and fine-tune the HUSH framework with the layout head using a pre-trained HUSH. Our layout head consists of a single cross-attention block and six linear layers that estimate the horizontal depth and room height from the depth and surface normals estimated by pre-trained HUSH. We train our HUSH framework for 200 epochs on the MatterportLayout [9] dataset while other recent layout estimation methods [4–7] trained for 1,000 or 2,000 epochs for optimization.

Layout estimation results. As shown in Table 3, our HUSH-based layout estimation framework shows comparable performance even with a few training epochs compared to other layout estimation methods. Although HUSH shows lower performance on 2DIoU and 3DIoU metrics, it achieves the best performance on the RMSE metric by fine-tuning the pre-trained HUSH. Additional qualitative results are shown in Fig. 5.

Method	Epoch	2DIoU(%)	3DIoU(%)	RMSE	δ_1
LED^2-Net [7]	1000	82.61	80.14	0.207	0.947
LGT-Net [4]	1000	83.52	81.11	0.204	**0.951**
DOPNet [5]	2000	84.11	81.70	0.197	0.950
Bi-Layout [6]	2000	**84.56**	**82.05**	-	-
Ours	200	81.72	78.30	**0.196**	0.930

Table 3. **Quantitative comparison of layout estimation on the MatterportLayout dataset**.

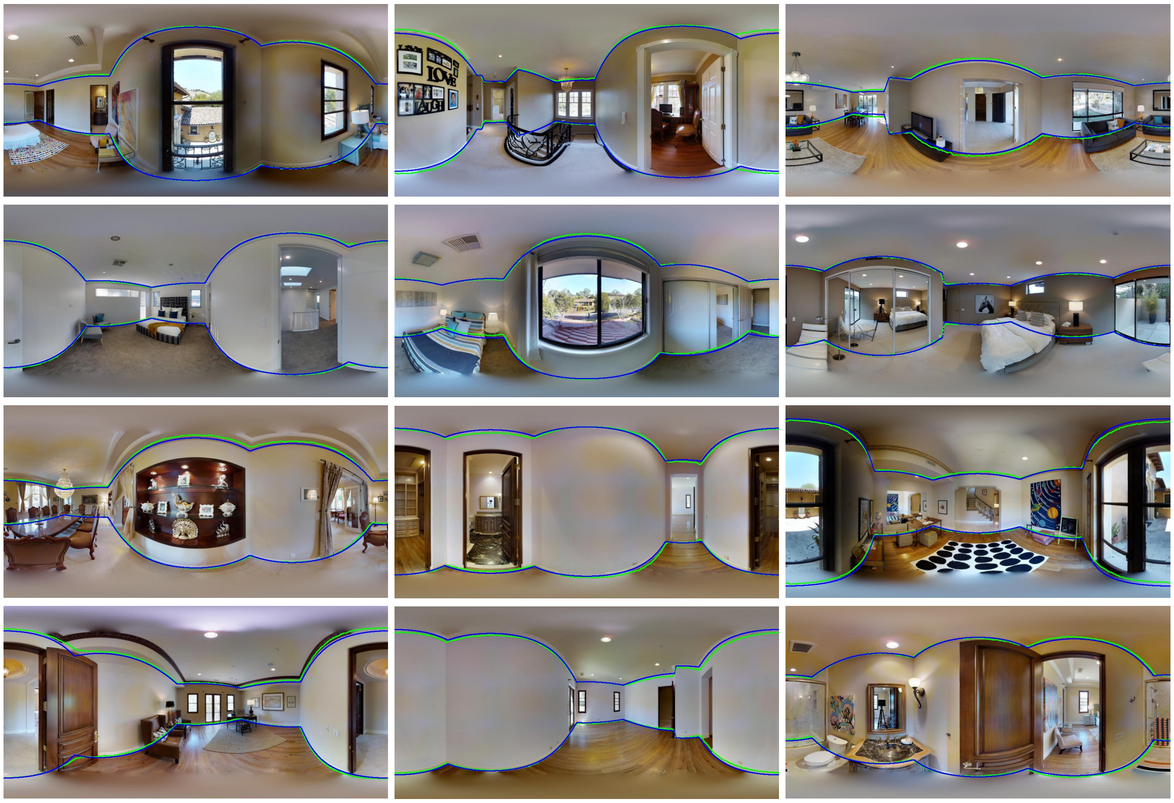

Figure 5. **HUSH framework-based layout estimation**. Blue and green lines represent ground truth labels and our predictions respectively.

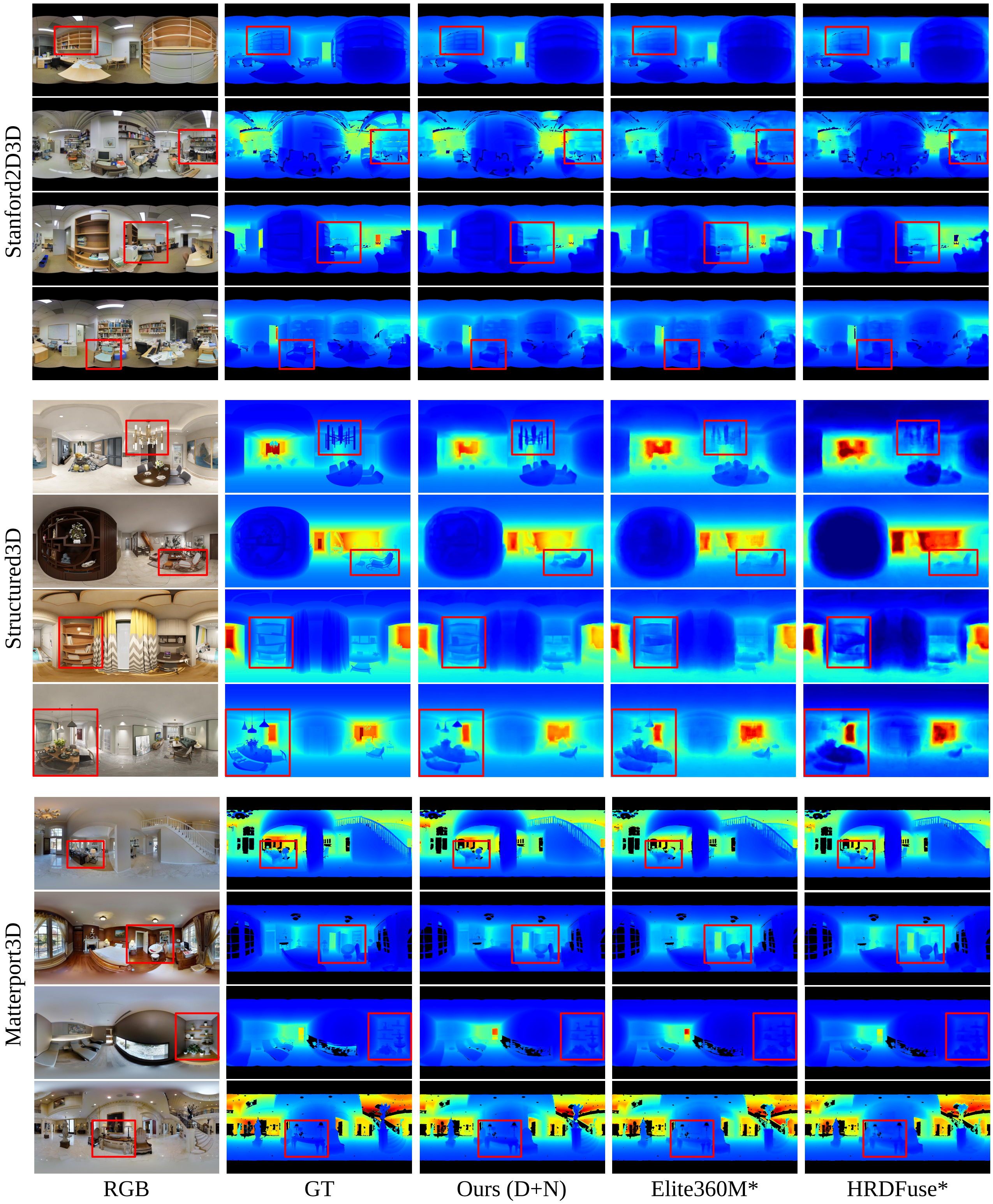

Figure 6. **Qualitative comparison of depth estimation**.

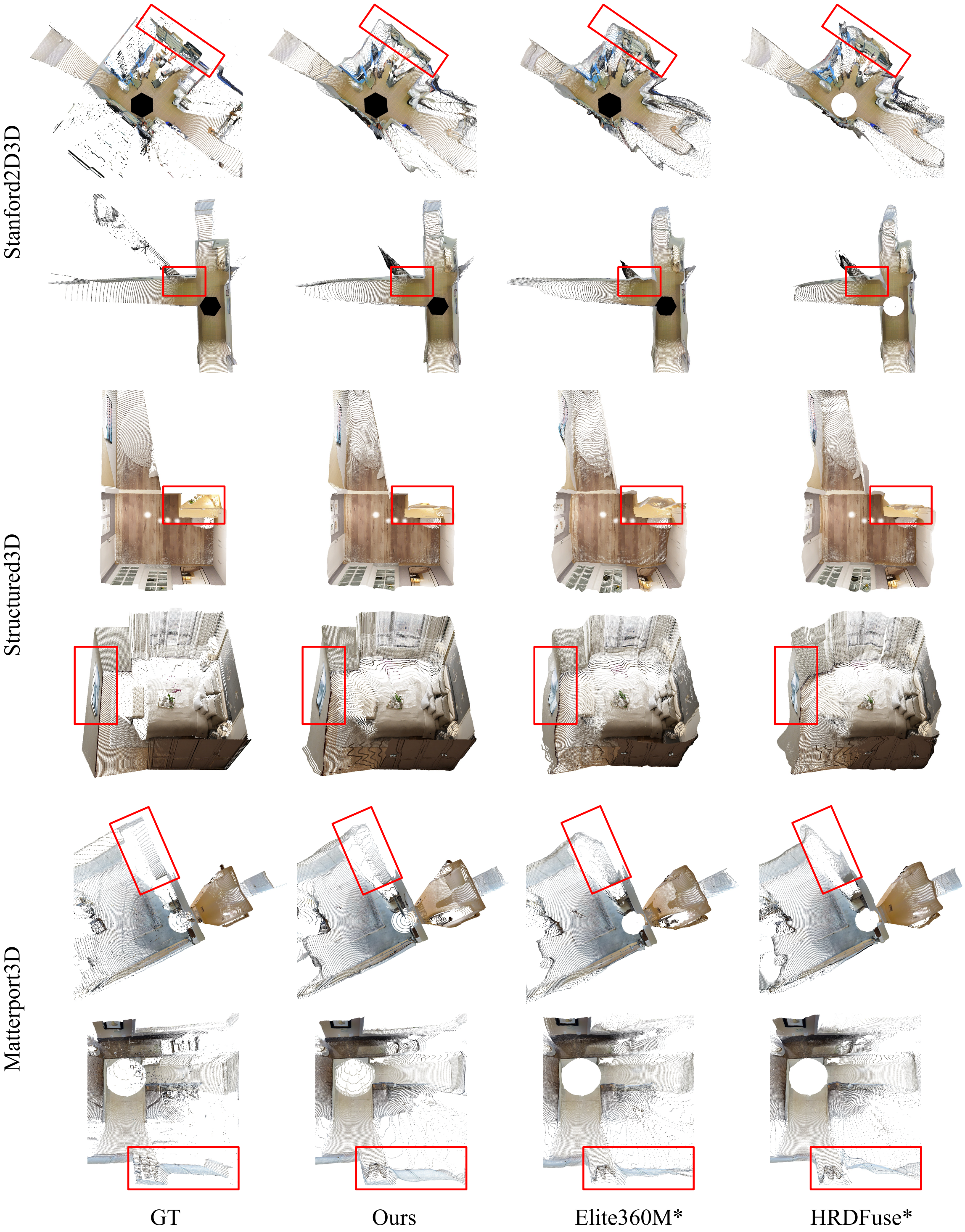

Figure 7. **Qualitative comparison of depth estimation on 3D space.**

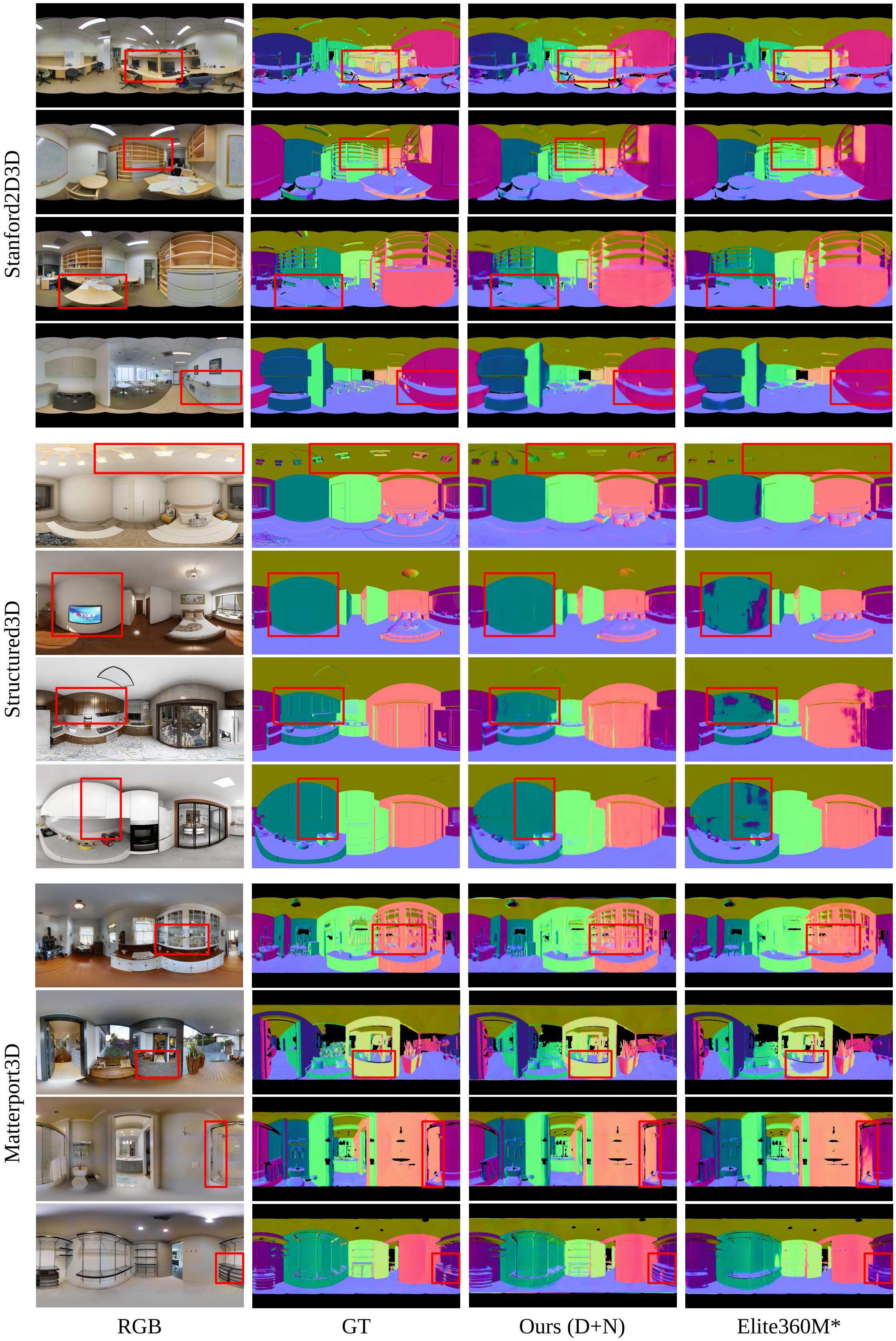

Figure 8. **Qualitative comparison of surface normal estimation**.

References

[1] Hao Ai and Lin Wang. Elite360M: Efficient 360 Multi-task Learning via Bi-projection Fusion and Cross-task Collaboration. *arXiv preprint arXiv:2408.09336*, 2024. 1

[2] Hao Ai, Zidong Cao, Yan-Pei Cao, Ying Shan, and Lin Wang. HRDFuse: Monocular 360deg Depth Estimation by Collaboratively Learning Holistic-With-Regional Depth Distributions. In *CVPR*, 2023. 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. 1

[4] Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *CVPR*, 2022. 5

[5] Zhijie Shen, Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Shuai Zheng, and Yao Zhao. Disentangling orthogonal planes for indoor panoramic room layout estimation with cross-scale distortion awareness. In *CVPR*, 2023. 5

[6] Yu-Ju Tsai, Jin-Cheng Jhang, Jingjing Zheng, Wei Wang, Albert YC Chen, Min Sun, Cheng-Hao Kuo, and Ming-Hsuan Yang. No more ambiguity in 360deg room layout via bi-layout estimation. In *CVPR*, 2024. 5

[7] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Led2-net: Monocular 360deg layout estimation via differentiable depth rendering. In *CVPR*, 2021. 5

[8] 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

[9] Chuhang Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung-Kuo Chu, and Derek Hoiem. Manhattan room layout reconstruction from a single 360 image: A comparative study of state-of-the-art methods. *IJCV*, 2021. 5