MCPDepth: Practical Omnidirectional Depth Estimation from Multiple Cylindrical Panoramas via Stereo Matching

Supplementary Material

1. Framework

1.1. Stereo Matching

Fig. 2 illustrates the architecture of the stereo matching network, incorporating the circular attention module. Meanwhile, Fig. 1 compares the attention heatmaps before and after applying circular attention. The circular attention module enhances focus on key regions, particularly the edges of objects and textured areas where rapid disparity changes occur. In contrast, it pays less attention to objects' backgrounds and central areas, where disparities tend to be smooth and consistent. This is because our circular attention module is capable of capturing the 360° feature and assigning similar weights to the same object.

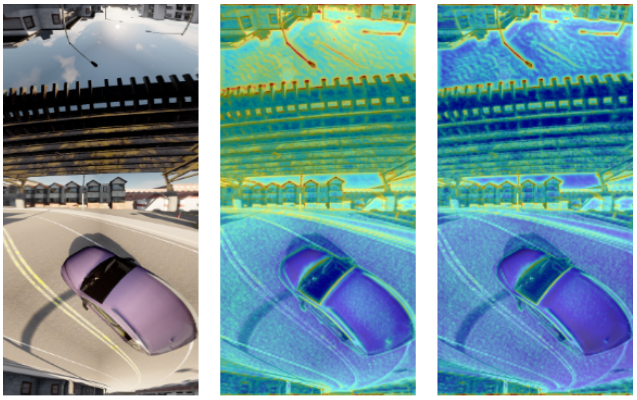

Figure 1. Comparison of heatmaps before and after circular attention.

1.2. Depth Fusion

We adopt the depth fusion architecture from MODE [1], which integrates depth maps, confidence maps, and reference panoramas for robust depth estimation. For the Deep360 dataset, 6 depth maps, 6 confidence maps, and 4 panoramas are fed into the fusion model, while for the 3D60 dataset, 3 depth maps, 3 confidence maps, and 3 panoramas are used. The architecture for Deep360 is depicted in Fig. 3.

2. Training Details

During the stereo matching stage, we use 2 NVIDIA A40 GPUs to train our models with a batch size of 4 for the Deep360 dataset. Training takes 158 hours; For 3D60, we use a single NVIDIA A6000 GPU with batch size 4, which takes 252 hours to train. The model is trained for 45 epochs with a learning rate of 0.001, followed by a decay of the learning rate to 0.0001 for an additional 10 epochs. In the depth fusion stage, we train the network for 150 epochs with a learning rate of 0.0001.

3. Visualizations

Fig. 4 and Fig. 5 show the performance of stereo matching on Deep360 and 3D60 test datasets. Our model demonstrates a strong ability to distinguish foreground objects from the background, even in regions with significant distortion. Additionally, Fig. 7 illustrates the depth estimation performance on the 3D60 test dataset. By leveraging more accurate disparity estimation, our method surpasses MODE in depth estimation performance, excelling even in areas where ground truth data is unavailable.

Fig. 6 compares the stereo matching and depth estimation performance of our model with MODE. Our model outperforms MODE in both stereo matching and depth estimation tasks, achieving higher accuracy and consistency across the entire dataset.

4. Attention Mechanism for ERP

We compare with EGFormer, which is designed for ERP. Tab. 1 demonstrates that EGFormer hurts cylindrical feature extraction. Our circular attention designed for cylindrical panorama also improves the extraction of spherical panoramas and cubic panoramas, demonstrating the generalization of our module.

Table 1. Comparison with EGFormer

Projection	Method	MAE	Px1 (%)	D1 (%)
Cylindrical	Baseline	0.2179	2.6489	1.0236
	EGFormer	0.5697	10.0792	3.7135
	Ours	**0.2112**	**2.5713**	**0.9828**

5. Inference Time of Stereo Matching Models

The inference time of different stereo matching models on Deep360 and 3D60 datasets are shown in Tab. 2. The models are tested in both PyTorch and ONNX formats on NVIDIA GeForce RTX 3090. The panoramas have dimensions of $H \times W = 1024 \times 512$ on Deep360 and $H \times W = 512 \times 256$ on 3D60 for both cylindrical and spherical projections. For Deep360, the maximum disparity is set to 272 for cylindrical projection and 192 for spherical projection, while on 3D60, the maximum disparity is set to 256 for all projections. The circular attention module exhibits inference times of 18.22 ms for Deep360 and 2.81

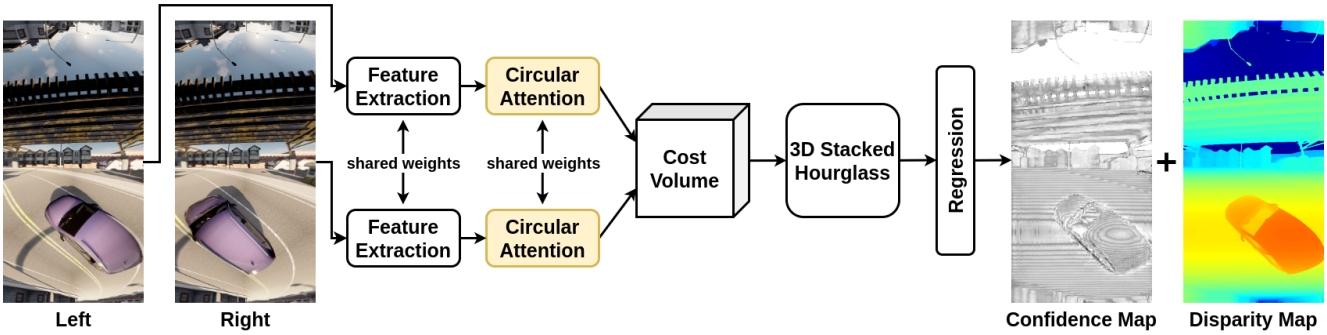

Figure 2. The architecture of the stereo matching network.

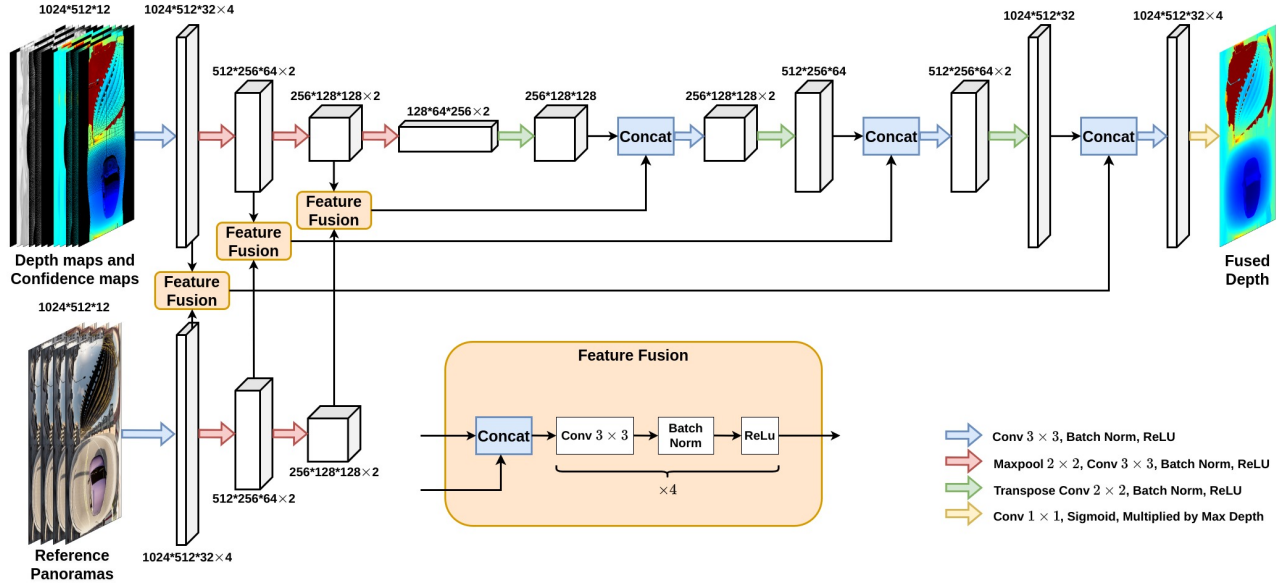

Figure 3. Depth Fusion Architecture.

ms for 3D60, demonstrating its efficiency across varying dataset resolutions and disparity ranges.

Our stereo matching model, even without the circular attention module, surpasses MODE in stereo matching performance. Moreover, it demonstrates a slightly shorter inference time than MODE when processing the same maximum disparity range, as shown in Table 2 for the 3D60 dataset. While our model is marginally slower than MODE in PyTorch format, it offers substantial benefits when exported to ONNX format. Specifically, our model in ONNX format achieves significantly faster inference times compared to MODE in PyTorch format when handling the same maximum disparity range.

Table 2. Inference time of stereo matching model. "CA" denotes Circular Attention.

Dataset	Methods	Projection	PyTorch (ms)	ONNX (ms)
Deep360 [1]	MODE [1]	Cassini	200.62	-
	Ours w/o CA	Cylindrical	238.61	197.63
	Ours	Cylindrical	266.27	226.30
3D60 [2]	MODE [1]	Cassini	66.74	-
	Ours w/o CA	Cylindrical	66.21	46.11
	Ours	Cylindrical	69.12	51.24

References

[1] Ming Li, Xueqian Jin, Xuejiao Hu, Jingzhao Dai, Sidan Du, and Yang Li. Mode: Multi-view omnidirectional depth estimation with 360○ cameras. In *European Conference on Computer Vision*, pages 197–213. Springer, 2022. 1, 2

[2] 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. 2

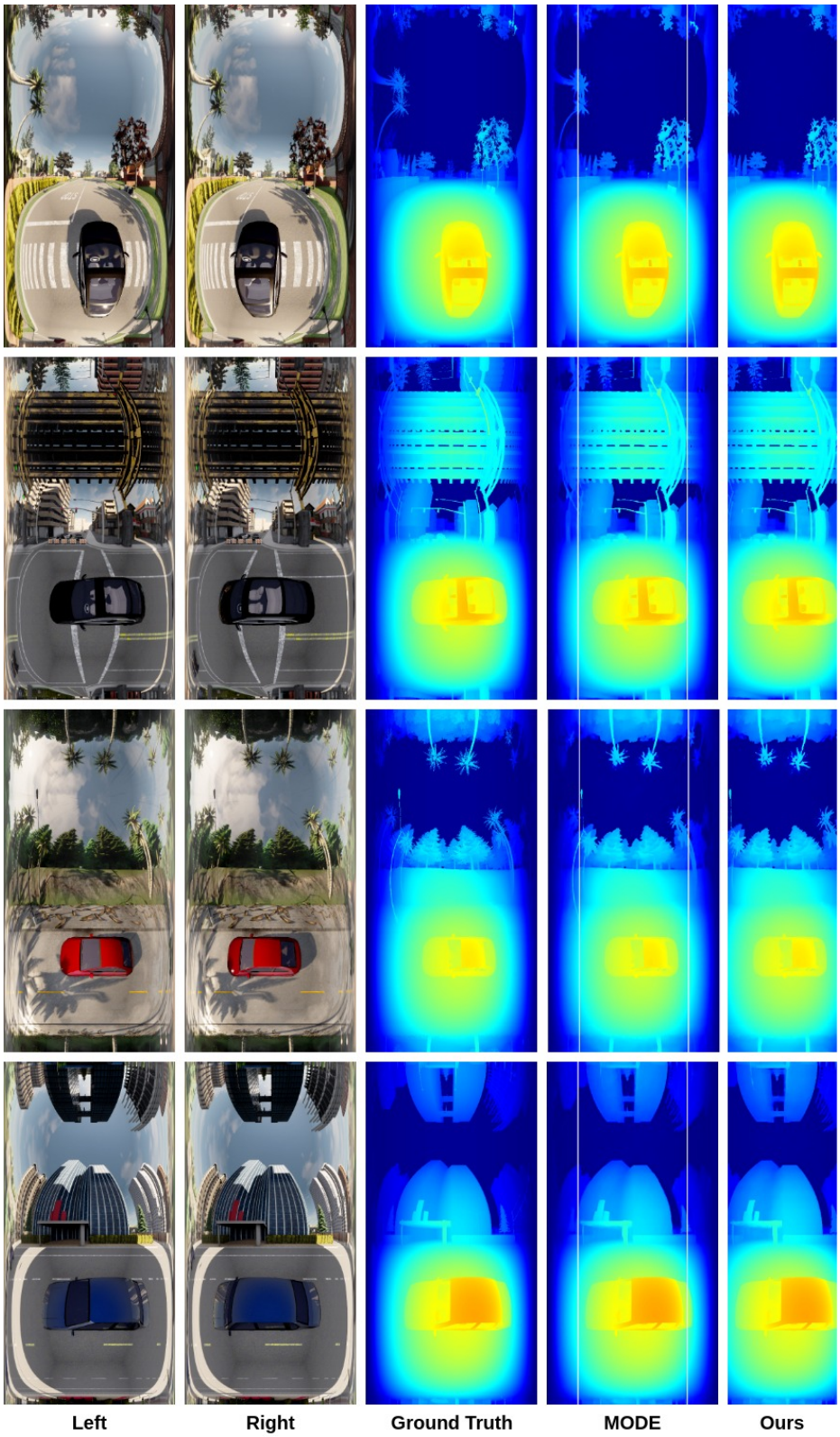

Figure 4. Disparity estimation results on the Deep360 test dataset compared to MODE.

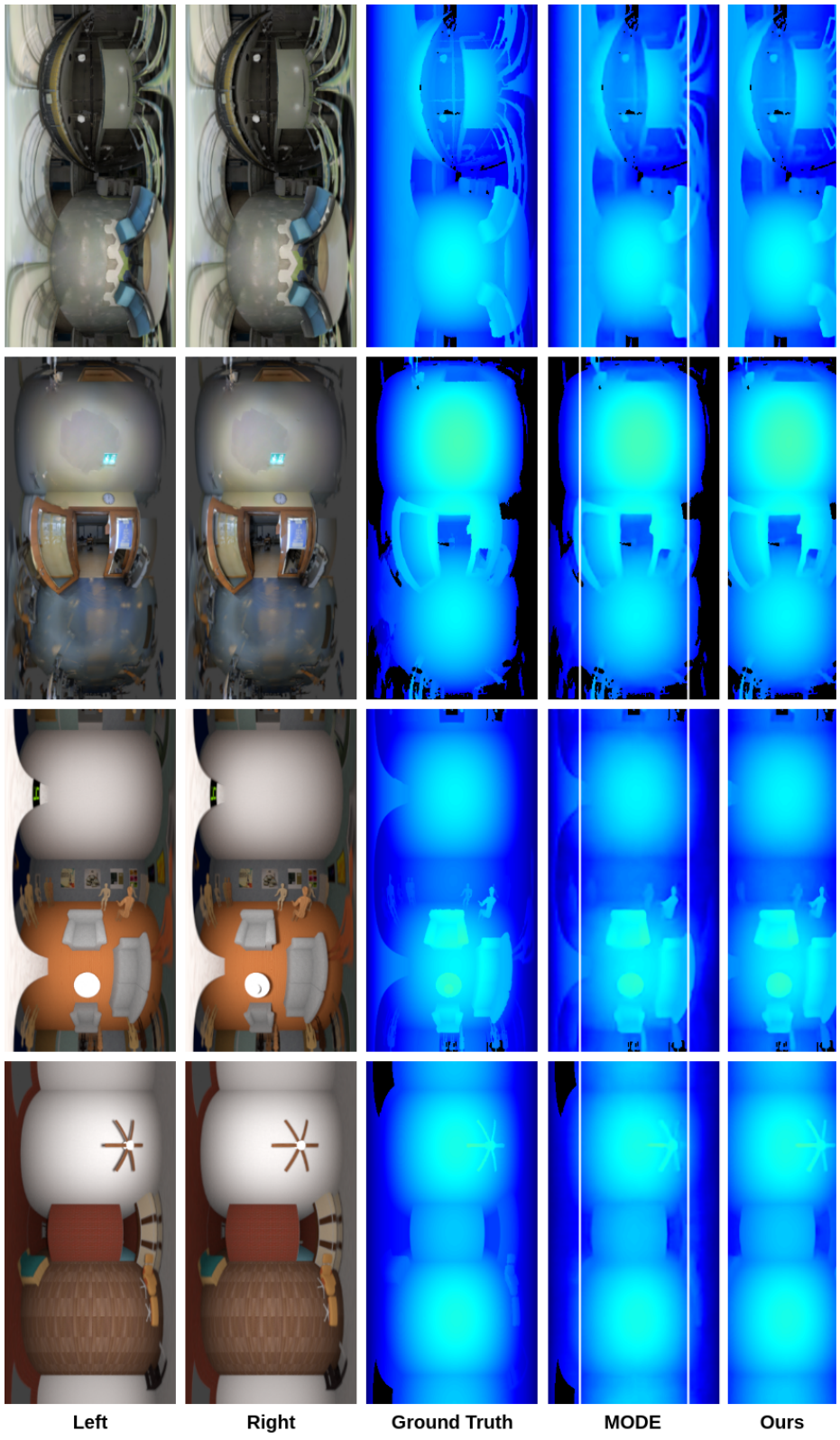

Figure 5. Disparity estimation results on the 3D60 test dataset compared to MODE.

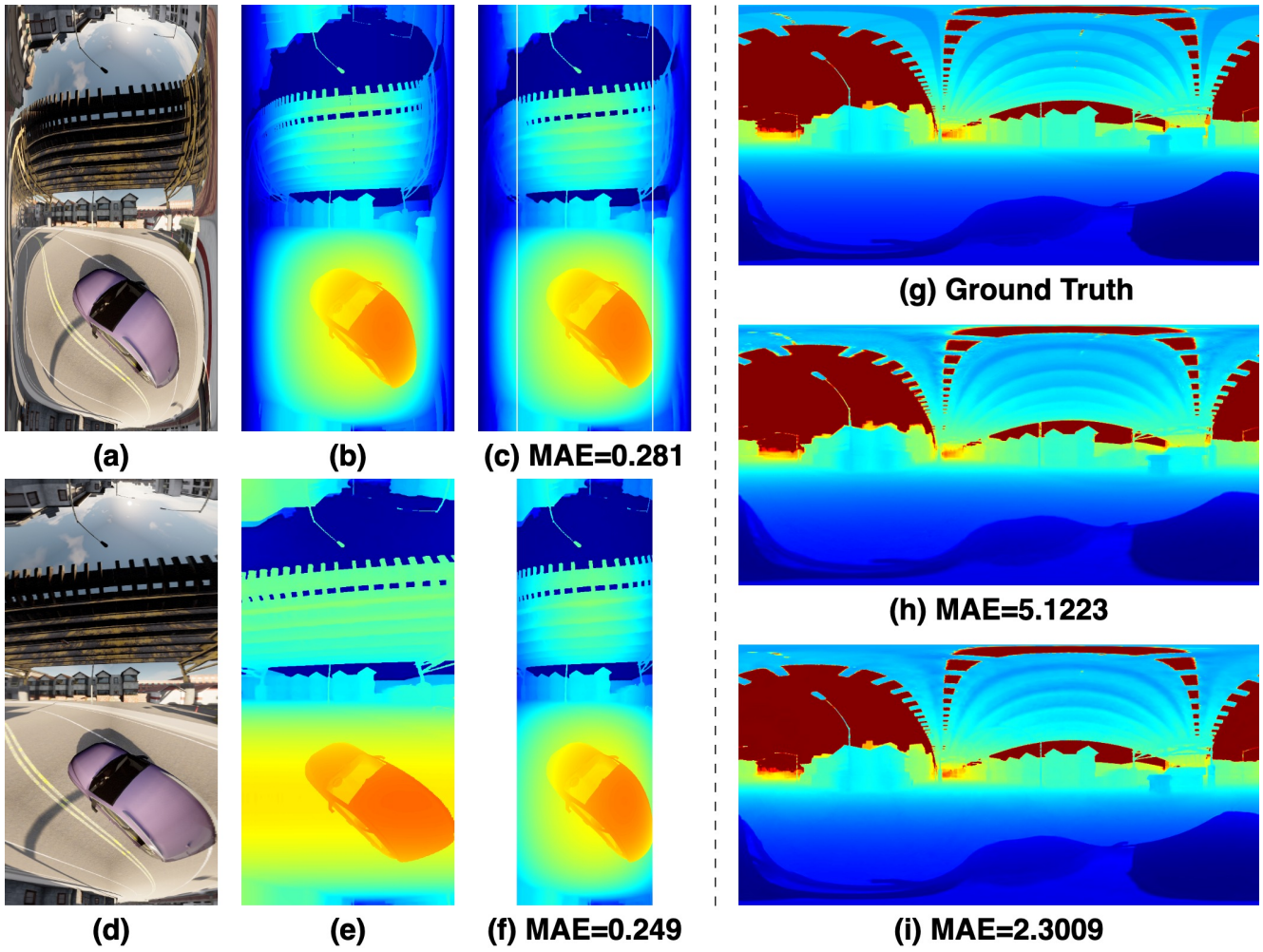

Figure 6. Qualitative comparison with MODE for stereo matching and depth estimation. (a) shows the left panorama in Cassini projection, (b) the ground truth disparity, and (c) the estimated disparity from MODE. (d) shows the left panorama in cylindrical projection, while (e), (f) depict the estimated disparity in cylindrical and Cassini projections from ours. (g), (h), (i) are the ground truth depth map, the estimation from MODE, and our estimated depth map, respectively.

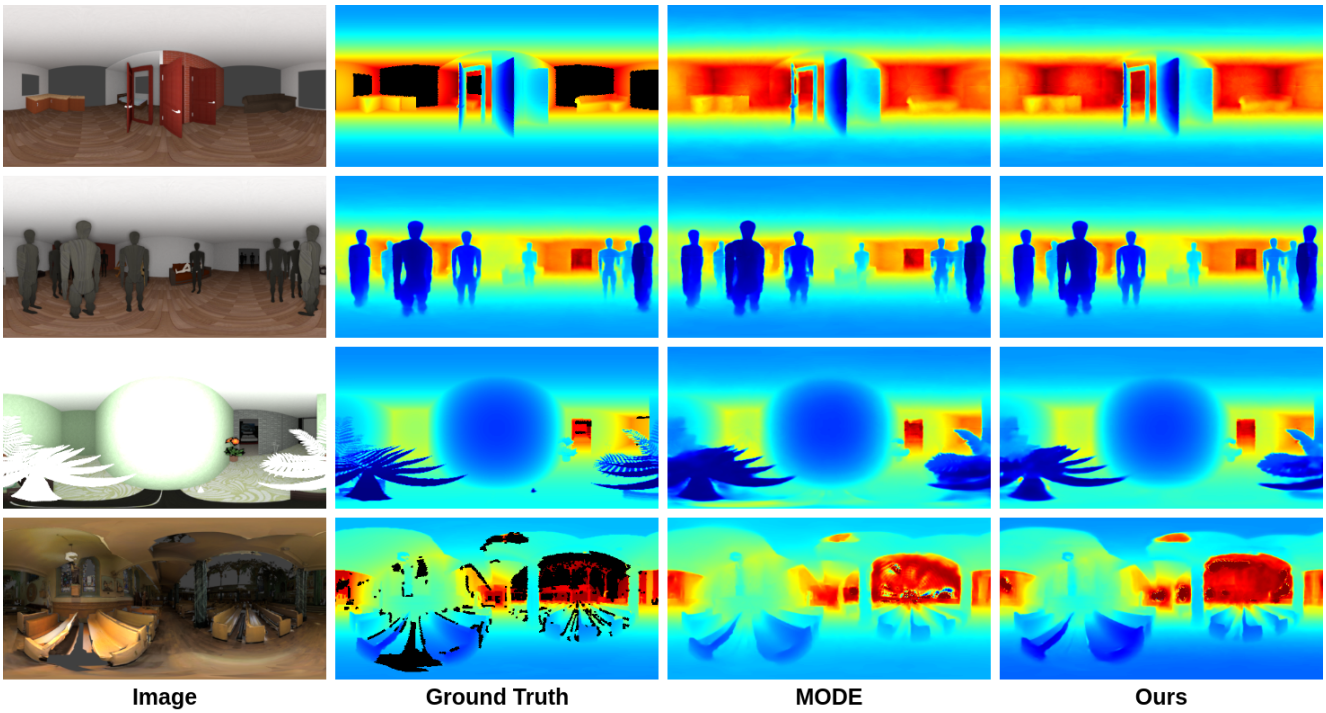

Figure 7. Qualitative comparisons of omnidirectional depth estimation methods on 3D60 compared with MODE.