OmniStereo: Real-time Omnidireactional Depth Estimation with Multiview Fisheye Cameras

Supplementary Material

6. Comparison with OmniVidar

There are two key reasons why we conducted a separate comparison with OmniVidar [30]: (1) evaluation protocols: OmniVidar [30] computes errors using the index of depth candidates derived from spherical sweeping [27], while OmniStereo evaluates errors using depth values. This discrepancy in evaluation protocols prevents a direct comparison under a unified metric. Moreover, OmniVidar's code is not available, further complicating direct comparisons. (2) evaluation regions: OmniVidar [30] generates incomplete depth maps in the vertical direction, whereas OmniStereo provides depth estimations over a wider FOV, as shown in Fig. 7. Consequently, ensuring that all methods are evaluated over identical regions is challenging, making the comparison less rigorous. Despite these limitations, we included experiments with OmniVidar [30] to provide a more comprehensive evaluation of OmniStereo performance.

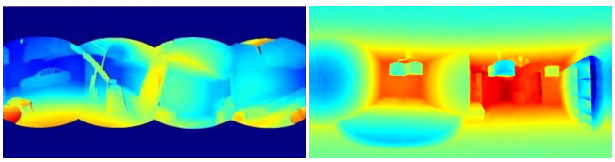

Figure 7. **Comparison with OmniVidar [30].** The depth map estimated by OmniVidar [30] is obtained from its paper. However, since the specific scene used in the paper is not detailed, the depth map used for comparison may not correspond to the same scene. Compared to OmniVidar [30], OmniStereo provides a wider vertical field of view for omnidirectional depth estimation, which is crucial for improving safety in real-world applications.

As OmniVidar code is not available, we rely on the data reported in its paper [30] for comparison. To ensure consistency, our experimental setup aligns with OmniVidar [30] evaluation protocol. Qualitative comparison results on the OmniHouse and OmniThings [26] are presented in Tab. 5.

As shown in Tab. 5, OmniStereo outperforms OmniVidar [30] in accuracy on both OmniHouse and OmniThing datasets [26]. In terms of efficiency, OmniVidar [30] achieves an inference latency of 66ms on RTX 2080 Ti GPU with 13.45 TFLOPS [30], while OmniStereo achieves a latency of 28.2ms on TITAN RTX with 16.31 TFLOPS, as detailed in Tab. 1. Despite only a 17.5% increase in computational power, OmniStereo reduces inference latency by 57.2%, demonstrating its superior efficiency over OmniVidar [26].

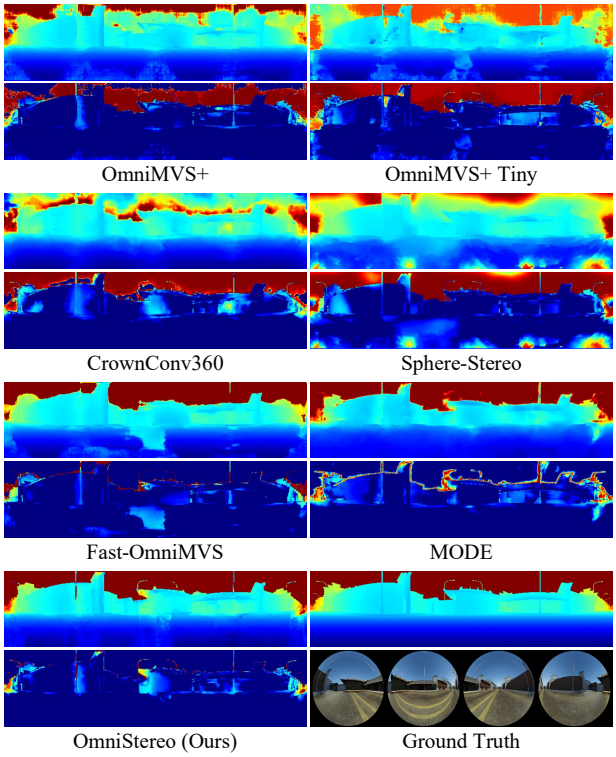

Figure 8. **Qualitative results for generalization on the unseen OmniTown dataset [25].** The red and yellow areas in error maps highlight the inaccurate estimation.

7. Qualitative comparison with SOTA methods

We performed qualitative comparisons with SOTA methods [11, 15, 19, 25, 28] on the Urban [27], OmniHouse [26] and OmniThings [26] datasets, as shown in Fig. 9. The results demonstrate that OmniStereo achieves the best performance across all three datasets, producing depth estimates closest to the ground truth. In the error maps, OmniStereo exhibits the smallest discrepancies and the fewest number of error-prone regions, affirming its SOTA accuracy.

Additionally, We conducted a generalization comparison against SOTA methods [11, 15, 19, 25, 28] on the unseen OmniTown dataset [25]. As shown in Fig. 8, OmniStereo produces the most accurate depth estimates, particularly in key features such as buildings, the sky, and the ground. These findings underscore the strong generalization capability of OmniStereo.

Dataset	OmniHouse [26]			OmniThings [26]		
Method	**bad 1.0(↓)**	**RMSE(↓)**	**MAE(↓)**	**bad 1.0(↓)**	**RMSE(↓)**	**MAE(↓)**
OmniVidar [30]	**4.886**	1.002	1.042	**11.199**	**0.958**	**0.461**
OmniMVS+ [28]	26.357	0.929	0.705	59.753	2.270	1.723
CrownConv360 [11]	85.559	11.077	6.729	93.946	9.545	6.518
Sphere-Stereo [19]	11.304	**0.671**	**0.410**	65.348	2.550	1.941
OmniStereo (Ours)	**0.647** (4.239)	**0.151** (0.52)	**0.061** (0.349)	**11.017** (0.182)	**0.818** (0.14)	**0.405** (0.056)

Table 5. **Quantitative comparison with OmniVidar [30].** In the table, Green indicates the method with the highest accuracy. **Bold** text denotes the second-highest accuracy. **Red** and **Blue** represent improvements and reductions relative to the second-ranked method.

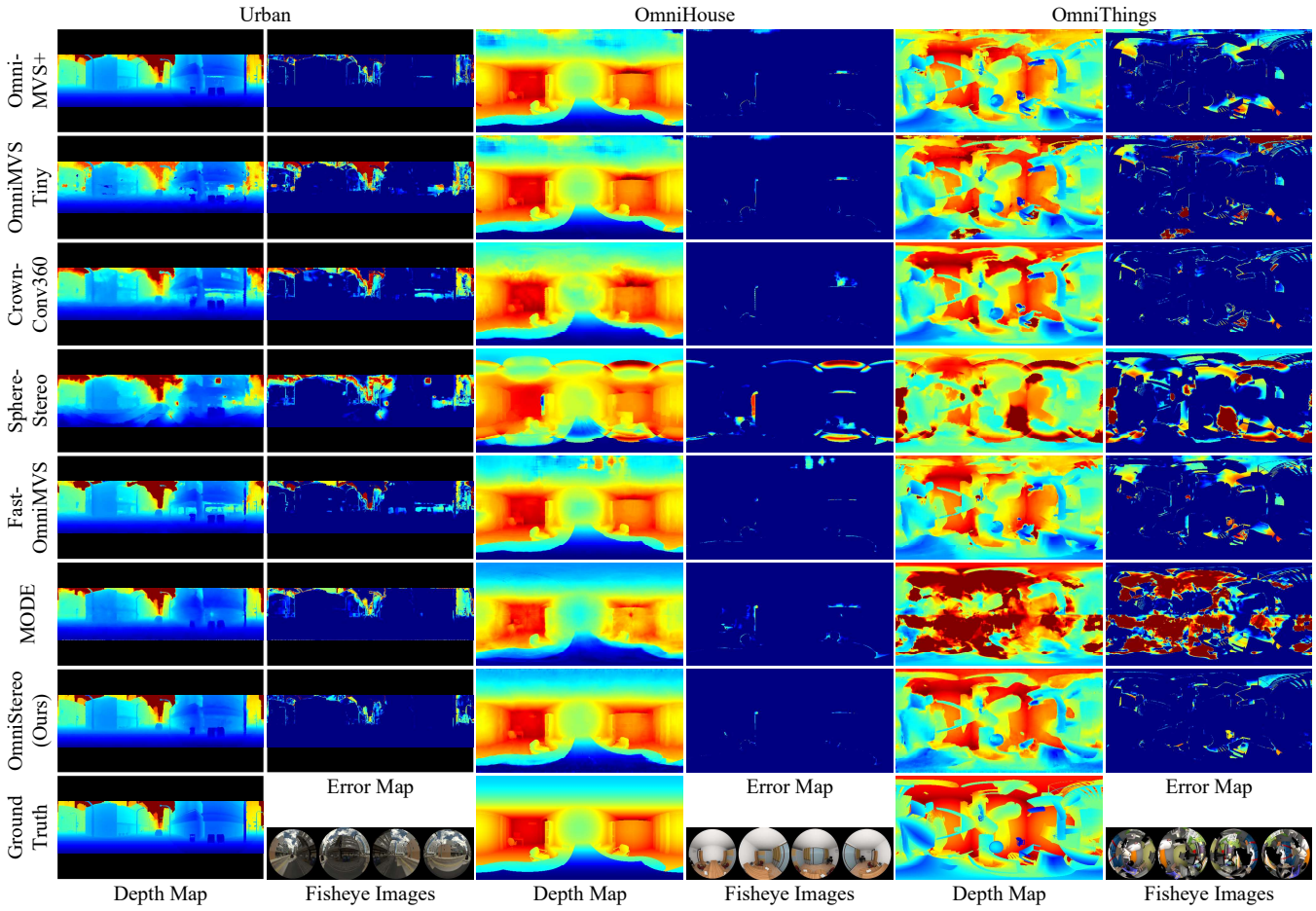

Figure 9. **Qualitative results with SOTA methods.** The red and yellow areas in error maps highlight the inaccurate estimation.

8. Qualitative results of Ablation Studies

We provide additional qualitative results of ablation studies. As illustrated in Fig. 10 (w/o Fusion, black box), directly fusing multiple depth maps using extrinsic parameters results in invalid pixels and depth discontinuities in the omnidirectional depth map. These issues are effectively resolved by the fusion module, as shown in Fig. 10 (Fusion). However, as highlighted in Fig. 11 (Fusion, red box), the fusion and multiple warping processes can lead to a loss of fine details due to filter-like operations and interpolation. For instance, details such as the streetlamp are almost entirely lost in the fusion stage. The refinement module effectively restores the streetlamp and other fine details, thereby enhancing overall accuracy, as shown in Fig. 11 (Fusion+Refine).

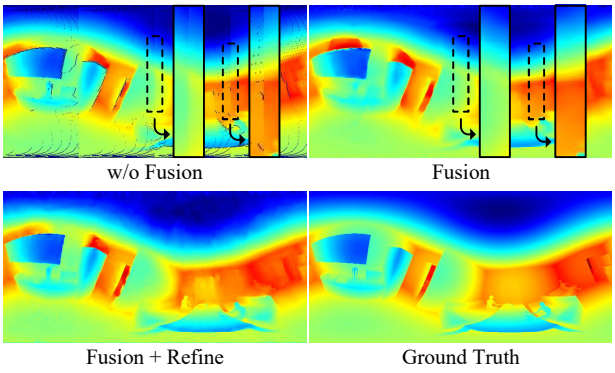

Figure 10. **Qualitative results of ablation study.** The black boxes in the depth maps highlight the depth discontinuities.

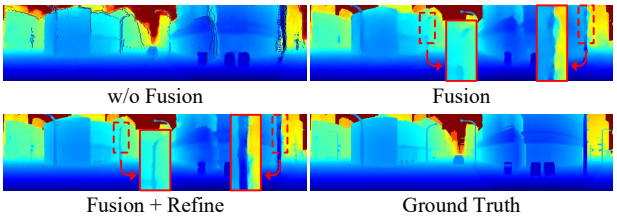

Figure 11. **Qualitative results of ablation study.** The red boxes in the depth maps highlight the detailed structures enhanced by the refinement.

9. Spatial correspondence between fisheye images and depth map

The spatial correspondence between fisheye images and the omnidirectional depth map is shown in Fig. 12. The image captured by cam1 is located at the center of the depth map, while the images from cam2 and cam4 appear on its right and left sides, respectively. The image from cam3 is split into two parts and displayed separately within the depth map.

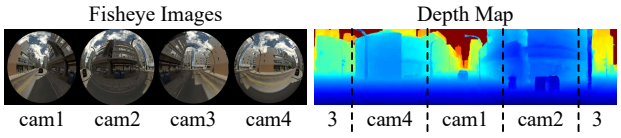

Figure 12. The spatial correspondence between fisheye images and the omnidirectional depth map. For clarity, the overlapping areas between adjacent fisheye cameras are not shown.