360MonoDepth: High-Resolution 360° Monocular Depth Estimation — Supplemental Document —

Manuel Rey-Area* Mingze Yuan* Christian Richardt
University of Bath

1. Metrics and evaluation procedure

Like MiDaS, our disparity estimates are ambiguous up to scale and offset. We therefore determine the optimal scale and offset to match the ground-truth disparity map (inverse depth) using least squares [3, Equation 14]. As all baselines predict depth and not disparity, we rescale them similarly but in depth space. In the following metrics, z and z^* represent the predicted and ground-truth depth, respectively:

- Absolute relative error (AbsRel): $\frac{1}{N}\sum_{i=1}^{N}\frac{|z_i - z_i^*|}{z_i^*}$
- Mean absolute error (MAE): $\frac{1}{N}\sum_{i=1}^{N}|z_i - z_i^*|$
- RMSE: $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\|z_i - z_i^*\|^2}$
- RMSE (log): $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\log_{10} z + i - \log_{10} z_i^*\|^2}$
- Accuracy $\delta < \tau$: % of z s.t. $\delta = \max\left(\frac{z_i}{z_i^*}, \frac{z_i^*}{z_i}\right) < \tau$

2. Runtime measurements

We measured the runtime of our method on a 2.1–3.2 GHz 16-core Xeon Silver 4216 processor with an NVIDIA RTX 3090 GPU. Table 1 list the runtime for preprocessing, including factorisation of the Poisson blending problem matrix, and the time required for each of the four stages of our method.

Table 1. Runtime measurements of our framework for different stages nd input resolutions ('Res.'), in seconds. For Poisson blending, we factorise the linear system in a preprocessing step once.

		once	per image			
Blending	Res.	Preproc.	Projection	MiDaS	Alignment	Blending
Frustum ^{M2}	2K	—	1.0	11.2	37.0	3.7
Frustum ^{M3}	2K	—	1.0	24.3	39.6	3.0
Poisson ^{M2}	2K	43.5	1.0	10.3	37.8	17.4
Poisson ^{M3}	2K	46.7	1.0	25.4	41.5	17.9
Frustum ^{M2}	4K	—	1.1	11.3	37.8	13.1
Frustum ^{M3}	4K	—	1.1	24.5	37.1	18.8

^{M2} Using MiDaS v2 [3] ^{M3} Using MiDaS v3 [2]

3. Extended discussion

Our method can fail if the tangent disparity estimates are incorrect, e.g. for large plain walls, saturated skies, or large photorealistic wallpapers, as shown in Figure 1 (left). As monocular depth estimates improve over time, our method can take advantage of them immediately. In some cases, the least-squares rescaling to fit the ground-truth disparity map pushes disparity values out of bounds, towards negative disparities. These negative disparities correspond to negative depth values that are incorrect (see Figure 1, right).

We also found inconsistencies in the reconstructed meshes of Matterport3D [1], such as windows and mirrors with depths labelled at their surface instead of corresponding

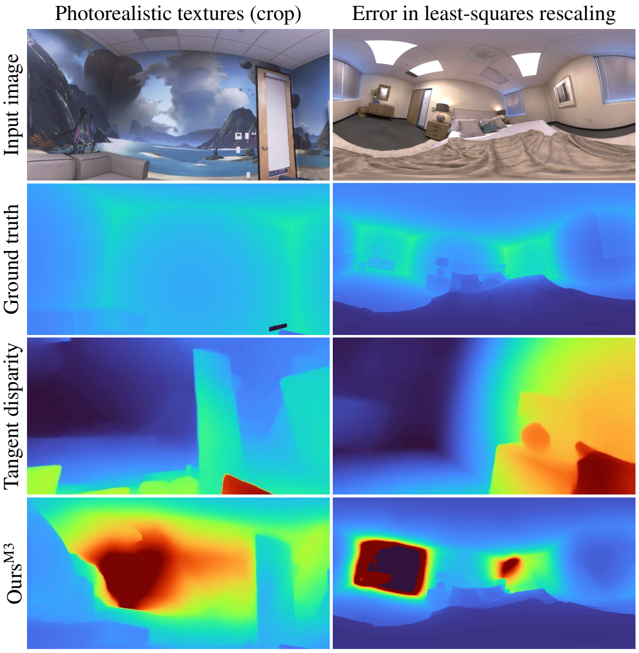
Photorealistic textures (crop)
Error in least-squares rescaling
Input image
Ground truth
Tangent disparity
Ours^{M3}

Figure 1. Failure cases for our method. **Left:** Our method cannot overcome incorrect tangent disparity estimates such as this photorealistic textured wall, which is treated as if it was an island view and not a wall. **Right:** In some cases, the least-squares rescaling to fit the ground-truth disparity range results in negative disparities, which produces incorrect, negative depth values (dark purple).

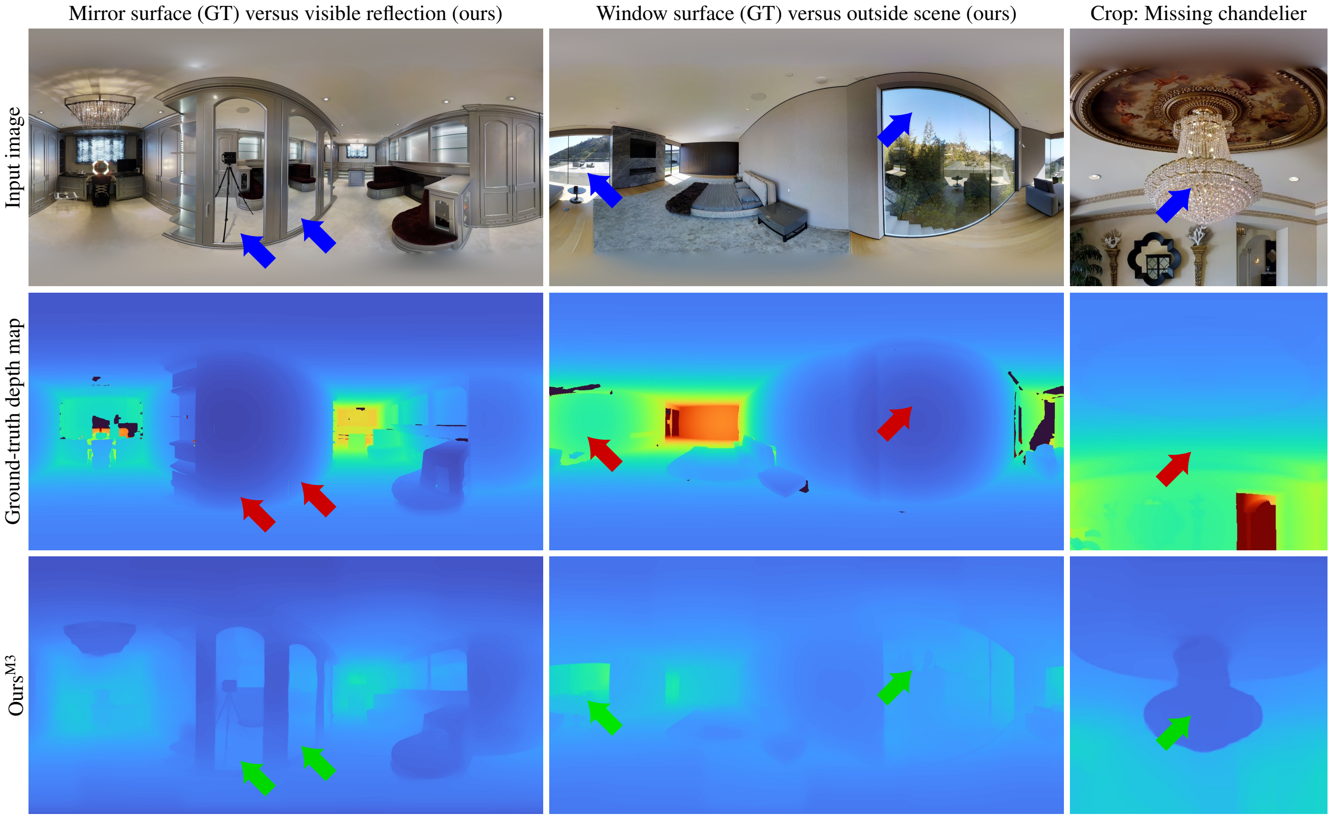

Figure 2. Inconsistent ground-truth depth maps in Matterport3D [1]. **Left:** The mesh geometry covers the surface of the mirrors instead of representing the reflection of the visible scene. **Centre:** The large windows in the room are treated as if they were opaque, instead of showing the depth of the environment outside or being masked out. **Right:** The chandelier is missing in the mesh but reconstructed by our method.

to the visible scene outside or being reflected, or missing lamps or chandeliers that are clearly visible in the image. We show examples in Figure 2, in which our method reconstructs arguably more plausible depth than the ground truth.

References

[1] 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. In *3DV*, pages 667–676, 2017.

[2] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021.

[3] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2021.