Supplementary Material: Exploiting Multiple Priors for Neural 3D Indoor Reconstruction

Federico Lincetto[1]
federico.lincetto@phd.unipd.it
Gianluca Agresti[2]
Gianluca.Agresti@sony.com
Mattia Rossi[2]
Mattia.Rossi@sony.com
Pietro Zanuttigh[1]
zanuttigh@dei.unipd.it

[1] Media Lab,
University of Padova,
Padova, IT

[2] Stuttgart Laboratory 1,
Sony Semiconductor Solutions Europe,
Sony Europe B.V.,
Stuttgart, GE

In this document we provide some additional details in order to better evaluate the performances of MP-SDF. Firstly, we present in detail the employed comparison metrics (Section 1). Then, a further analysis is shown to justify the use of a Multi-View Stereo (MVS) method rather than a monocular one for depth priors (Section 2). Moreover, we discuss the exposure compensation effectiveness showing the results achieved on scenes acquired with variable or fixed exposure settings and then reconstructed with or without exposure compensation enabled (Section 3). Finally, we present more qualitative results showing the reconstructed scenes (Section 4).

1 Quantitative Metrics

In this section, we recall the definition of the quantitative metrics employed in this work. In particular, we used the F-score and the Chamfer-L_1 distance. These two metrics compare the ground truth point cloud against the predicted one, extracted considering the set of vertexes of the respective mesh. The predicted mesh is produced by MP-SDF querying the SDF network on a three dimensional grid defined at a chosen resolution, then running the marching cubes algorithm [3]. In our experiments, aligned to what done in MonoSDF, the meshes are extracted from a grid at a resolution of $512 \times 512 \times 512$.

F-score To define F-score, it is helpful to introduce precision p and recall r first. Precision is defined as follows:

$$p = \frac{1}{|\mathcal{P}|} \sum_{\mathbf{p} \in \mathcal{P}} \mathbf{1}_T \left(\min_{\mathbf{p}^* \in \mathcal{P}^*} \|\mathbf{p} - \mathbf{p}^*\| \right) \quad \text{with} \quad \mathbf{1}_T(x) := \begin{cases} 1 \text{ if } x < T \\ 0 \text{ if } x \geq T \end{cases} \tag{1}$$

where $\mathcal{P}$ is the predicted point cloud, $\mathcal{P}^*$ the ground truth one and $T = 0.05$. Recall is defined as follows:

$$r = \frac{1}{|\mathcal{P}^*|} \sum_{\mathbf{p}^* \in \mathcal{P}^*} \mathbf{1}_T \left(\min_{\mathbf{p} \in \mathcal{P}} \|\mathbf{p} - \mathbf{p}^*\| \right). \tag{2}$$

At this point, it is possible to define the F-score as follows:

$$\text{F-score} = \frac{2 \cdot p \cdot r}{p + r}. \tag{3}$$

The F-score values presented in the main article are multiplied by a factor 100, to express them as percentages.

Chamfer-L_1 distance To define the Chamfer-L_1 distance it is convenient to introduce the accuracy a and the completeness c first. The accuracy is defined as follows:

$$a = \frac{1}{|\mathcal{P}|} \sum_{\mathbf{p} \in \mathcal{P}} \left(\min_{\mathbf{p}^* \in \mathcal{P}^*} \|\mathbf{p} - \mathbf{p}^*\| \right) \tag{4}$$

where $\mathcal{P}$ is the predicted point cloud and $\mathcal{P}^*$ the ground truth one. The completeness is defined as follows:

$$c = \frac{1}{|\mathcal{P}^*|} \sum_{\mathbf{p}^* \in \mathcal{P}^*} \left(\min_{\mathbf{p} \in \mathcal{P}} \|\mathbf{p} - \mathbf{p}^*\| \right). \tag{5}$$

At this point, it is possible to define the Chamfer-L_1 distance as:

$$\text{Chamfer-}L_1 = \frac{a + c}{2}. \tag{6}$$

The F-score assigns a fixed weight to points with a distance larger than the selected threshold, independently on the actual error. On the contrary, the Chamfer-L_1 distance is computed considering every distance value, which makes it more sensible to outliers.

2 Depth Prior Evaluation

In this section we propose a further investigation on the use of external priors. MonoSDF [7] and NeuRIS [6] exploit monocular depth and normal maps produced by Omnidata [1] to supervise the training. Instead, MP-SDF exploits multi-view stereo depth maps obtained from IterMVS [5].

We argue that geometric monocular cues can lead to wrong reconstructions: monocular approaches often produce predictions at a wrong scale and not consistent over subsequent frames. Moreover, in presence of flat but perspective coherent backgrounds, they may fail considering the 2D surface as part of the 3D volume. In the case of Replica `scan4`, presented in Fig. 1, the office wall is painted with a natural background. The monocular depth and normal estimations predict a depth and a normal maps that include geometric details of the background, even if the real surface is flat. Therefore, the reconstructed mesh is biased by these priors and exhibits artifacts on the wall. Considering Replica `scan5` in Fig. 3, it is clear how Omnidata predicts wrong normals for the office walls. This strongly affects the reconstruction producing wall deformations.

Considering the same scenes reconstructed by MP-SDF supervised by IterMVS, it is possible to observe in Fig. 2 and Fig. 4 that they do not exhibit the artifacts mentioned above. Indeed, the MVS depth maps are more reliable and consistent with the real geometry of the scene. In particular, in `scan4` the painting does not affect the depth estimation, which is consistent with the planar wall. Finally, in both the considered scenes, the IterMVS confidence maps show high reliability, thus enforcing the quality of MVS predictions.

3 Exposure compensation effectiveness

In Fig. 5 we show the effect produced by the exposure compensation. In the first row, it is possible to see some of the input RGB images of the two datasets: the Replica `Scan 1` is acquired with fixed exposure while the Tanks and Temples `Meetingroom` is captured with variable exposure. In the second and third rows, the normal maps of the reconstructions obtained with and without the exposure compensation are shown, respectively, while in the fourth row the difference map between normals is presented. As expected, its contribution on Replica is small, since input images were rendered with fixed exposure. Instead, it is very relevant on the Tanks and Temples scene, as it was acquired with variable exposure settings. The exposure compensation helps in reconstructing more accurately the structures on the ceiling and the challenging table and chair area in the middle of the room. As confirmed in Tab. 2 of the paper, the results achieved by the model without the exposure compensation are severely affected on the Tanks and Temples `Meetingroom`, while on Replica `Scan 1` the ablation has almost no effect.

4 Qualitative results

Fig. 6 presents some qualitative comparisons between MonoSDF [7], NeuRIS [5] and our approach (MP-SDF) on the Replica dataset [4] and on the `Meetingroom` scene from Tanks and Temples [2]. As discussed in the main article, on Replica our method outperforms the current state-of-the-art methods for indoor reconstruction, namely MonoSDF and NeuRIS. Differently from them, the visual quality of the reconstruction provided by MP-SDF is consistent across the different scenes in the figure, as also suggested by the quantitative results presented in the article. Finally, MP-SDF clearly outperforms the other methods also on Tanks and Temples `Meetingroom` in terms of visual quality, also in this case accordingly to the quantitative evaluation in the main article.

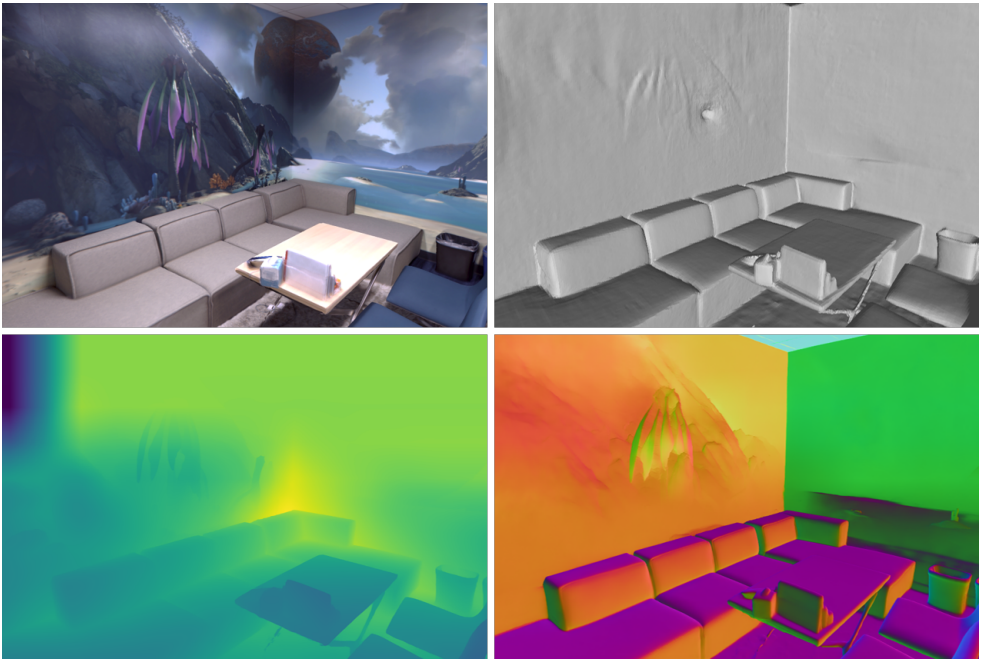

Figure 1: Replica `scan4`. From left to right and from top to bottom: RGB ground truth, MonoSDF reconstructed mesh, Omnidata depth map, Omnidata normal map.

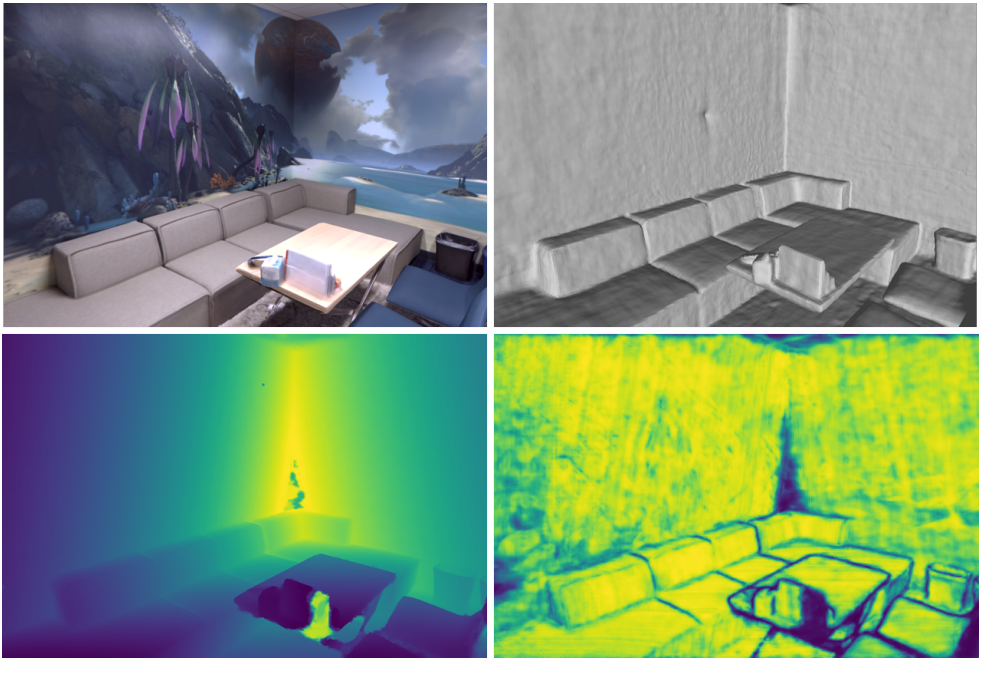

Figure 2: Replica `scan4`. From left to right and from top to bottom: RGB ground truth, MP-SDF (Ours) reconstructed mesh, IterMVS depth map, IterMVS confidence map.

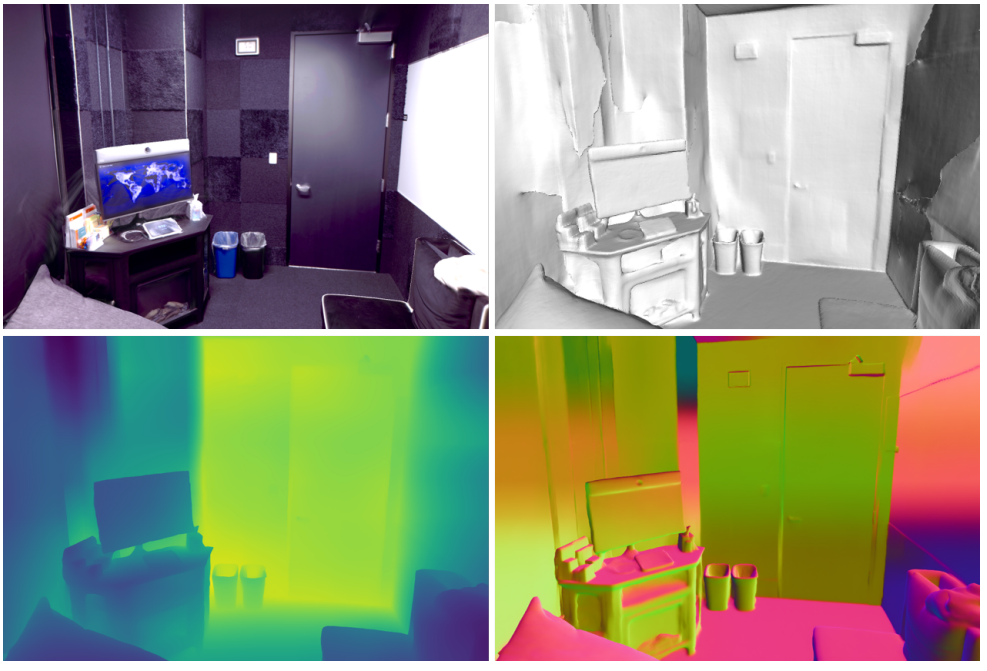

Figure 3: Replica `scan5`. From left to right and from top to bottom: RGB ground truth, MonoSDF reconstructed mesh, Omnidata depth map, Omnidata normal map.

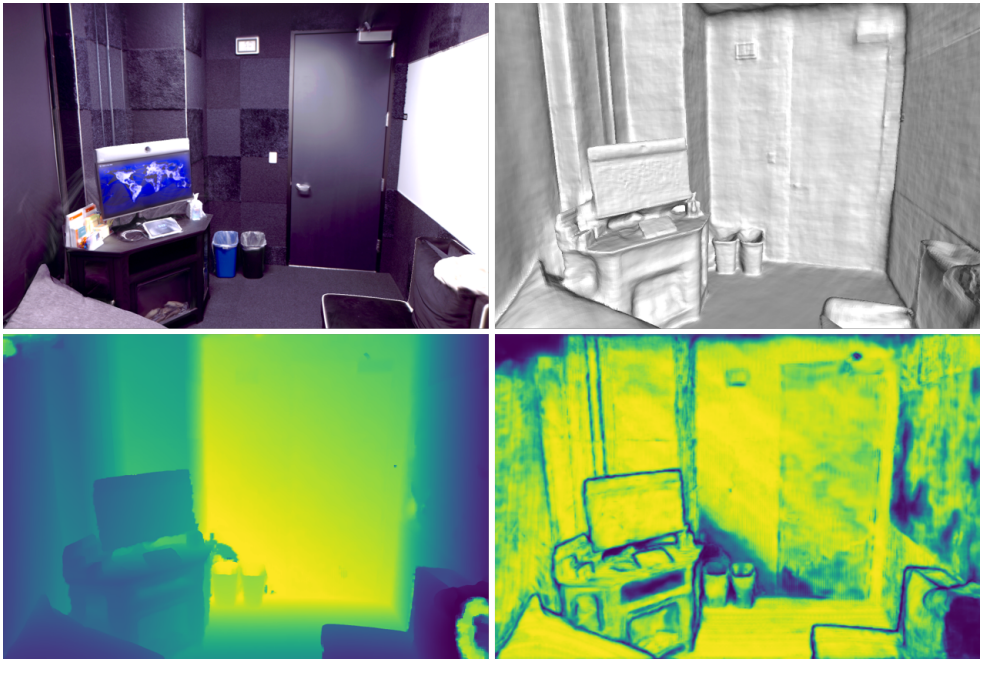

Figure 4: Replica `scan5`. From left to right and from top to bottom: RGB ground truth, MP-SDF (Ours) reconstructed mesh, IterMVS depth map, IterMVS confidence map.

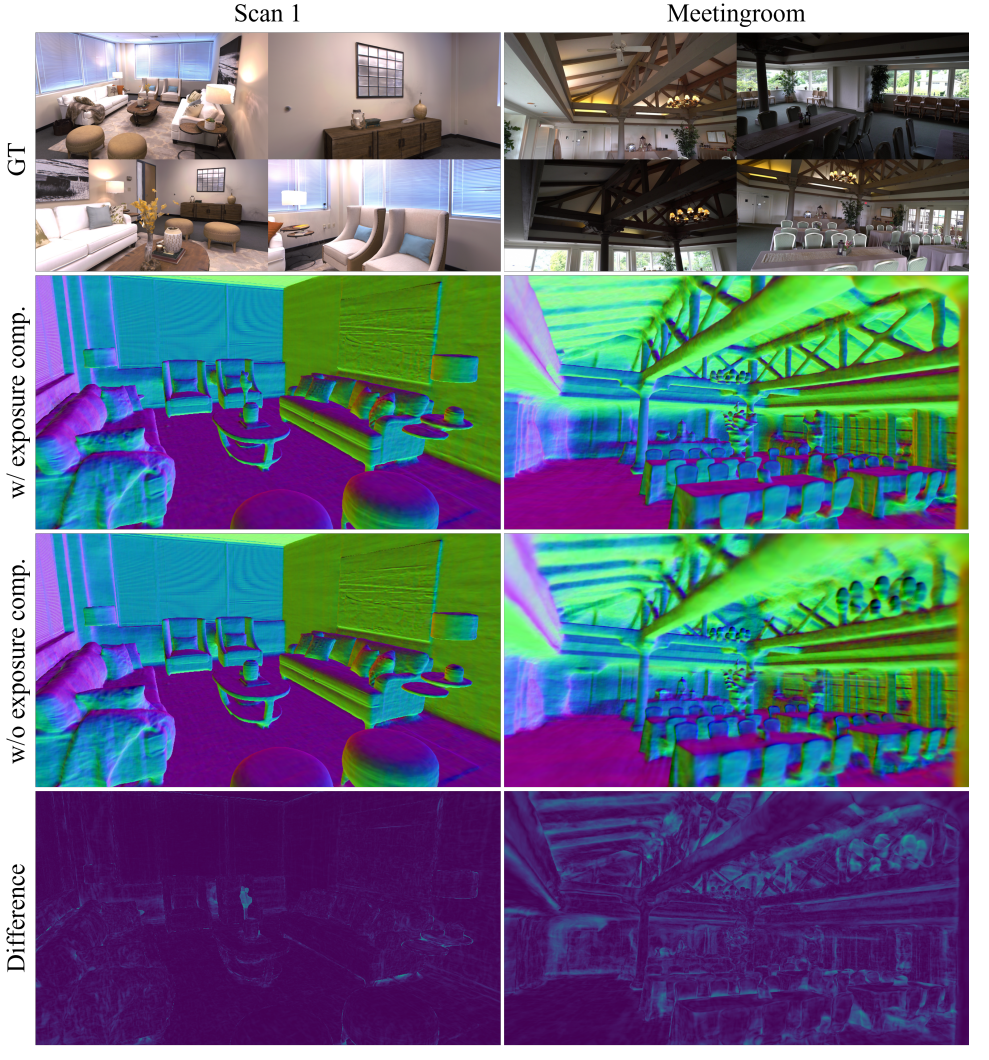

Figure 5: Exposure compensation effectiveness.

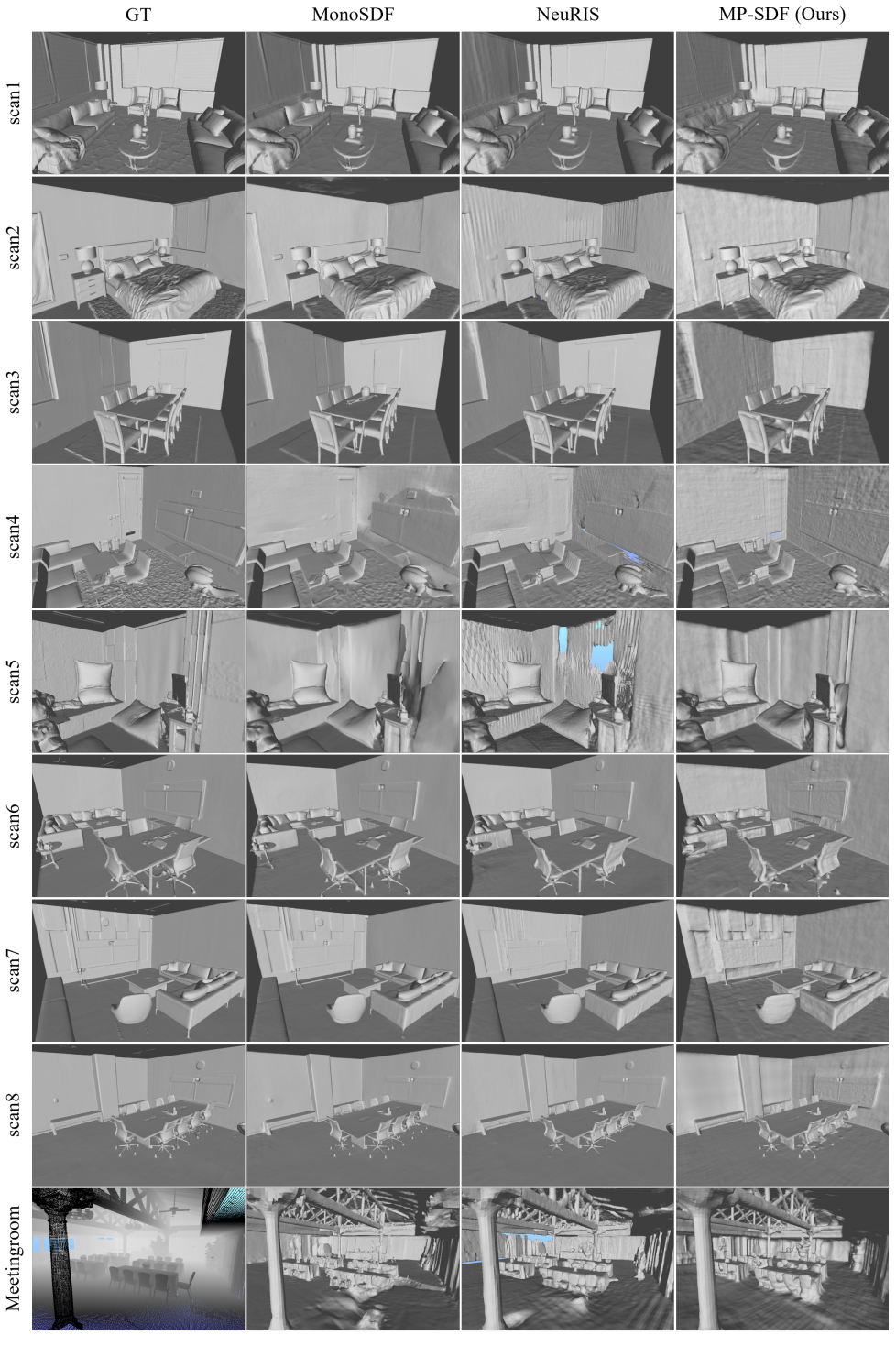

Figure 6: Qualitative evaluation on Replica and Tanks and Temples `Meetingroom`.

References

[1] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the International Conference on Computer Vision*, pages 10786–10796, 2021.

[2] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

[3] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *Proceedings of the ACM International Conference on Computer Graphics and Interactive Techniques*, 21(4):163–169, 1987.

[4] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[5] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Itermvs: iterative probability estimation for efficient multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8606–8615, 2022.

[6] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *Proceedings of the European Conference on Computer Vision*, pages 139–155. Springer, 2022.

[7] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in Neural Information Processing Systems*, 2022.