

## A 3D RECONSTRUCTION

We demonstrate that our method is applicable to various omnidirectional downstream tasks, including pose estimation and 3D reconstruction. From the dense correspondences and the certainty map produced by EDM, we can estimate the essential matrix and the relative pose. Using this predicted relative pose and dense correspondences between a pair of omnidirectional images, we can construct the dense 3D reconstruction through spherical triangulation. To address spherical triangulation, we simply solve the closed-form expression (Eising, 2022),

$$\mathbf{S} \times (R(\mathbf{X} - \mathbf{C})) = \mathbf{0}, \quad (1)$$

where  $\mathbf{S} = (S^x, S^y, S^z)$  is the 3D Cartesian coordinates,  $R \in SO(3)$  denotes the orientation of the camera,  $\mathbf{X}$  represents the target 3D point, and  $\mathbf{C}$  indicates the camera position. The cross product can be expressed using a skew-symmetric matrix, leading to the following equation,

$$\begin{aligned} S^x \mathbf{r}^{3T}(\mathbf{X} - \mathbf{C}) - S^z \mathbf{r}^{1T}(\mathbf{X} - \mathbf{C}) &= 0, \\ S^y \mathbf{r}^{3T}(\mathbf{X} - \mathbf{C}) - S^z \mathbf{r}^{2T}(\mathbf{X} - \mathbf{C}) &= 0, \\ S^x \mathbf{r}^{2T}(\mathbf{X} - \mathbf{C}) - S^y \mathbf{r}^{1T}(\mathbf{X} - \mathbf{C}) &= 0, \end{aligned} \quad (2)$$

where  $\mathbf{r}^{iT}$  denotes the  $i$ th row of  $R$ . To determine the target 3D point  $\mathbf{X}$ , we can estimate the two-view geometry using the linear equation  $A\mathbf{X} = \mathbf{b}$ . This equation can be solved by the pseudo-inverse method, considering two omnidirectional cameras  $\mathcal{M}$  and  $\mathcal{N}$ ,

$$A = \begin{pmatrix} S_{\mathcal{M}}^x \mathbf{r}_{\mathcal{M}}^{3T} - S_{\mathcal{M}}^z \mathbf{r}_{\mathcal{M}}^{1T} \\ S_{\mathcal{M}}^y \mathbf{r}_{\mathcal{M}}^{3T} - S_{\mathcal{M}}^z \mathbf{r}_{\mathcal{M}}^{2T} \\ S_{\mathcal{N}}^x \mathbf{r}_{\mathcal{N}}^{3T} - S_{\mathcal{N}}^z \mathbf{r}_{\mathcal{N}}^{1T} \\ S_{\mathcal{N}}^y \mathbf{r}_{\mathcal{N}}^{3T} - S_{\mathcal{N}}^z \mathbf{r}_{\mathcal{N}}^{2T} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} (S_{\mathcal{M}}^x \mathbf{r}_{\mathcal{M}}^{3T} - S_{\mathcal{M}}^z \mathbf{r}_{\mathcal{M}}^{1T}) \mathbf{C}_{\mathcal{M}} \\ (S_{\mathcal{M}}^y \mathbf{r}_{\mathcal{M}}^{3T} - S_{\mathcal{M}}^z \mathbf{r}_{\mathcal{M}}^{2T}) \mathbf{C}_{\mathcal{M}} \\ (S_{\mathcal{N}}^x \mathbf{r}_{\mathcal{N}}^{3T} - S_{\mathcal{N}}^z \mathbf{r}_{\mathcal{N}}^{1T}) \mathbf{C}_{\mathcal{N}} \\ (S_{\mathcal{N}}^y \mathbf{r}_{\mathcal{N}}^{3T} - S_{\mathcal{N}}^z \mathbf{r}_{\mathcal{N}}^{2T}) \mathbf{C}_{\mathcal{N}} \end{pmatrix}. \quad (3)$$

The results of 3D reconstruction are shown in Fig. 1 and Fig. 2.

## B FURTHER QUALITATIVE RESULTS

### B.1 MATTERPORT3D

We provide additional qualitative results for Matterport3D, as shown in Fig. 3 and Fig. 4. In Fig. 3, we present the results of RoMa (Edstedt et al., 2023b) instead of DKM, differing from the main paper.

### B.2 STANFORD2D3D

There are many occluded regions due to narrow corridors in the scenes. However, EDM, which is trained on Matterport3D, has the capability to handle these regions with certainty estimation, as shown in Fig. 5.

### B.3 EGO-NeRF AND OMNI-Photos

As the environments of EgoNeRF and OmniPhotos differ significantly from the Matterport3D dataset, there is a slight performance degradation. However, comparable performance maintained with certainty estimation, as shown in Fig. 6 and 7.

054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
091  
092  
093  
094  
095  
096  
097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107



Figure 1: 3D geometry of Matterport3D using matches and certainties produced by EDM. These point clouds result from spherical triangulation with estimated poses between two omnidirectional images.

108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161

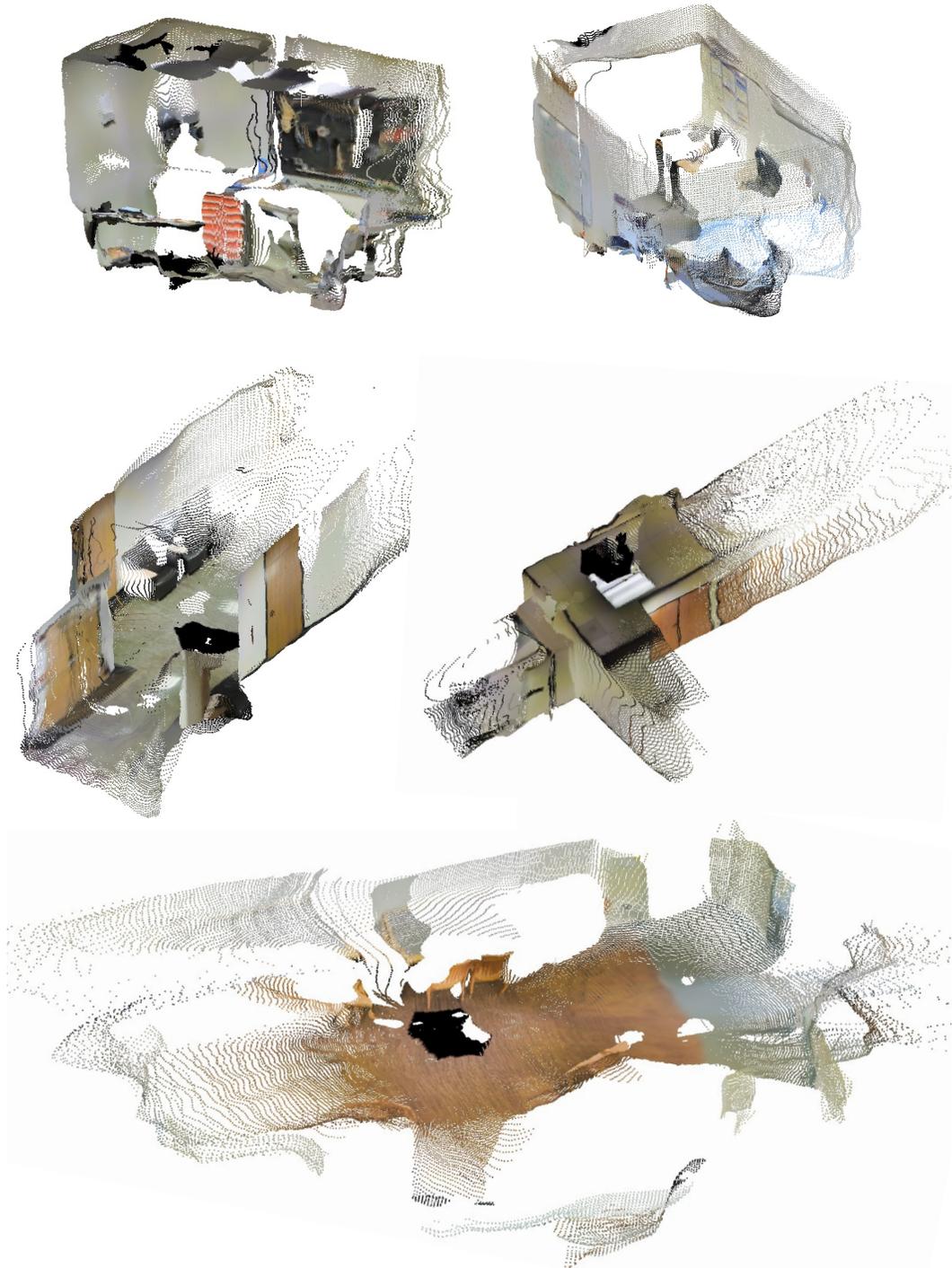
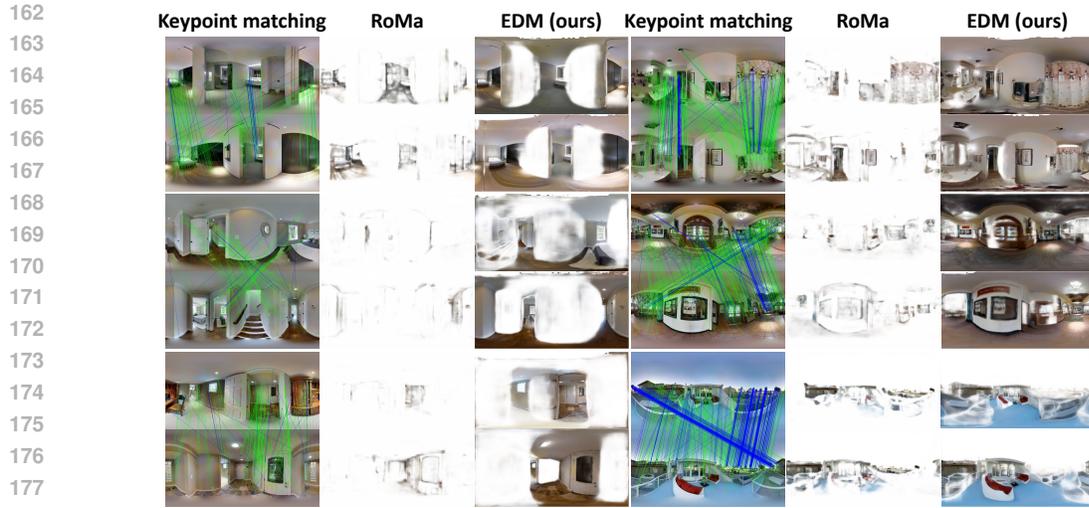


Figure 2: 3D geometry of Stanford2D3D using matches and certainties produced by EDM. These point clouds result from spherical triangulation with estimated poses between two omnidirectional images.



179 Figure 3: Qualitative results on Matterport3D. The blue lines represent the results of matching points from SPHORB (Zhao et al., 2015); the green lines correspond to SphereGlue (Gava et al., 2023). EDM demonstrates  
180 more robust performance compared to other methods.  
181



216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269

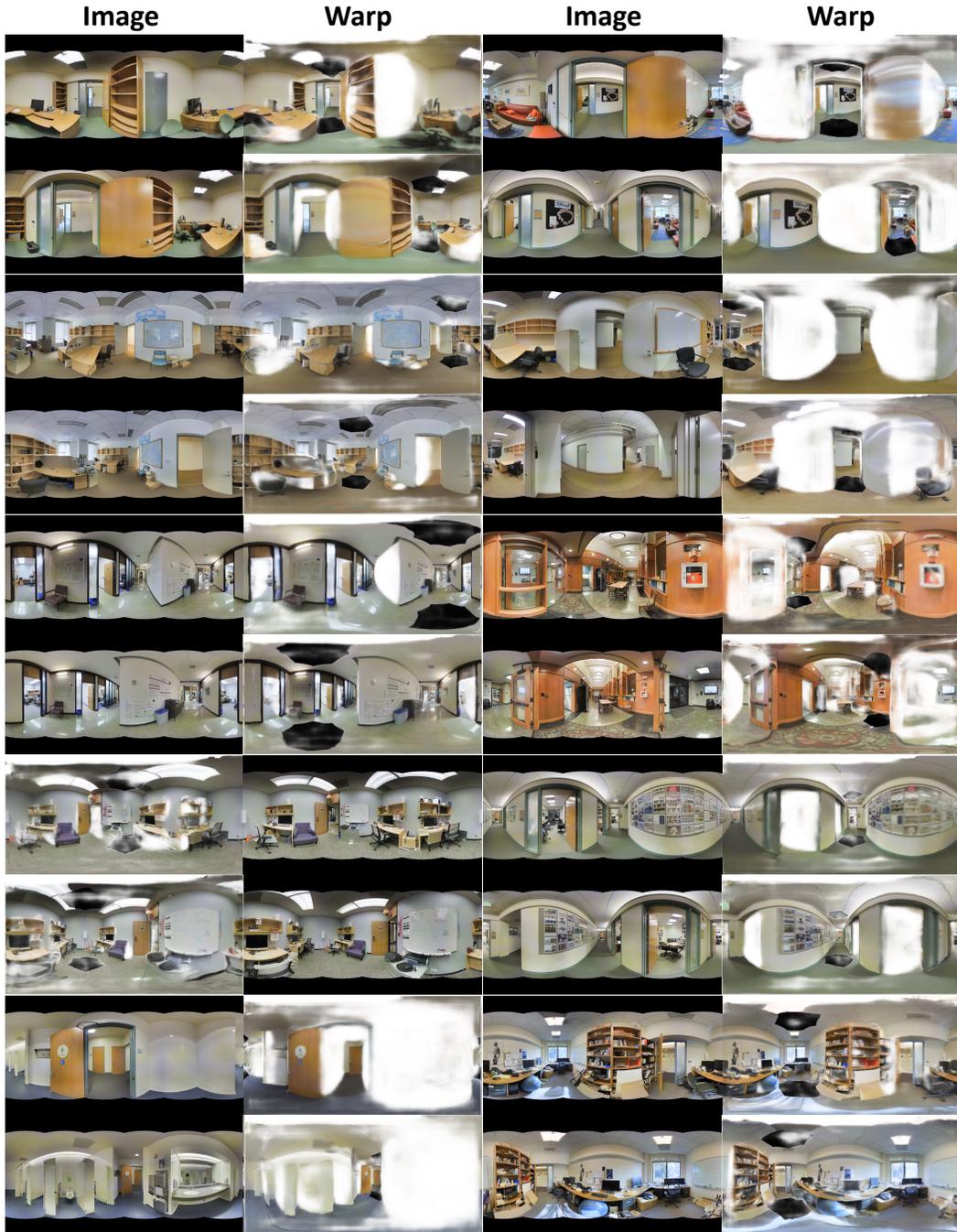


Figure 5: Qualitative results on Stanford2D3D.

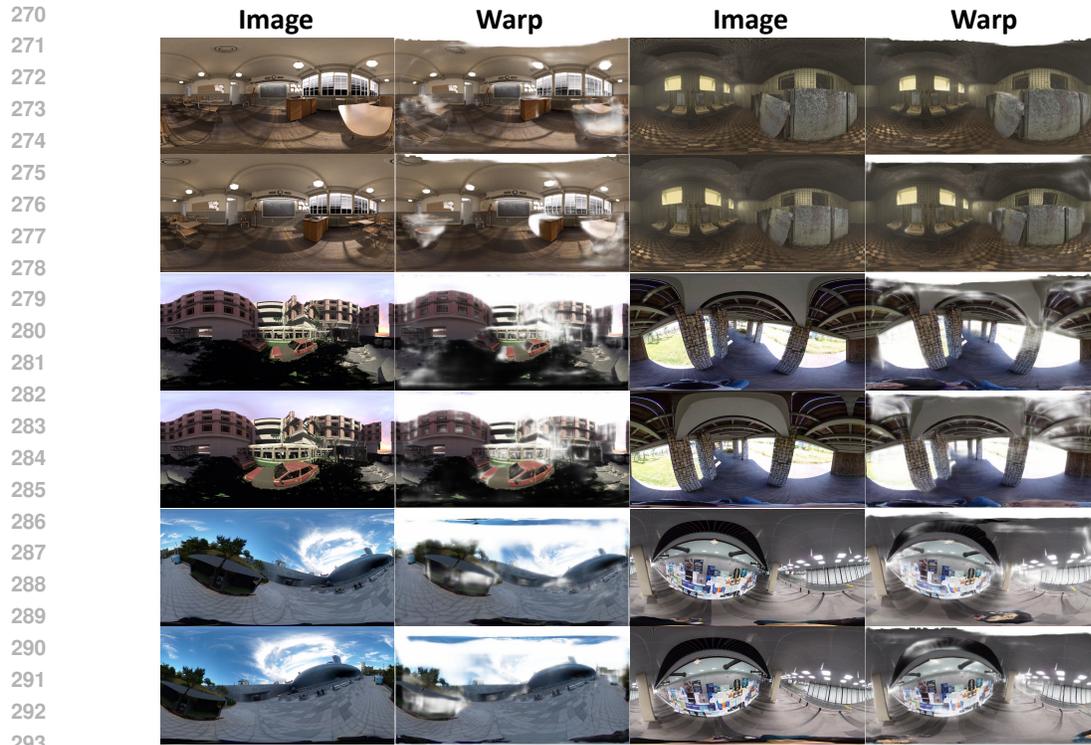


Figure 6: Qualitative results on EgoNeRF.

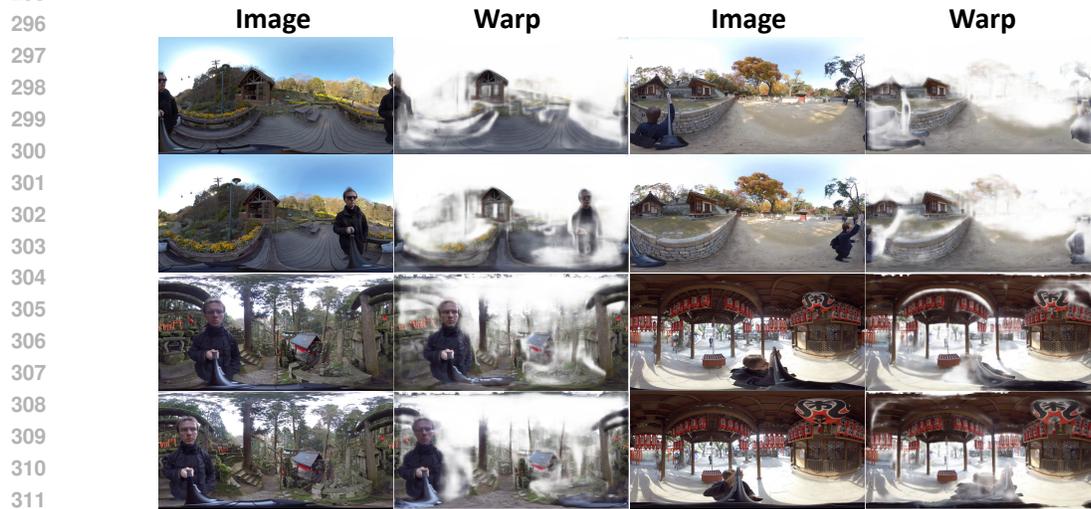


Figure 7: Qualitative results on OmniPhotos.

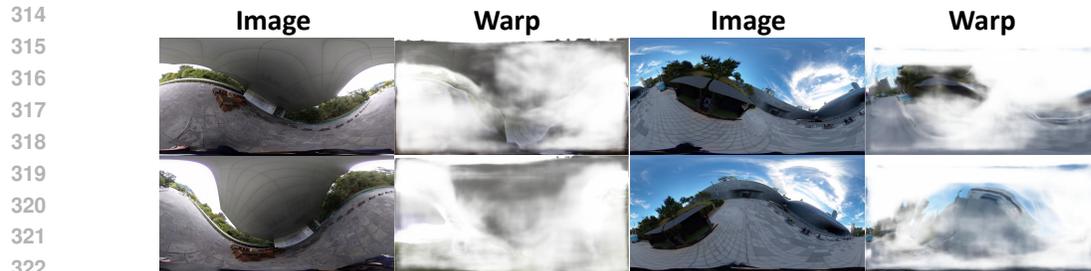


Figure 8: Failure cases.

## C THOROUGH DISCUSSION ON LIMITATIONS AND FUTURE WORK

In this section, we provide a thorough discussion of limitations and future work associated with our study. As our work is the first to develop a dense feature matching method for omnidirectional images, we believe this discussion will advance this research direction and offer deeper insights for the 360° imaging research community.

### C.1 RUNTIME EVALUATION

EDM’s runtime is almost the same as the DKM (Edstedt et al., 2023a) method because EDM includes an additional coordinate transformation between layers without requiring extra learning parameters. Both DKM and EDM take approximately 0.24 seconds per frame pair on a 3090 GPU. Comparing the runtime between sparse matching, such as SphereGlue (Gava et al., 2023) and dense matching is somewhat challenging due to differences in feature extraction and the number of matches. Sparse matching requires feature extraction before matching, and SphereGlue involves a local planar approximation to create multiple tangential images (perspective images) during feature extraction, which takes about 3.2 seconds. The inference speed for matching itself depends on the number of extracted features. In most cases, the number of features is much smaller than in dense matching, making it faster than 0.2 seconds.

### C.2 ROTATIONAL DIVERSITY IN TRAINING DATA

Our primary training dataset, Matterport3D (Chang et al., 2017), consists of indoor scenes captured with vertically fixed cameras. As a result, images with extreme rotations do not perform well in EDM, as shown in Fig. 8. We believe this problem can be mitigated by collecting more diverse training data, including images with various rotational angles, and by applying additional rotational augmentation techniques during the training process. These steps would enhance the model’s ability to handle a wider range of image orientations effectively.

### C.3 ENCODER CHOICE AND DISTORTION COMPENSATION

In this paper, we use a ResNet encoder for multi-scale feature extraction. While distortion-aware approaches (Jiang et al., 2021; Wang et al., 2020; Shen et al., 2022) exist, these methods did not yield satisfactory results in our experiments and required significant computational resources. Consequently, we employed ResNet with spherical positional embeddings to compensate for distortion without adding extra trainable layers. This approach demonstrates promising results, however, feature extraction does not fully address distortion issues. In the future, we will extend our work to develop more efficient encoders capable of handling distortions.

### C.4 UTILIZATION OF FOUNDATION MODELS

In dense matching tasks for perspective images, leveraging foundation models for coarse features (Edstedt et al., 2023b) has shown better performance compared to sharing coarse-fine features using a ResNet encoder (Edstedt et al., 2023a). In this paper, our primary goal is to demonstrate the potential of a dense matching method for omnidirectional images. We believe that adopting different foundational models, as Edstedt et al. (2023b) did, could improve our framework. We plan to train foundation models such as DINOv2 (Oquab et al., 2023) or CroCo (Weinzaepfel et al., 2022) on omnidirectional images and integrate these into our approach.

## REFERENCES

- 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. 7
- Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *CVPR*, 2023a. 7

- 378 Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Re-  
379 visiting robust losses for dense feature matching. *arXiv preprint arXiv:2305.15404*, 2023b. 1,  
380 7
- 381 Ciarán Eising. Direct triangulation with spherical projection for omnidirectional cameras. *arXiv*  
382 *preprint arXiv:2206.03928*, 2022. 1
- 384 Christiano Gava, Vishal Mukunda, Tewodros Habtegebrial, Federico Raue, Sebastian Palacio, and  
385 Andreas Dengel. Spherglue: Learning keypoint matching on high resolution spherical images.  
386 In *CVPR Workshops*, 2023. 4, 7
- 387 Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion  
388 for 360 panorama depth estimation. *IEEE Robotics and Automation Letters*, 6(2):1519–1526,  
389 2021. 7
- 391 Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov,  
392 Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning  
393 robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7
- 394 Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer:  
395 Panorama transformer for indoor 360° depth estimation. In *ECCV*. Springer, 2022. 7
- 396 Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360  
397 depth estimation via bi-projection fusion. In *CVPR*, 2020. 7
- 399 Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav  
400 Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-  
401 supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Infor-*  
402 *mation Processing Systems*, 35:3502–3516, 2022. 7
- 403 Qiang Zhao, Wei Feng, Liang Wan, and Jiawan Zhang. Sphorb: A fast and robust binary feature on  
404 the sphere. *International journal of computer vision*, 113:143–159, 2015. 4
- 405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431