EDM: Equirectangular Projection-Oriented Dense Kernelized Feature Matching

Supplementary Material

1. Two-view Geometry

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 [4],

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

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,

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

where $\mathbf{r}^{i\mathrm{T}}$ denotes the ith 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 $\mathcal{N}$,

$$\begin{aligned} A = \begin{pmatrix} S^x_{\mathcal{M}} \mathbf{r}^{3\mathrm{T}}_{\mathcal{M}} - S^z_{\mathcal{M}} \mathbf{r}^{1\mathrm{T}}_{\mathcal{M}} \\ S^y_{\mathcal{M}} \mathbf{r}^{3\mathrm{T}}_{\mathcal{M}} - S^z_{\mathcal{M}} \mathbf{r}^{2\mathrm{T}}_{\mathcal{M}} \\ S^x_{\mathcal{N}} \mathbf{r}^{3\mathrm{T}}_{\mathcal{N}} - S^z_{\mathcal{N}} \mathbf{r}^{1\mathrm{T}}_{\mathcal{N}} \\ S^y_{\mathcal{N}} \mathbf{r}^{3\mathrm{T}}_{\mathcal{N}} - S^z_{\mathcal{N}} \mathbf{r}^{2\mathrm{T}}_{\mathcal{N}} \end{pmatrix}, \\ \mathbf{b} = \begin{pmatrix} (S^x_{\mathcal{M}} \mathbf{r}^{3\mathrm{T}}_{\mathcal{M}} - S^z_{\mathcal{M}} \mathbf{r}^{1\mathrm{T}}_{\mathcal{M}})\mathbf{C}_{\mathcal{M}} \\ (S^y_{\mathcal{M}} \mathbf{r}^{3\mathrm{T}}_{\mathcal{M}} - S^z_{\mathcal{M}} \mathbf{r}^{2\mathrm{T}}_{\mathcal{M}})\mathbf{C}_{\mathcal{M}} \\ (S^x_{\mathcal{N}} \mathbf{r}^{3\mathrm{T}}_{\mathcal{N}} - S^z_{\mathcal{N}} \mathbf{r}^{1\mathrm{T}}_{\mathcal{N}})\mathbf{C}_{\mathcal{N}} \\ (S^y_{\mathcal{N}} \mathbf{r}^{3\mathrm{T}}_{\mathcal{N}} - S^z_{\mathcal{N}} \mathbf{r}^{2\mathrm{T}}_{\mathcal{N}})\mathbf{C}_{\mathcal{N}} \end{pmatrix}. \end{aligned} \tag{3}$$

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

2. Further Qualitative Results

2.1. 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 [3] instead of DKM, differing from the main paper.

2.2. 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.

2.3. EgoNeRF and OmniPhotos

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.

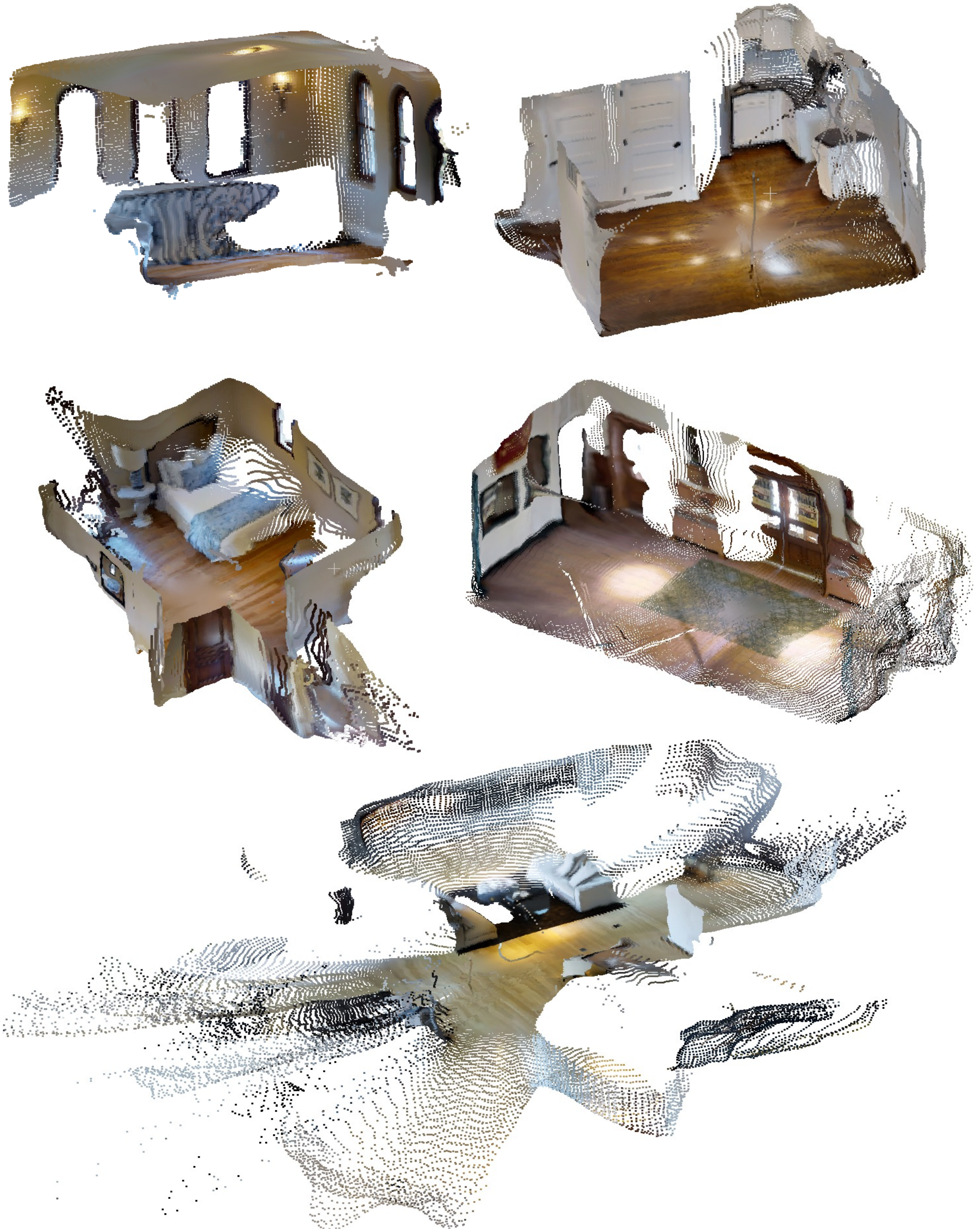

Figure 1. 3D geometry of Matterport3D using matches and certainties produced by EDM. These point clouds result from spherical triangulation with estimated poses between two omnidirectional images.

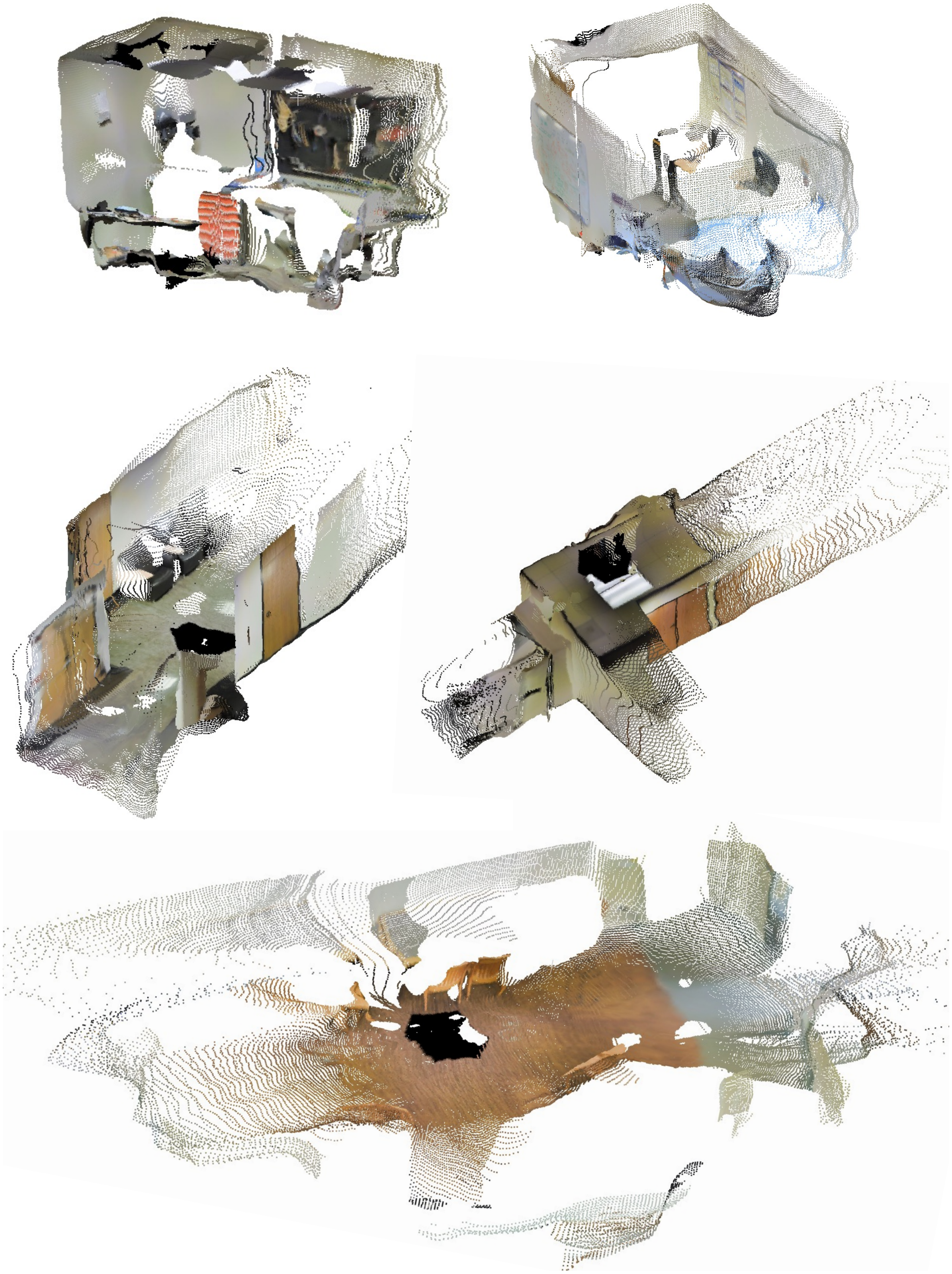

Figure 2. 3D geometry of Stanford2D3D using matches and certainties produced by EDM. These point clouds result from spherical triangulation with estimated poses between two omnidirectional images.

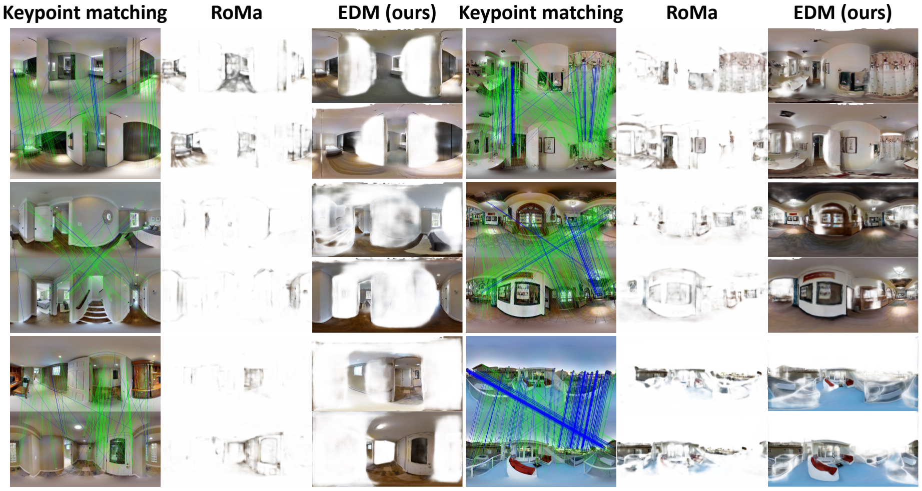

Figure 3. Qualitative results on Matterport3D. The blue lines represent the results of matching points from SPHORB [11]; the green lines correspond to SphereGlue [5]. EDM demonstrates more robust performance compared to other methods.

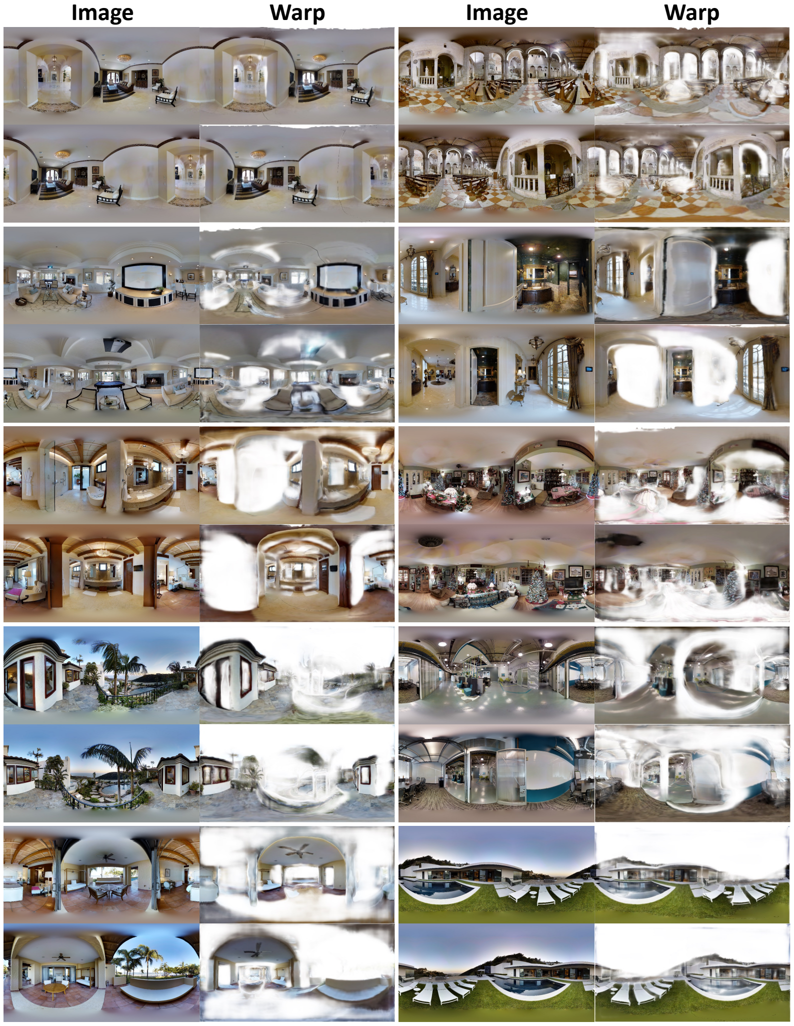

Figure 4. Qualitative results on Matterport3D.

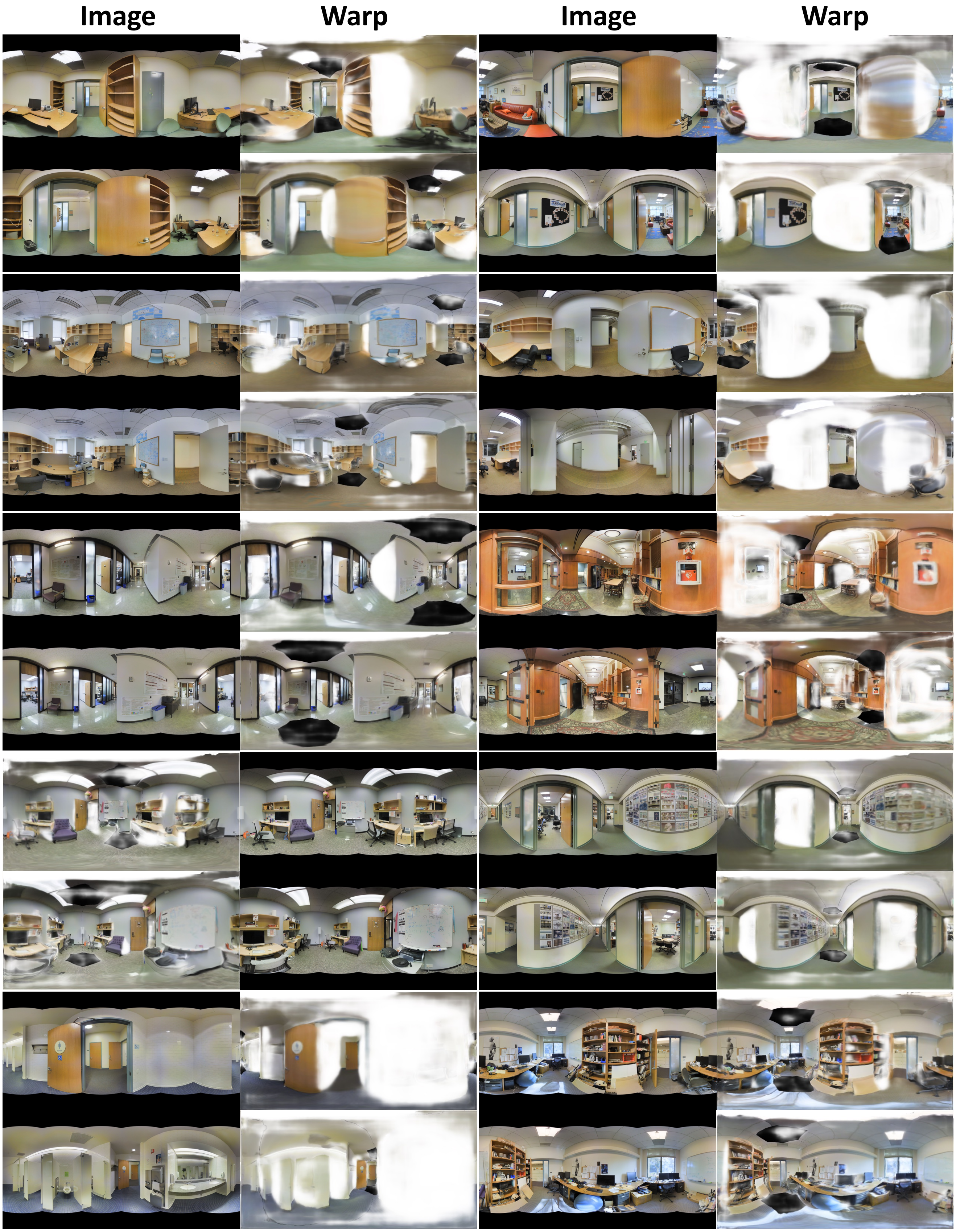

Figure 5. Qualitative results on Stanford2D3D.

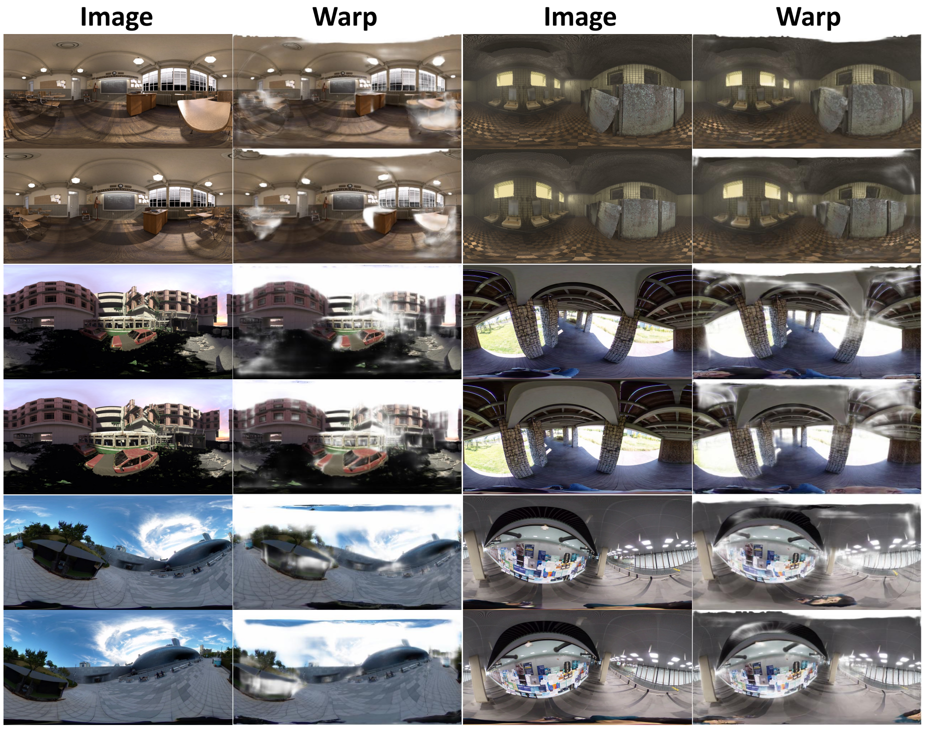

Figure 6. Qualitative results on EgoNeRF.

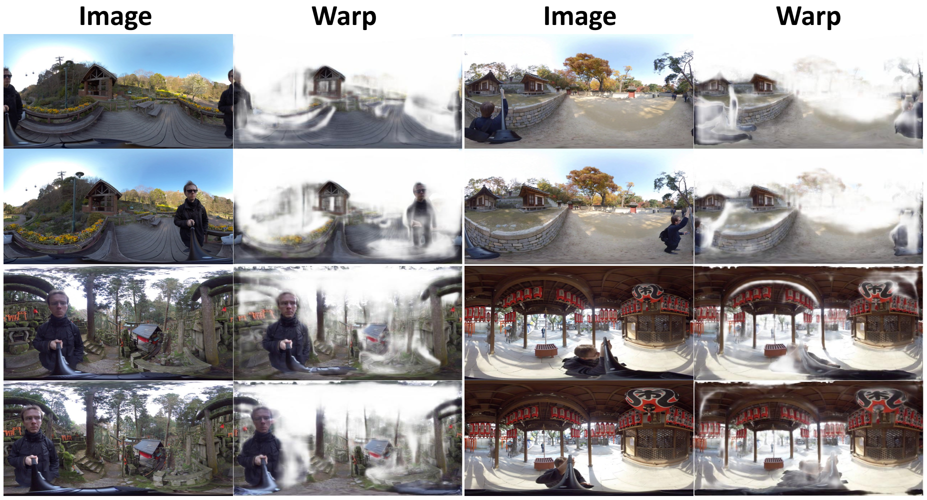

Figure 7. Qualitative results on OmniPhotos.

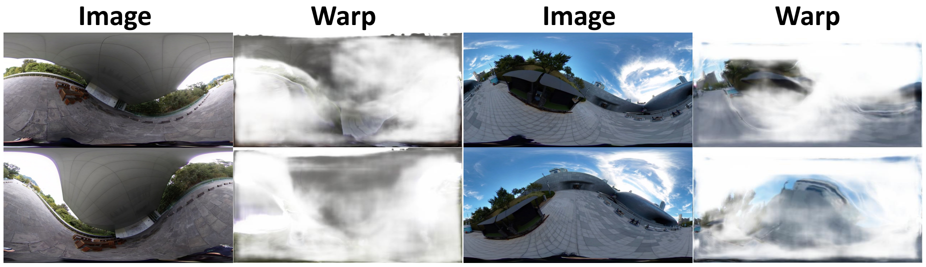

Figure 8. Failure cases.

3. 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.

3.1. Runtime Evaluation

EDM's runtime is almost the same as the DKM [2] 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 [5] 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.

3.2. Rotational Diversity in Training Data

Our primary training dataset, Matterport3D [1], 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.

3.3. Encoder Choice and Distortion Compensation

In this paper, we use a ResNet encoder for multi-scale feature extraction. While distortion-aware approaches [6, 8, 9] 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.

3.4. Utilization of Foundation Models

In dense matching tasks for perspective images, leveraging foundation models for coarse features [3] has shown better performance compared to sharing coarse-fine features using a ResNet encoder [2]. 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. [3] did, could improve our framework. We plan to train foundation models such as DINOv2 [7] or CroCo [10] on omnidirectional images and integrate these into our approach.

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. *arXiv preprint arXiv:1709.06158*, 2017. 7

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

[3] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Revisiting robust losses for dense feature matching. *arXiv preprint arXiv:2305.15404*, 2023. 1, 7

[4] Ciarán Eising. Direct triangulation with spherical projection for omnidirectional cameras. *arXiv preprint arXiv:2206.03928*, 2022. 1

[5] Christiano Gava, Vishal Mukunda, Tewodros Habtegebrial, Federico Raue, Sebastian Palacio, and Andreas Dengel. Sphereglue: Learning keypoint matching on high resolution spherical images. In *CVPR Workshops*, 2023. 4, 7

[6] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE Robotics and Automation Letters*, 6 (2):1519–1526, 2021. 7

[7] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7

[8] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360° depth estimation. In *ECCV*. Springer, 2022. 7

[9] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *CVPR*, 2020. 7

[10] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022. 7

[11] Qiang Zhao, Wei Feng, Liang Wan, and Jiawan Zhang. Sphorb: A fast and robust binary feature on the sphere. *International journal of computer vision*, 113:143–159, 2015. 4