This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

IM360: Large-scale Indoor Mapping with 360 Cameras

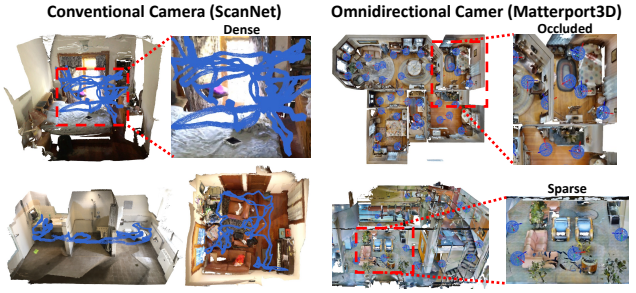
Dongki Jung* Jaehoon Choi* Yonghan Lee Dinesh Manocha
University of Maryland, College Park

Abstract

We present a novel 3D mapping pipeline for large-scale indoor environments. To address the significant challenges in large-scale indoor scenes, such as prevalent occlusions and textureless regions, we propose IM360, a novel approach that leverages the wide field of view of omnidirectional images and integrates the spherical camera model into the Structure-from-Motion (SfM) pipeline. Our SfM utilizes dense matching features specifically designed for 360° images, demonstrating superior capability in image registration. Furthermore, with the aid of mesh-based neural rendering techniques, we introduce a texture optimization method that refines texture maps and accurately captures view-dependent properties by combining diffuse and specular components. We evaluate our pipeline on large-scale indoor scenes, demonstrating its effectiveness in real-world scenarios. In practice, IM360 demonstrates superior performance, achieving a 3.5 PSNR increase in textured mesh reconstruction. We attain state-of-the-art performance in terms of camera localization and registration on Matterport3D and Stanford2D3D. Project page: https://jdk9405.github.io/IM360/

1. Introduction

Indoor 3D mapping and photorealistic rendering are core technologies for various applications in computer vision, robotics, and graphics. High-fidelity digitization of the real world enables immersive experiences in AR/VR and helps bridge the sim-to-real gap in robotic applications. However, the conventional image acquisition process [2, 4, 12] is often labor-intensive and time-consuming due to the limited field-of-view and sparse scene coverage, as shown in Fig 1. Previous studies [27, 29] have demonstrated that omnidirectional cameras can significantly reduce acquisition time, which is crucial for capturing various indoor environments. Nevertheless, most recent research [33, 45, 46, 54, 64] relies on conventional cameras with limited fields of view, which results in inadequate scene coverage, susceptibility to motion blur, and substantial time requirements for capturing large-scale indoor environments.



Dataset	# Scenes	# Images	Floor Space (m^2)	View Density (images/m^2)
Stanford2D3D [1]	6	1,413 (Omnidirectional)	6020	**0.23**
Matterport3D [5]	90	10,800 (Omnidirectional)	46,561	**0.23**
ScanNet [12]	707	2,492,518 (Conventional)	34,453	72.35

Figure 1. While the wide field of view of the omnidirectional camera significantly reduces the data acquisition time, it introduces challenges such as sparse views and occlusions. According to the Table, Matterport3D [5] and Stanford2D3D [1] are captured with significantly lower view density compared to ScanNet [12]. Our approach provides an effective solution to handle issues in sparsely scanned indoor environments and enables efficient reconstruction and rendering using omnidirectional cameras.

Unlike conventional cameras, which benefit from well-established photogrammetric software such as COLMAP [54] and OpenMVG [46], research on omnidirectional cameras remains fragmented across various domains and lacks a robust pipeline for Structure-from-Motion (SfM) and photorealistic rendering. Notably, there has been limited academic exploration of sparse scanning scenarios aimed at significantly reducing image acquisition time. There are two key challenges in developing 3D mapping and rendering for omnidirectional cameras.

First, the SfM pipeline in sparsely scanned indoor environments suffer from the prevalence of large textureless regions, which degrade the accuracy of both feature detection and matching. Moreover, frequent occlusions in indoor environments, combined with sparse input views as shown in Fig. 1, make the feature matching process even more challenging. In such scenarios, conventional perspective cameras further worsen the problem due to their limited field of view, which reduces the overlap between views.

*These two authors contributed equally.

This lack of overlap complicates the estimation of accurate camera poses, making the reconstruction task considerably more difficult. Recently, detector-free or dense matchers [15, 16, 44, 58, 61, 62] have emerged as a promising alternative to detector-based approaches [13, 50, 63], particularly in handling repetitive or indiscriminate regions where keypoint detection performance tends to degrade.

Additionally, in sparsely scanned scenarios, recent neural rendering methods [33, 45] often face challenges in producing high-quality novel-view images. It is widely recognized that both neural radiance fields (NeRF) [45] and 3D Gaussian splatting (3DGS) [33] are optimized for densely captured scenes with substantial overlap between viewpoints. Prior studies [22, 37, 43, 68] have demonstrated that neural rendering algorithms often exhibit overfitting to the training views and a poor rendering quality on novel-view images far from training views.

Main Results We present a novel indoor 3D mapping and rendering pipeline designed for omnidirectional images and address the challenges of sparsely scanned indoor environments. We tackle the problem of extensive textureless regions by employing detector-free matching methods [32]. To mitigate the frequent occlusions encountered in sparsely scanned images, we utilize an spherical camera model with a wide field of view, increasing the overlap between images and enhancing robustness in multi-view matching. While traditional methods, such as OpenMVG [46] and COLMAP [54], fail to register nearly half of the images in challenging datasets like Matterport3D [5] and Stanford2D3D [1], our method achieves unparalleled registration performance with high pose accuracy, thanks to the tight integration of a wide-view spherical model. To the best of our knowledge, ours is the first work to present a complete SfM pipeline dedicated to spherical camera models, where entire step—feature matching, two-view geometry estimation, track triangulation, and bundle adjustment—is conducted on a spherical manifold.

We extend our approach to mesh reconstruction and texturing and provide a comprehensive solution for indoor 3D mapping and rendering. For mesh reconstruction, we leverage neural surface reconstruction [67] by training a signed distance field and extracting a mesh. Classical texture mapping [64] is then combined with texture optimization through differentiable rendering. We optimize textures and train small multi-layer perceptrons (MLPs) to represent specular color. By jointly optimizing diffuse and specular textures along with the MLPs via differentiable rendering [36] under image supervision, our approach achieves superior rendering quality. Additionally, it demonstrates robustness in sparsely scanned indoor environments, outperforming recent neural rendering methods. The proposed method demonstrates significant improvements over traditional approaches, particularly in challenging indoor environments with sparse and textureless scenes.

We evaluate our approach using the Matterport3D dataset [5], which offers a larger physical scale than typical indoor datasets like ScanNet, according to [49]. We also evaluate the performance on the Stanford2D3D datasets [1]. The novel contributions of our work include:

- We introduce a unified framework for reconstructing textured meshes using omnidirectional cameras in sparsely scanned, large-scale indoor scenes.
- We propose the first spherical Structure-from-Motion approach utilizing dense matching features specifically designed for omnidirectional cameras.
- We integrate classical texture mapping with neural texture fine-tuning through differentiable rendering, demonstrating improved rendering quality, in sparsely scanned, in-the-wild indoor environments.

We compare our method with SOTA and observe higher accuracy in terms of camera localization and rendering. In particular our approach can render higher frequency details and results in lower noise in the reconstructed models.

2. Related Work

2.1. Structure from Motion

Structure from Motion (SfM) is a fundamental problem in computer vision. Traditional SfM frameworks typically begin with the detection of keypoints and descriptors, which serve as a core component, ranging from classical algorithms [41] to learning-based techniques [13, 50, 52]. Then, they match these keypoints by either nearest neighbor [54] or learning-based methods [52]. The matching pair is verified by recovering the two-view geometry [24] with RANSAC [18]. Based on these matching pairs, popular SfM methods [46, 54, 59] follow incremental approach, sequentially registering new images and reconstructing their 3D structures through triangulation [24] and bundle adjustment [60]. Alternatively, a recent work [48] employs a global approach, recovering the camera geometry for all input images simultaneously. The 2D correspondence matching significantly impact the overall performance of SfM pipeline. In textureless region of indoor environment, these methods suffer from poor feature detection [25]. Detector-free or dense matching methods [15, 16, 58] is proposed to solve this issue by estimating dense feature matches at pixel level. DetectorFreeSfM [25] builds these detector-free matches [58] and refines the tracks and geometry of the coarse SfM model by enforcing multi-view consistency.

Conversely, there are relatively few open-source software packages [30, 46] that support omnidirectional cameras. These frameworks typically employ an incremental SfM workflow based on a spherical camera model [30]. However, existing methods lack a comprehensive pipeline

specifically designed for omnidirectional images. Recently, the first dense matching method for omnidirectional images, EDM [32], has been introduced. Building upon this feature matching and two-view geometry estimation [57], we propose a complete pipeline for spherical SfM to address the challenges inherent in indoor scene reconstruction.

2.2. Textured Mesh Reconstruction

Textured mesh is the basic component of photorealistic rendering. Traditional computer vision has developed effective techniques for mapping between texture and geometry. Classical texture mapping [19, 38, 64] employs a Markov Random Field to select the optimal color image for each mesh face and applies global color adjustment for consistency [64]. Recently, numerous neural rendering techniques have been developed for surface reconstruction [10, 65, 67, 69] and rendering [2, 3, 33, 45]. Notably, some approaches leverage differentiable rendering [21, 28, 31, 36, 47] to learn neural representations of texture, lighting, and geometry. Additionally, certain NeRF variants [8, 11, 26, 39, 42] employ differentiable volumetric rendering with mesh primitives, which is similar to rasterization-based rendering. Some neural rendering methods are tailored for omnidirectional cameras. EgoNeRF [9] utilizes a spherical feature grid to represent a distant environment for rendering. OmniSDF [34] introduces an adaptive spherical binoctree method for surface reconstruction. However, in sparsely scanned indoor environment, which contains lots of textureless regions and sparse viewpoints, most neural rendering methods struggle with severe artifacts. To address this limitation, we propose a hybrid approach that combines classical texture mapping with neural texture optimization.

3. Our Method

Given a set of equirectangular projection (ERP) images, our goal is to reconstruct the geometry and texture of real-world scenes. In Fig. 2, our pipeline is composed of three primary components: Spherical Structure from Motion (SfM), geometry reconstruction, and texture optimization. To address challenges associated with limited overlap and frequent occlusions in indoor environments, our method leverages the wide field of view inherent in spherical imagery. We perform SfM utilizing a spherical camera model and dense features to estimate camera poses from sparsely scanned images in textureless regions. However, since a naive combination of SfM and dense feature matching is not directly feasible, we refine this approach to ensure compatibility. For 3D geometric reconstruction accurately, we converted ERP images into cubemaps and conducted training on the resulting perspective images. Then, we employ a neural model [67] that represents a signed distance function (SDF). After training the model, we extract the mesh using the

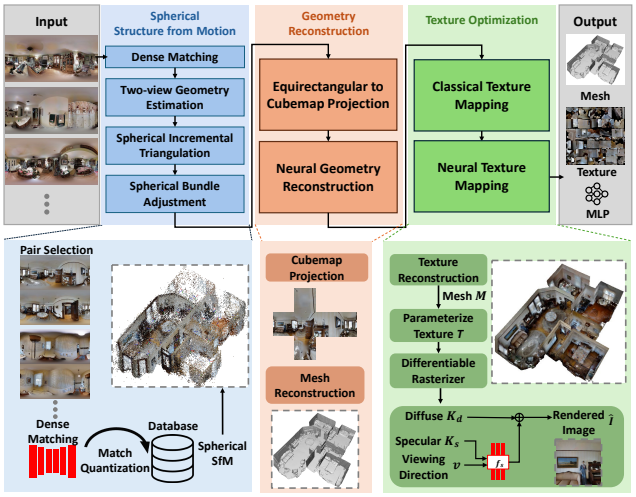


Figure 2. **Overview of Our Pipeline**. It is comprised of three steps: Spherical Structure from Motion (Sec. 3.1), Geometry Reconstruction (Sec. 3.2), and Texture Optimization (Sec. 3.3). In Section 3.1, spherical SfM is utilized to estimate the camera poses for omnidirectional images in challenging indoor environments. Then, in Section 3.2, our method apply cubemap projection and estimate monocular depth and normal to train geometry f_{sdf} representation using volumetric rendering. In Section 3.3, we initlize texture using classical texture mapping and jointly optimize diffuse, specular texture along with small MLP for modeling appearance.

Marching Cubes algorithm [40] and apply classical texture mapping [64] to initialize texture. Finally, we optimize the texture for diffuse, specular and small MLP components using differentiable rendering [36].

3.1. Dense Matching and Structure from Motion

Large-scale indoor scenes often pose significant challenges for camera localization using conventional SfM methods such as COLMAP [54] or OpenMVG [46], due to numerous occlusions and textureless regions. To effectively reconstruct sparsely captured indoor scenes, we integrate a spherical camera model into every step of SfM pipeline, combining spherical dense matching, two-view geometry estimation, and optimization methods into a unified, comprehensive 3D reconstruction framework. We construct our pipeline based on the unit bearing vector $u \in \mathbb{S}^2$,

$$u = (\sin(\theta)\cos(\phi),\ \sin(\phi),\ \cos(\theta)\cos(\phi)) \qquad (1)$$

where $\theta \in [-\pi, \pi]$ and $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

Spherical Dense Matching Performing spherical SfM requires feature matching as a preliminary step. Several feature matching methods specifically designed for 360°images have been proposed, such as SPHORB [71], SphereGlue [20], and EDM [32]. Alternatively, feature matching can also be performed independently on each face

of a cubemap generated from a 360°image. Among these approaches, dense feature matching [32] recently demonstrated superior feature matching performance for 360° images, especially in sparse scenes with large viewpoint changes. However, since dense matching methods do not explicitly provide feature descriptors, constructing feature tracks of multiple views for SfM becomes challenging. Inspired by [25, 51], we apply the additional strategies for match quantization by rounding [6] and merging [56] techniques, to ensure multiview match consistency.

Spherical Two-view Geometry Estimation It can be formally proven that the traditional two-view epipolar constratint for normalized coordinates extends to unit bearing vectors in spherical images [53],

$$u_1^T E u_2 = 0 \text{ with } E = R[t]_\times, \tag{2}$$

where $E \in \mathbb{R}^{3\times3}$ is the essential matrix, which can be decomposed into the camera rotation $R \in$ SO(3) and the translation $t \in \mathbb{R}^3$. Solarte *et al.* [57] employed the 8-point algorithm to solve spherical two-view geometry using the Direct Linear Transform (DLT) framework [24]. We adopt this two-view geometry estimation method with additional normalization [57] to enhance numerical stability during Singular Value Decomposition (SVD) computation. Using those initial pose with rotation R and translation t, we activate incremental triangulation and bundle adjustment under the sphere camera model.

Spherical Bundle Adjustment By representing each feature track observation as a bearing vector u, conventional Bundle Adjustment (BA) can be reformulated in a manner similar to the perspective case, as follows:

$$L = \sum_i \sum_j \rho(||\Pi(P_j; R_i, t_i) - u_{ij}||^2), \tag{3}$$

where ρ denotes the robust loss with a soft L1 function in local bundle adjustment, while global bundle adjustment is performed without this robustifier. $\Pi(\cdot)$ represents the spherical projection $\Pi : P \to u$, $\mathbb{R}^3 \mapsto \mathbb{S}^2$, which maps the 3D points X_j from triangulation to unit bearing vectors, given the camera poses $R_j \in$ SO(3) and $t_j \in \mathbb{R}^3$.

3.2. Geometric Reconstruction

To address visual localization and mapping in textureless and highly occluded sparsely scanned indoor scenes, we use neural rendering techniques [67, 69] for surface reconstruction instead of traditional 3D reconstruction methods such as multi-view stereo (MVS) [55]. Many methods leverage monocular geometric priors [67, 69] and demonstrate higher robustness in textureless and sparsely covered areas. However, in terms of ERP images there are no robust methods for zero-shot monocular depth and normal estimation. As shown in Fig. 3, we also observed that OmniSDF [34], which learns signed distance functions (SDF) from equirectangular projection (ERP) images, failed to converge when applied to large-scale indoor datasets. Despite incorporating geometric priors [17], Gaussian Splatting-based surface reconstruction [7, 23] still fails to produce reasonable geometry, resulting in a polygonal soup.

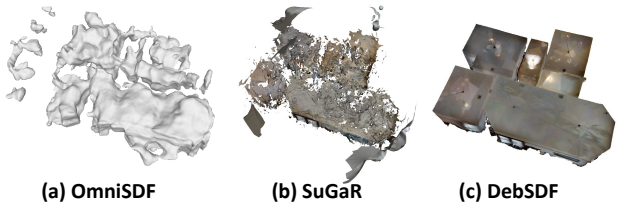


Figure 3. Our method, which applies cubemap projection and trains neural surface reconstruction using multiple perspective images, achieves highly accurate geometry reconstruction. In contrast, neural surface reconstruction designed for ERP images (OmniSDF [34]) and Gaussian Splatting-based surface reconstruction (SuGaR [23] and VCR-GauS [7]) fail to generate a clean mesh.

To address these limitations, we first project ERP images onto a cubemap representation and then apply a volumetric rendering approach to estimate the signed distance function (SDF). Since most existing mesh rendering techniques—except for volumetric methods leveraging monocular geometric priors [67, 69]—struggle in sparse indoor scenes, we employ DebSDF [67] for geometric mesh reconstruction and integrate the Marching Cubes algorithm [40] to generate a high-fidelity mesh.

3.3. Texture Optimization

In this stage, we project textures onto the 3D mesh M generated in the preceding step. In comparison to recent neural rendering methods such as NeRF [45] and 3DGS [33], it is important to highlight that texture mapping provides a more robust solution for novel view synthesis in sparsely scanned indoor environments. Utilizing images obtained from cubemap projections along with their corresponding camera poses, we initially employ TexRecon [64] to construct the texture map T. However, this conventional texture mapping approach suffers from visible seams between texture patches and is prone to geometric inaccuracies.

Inspired by TMO [10], we parameterize the texture map and leverage a differentiable rasterizer $\mathcal{R}$ [36] to render an image from a given viewpoint P. While TMO [10] is constrained to representing only diffuse textures and fails to capture view-dependent effects, our approach surpasses these limitations by enabling a more comprehensive and physically accurate representation. The diffuse texture $K_d \in \mathbb{R}^3$ is directly converted into an RGB image. Additionally, we initialize a specular feature $K_s \in \mathbb{R}^3$ and employ a small MLP f_s as a fragment shader. This MLP takes the specular feature K_s and viewing direction v as inputs to

compute the specular color as follows:

$$\hat{I}_d = \mathcal{R}(M, K_d, P),\ \hat{I}_s = f_s(\mathcal{R}(M, K_s, P), v) \quad (4)$$

The final rendered color $\hat{I} \in \mathbb{R}^3$ is obtained by combining the diffuse and specular components, $\hat{I} = \hat{I}_d + \hat{I}_s$. To optimize the diffuse, specular features, and an MLP, we employ a combination of L1 and SSIM [66] as the photometric loss in image space. This loss is calculated between the final rendered image and the ground truth image from the corresponding viewpoint, as follows:

$$L_{photo} = (1-\alpha) \parallel \hat{I} - I \parallel + \alpha\,(1 - SSIM(\hat{I}, I)) \quad (5)$$

where α is set to 0.2 and the weight to balance the losses.

4. Implementation and Performance

4.1. Experiment Settings

Matterport3D Dataset The Matterport3D dataset [5] consists of 90 indoor scenes encompassing a total of 10,800 panoramic images. Following the EDM approach [32], we define the ground truth camera poses of these images. From the validation and test sets of the official benchmark split, we selected three scenes each, ensuring that these scenes are entirely indoors or composed of levels no higher than the second floor. Since dense matching requires overlapped image pairs as input, we assume that during 360-degree camera scanning, the information about which room each camera belongs to is known. To utilize this information effectively, we leveraged the *house_segmentation* labels available in the Matterport3D dataset. Images taken within the same room were selected as matching pairs, while for transitions between different rooms, only one image pair with the highest overlap was linked. For EDM [32], the ERP images were resized to 640×320, whereas images for other feature matching methods were resized to 1024×512.

Stanford2D3D Datatset The Stanford2D3D dataset [1] comprises scenes from six large-scale indoor environments collected from three different buildings. We utilize this dataset to demonstrate that our spherical SfM method outperforms existing SfM approaches in terms of registration performance and accurate pose estimation. Specifically, we selected three scenes containing a moderate number of images. Assuming that overlapping image pairs were predefined during the scanning process, we determined covisibility and defined matching pairs using geometric meshes and ground truth poses. Additionally, we resized the equirectangular projection (ERP) images following the same methodology employed in the Matterport3D dataset.

Implementation Details We construct the Spherical Structure from Motion based on SphereSfM [30] and COLMAP [54]. To address challenges in textureless or occluded regions, we leverage the dense matching algorithm EDM [32] for enhanced feature extraction. Following the DebSDF [67], the SDF and color networks are implemented using an 8-layer MLP and a 2-layer MLP, respectively, each with a hidden dimension of 256. The networks are optimized for geometric reconstruction using the Adam optimizer [35], starting with an initial learning rate of 5×10^{-4} and applying exponential decay at each iteration. Input images are resized to 384×384, and geometric cues are provided by the pretrained Omnidata model [17]. In Texture Optimization, the specular features are initialized to zero values, matching the resolution of the diffuse texture image. The small MLP consists of 2 layers with 32 hidden dimensions. We employ the Adam optimizer [35] with a learning rate of 5×10^{-4} and utilize a learning rate scheduler. The model is trained for 7,000 iterations. We render images at a resolution of 512×512.

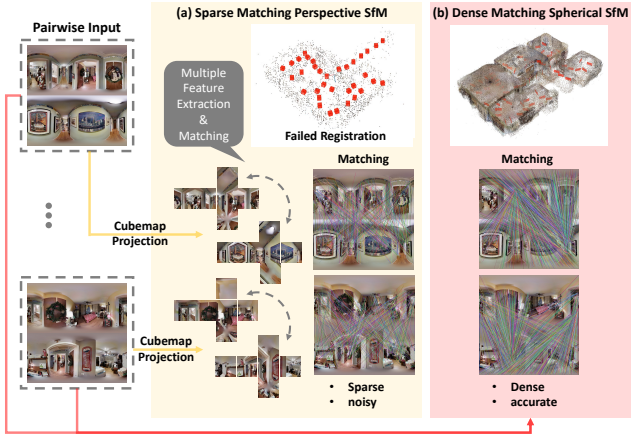


Figure 4. We provide a visual comparison between sparse matching perspective SfM and our proposed dense matching spherical SfM. In (a), ERP images are converted into a cubemap representation, after which feature matching is performed across all 36 possible image pairs, resulting in sparse and noisy correspondence matches. (b) demonstrates our approach, which directly finds dense and accurate correspondences on ERP images, thereby facilitating the construction of a detailed 3D structure.

4.2. Experimental Results

Camera Pose Estimation Accurate geometric 3D reconstruction depends on precise camera pose estimation. We compare our proposed method, which leverages dense matching [32] and Spherical SfM, against four different methods: 1) **OpenMVG:** An open-source SfM pipeline that supports spherical camera models [46]. 2) **SPSG COLMAP:** This approach employs SuperPoint [13] and SuperGlue [52] for feature detection and matching on perspective images. ERP images are transformed into a cubemap representation, yielding six perspective views per ERP

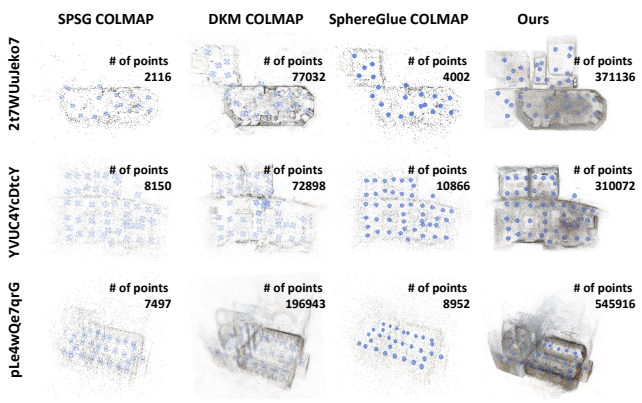


Figure 5. **Qualitative Comparison of SfM results on Matterport3D.** By leveraging dense features on equirectangular projection (ERP) images, our method effectively finds correspondences in textureless regions.

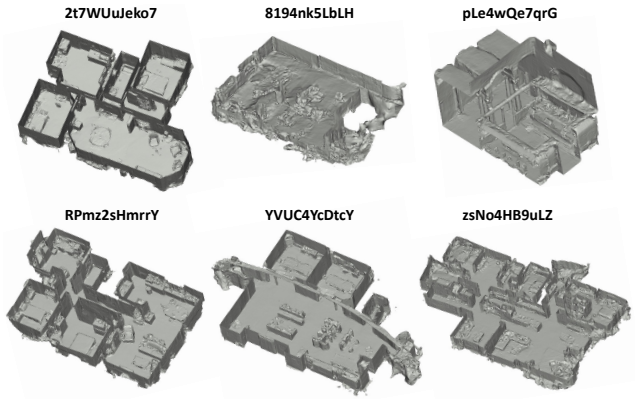


Figure 6. We visualize the shading meshes extracted without texture on Matterport3D. Leveraging the estimated poses from IM360, we achieve accurate surface reconstruction, demonstrating the effectiveness of our SfM method in generating decent geometric meshes from omnidirectional data.

image. Feature matching is performed in a brute-force manner across all 36 possible image pairs. Subsequently, incremental triangulation and bundle adjustment are performed under the rig constraints for each 360 camera node. 3) **DKM COLMAP:** This method leverages DKM [15] to establish dense correspondences without relying on explicit feature extraction. Since DKM is designed for perspective images, it also follows the cubemap projection and exhaustive pairwise matching, and rig-based optimization. 4) **SphereGlue COLMAP:** Building on COLMAP, SphereSfM [30] incorporates spherical camera models and spherical bundle adjustment to handle ERP images. For feature extraction and matching, SuperPoint [13] with a local planar approximation [14] and SphereGlue [20] are utilized to mitigate distortion in ERP images. To estimate initial poses for incremental triangulation, we follow the normalization strategy and non-linear optimization proposed by Solarte et al. [57].

Table 1 presents the quantitative results of camera pose estimation on the Matterport3D dataset. While previous methods also performed well in smaller scenes with significant image overlap, our proposed method demonstrated superior robustness across diverse scenes. By leveraging a wide field of view (FoV) for matching, our approach proved particularly effective on datasets with sparse input views. Notably, it achieved the most accurate registrations in scenes with frequent occlusions. Figure 5 visualizes the estimated camera poses and triangulated 3D points, demonstrating that our method estimated both the largest number of 3D points and the most camera poses.

Surface Reconstruction Recently, neural surface reconstruction methods based on volumetric rendering techniques [65, 67, 69] have shown strong performance in dense 3D reconstruction. In particular, several approaches that leverage monocular geometry priors [67, 69] demonstrate robustness in handling large textureless regions and sparse input views. Therefore, we adopt DebSDF [67] to estimate the signed distance function and apply the marching cubes algorithm [40] to convert the neural implicit representation into triangle meshes. Figure 6 presents shading meshes extracted without texture, the most effective way to visualize geometric differences. This highlights the effectiveness of our SfM method in producing decent geometric meshes from omnidirectional data.

Texture Map Optimization We now evaluate the results of our view synthesis method by comparing it against four different approaches: 1) **ZipNeRF** [3], the state-of-the-art neural radiance field method; 2) **3DGS** [33], which employs 3D Gaussian splats rendered through rasterization; 3) **SparseGS** [68], designed to address sparse-view challenges in 3D Gaussian splatting; and 4) **TexRecon** [64], a classical texture mapping method for large-scale scenes. Table 3 reports the rendering quality with the PSNR, SSIM [66], and LPIPS [70] on Matterport3D dataset. We consistently observe that neural rendering methods, including NeRF [3] and 3DGS [33], perform inferior to traditional texture reconstruction techniques [64]. The significant discrepancy between training PSNR and novel view PSNR indicates that this gap arises due to the sparsity of the scanned images. To address this, we evaluated the state-of-the-art SparseGS [68] approach, specifically designed for sparse-view scenarios. While SparseGS [68] achieves a 1 PSNR improvement in rendering quality, this gain indicates that sparse scanning remains a significant challenge. Our method achieves improved rendering quality across all scenes, demonstrating a 3.5 PSNR increase over the TexRecon [64] method.

In Fig. 8, we visually compare our method to sev-

Method	2t7WUuJeko7				8194nk5LbLH				pLe4wQe7qrG			
	# Registered	AUC @3°	AUC @5°	AUC @10°	# Registered	AUC @3°	AUC @5°	AUC @10°	# Registered	AUC @3°	AUC @5°	AUC @10°
OpenMVG	5 / 37	0.82	1.10	1.30	2 / 20	0.27	0.37	0.45	25 / 31	45.11	52.82	58.67
SPSG COLMAP	16 / 37	14.30	15.79	16.90	12 / 20	26.64	29.88	32.31	**31 / 31**	75.02	85.01	92.51
DKM COLMAP	22 / 37	25.14	27.70	29.62	9 / 20	14.64	16.36	17.66	**31 / 31**	**75.94**	**85.56**	**92.78**
SphereGlue COLMAP	21 / 37	23.95	26.98	29.26	12 / 20	23.11	27.76	31.24	**31 / 31**	66.83	80.06	90.03
IM360 (Ours)	**37 / 37**	**49.16**	**69.05**	**84.53**	**20 / 20**	**34.91**	**44.16**	**66.87**	**31 / 31**	73.95	84.37	92.18
Method	RPmz2sHmrrY				YVUC4YcDtcY				zsNo4HB9uLZ			
	# Registered	AUC @3°	AUC @5°	AUC @10°	# Registered	AUC @3°	AUC @5°	AUC @10°	# Registered	AUC @3°	AUC @5°	AUC @10°
OpenMVG	22 / 59	9.86	11.31	12.41	8 / 46	2.22	2.41	2.56	2 / 53	–	–	–
SPSG COLMAP	34 / 59	22.93	26.87	29.83	**46 / 46**	**84.21**	**90.53**	**95.26**	19 / 53	7.84	9.22	10.55
DKM COLMAP	51 / 59	41.56	53.56	62.58	**46 / 46**	56.33	61.00	64.67	19 / 53	7.87	8.71	9.41
SphereGlue COLMAP	34 / 59	22.23	26.45	29.62	**46 / 46**	78.32	86.97	93.48	22 / 53	6.25	7.56	9.46
IM360 (Ours)	**59 / 59**	**53.44**	**71.87**	**85.93**	**46 / 46**	72.40	83.40	91.71	**53 / 53**	**52.35**	**71.14**	**85.49**

Table 1. **Quantitative Evaluation of Camera Localization Performance** on the Matterport3D dataset [5]. Our method demonstrates superior results in terms of registration accuracy and achieves comparable performance in pose estimation across most scenes.

Method	area 3				area 4				area 5a			
	# Registered	AUC @3°	AUC @5°	AUC @10°	# Registered	AUC @3°	AUC @5°	AUC @10°	# Registered	AUC @3°	AUC @5°	AUC @10°
OpenMVG	6 / 85	0.35	0.38	0.40	17 / 258	0.34	0.37	0.39	8 / 143	0.23	0.25	0.26
SPSG COLMAP	28 / 85	6.71	7.47	8.29	73 / 258	4.75	5.77	6.55	54 / 143	8.84	10.93	12.51
DKM COLMAP	45 / 85	13.10	14.92	16.32	79 / 258	3.39	3.96	4.40	78 / 143	15.27	18.71	21.38
SphereGlue COLMAP	30 / 85	5.87	7.38	8.75	116 / 258	8.62	11.64	14.13	69 / 143	5.10	8.06	11.73
IM360 (Ours)	**85 / 85**	**32.20**	**45.12**	**58.23**	**258 / 258**	**28.70**	**50.50**	**71.29**	**138 / 143**	**19.52**	**32.54**	**51.72**

Table 2. **Quantitative evaluation of camera Localization Performance** on the Stanford2D3D dataset [1]. Our method outperforms all other approaches in terms of registration accuracy and pose estimation.

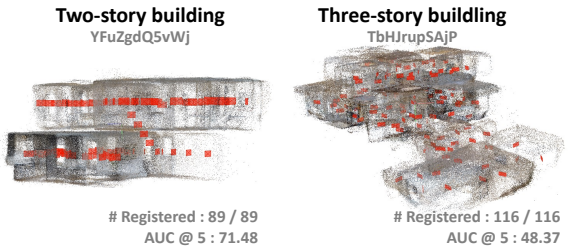


Figure 7. Localization results in multi-floor scenes of Matterport3D. Multi-floor scenes pose significant challenges for visual localization due to narrow corridors and staircases connecting different floors. Despite these difficulties, IM360 demonstrates the ability to estimate poses, ensuring robust localization in large-scale indoor environments.

eral recent rendering approaches: TexRecon [64], SparseGS [68], and ZipNeRF [3]. We visualize only SparseGS instead of 3DGS, as both methods employ similar rendering techniques, but SparseGS demonstrates superior rendering quality in sparsely scanned indoor environments. TexRecon shows the visual seams between face texture and color misalignment. ZipNeRF tends to generate ambiguous pixels, resulting in severe artifacts. SparseGS contains obvious floating artifacts in ground and floor regions. In contrast, our method produces much higher quality textures and consistently demonstrate improvement across various scenes.

4.3. Additional Experiments and Ablation Study

We observed that when initial poses are determined using two-view geometry without applying the method proposed by Solarte et al. [57], registration performance significantly deteriorates, resulting in fragmented SfM results. Furthermore, as shown in Table 1 and 2, by modifying the feature matching in COLMAP-based methods and comparing the registration and pose accuracy, we confirmed that our approach demonstrates superior robustness. We also evaluate our SfM method on multi-floor scenes, as illustrated in Fig. 7. Our spherical SfM exhibits robust performance in multi-floor scenes, highlighting the advantages of leveraging spherical cameras within our proposed pipeline for large-scale indoor environments.

In table 3, IM360* only finetunes diffuse texture without utilizing specular features. Without texture optimization stage outlined in Sec 3.3, our rendering quality is equivalent to that of TexRecon [64]. Although we apply texture optimization soley to the diffuse texture map, our method achieves a substantial improvement in rendering quality, with a 2.7 PSNR increase. Additionally, our final approach, which combines both diffuse and specular colors, provides a further 0.8 PSNR gain by accounting for view-dependent components.

5. Conclusion, Limitations and Future Work

We introduce a novel framework for textured mesh reconstruction for omnidirectional cameras in sparsely scanned large-scale indoor environments. To tackle the challenges posed by textureless regions and sparse viewpoints, we propose a spherical Structure-from-Motion (SfM) that incorporates dense feature matching and integrates spherical camera models. Additionally, we reconstruct texture maps and leverage differentiable rendering to enhance its quality, improving the rendering performance in sparse-view scenarios. One limitation of our approach is its reliance on hu-

Method	Rendering	PSNR ↑ / SSIM ↑ / LPIPS ↓						
		2t7WUuJeko7	8194nk5LbLH	pLe4wQe7qrG	RPmz2sHmrrY	YVUC4YcDtcY	zsNo4HB9uLZ	Mean
ZipNeRF [3]	Volume	15.1 / 0.51 / 0.69	11.9 / 0.44 / 0.74	14.1 / 0.43 / 0.70	13.1 / 0.50 / 0.69	15.3 / 0.53 / 0.68	14.1 / 0.65 / 0.68	13.9 / 0.51 / 0.68
3DGS [33]	Splat	14.4 / 0.47 / 0.53	11.9 / 0.38 / 0.60	13.5 / 0.37 / 0.56	12.9 / 0.51 / 0.55	14.2 / 0.49 / 0.52	13.7 / 0.60 / 0.50	13.4 / 0.47 / 0.55
SparseGS [68]	Splat	15.4 / 0.48 / 0.50	12.8 / 0.36 / 0.58	13.8 / 0.37 / 0.52	13.6 / 0.46 / 0.57	15.5 / 0.50 / 0.49	14.4 / 0.58 / 0.55	14.3 / 0.46 / 0.53
TexRecon [64]	Mesh	17.1 / 0.51 / 0.43	14.2 / 0.47 / 0.49	15.9 / 0.45 / 0.46	16.7 / 0.60 / 0.38	15.0 / 0.59 / 0.42	17.0 / 0.51 / 0.42	15.9 / 0.54 / 0.43
IM360* (Ours)	Mesh	19.6 / 0.60 / 0.37	16.7 / 0.60 / 0.43	18.5 / 0.65 / 0.36	18.9 / 0.65 / 0.38	18.3 / 0.60 / 0.42	19.8 / 0.68 / 0.41	18.6 / 0.63 / 0.39
IM360 (Ours)	Mesh	**19.8 / 0.61 / 0.38**	**16.9 / 0.61 / 0.43**	**18.8 / 0.66 / 0.35**	**20.6 / 0.75 / 0.32**	**19.4 / 0.66 / 0.32**	**21.0 / 0.74 / 0.37**	**19.4 / 0.67 / 0.37**

Table 3. **Quantitative Comparison of Rendering Performance** on the Matterport3D Dataset. Our method shows higher rendering quality of our method, IM360,across all scenes, achieving a 3.5 PSNR over the TexRecon [64]. The best-performing algorithms for each metric are shown in boldface.

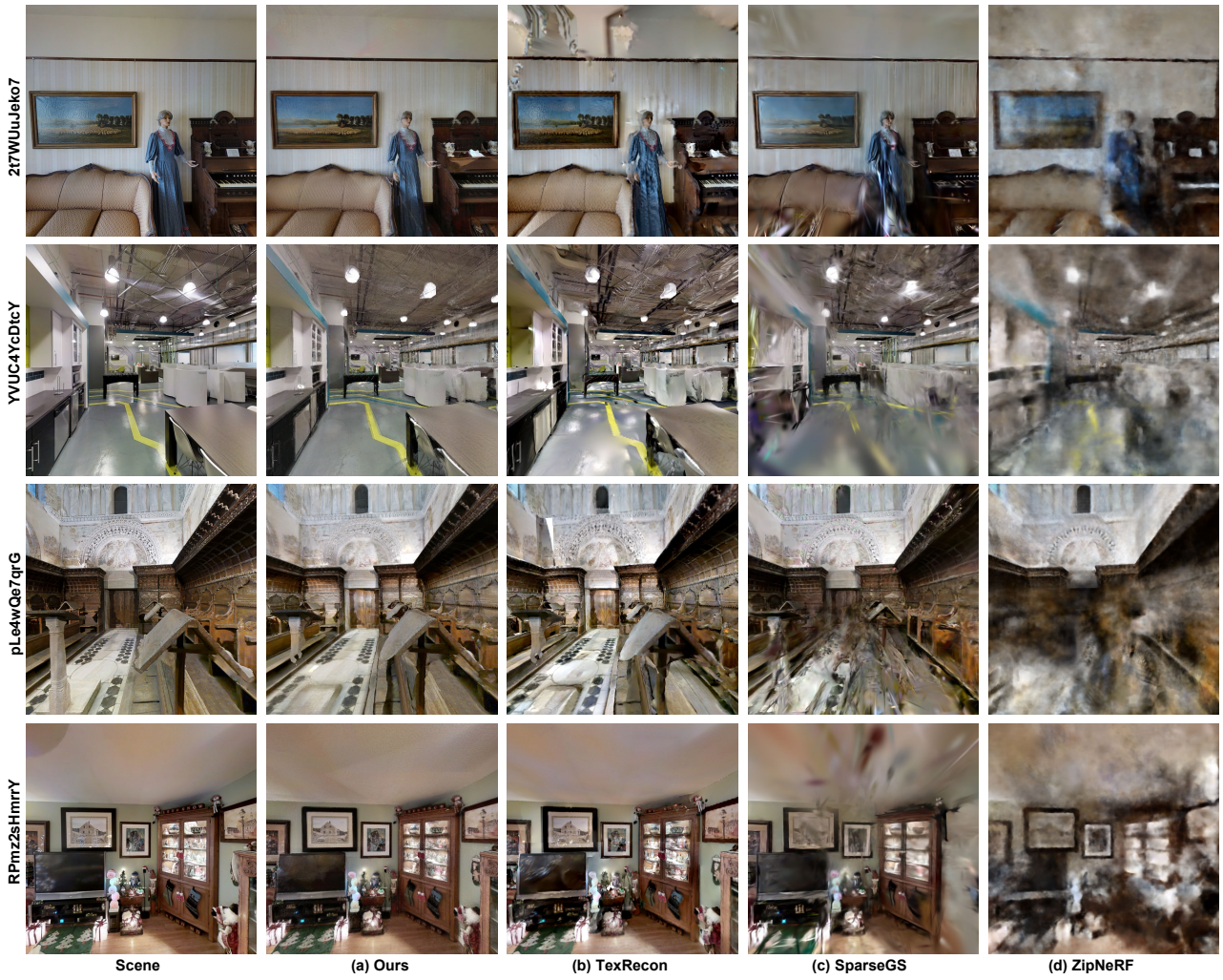


Figure 8. **Qualitative Comparisons with Existing Methods.** We visually compare our proposed approach with TexRecon [64], SparseGS [68], and ZipNeRF [3]. Our method can render high frequency details and results in lower noise.

man annotation for determining matching pairs. While 360° image scanning scenarios typically involve a small number of captures, making manual annotation feasible, integrating an automated image retrieval system would be a more scalable solution. For geometric reconstruction, we employ cubemap projection to transform 360° images into perspective views and utilize state-of-the-art volumetric rendering methods. However, this approach has limited capacity for reconstruction, as the quality deteriorates when the scene exceeds a certain scale. In future work, we aim to further refine geometric reconstruction and enhance the overall pipeline for large-scale 360° mapping.

Acknowledgements This work was supported in part by ARO Grants W911NF2310046, W911NF2310352, and U.S. Army Cooperative Agreement W911NF2120076.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 1, 2, 5, 7

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 1, 3

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023. 3, 6, 7, 8

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 1

[5] 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. 1, 2, 5, 7

[6] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *European Conference on Computer Vision*, pages 20–36. Springer, 2022. 4

[7] Hanlin Chen, Fangyin Wei, Chen Li, Tianxin Huang, Yunsong Wang, and Gim Hee Lee. Vcr-gaus: View consistent depth-normal regularizer for gaussian surface reconstruction. *Advances in Neural Information Processing Systems*, 37: 139725–139750, 2025. 4

[8] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16569–16578, 2023. 3

[9] Changwoon Choi, Sang Min Kim, and Young Min Kim. Balanced spherical grid for egocentric view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16590–16599, 2023. 3

[10] Jaehoon Choi, Dongki Jung, Taejae Lee, Sangwook Kim, Youngdong Jung, Dinesh Manocha, and Donghwan Lee. Tmo: Textured mesh acquisition of objects with a mobile device by using differentiable rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16674–16684, 2023. 3, 4

[11] Jaehoon Choi, Rajvi Shah, Qinbo Li, Yipeng Wang, Ayush Saraf, Changil Kim, Jia-Bin Huang, Dinesh Manocha, Suhib Alsisan, and Johannes Kopf. Ltm: Lightweight textured mesh extraction and refinement of large unbounded scenes for efficient storage and real-time rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5053–5063, 2024. 3

[12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1

[13] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2, 5, 6

[14] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12426–12434, 2020. 6

[15] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023. 2, 6

[16] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 2

[17] 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 IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 4, 5

[18] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[19] Yanping Fu, Qingan Yan, Long Yang, Jie Liao, and Chunxia Xiao. Texture mapping for 3d reconstruction with rgb-d sensor. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4645–4653, 2018. 3

[20] Christiano Gava, Vishal Mukunda, Tewodros Habtegebrial, Federico Raue, Sebastian Palacio, and Andreas Dengel. Sphereglue: Learning keypoint matching on high resolution spherical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6134–6144, 2023. 3, 6

[21] Shubham Goel, Georgia Gkioxari, and Jitendra Malik. Differentiable stereopsis: Meshes from multiple views using differentiable rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8635–8644, 2022. 3

[22] Guangcong, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[23] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2024. 4

[24] Richard Hartley. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2, 4

[25] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21594–21603, 2024. 2, 4

[26] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021. 3

[27] Sorin Herban, Domenica Costantino, Vincenzo Saverio Alfio, and Massimiliano Pepe. Use of low-cost spherical cameras for the digitisation of cultural heritage structures into 3d point clouds. *Journal of Imaging*, 8(1):13, 2022. 1

[28] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. https://mitsuba-renderer.org. 3

[29] Mateusz Janiszewski, Masoud Torkan, Lauri Uotinen, and Mikael Rinne. Rapid photogrammetry with a 360-degree camera for tunnel mapping. *Remote Sensing*, 14(21):5494, 2022. 1

[30] San Jiang, Kan You, Yaxin Li, Duojie Weng, and Wu Chen. 3d reconstruction of spherical images: a review of techniques, applications, and prospects. *Geo-spatial Information Science*, pages 1–30, 2024. 2, 5, 6

[31] Justin Johnson, Nikhila Ravi, Jeremy Reizenstein, David Novotny, Shubham Tulsiani, Christoph Lassner, and Steve Branson. Accelerating 3d deep learning with pytorch3d. In *SIGGRAPH Asia 2020 Courses*, pages 1–1. 2020. 3

[32] Dongki Jung, Jaehoon Choi, Yonghan Lee, Somi Jeong, Taejae Lee, Dinesh Manocha, and Suyong Yeon. Edm: Equirectangular projection-oriented dense kernelized feature matching. *arXiv preprint arXiv:2502.20685*, 2025. 2, 3, 4, 5

[33] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1, 2, 3, 4, 6, 8

[34] Hakyeong Kim, Andreas Meuleman, Hyeonjoong Jang, James Tompkin, and Min H Kim. Omnisdf: Scene reconstruction using omnidirectional signed distance functions and adaptive binoctrees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20227–20236, 2024. 3, 4

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[36] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 2, 3, 4

[37] Yonghan Lee, Jaehoon Choi, Dongki Jung, Jaeseong Yun, Soohyun Ryu, Dinesh Manocha, and Suyong Yeon. Modegs: Monocular depth guided anchored 3d gaussian splatting for robust ground-view scene rendering. *arXiv preprint arXiv:2410.04646*, 2024. 2

[38] Victor Lempitsky and Denis Ivanov. Seamless mosaicing of image-based texture maps. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–6. IEEE, 2007. 3

[39] Jeffrey Yunfan Liu, Yun Chen, Ze Yang, Jingkang Wang, Sivabalan Manivasagam, and Raquel Urtasun. Real-time neural rasterization for large scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8416–8427, 2023. 3

[40] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 3, 4, 6

[41] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2

[42] Fan Lu, Yan Xu, Guang Chen, Hongsheng Li, Kwan-Yee Lin, and Changjun Jiang. Urban radiance field representation with deformable neural mesh primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 465–476, 2023. 3

[43] T. Berriel Martins and Javier Civera. Feature splatting for better novel view synthesis with low overlap. 2024. 2

[44] Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, and Juho Kannala. Dgc-net: Dense geometric correspondence network. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1034–1042. IEEE, 2019. 2

[45] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2, 3, 4

[46] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. Openmvg: Open multiple view geometry. In *Reproducible Research in Pattern Recognition: First International Workshop, RRPR 2016, Cancún, Mexico, December 4, 2016, Revised Selected Papers 1*, pages 60–74. Springer, 2017. 1, 2, 3, 5

[47] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Mueller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. *arXiv:2111.12503*, 2021. 3

[48] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[49] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 2

[50] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019. 2

[51] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12716–12725, 2019. 4

[52] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2, 5

[53] Davide Scaramuzza and Friedrich Fraundorfer. Visual odometry [tutorial]. *IEEE robotics & automation magazine*, 18(4): 80–92, 2011. 4

[54] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 2, 3, 5

[55] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016. 4

[56] Zehong Shen, Jiaming Sun, Yuang Wang, Xingyi He, Hujun Bao, and Xiaowei Zhou. Semi-dense feature matching with transformers and its applications in multiple-view geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7726–7738, 2022. 4

[57] Bolivar Solarte, Chin-Hsuan Wu, Kuan-Wei Lu, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. Robust 360-8pa: Redesigning the normalized 8-point algorithm for 360-fov images. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11032–11038. IEEE, 2021. 3, 4, 6, 7

[58] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 2

[59] Chris Sweeney. Theia multiview geometry library: Tutorial & reference, 2015. 2

[60] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*, pages 298–372. Springer, 2000. 2

[61] Prune Truong, Martin Danelljan, and Radu Timofte. Glunet: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6258–6268, 2020. 2

[62] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5714–5724, 2021. 2

[63] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 2

[64] Michael Waechter, Nils Moehrle, and Michael Goesele. Let there be color! large-scale texturing of 3d reconstructions. In *European conference on computer vision*, pages 836–850. Springer, 2014. 1, 2, 3, 4, 6, 7, 8

[65] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems*, 2021. 3, 6

[66] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5, 6

[67] Yuting Xiao, Jingwei Xu, Zehao Yu, and Shenghua Gao. Debsdf: Delving into the details and bias of neural indoor scene reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 3, 4, 5, 6

[68] Haolin Xiong, Sairisheek Muttukuru, Rishi Upadhyay, Pradyumna Chari, and Achuta Kadambi. Sparsegs: Real-time 360° sparse view synthesis using gaussian splatting. *Arxiv*, 2023. 2, 6, 7, 8

[69] 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*, 35:25018–25032, 2022. 3, 4, 6

[70] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[71] 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. 3