Wid3R: Wide Field-of-View 3D Reconstruction via Camera Model Conditioning

Dongki Jung[1,2], Jaehoon Choi[1], Adil Qureshi[1], Somi Jeong[2], Dinesh Manocha[1], and Suyong Yeon[2]

[1]University of Maryland, College Park and [2]NAVER LABS

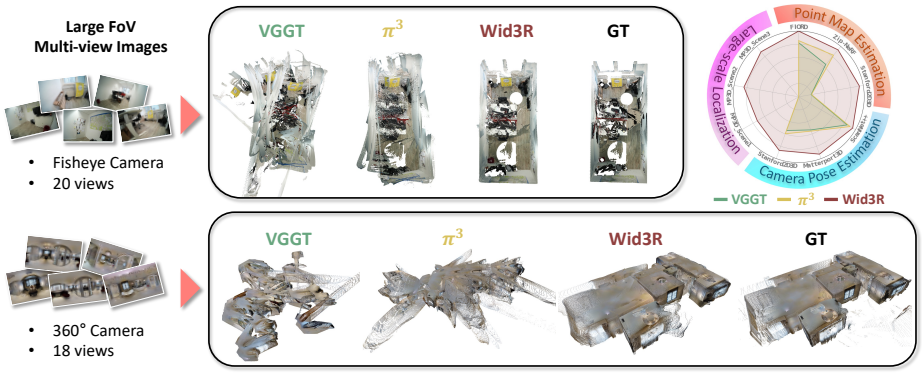

Fig. 1: Wid3R surpasses VGGT [76] and π^3 [82] in feed-forward reconstruction of fisheye and 360° imagery, demonstrating strong robustness to camera distortion. It achieves superior performance not only in pose and point map estimation but also in large-scale localization and mapping, enabled by its ray-based modules and camera model conditioning.

Abstract. We present Wid3R, a feed-forward neural network for multi-view visual geometry reconstruction that supports wide field-of-view camera models. Unlike existing methods that assume rectified or pinhole inputs, Wid3R directly models wide-angle imagery without explicit calibration or undistortion. Our approach leverages a ray-based representation with spherical harmonics and introduces a novel camera model token to enable distortion-aware reconstruction. To the best of our knowledge, Wid3R is the first multi-frame feed-forward 3D reconstruction method that supports 360° imagery. Moreover, we show that conditioning on diverse camera types improves generalization to 360° scenes and alleviates data sparsity issues. Wid3R achieves significant performance gains, improving AUC@30° by up to +33.67 on Zip-NeRF (fisheye) and +77.33 on Stanford2D3D (360). Project Page: https://jdk9405.github.io/Wid3R/

1 Introduction

Modern computer vision systems increasingly rely on accurate 3D geometric understanding across a wide variety of imaging devices. While most learning-based multi-view geometry methods [6, 43, 76, 80, 82] have achieved impressive

Table 1: We present Wid3R, the first multi-frame feed-forward network capable of handling distorted images. Wid3R enables accurate 3D point map and camera pose estimation, even in the presence of image distortion.

Method	Input	Model	Output	
	Number of Frames	Distortion Awareness	3D Point	Pose
DAC [23]	Single	✓	✓	✗
Unik3D [56]	Single	✓	✓	✗
Dust3R [80]	Pair	✗	✓	✓
VGGT [76]	Multi	✗	✓	✓
π^3 [82]	Multi	✗	✓	✓
Wid3R (Ours)	Multi	✓	✓	✓

progress, they overwhelmingly assume that input images are captured with pinhole cameras or have been rectified through calibration and undistortion. These assumptions are deeply embedded in both the architectural design of these models and the datasets on which they are trained. As a result, existing approaches are inherently biased toward perspective images and struggle to generalize when handling distorted images from wide field-of-view cameras, as shown in Fig. 1.

However, real-world scenarios rarely adhere to these limitations, which often involve wide-angle cameras. Robotics [9], autonomous navigation [40], AR/VR [16], and large-scale mapping systems [31,33,34] frequently employ fisheye [12,84] or spherical cameras [4,7] to maximize coverage and minimize blind spots. These cameras introduce complex non-linear projection effects, and their imaging geometry deviates significantly from the pinhole model. Consequently, current learning-based techniques [27, 76, 79, 80], which are trained solely on pinhole images, often fail to produce reliable depth, pose, or 3D reconstruction outputs in such settings without substantial preprocessing, including manual calibration and rectification. These preprocessing steps not only introduce engineering overhead but can also degrade geometric fidelity due to information loss during rectification [41, 48, 56].

More recently, researchers have explored learning-based geometry estimation for large field-of-view imagery, either by projecting different camera models into a common equirectangular space [23] or by adopting ray-based formulations that implicitly capture projection effects within the network [56]. However, these methods primarily focus on single-view depth or point map estimation and do not directly address multi-view 3D reconstruction. To the best of our knowledge, existing work does not provide a generalizable solution for multi-view 3D geometric estimation across cameras with wide fields of view.

Main Results As shown in Table 1, we present Wid3R, the first feed-forward neural network for multi-view 3D reconstruction from wide field-of-view images captured with fisheye or spherical cameras. Given a set of distorted multi-view images, Wid3R jointly estimates dense 3D point maps and camera poses without requiring explicit rectification or camera-specific preprocessing. Wid3R utilizes a ray-based geometric representation that models each pixel as a camera ray and encodes ray directions using spherical harmonics [56]. This formulation enables the network to reason directly about scene geometry across diverse projection models and distortion characteristics within a unified network. In addition, we

introduce a camera model token as a lightweight conditioning mechanism that enables the network to adapt to different camera models. Together, these components establish Wid3R as a multi-view geometric foundation model that operates in ray space with camera-aware conditioning, enabling robust 3D reconstruction across diverse datasets captured by wide-angle cameras [4, 7, 12, 22, 84].

In summary, the novel contributions of our work include:

- We introduce the first multi-view feed-forward network that operates directly on distorted images captured with wide field-of-view cameras.
- We propose a unified ray-based representation for multi-view 3D reconstruction that encodes camera geometry with spherical harmonics, enabling the model to robustly handle distortion.
- We design camera model tokens as an explicit conditioning mechanism that informs the network of projection characteristics specific to each camera model within a shared architecture.

We validate Wid3R through comprehensive evaluations across multiple datasets, achieving up to 3.38% improvement in $\delta_{1.25}$ for depth estimation on Matterport3D [7], 82.23% improvement in AUC@30° for camera pose estimation on Zip-NeRF [12], and 48.2% accuracy improvement in point map reconstruction on Stanford2D3D [4].

2 Related Work

2.1 Conventional 3D Reconstruction

Structure from Motion (SfM) [66] is a fundamental problem in computer vision that aims to recover camera poses and 3D structure from a collection of images. Traditional SfM pipelines consist of multiple sequential stages. These typically begin with keypoint detection and descriptor extraction [10, 47, 61, 64], followed by feature matching across images [64]. Based on the established correspondences, two-view geometry is estimated using robust methods such as RANSAC [17] to recover relative camera poses [24]. SfM systems then proceed with incremental triangulation, where new images are registered and additional 3D points are reconstructed. The estimated camera parameters and 3D structure are further refined through bundle adjustment [71]. The resulting sparse reconstruction can subsequently be used as input for multi-view stereo [18,67] or neural rendering methods [37, 50] to obtain dense scene representations. Several works have proposed improvements to this classical SfM pipeline. GLOMAP [54] adopts a global optimization strategy that estimates geometry for all input images simultaneously, improving robustness to local errors. DetectorFreeSfM [26] shows that detector-free or dense matching approaches [13, 14, 68] can significantly improve reconstruction quality in texture-poor environments where keypoint-based methods struggle. While most prior work assumes rectified or undistorted pinhole images, IM360 [33] shows that spherical cameras can be effective for sparse indoor reconstruction. By leveraging detector-free matching [31] on spherical imagery, this approach extends SfM to wide field-of-view camera models. Despite

these advances, existing SfM methods continue to rely on multi-stage pipelines that can introduce computational bottlenecks, such as iterative optimization and dense matching across large numbers of image pairs.

2.2 Feed-Forward 3D Reconstruction

Recent advances in feed-forward reconstruction models have shown that scene-level 3D structure can be estimated directly from image inputs without relying on iterative optimization at test time. DPT [60] demonstrates that Vision Transformers [11] can significantly improve dense prediction quality for monocular depth estimation by leveraging global image context. Metric3D [85] introduces a canonical camera transformation module that enables monocular metric depth estimation by normalizing camera geometry, while MoGe [79] proposes affine-invariant point map estimation coupled with optimal alignment to recover consistent geometric structure.

To address consistency across views, DUSt3R [80] formulates feed-forward pairwise reconstruction, where an image pair is jointly processed to predict dense point clouds expressed in the coordinate frame of one camera. Building on this formulation, Spann3R [75] incorporates an external spatial memory to support incremental reconstruction over multiple views, while FLARE [87] modifies the architecture to progressively predict 3D points under the estimated camera pose. VGGT [76] further improves robustness and accuracy by leveraging large-scale training data and multi-task supervision to estimate geometry in a global world reference frame. π^3 [82] addresses sensitivity to reference view selection by introducing a permutation-equivariant architecture that reduces performance variance across different view orderings. Recently, Depth Anything 3 [45] and MapAnything [36] utilized ray maps to represent camera parameters. However, these feed-forward reconstruction methods still assume distortion-free perspective images and pinhole camera models, resulting in degraded performance on wide field-of-view imagery. We overcome this limitation by operating directly in ray space with camera-aware conditioning, enabling robust multi-view reconstruction in the presence of lens distortion.

2.3 Wide Field-of-View Camera Models

Camera calibration plays a critical role in recovering 3D structure from the 2D image plane by estimating intrinsic parameters such as focal length, principal point, and lens distortion. A variety of camera models have been proposed to capture different imaging geometries, including the pinhole model, the Kannala–Brandt model [35], the Mei model [49], the Unified Camera Model [21,38], the Double Sphere model [73], and omnidirectional camera models [65]. However, approaches based on explicit camera modeling are sensitive to data acquisition conditions, often leading to variability in the estimated calibration parameters [1,59]. Moreover, as camera models become more complex with increased parameterization, the calibration process may introduce additional sources of error and instability [8,46].

As an alternative to explicit camera modeling, learning-based methods have emerged that perform monocular point estimation by representing images as pencils of rays in a spherical formulation [56, 57]. In particular, UniK3D [56] removes explicit camera assumptions by modeling scene geometry directly in ray space, leveraging a spherical harmonics basis. Recent work adapts pretrained monocular depth estimation models to fisheye cameras by introducing additional calibration tokens [19]. In contrast, we encode camera models through dedicated tokens within a ray-based geometric representation and extend this formulation to multi-view reconstruction, enabling adaptation to wide-angle camera types, including fisheye and 360° imagery.

3 Our Approach: Wid3R

Generalizable 3D geometry estimation across diverse camera configurations is essential for real-world applications. We introduce a novel method that enables multi-view 3D reconstruction from wide field-of-view camera images. We first present the problem formulation in Section 3.1, followed by a discussion of the local coordinate representation in Section 3.2, and finally describe the camera configuration modeled through a ray-based representation in Section 3.3.

3.1 Problem Definition

Given a sequence of N RGB images $\{I_i\}_{i=1}^N$, where each image $I_i \in \mathbb{R}^{3\times H\times W}$, the Wid3R network $f(\cdot)$ maps the multi-view image sequence into a unified 3D geometric representation:

$$f\big(\{\mathbf{I}_i, \mathbf{C}_i\}_{i=1}^N\big) = \{\mathbf{T}_i, \boldsymbol{\mathcal{R}}_i, \mathbf{D}_i, \mathbf{U}_i\}_{i=1}^N, \tag{1}$$

where $\mathbf{C}_i$ denotes a camera model token associated with view i. $\mathbf{T}_i \in \mathbb{R}^{4\times4}$ represents the camera pose, $\boldsymbol{\mathcal{R}}_i \in \mathbb{R}^{3\times H\times W}$ denotes the pencil of rays defined at each image location, $\mathbf{D}_i \in \mathbb{R}^{H\times W}$ corresponds to the radial distance along each ray, and $\mathbf{U}_i \in \mathbb{R}^{H\times W}$ represents the associated uncertainty. Here, we incorporate the camera model token as an input condition for geometric prediction, enabling robust reconstruction from wide field-of-view images without explicit calibration.

3.2 Per-View Local Geometry Reconstruction

Wide field-of-view images provide extensive scene coverage, enabling large portions of a scene to be captured with a relatively small number of images. For example, 360° cameras observe the entire surrounding environment at each viewpoint, which often leads to sparsely distributed scene captures along a trajectory [33]. In such settings, consecutive views may exhibit limited visual overlap, particularly in long-term sequences or large-scale environments. Under such conditions, methods that rely on view-dependent assumptions [76] may struggle when reference and source views lack sufficient overlap.

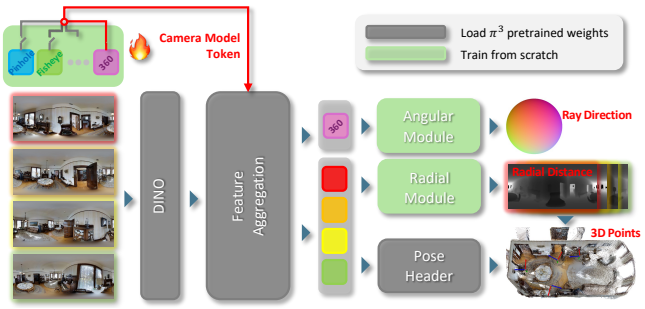

Fig. 2: Overview of the architecture. Multi-view wide field-of-view images are processed by DINO and aggregated through a feature aggregation module. An appropriate camera model token is selected to condition the network on camera-specific priors. Instead of directly regressing point or depth maps, Wid3R decomposes 3D reconstruction into angular and radial components, predicting ray directions and radial distances that are robust to camera projection distortion. A pose header estimates camera poses, enabling the reconstruction of global 3D points. Pretrained π^3 [82] weights are loaded where applicable to accelerate training convergence, while the remaining components are trained from scratch.

To address this challenge, we adopt a permutation-equivariant network design inspired by [82], in which scene geometry is represented in per-view local coordinate frames rather than a shared global reference frame. For each input image $\mathbf{I}_i$, our network estimates a pixel-aligned 3D point map $\hat{\mathbf{P}}_i = \hat{\boldsymbol{\mathcal{R}}}_i \odot \hat{\mathbf{D}}_i$, obtained via element-wise multiplication between the ray directions $\hat{\boldsymbol{\mathcal{R}}}_i$ and the corresponding radial distances $\hat{\mathbf{D}}_i$. Since each point map is expressed in its own local camera coordinate system, scale ambiguity arises across the N images. Some prior works in monocular depth estimation address this issue by estimating scale differences either with respect to the ground truth [79] or across the predicted point maps [30]. Rather than resolving scale on a per-image basis, π^3 [82] assumes that all predicted point maps share a single, unknown but consistent scale factor across the multi-view images. We follow this scaling and align the predicted point maps $\{\hat{\mathbf{P}}_i\}_{i=1}^N$ to the corresponding ground-truth set $\{\mathbf{P}_i\}_{i=1}^N$ by solving for a single optimal scale factor s^* [79] as follows,

$$s^* = \arg\min_s \sum_{i=1}^{N} \sum_{j=1}^{HW} \frac{1}{\mathbf{D}_{i,j}} \|s\hat{\mathbf{P}}_{i,j} - \mathbf{P}_{i,j}\|_1, \tag{2}$$

where $\hat{\mathbf{P}}_{i,j}, \mathbf{P}_{i,j} \in \mathbb{R}^3$ denote the predicted and ground-truth 3D points at index j of the point maps $\hat{\mathbf{P}}_i$ and $\mathbf{P}_i$, respectively. Here, instead of using perpendicular depth as the weight in the ℓ_1 loss, we utilize radial distance $\mathbf{D}_{i,j}$ to accommodate arbitrary camera models. The resulting weighted ℓ_1 loss over the entire image sequence ensures consistent scale alignment across all views. Using this optimal scale term, the point map loss is defined as:

$$\mathcal{L}_{\text{points}} = \frac{1}{3NHW} \sum_{i=1}^{N} \sum_{j=1}^{HW} \frac{1}{\mathbf{D}_{i,j}} \|s^*\hat{\mathbf{P}}_{i,j} - \mathbf{P}_{i,j}\|_1. \tag{3}$$

We incorporate a normal loss to enforce local surface smoothness by minimizing the angular difference between predicted normals $\hat{\mathbf{n}}_{i,j}$ and ground truth normals $\mathbf{n}_{i,j}$. Normals are computed using an 8-neighbor convention on the point map [32, 83], by taking the cross products of vectors formed with neighboring points on the image grid,

$$\mathcal{L}_{\text{normal}} = \sum_{i=1}^{N} \sum_{j=1}^{HW} \arccos(\hat{\mathbf{n}}_{i,j} \cdot \mathbf{n}_{i,j}). \tag{4}$$

Our network predicts the camera pose $\hat{\mathbf{T}}_i$ corresponding to each image $\mathbf{I}_i$. The relative pose is computed between the different frames:

$$\hat{\mathbf{T}}_{i \leftarrow j} = \hat{\mathbf{T}}_i^{-1} \hat{\mathbf{T}}_j, \tag{5}$$

where $\hat{\mathbf{T}}_{i \leftarrow j}$ consists of a rotation $\hat{\mathbf{R}}_{i \leftarrow j} \in SO(3)$ and a translation $\hat{\mathbf{t}}_{i \leftarrow j} \in \mathbb{R}^3$. The camera pose loss is defined as a weighted average of a rotation loss term and a translation loss term:

$$\mathcal{L}_{\text{pose}} = \frac{1}{N(N-1)} \sum_{i \neq j} \big(\mathcal{L}_{\text{rot}}(i,j) + \lambda_{\text{trans}} \mathcal{L}_{\text{trans}}(i,j) \big). \tag{6}$$

The rotation loss minimizes the geodesic distance between the predicted relative rotation $\hat{\mathbf{R}}_{i \leftarrow j}$ and its ground truth counterpart $\mathbf{R}_{i \leftarrow j}$. To align the predicted 3D point maps $\{\hat{\mathbf{P}}_i\}_{i=1}^N$ and camera poses $\{\hat{\mathbf{T}}_i\}_{i=1}^N$ in a shared global coordinate, we apply the optimal scale factor s^* to the predicted relative camera translations $\hat{\mathbf{t}}_{i \leftarrow j}$ so that they are aligned with the ground truth $\mathbf{t}_{i \leftarrow j}$.

$$\begin{cases} \mathcal{L}_{\text{rot}}(i,j) = \arccos\left(\dfrac{\text{Tr}\big((\mathbf{R}_{i \leftarrow j})^\top \hat{\mathbf{R}}_{i \leftarrow j}\big) - 1}{2} \right) \\ \mathcal{L}_{\text{trans}}(i,j) = \ell_\delta(s^* \hat{\mathbf{t}}_{i \leftarrow j} - \mathbf{t}_{i \leftarrow j}) \end{cases} \tag{7}$$

where $\ell_\sigma(\cdot)$ denotes the Huber loss, which is used to reduce the influence of outliers.

3.3 Ray-Based Feed-Forward 3D Reconstruction

All existing feed-forward 3D reconstruction methods [76, 82] adopt point maps as their output representation, and accordingly employ a dedicated point-map head in the network architecture. While effective for standard perspective cameras, this design choice limits the ability to generalize across diverse camera models, especially those with wide fields of view. To address this limitation, we introduce a multi-frame feed-forward 3D reconstruction method tailored for wide field-of-view imagery. Building on UniK3D [56], we represent different camera models using spherical harmonics (SH) within a pencil-of-rays formulation, enabling a compact and accurate angular parameterization well suited for wide field-of-view

imagery. This formulation establishes a bijective mapping between the polar angle θ and azimuthal angle ϕ in spherical coordinates and 3D Cartesian directions. By leveraging a predefined SH basis $Y^{lm}(\cdot,\cdot)$, arbitrary camera geometries can be implicitly approximated using a compact set of SH coefficients. Inspired by this representation, we replace the conventional point-map head with two decoupled components: an Angular Module and a Radial Module. Furthermore, we introduce a novel trainable camera token $\mathbf{C}_i$, which acts as a prompt to explicitly differentiate camera models while being implicitly integrated into the network architecture. The Angular Module predicts SH coefficients $\hat{\boldsymbol{k}}_i^{lm}(\cdot)$ conditioned on the camera token $\mathbf{C}_i$. Given these coefficients and the basis, the pencil of rays $\hat{\boldsymbol{\mathcal{R}}}_i(\cdot,\cdot)$ is computed as:

$$\hat{\boldsymbol{\mathcal{R}}}_i(\theta,\phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \hat{\boldsymbol{k}}_i^{lm}(\mathbf{C}_i)\; Y^{lm}(\theta,\phi), \tag{8}$$

where l and m denote the degree and order of the harmonics, respectively. The Radial Module then estimates the radial distance $\hat{\mathbf{D}}_i$ and its associated uncertainty $\hat{\mathbf{U}}_i$ along each ray $\hat{\boldsymbol{\mathcal{R}}}_i$. Using these two modules, the predicted 3D point map can be expressed as $\hat{\mathbf{P}}_i = \hat{\boldsymbol{\mathcal{R}}}_i \odot \hat{\mathbf{D}}_i$ with these separate modules. For the ray loss, to compensate for the limited number of wide field-of-view images during training, we adopt an asymmetric formulation following [56],

$$\begin{aligned} \mathcal{L}_{\text{ray}} &= \beta \mathcal{L}_\theta^{0.7} + (1-\beta)\mathcal{L}_\phi^{0.5}, \\ \text{with} \quad \mathcal{L}_\omega^\alpha &= \alpha \sum_{\hat{\omega}>\omega^*} |\hat{\omega}-\omega| + (1-\alpha) \sum_{\hat{\omega}\le\omega} |\hat{\omega}-\omega^*|\,, \end{aligned} \tag{9}$$

where $\omega \in \{\theta,\phi\}$ and $\beta = 0.75$. The radial loss is defined as the L1 loss between the predicted radial distance $\hat{\mathbf{D}}_i$ and the ground-truth distance $\mathbf{D}_i$. The uncertainty loss is the L1 loss between the radial loss and the predicted uncertainty,

$$\mathcal{L}_{\text{rad}} = \|\hat{\mathbf{D}}_i - \mathbf{D}_i\|, \tag{10}$$

$$\mathcal{L}_{\text{uncer}} = \left\| |\hat{\mathbf{D}}_i - \mathbf{D}_i| - \hat{\mathbf{U}}_i \right\|. \tag{11}$$

3.4 Training

Training Losses All loss terms are combined using a weighted sum, and the model is trained end-to-end by minimizing the total loss function. We set $\lambda_{\text{normal}} = 10$, $\lambda_{\text{pose}} = 0.1$, $\lambda_{\text{ray}} = 1.0$, $\lambda_{\text{rad}} = 1.0$, and $\lambda_{\text{uncer}} = 0.1$ in all experiments,

$$\begin{aligned} \mathcal{L}_{\text{total}} \;=\; & \mathcal{L}_{\text{points}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} \\ & + \lambda_{\text{ray}}\mathcal{L}_{\text{ray}} + \lambda_{\text{rad}}\mathcal{L}_{\text{rad}} + \lambda_{\text{uncer}}\mathcal{L}_{\text{uncer}}. \end{aligned} \tag{12}$$

Training Datasets We train our model using a large and diverse collection of datasets that span a wide field-of-view camera configurations, including TartanAirV2 [81], ASE [16], Hypersim [63], KITTI-360 [44], 360Loc [28], Matterport3D [7], ScanNet++ [84], EDEN [42], and VKITTI [5]. Collectively, these

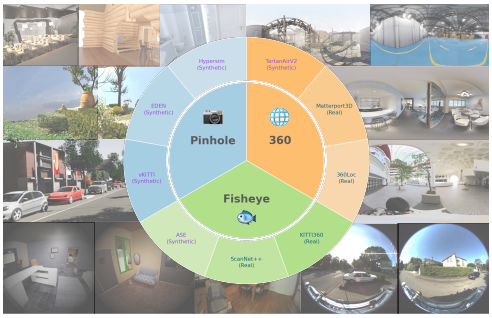

Fig. 3: Composition of camera models and training datasets. The diagram illustrates diverse camera model configurations, including pinhole, fisheye, and 360° cameras, along with representative scenes. Wid3R unifies these camera models within a single network, enabling robust reconstruction under wide field-of-view settings.

datasets cover three camera categories (pinhole, fisheye, and 360° cameras), with multiple datasets available for each configuration, enabling training across diverse camera projection models, as shown in Fig. 3. Among these datasets, five are synthetic and four are captured from real-world environments, providing a balance between large-scale, noise-free supervision and realistic scene complexity.

Since most of the training data consists of large-scale scenes, it is important to sample input images within each batch such that sufficient visual overlap is maintained. To this end, we compute pairwise distances between all images based on their camera positions, forming a 2D distance matrix. We then apply a softmax operation to the negative distances to obtain a probability matrix, which is used to sample input images with higher probability for closer viewpoints (see the Supplementary Material for details). Notably, training data for 360° cameras is significantly more limited than that for other camera types, accounting for less than 1% of the total number of images. To better understand this imbalance, we further analyze how incorporating data from different camera models influences performance under 360° camera settings. The results of this analysis are reported in Table 7.

Augmentations During training, we randomly apply a set of data augmentations that include both geometric and appearance transformations, such as random resizing and cropping, horizontal flipping, and photometric adjustments of brightness, contrast, saturation, gamma, and hue. In addition, to increase data diversity for wide field-of-view images, we adopt a camera augmentation strategy [56] that transforms pinhole camera images into fisheye camera models. Specifically, we first unproject the 2D depth map obtained under a pinhole camera model into a 3D point cloud, and then reproject the points onto the 2D image plane of a randomly sampled distorted camera model. To preserve fine details during this process, we also apply softmax-based splatting [52] to warp both images and depth maps. For equirectangular projection (ERP) images, we further exploit the fact that a single ERP image represents the full 360° viewing sphere [31]. We apply rotation-based augmentation by randomly sampling the azimuth angle $\in [0, 2\pi]$ and the elevation angle $\in [0, \pi]$, and rotating the image, depth and pose accordingly. To maintain geometric consistency, the same augmentations are applied consistently to all multi-view input images from the same scene within a batch.

Table 2: Zero-shot pose estimation on wide field-of-view images. Metrics report the ratio of samples with rotation and translation errors within 30° (higher is better). Ours denotes models trained with our composed wide-angle training data, implemented on the π^3 baseline (Ours (π^3)) and our ray-based modules (Ours (Wid3R)).

Method	FIORD [22]			Zip-NeRF [12]			Stanford2D3D [4]		
	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑
VGGT [76]	82.34	84.01	44.63	64.42	65.29	20.68	19.30	33.14	2.60
π^3 [82]	85.48	87.42	48.35	82.65	86.66	40.94	19.94	31.99	2.06
Ours (π^3 [82])	99.87	**96.39**	**72.56**	83.04	89.48	64.44	82.91	87.83	62.14
Ours (Wid3R)	**100.0**	93.38	68.20	**93.34**	**95.15**	**74.61**	**94.05**	**93.29**	**79.93**

Table 3: Zero-shot pose estimation on wide field-of-view images. Metrics measure rotation and translation distance errors (lower is better).

Method	FIORD [22]			Zip-NeRF [12]			Stanford2D3D [4]		
	ATE ↓	RPE trans ↓	RPE rot ↓	ATE ↓	RPE trans ↓	RPE rot ↓	ATE ↓	RPE trans ↓	RPE rot ↓
VGGT [76]	0.54	0.99	15.02	1.83	1.73	34.41	3.07	5.01	98.83
π^3 [82]	0.53	0.90	13.45	0.85	0.87	18.55	2.83	5.01	91.13
Ours (π^3 [82])	**0.32**	**0.51**	3.99	0.85	0.67	12.93	1.43	2.21	32.94
Ours (Wid3R)	0.44	0.71	**3.27**	**0.49**	**0.47**	**7.72**	**1.11**	**1.83**	**18.99**

4 Implementation and Performance

4.1 Experiment Settings

Implementation Details The networks are optimized for geometric reconstruction using the Adam optimizer [39] on RTX A6000 or H100 GPUs, starting with an initial learning rate of 5×10^{-5} and applying exponential decay at each iteration. Input images are resized to 518×336. We initialize the π^3 encoder, feature aggregation module, and pose header from pretrained weights and train all components end-to-end for faster convergence.

4.2 Camera Pose Estimation

Pose Evaluation We perform zero-shot camera pose evaluation on wide field-of-view cameras, including fisheye cameras on FIORD [22] and Zip-NeRF [12], and 360° cameras on Stanford2D3D [4]. For Stanford2D3D, we randomly select 10-30 images from each scene in *area_ 5a* and *area_ 5b*, repeating this sampling process 10 times per scene to construct a total of 20 cases. For FIORD, we randomly select an initial frame and then sample frames at an interval of five, resulting in 4–24 images per sequence from the *Kitchen_ In*, *meetingroom*, and *parakennus* scenes. For Zip-NeRF, we follow the official test split provided in [55]. Following [82], we evaluate camera pose accuracy using Relative Rotation Accuracy (RRA) and Relative Translation Accuracy (RTA) at predefined thresholds. We further report the Area Under the Curve (AUC) of the $\min(\text{RRA}, \text{RTA})$ threshold curve as a unified performance metric, as summarized in Table 2. In addition, we report Absolute Trajectory Error (ATE), Relative Pose Error for translation (RPE trans), and Relative Pose Error for rotation (RPE rot) in Table 3. Compared

Table 4: Quantitative evaluation of large-scale localization performance on the Matterport3D dataset [7]. Our method demonstrates superior registration accuracy and achieves comparable performance in large-scale scene pose estimation with significantly reduced runtime.

Method	Matterport3D Scene1				Matterport3D Scene2				Matterport3D Scene3				Time
	# Registered	AUC @3°	AUC @5°	AUC @10°	# Registered	AUC @3°	AUC @5°	AUC @10°	# Registered	AUC @3°	AUC @5°	AUC @10°	
OpenMVG [51]	5 / 37	0.82	1.10	1.30	2 / 20	0.27	0.37	0.45	25 / 31	45.11	52.82	58.67	~ 3min
SPSG [10,64] + COLMAP [66]	16 / 37	14.30	15.79	16.90	12 / 20	26.64	29.88	32.31	**31 / 31**	75.02	85.01	92.51	~ 5min
SphereGlue [20] + COLMAP [66]	21 / 37	23.95	26.98	29.26	12 / 20	23.11	27.76	31.24	**31 / 31**	66.83	80.06	90.03	~ 5min
DKM [13] + COLMAP [66]	22 / 37	25.14	27.70	29.62	9 / 20	14.64	16.36	17.66	**31 / 31**	**75.94**	**85.56**	**92.78**	~ 30min
EDM [31] + IM360 [33]	**37 / 37**	**49.16**	**69.05**	**84.53**	**20 / 20**	34.91	44.16	66.87	**31 / 31**	73.95	84.37	92.18	~ 30min
Ours (Wid3R)	**37 / 37**	31.64	51.02	72.81	**20 / 20**	**51.22**	**69.66**	**84.77**	**31 / 31**	64.74	78.77	89.40	**~ 3s**

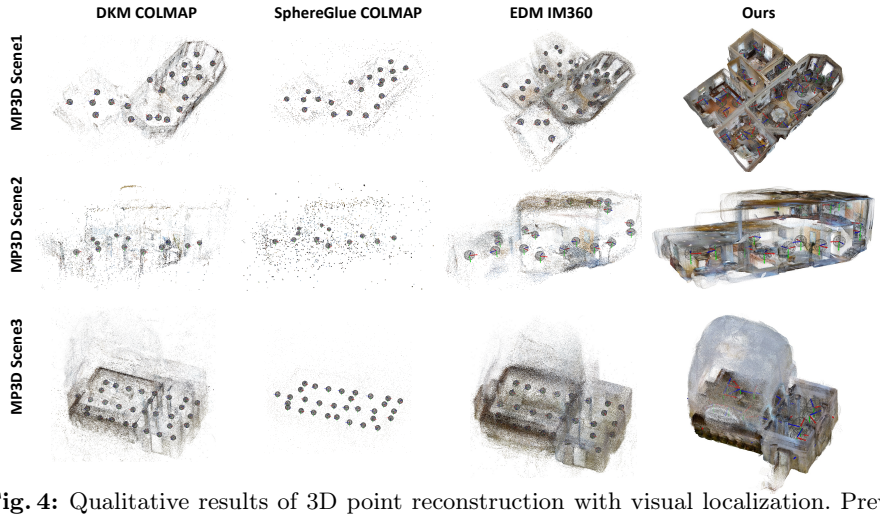

Fig. 4: Qualitative results of 3D point reconstruction with visual localization. Previous methods estimate 3D points through triangulation, whereas our method directly predicts them in a feed-forward manner. Our approach produces more complete and consistent reconstructions across large-scale Matterport3D (MP3D) [7] scenes.

to other state-of-the-art methods [76,82], Wid3R significantly outperforms prior approaches in camera pose estimation, demonstrating superior accuracy and robustness for wide field-of-view camera settings.

Large-scale Localization Following [33], we report AUC scores computed from the maximum of the relative rotation and translation errors across all image pairs. We compare our results with state-of-the-art methods in Table 4 on three scenes from the Matterport3D dataset. Existing localization approaches for 360° imagery predominantly rely on structure-from-motion (SfM) pipelines involving triangulation [24] and bundle adjustment [70]. While IM360 [33] achieves accurate localization performance through SfM, it requires dense matching EDM [31] across all image pairs, which introduces significant computational bottlenecks and leads to long processing times. In contrast, Wid3R adopts a feed-forward formulation and can produce 3D maps within three seconds even in large-scale indoor environments, demonstrating both computational efficiency and scalability. Figure 4 demonstrates a qualitative comparison of the visual results.

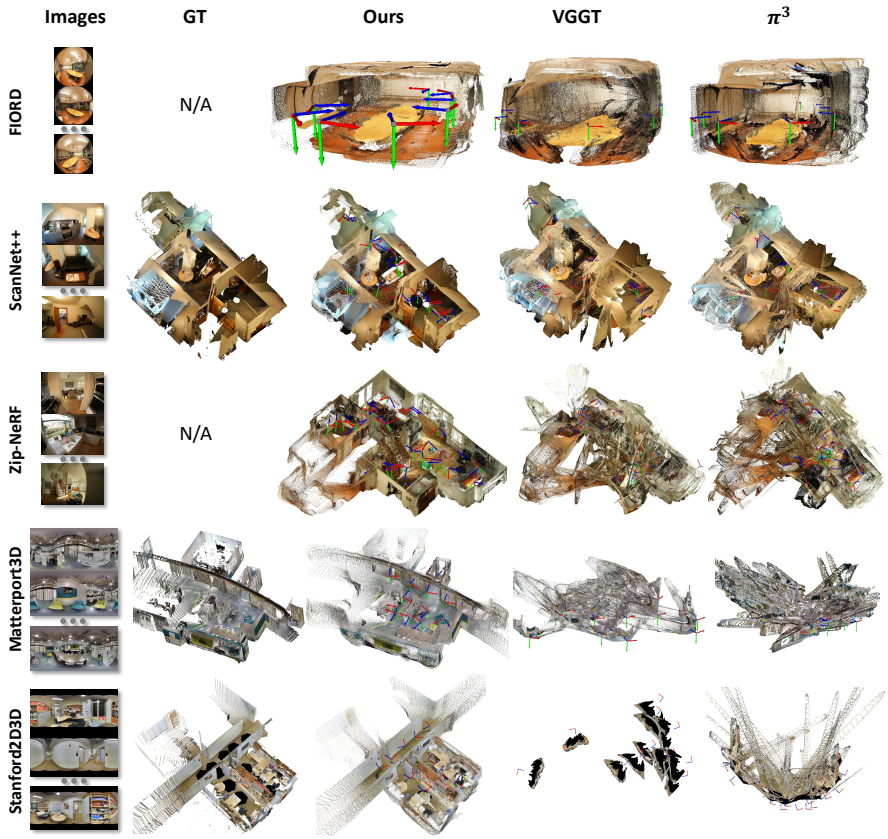

Fig. 5: Visualization of feed-forward 3D reconstruction on wide field-of-view images. Our method shows robust performance on fisheye images from FIORD [22], Zip-NeRF [12], ScanNet++ [84], and 360° images from Matterport3D [7] and Stanford2D3D [4].

4.3 Point Map Estimation

To evaluate point map estimation under wide field-of-view cameras, we conduct experiments on the ScanNet++ [84], Matterport3D [7], and Stanford2D3D [4] datasets. For ScanNet++, we use the official *nvs_sem_val* split and randomly sample 20–30 images per scene from 50 scenes. For Matterport3D, we follow the official test split consisting of 18 scenes and sample images to ensure sufficient visual overlap using the probability matrix described in Section 3.4. For Stanford2D3D, we use the same test cases as those described in Section 4.2. For point cloud alignment, we sequentially apply Umeyama alignment [72], optimal point alignment [79], followed by Iterative Closest Point. Following [78, 82], we report Accuracy (Acc.), Completion (Comp.), and Normal Consistency (N.C.) as evaluation metrics. Table 5 demonstrates that Wid3R achieves superior performance, particularly under 360° camera settings. We further compare Wid3R with Ours (π^3 [82]), which is designed based on the π^3 architecture without the ray representation and camera model tokens, showing that our method provides clear

Table 5: Point map estimation on wide field-of-view images. Ours denotes models trained with our composed wide-angle training data, implemented on the π^3 baseline (Ours (π^3)) and our ray-based module (Ours (Wid3R)).

Method	ScanNet++ [84]						Matterport3D [7]						Stanford2D3D [4]					
	Acc. ↓		Comp. ↓		N.C. ↑		Acc. ↓		Comp. ↓		N.C. ↑		Acc. ↓		Comp. ↓		N.C. ↑	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
VGGT [76]	0.135	0.093	0.068	0.031	0.704	0.855	0.327	0.258	1.756	1.494	0.530	0.536	0.387	0.362	2.115	1.812	0.536	0.546
π^3 [82]	0.086	0.052	0.037	0.019	0.739	0.904	0.315	0.270	1.308	1.092	0.539	0.555	0.380	0.316	0.745	0.574	0.557	0.598
Ours (π^3 [82])	0.027	0.011	0.014	0.007	0.781	0.930	0.161	0.080	0.142	0.061	0.610	0.702	0.211	0.118	0.273	0.167	0.596	0.686
Ours (Wid3R)	**0.018**	**0.008**	**0.012**	**0.006**	**0.803**	**0.951**	**0.094**	**0.049**	**0.087**	**0.035**	**0.790**	**0.928**	**0.197**	**0.099**	**0.172**	**0.101**	**0.729**	**0.868**

Table 6: Quantitative comparison of monocular depth estimation performance on the Matterport3D (M) test sets. (a) denotes models trained directly on the Matterport3D training set, and (b) denotes methods based on foundation models. Among foundation model based methods, DAC and UniK3D require training the model itself, while 360MonoDepth and RPG360 estimates depth via post optimization. Although all of these methods are designed for single view inputs, Wid3R achieves comparable performance in single view 360° depth estimation to state of the art methods.

	Method	Backbone	360	Train.	Opti.	Dataset Train → Test	Lower is better Abs Rel	RMSE	Higher is better $\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
(a)	SliceNet [58]	ResNet50 [25]	✓	✗		M → M	0.176	0.613	0.872	0.948	0.972
	UniFuse [29]	ResNet34 [25]	✓	✗		M → M	0.106	0.494	0.890	0.962	0.983
	EGFormer [86]	Transformer [11]	✓	✗		M → M	0.147	0.603	0.816	0.939	0.974
	ACDNet [88]	ResNet50 [25]	✓	✗		M → M	0.101	0.463	0.900	0.968	0.988
	BiFuse++ [74]	ResNet34 [25]	✓	✗		M → M	0.112	0.485	0.881	0.966	0.987
	HRDFuse [2]	ResNet34 [25]	✓	✗		M → M	0.117	0.503	0.867	0.962	0.985
	Elite360D [3]	EfficientNet-B5 [69]	✓	✗		M → M	0.105	**0.452**	0.899	0.971	**0.991**
	Depth Anywhere [77]	BiFuse++ [74]	✓	✗		M → M	**0.085**	-	0.917	**0.976**	**0.991**
(b)	360MonoDepth [62]	MiDaS v2 [60]	✗	✓		→ M	0.264	0.916	0.612	0.854	0.941
	DAC [23]	ResNet101 [25]	✓	✗		→ M	0.156	0.619	0.773	0.956	0.982
	UniK3D [56]	DINOv2 [53]	✓	✗		→ M	0.315	0.649	0.858	0.957	0.977
	RPG360 [34]	Metric3D v2 [27]	✗	✓		→ M	0.203	0.667	0.859	0.953	0.977
	Ours (Wid3R)	DINOv2 [53]	✓	✗		→ M	0.227	0.562	**0.948**	**0.976**	0.984

advantages. Qualitative results are presented in Fig. 5, illustrating the robustness of Wid3R across diverse camera types. Notably, while prior methods exhibit performance degradation on distorted imagery, our approach enables accurate estimation even for 360° images.

4.4 Monocular 360 Depth Estimation

We conduct monocular panoramic depth estimation evaluation on Matterport3D [7], as in [15, 33]. Table 6 presents the quantitative results. Wid3R achieves comparable or superior performance to recent monocular 360° depth estimation methods, even though it is trained in a multi-view setting and uses not only 360° cameras but also pinhole and fisheye cameras.

4.5 Ablation Study

To validate the effectiveness of our proposed method, we conduct an ablation study using point map estimation metrics on Stanford2D3D [4], with results reported in Table 7. We first analyze the impact of training data composition. Compared to other camera types, 360° datasets contain significantly fewer geometry-labeled samples, accounting for less than 1% of the total training data.

Table 7: Ablation study on Stanford2D3D [4] under the 360° setting, which is not included in the training data. We analyze the effects of camera model tokens, training data composition (pinhole, fisheye, 360), and the number of frames per batch using point map estimation metrics. Notably, this study demonstrates that training with diverse camera models improves performance on 360° images.

Study	Frames per batch	Camera Token Condition	Training Data			Acc. ↓		Comp. ↓		N.C. ↑		Acc. std. ↓		Comp. std. ↓		N.C. std. ↓	
			Pinhole	Fisheye	Sphere	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
(a)	$2 \sim 12$	✓	✗	✗	✓	0.220	0.142	0.355	0.235	0.710	0.827	0.187	0.135	0.416	0.311	0.130	0.197
	$2 \sim 12$	✓	✓	✗	✓	0.254	0.128	0.166	0.108	0.698	0.839	0.099	0.107	0.243	0.206	0.099	0.149
	$2 \sim 12$	✓	✗	✓	✓	0.593	0.126	**0.120**	**0.056**	0.686	0.822	0.101	0.109	**0.136**	**0.071**	0.100	0.173
	$2 \sim 12$	✓	✓	✓	✓	0.248	0.131	0.146	0.074	0.706	0.842	**0.089**	0.103	0.164	0.090	**0.084**	**0.134**
(b)	$2 \sim 12$	✗	✓	✓	✓	0.284	0.145	0.152	0.089	0.680	0.803	0.103	0.095	0.178	0.147	0.091	0.152
(c)	$2 \sim 24$	✓	✓	✓	✓	**0.197**	**0.099**	0.172	0.101	**0.729**	**0.868**	0.113	**0.089**	0.257	0.216	0.095	0.143

To address this imbalance, prior work has proposed generating pseudo-labels using large-scale perspective models [77] or applying optimization-based post-processing methods [34, 62]. In contrast, we investigate whether incorporating data from other camera models can improve performance on 360° cameras. As shown in Table 7(a), jointly training with additional camera model datasets improves geometry estimation performance on 360° data while reducing performance variance, compared to training solely on 360° images. This indicates that, while camera model tokens provide explicit conditioning, exposing the shared network parameters to more diverse training data mitigates overfitting caused by limited 360° supervision and leads to more stable and robust learning. We further observe in Table 7(b) that incorporating camera model tokens as prompts consistently improves performance over models trained without them, highlighting the benefit of explicitly encoding camera configuration. Finally, we examine the effect of the maximum number of frames per batch in Table 7(c). While hardware constraints limit the maximum number to 12 frames on an RTX A6000 GPU, up to 24 frames can be used on an H100 GPU. Increasing the number of multi-view images per batch improves Acc. and N.C., indicating that larger numbers of frames during training enable the model to learn richer multi-view attention patterns and yield more accurate geometry estimation.

5 Conclusion, Limitations, and Future Work

We introduce a novel feed-forward 3D reconstruction method designed for wide field-of-view camera models. To address the inherent distortions of large field-of-view imagery, our approach incorporates ray-based geometry representation modules and camera model tokens as explicit conditioning prompts. We further demonstrate that our method achieves superior accuracy in monocular depth estimation, visual localization, and 3D reconstruction, with particularly strong performance on 360° datasets, which are challenging for large-scale model training due to the limited availability of labeled benchmarks. One limitation of our approach is that, while it supports wide field-of-view inputs and large-scale environments, it does not explicitly model dynamic scene elements. As future work, we plan to extend the framework by incorporating datasets with dynamic content and designing additional modules to handle scene dynamics.

Acknowledgment

This work is partially supported by NVIDIA Academic Grant. This work was supported by the Korea Planning & Evaluation Institute of Industrial Technology(KEIT) and the Ministry of Trade, Industry & Resources(MOTIR) of the Republic of Korea (RS-2024-00417108).

References

1. Abbas, S.M., Aslam, S., Berns, K., Muhammad, A.: Analysis and improvements in apriltag based state estimation. Sensors **19**(24), 5480 (2019) 4
2. Ai, H., Cao, Z., Cao, Y.P., Shan, Y., Wang, L.: Hrdfuse: Monocular 360deg depth estimation by collaboratively learning holistic-with-regional depth distributions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13273–13282 (2023) 13
3. Ai, H., Wang, L.: Elite360d: Towards efficient 360 depth estimation via semantic- and distance-aware bi-projection fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9926–9935 (2024) 13
4. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017) 2, 3, 10, 12, 13, 14
5. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2. arXiv preprint arXiv:2001.10773 (2020) 8
6. Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., Leroy, V.: Must3r: Multi-view network for stereo 3d reconstruction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1050–1060 (2025) 1
7. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158 (2017) 2, 3, 8, 11, 12, 13
8. Clarke, T.A., Fryer, J.G.: The development of camera calibration methods and models. The Photogrammetric Record **16**(91), 51–66 (1998) 4
9. Courbon, J., Mezouar, Y., Eckt, L., Martinet, P.: A generic fisheye camera model for robotic applications. In: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1683–1688. IEEE (2007) 2
10. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 224–236 (2018) 3, 11
11. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4, 13
12. Duckworth, D., Hedman, P., Reiser, C., Zhizhin, P., Thibert, J.F., Lučić, M., Szeliski, R., Barron, J.T.: Smerf: Streamable memory efficient radiance fields for real-time large-scene exploration (2023) 2, 3, 10, 12
13. Edstedt, J., Athanasiadis, I., Wadenbäck, M., Felsberg, M.: Dkm: Dense kernelized feature matching for geometry estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17765–17775 (2023) 3, 11
14. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19790–19800 (2024) 3

15. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014) 13
16. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., et al.: Project aria: A new tool for egocentric multi-modal ai research. arXiv preprint arXiv:2308.13561 (2023) 2, 8
17. Fischler, M.A., Bolles, R.C.: 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) 3
18. Furukawa, Y., Hernández, C., et al.: Multi-view stereo: A tutorial. Foundations and trends® in Computer Graphics and Vision **9**(1-2), 1–148 (2015) 3
19. Gangopadhyay, S., Kim, J.H., Chen, X., Rim, P., Park, H., Wong, A.: Extending foundational monocular depth estimators to fisheye cameras with calibration tokens. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5198–5209 (2025) 5
20. Gava, C., Mukunda, V., Habtegebrial, T., Raue, F., Palacio, S., Dengel, A.: Sphereglue: Learning keypoint matching on high resolution spherical images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6134–6144 (2023) 11
21. Geyer, C., Daniilidis, K.: A unifying theory for central panoramic systems and practical implications. In: European conference on computer vision. pp. 445–461. Springer (2000) 4
22. Gunes, U., Turkulainen, M., Ren, X., Solin, A., Kannala, J., Rahtu, E.: Fiord: A fisheye indoor-outdoor dataset with lidar ground truth for 3d scene reconstruction and benchmarking. In: Scandinavian Conference on Image Analysis. pp. 3–17. Springer (2025) 3, 10, 12
23. Guo, Y., Garg, S., Miangoleh, S.M.H., Huang, X., Ren, L.: Depth any camera: Zero-shot metric depth estimation from any camera. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26996–27006 (2025) 2, 13
24. Hartley, R.: Multiple view geometry in computer vision, vol. 665. Cambridge university press (2003) 3, 11
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 13
26. He, X., Sun, J., Wang, Y., Peng, S., Huang, Q., Bao, H., Zhou, X.: Detector-free structure from motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21594–21603 (2024) 3
27. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. arXiv preprint arXiv:2404.15506 (2024) 2, 13
28. Huang, H., Liu, C., Zhu, Y., Cheng, H., Braud, T., Yeung, S.K.: 360loc: A dataset and benchmark for omnidirectional visual localization with cross-device queries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22314–22324 (2024) 8
29. Jiang, H., Sheng, Z., Zhu, S., Dong, Z., Huang, R.: Unifuse: Unidirectional fusion for 360 panorama depth estimation. IEEE Robotics and Automation Letters **6**(2), 1519–1526 (2021) 13
30. Jung, D., Choi, J., Lee, Y., Eum, S., Kwon, H., Manocha, D.: More: Monocular geometry refinement via graph optimization for cross-view consistency. arXiv preprint arXiv:2510.07119 (2025) 6

31. Jung, D., Choi, J., Lee, Y., Jeong, S., Lee, T., Manocha, D., Yeon, S.: Edm: Equirectangular projection-oriented dense kernelized feature matching. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6337–6347 (2025) 2, 3, 9, 11
32. Jung, D., Choi, J., Lee, Y., Kim, D., Kim, C., Manocha, D., Lee, D.: Dnd: Dense depth estimation in crowded dynamic indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12797–12807 (2021) 7
33. Jung, D., Choi, J., Lee, Y., Manocha, D.: Im360: Large-scale indoor mapping with 360 cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 29040–29050 (2025) 2, 3, 5, 11, 13
34. Jung, D., Choi, J., Lee, Y., Manocha, D.: Rpg360: Robust 360 depth estimation with perspective foundation models and graph optimization. arXiv preprint arXiv:2509.23991 (2025) 2, 13, 14
35. Kannala, J., Brandt, S.S.: A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. IEEE transactions on pattern analysis and machine intelligence **28**(8), 1335–1340 (2006) 4
36. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025) 4
37. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023) 3
38. Khomutenko, B., Garcia, G., Martinet, P.: An enhanced unified camera model. IEEE Robotics and Automation Letters **1**(1), 137–144 (2015) 4
39. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 10
40. Kumar, V.R., Eising, C., Witt, C., Yogamani, S.K.: Surround-view fisheye camera perception for automated driving: Overview, survey & challenges. IEEE Transactions on Intelligent Transportation Systems **24**(4), 3638–3659 (2023) 2
41. Kumar, V.R., Yogamani, S., Bach, M., Witt, C., Milz, S., Mäder, P.: Unrectdepthnet: Self-supervised monocular depth estimation using a generic framework for handling common camera distortion models. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8177–8183. IEEE (2020) 2
42. Le, H.A., Mensink, T., Das, P., Karaoglu, S., Gevers, T.: Eden: Multimodal synthetic dataset of enclosed garden scenes. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1579–1589 (2021) 8
43. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European Conference on Computer Vision. pp. 71–91. Springer (2024) 1
44. Liao, Y., Xie, J., Geiger, A.: Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3292–3310 (2022) 8
45. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025) 4
46. Lochman, Y., Liepieshov, K., Chen, J., Perdoch, M., Zach, C., Pritts, J.: Babelcalib: A universal approach to calibrating central cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15253–15262 (2021) 4

47. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**, 91–110 (2004) 3
48. Luan, W., Lu, S., Zheng, Y., Xu, W., Nie, L., Zhou, Z., Liao, K.: Lifting the structural morphing for wide-angle images rectification: Unified content and boundary modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25529–25538 (2025) 2
49. Mei, C., Rives, P.: Single view point omnidirectional camera calibration from planar grids. In: Proceedings 2007 IEEE International Conference on Robotics and Automation. pp. 3945–3950. IEEE (2007) 4
50. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 3
51. Moulon, P., Monasse, P., Perrot, R., Marlet, R.: Openmvg: Open multiple view geometry. In: International Workshop on Reproducible Research in Pattern Recognition. pp. 60–74. Springer (2016) 11
52. Niklaus, S., Liu, F.: Softmax splatting for video frame interpolation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5437–5446 (2020) 9
53. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 13
54. Pan, L., Baráth, D., Pollefeys, M., Schönberger, J.L.: Global structure-from-motion revisited. In: European Conference on Computer Vision (ECCV) (2024) 3
55. Pataki, Z., Sarlin, P.E., Schönberger, J.L., Pollefeys, M.: Mp-sfm: Monocular surface priors for robust structure-from-motion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21891–21901 (2025) 10
56. Piccinelli, L., Sakaridis, C., Segu, M., Yang, Y.H., Li, S., Abbeloos, W., Van Gool, L.: Unik3d: Universal camera monocular 3d estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1028–1039 (2025) 2, 5, 7, 8, 9, 13
57. Piccinelli, L., Yang, Y.H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F.: Unidepth: Universal monocular metric depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10106–10116 (2024) 5
58. Pintore, G., Agus, M., Almansa, E., Schneider, J., Gobbetti, E.: Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11536–11545 (2021) 13
59. Pollefeys, M., Van Gool, L., ESAT-PSI, K.: Some issues on self-calibration and critical motion sequences. Asian Confer (2000) 4
60. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021) 4, 13
61. Revaud, J., De Souza, C., Humenberger, M., Weinzaepfel, P.: R2d2: Reliable and repeatable detector and descriptor. Advances in neural information processing systems **32** (2019) 3
62. Rey-Area, M., Yuan, M., Richardt, C.: 360monodepth: High-resolution 360deg monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3762–3772 (2022) 13, 14

63. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10912–10922 (2021) 8
64. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020) 3, 11
65. Scaramuzza, D.: Omnidirectional camera. In: Computer vision: A reference guide, pp. 900–909. Springer (2021) 4
66. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016) 3, 11
67. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: European conference on computer vision. pp. 501–518. Springer (2016) 3
68. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8922–8931 (2021) 3
69. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019) 13
70. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment—a modern synthesis. In: International workshop on vision algorithms. pp. 298–372. Springer (1999) 11
71. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment—a modern synthesis. In: Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings. pp. 298–372. Springer (2000) 3
72. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on pattern analysis and machine intelligence **13**(4), 376–380 (2002) 12
73. Usenko, V., Demmel, N., Cremers, D.: The double sphere camera model. In: 2018 International Conference on 3D Vision (3DV). pp. 552–560. IEEE (2018) 4
74. Wang, F.E., Yeh, Y.H., Tsai, Y.H., Chiu, W.C., Sun, M.: Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation. IEEE transactions on pattern analysis and machine intelligence **45**(5), 5448–5460 (2022) 13
75. Wang, H., Agapito, L.: 3d reconstruction with spatial memory. arXiv preprint arXiv:2408.16061 (2024) 4
76. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025) 1, 2, 4, 5, 7, 10, 11, 13
77. Wang, N.H., Liu, Y.L.: Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. arXiv preprint arXiv:2406.12849 (2024) 13, 14
78. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10510–10522 (2025) 12
79. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal

training supervision. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5261–5271 (2025) 2, 4, 6, 12
80. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024) 1, 2, 4
81. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4909–4916. IEEE (2020) 8
82. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: Pi3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025) 1, 2, 4, 6, 7, 10, 11, 12, 13
83. Yang, Z., Wang, P., Xu, W., Zhao, L., Nevatia, R.: Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence (2018) 7
84. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023) 2, 3, 8, 12, 13
85. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9043–9053 (2023) 4
86. Yun, I., Shin, C., Lee, H., Lee, H.J., Rhee, C.E.: Egformer: Equirectangular geometry-biased transformer for 360 depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6101–6112 (2023) 13
87. Zhang, S., Wang, J., Xu, Y., Xue, N., Rupprecht, C., Zhou, X., Shen, Y., Wetzstein, G.: Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21936–21947 (2025) 4
88. Zhuang, C., Lu, Z., Wang, Y., Xiao, J., Wang, Y.: Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3653–3661 (2022) 13