CAM3R: Camera-Agnostic Model for 3D Reconstruction

Namitha Guruprasad[1], Abhay Yadav[1], Rama Chellappa[1], and Cheng Peng[2]

[1] Johns Hopkins University, Baltimore, MD 21218, USA
{ngurupr1, ayadav13, rchella4}@jhu.edu
[2] University of Virginia, Charlottesville, VA 22904, USA
xuz7wn@virginia.edu

Project Page

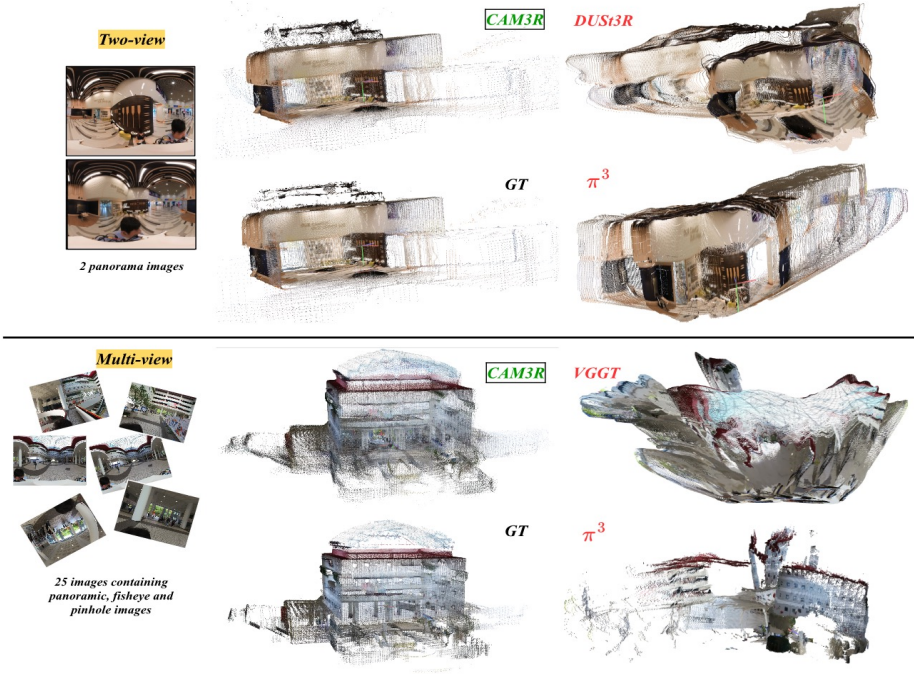

Fig. 1: CAM3R provides a robust, feed-forward 3D reconstruction for two-view and multi-view scenarios across disparate camera models, including pinhole, fisheye, and panorama, where recent 3D foundation models fail. Above, we highlight CAM3R's performance on unseen scenes from the 360Loc dataset, visualizing both raw two-view predictions and multi-view reconstructions.

Abstract. Recovering dense 3D geometry from unposed images remains a foundational challenge in computer vision. Current state-of-the-art models are predominantly trained on perspective datasets, which implicitly constrains them to a standard pinhole camera geometry. As a result, these models suffer from significant geometric degradation when applied to wide-angle imagery captured via non-rectilinear optics, such as fisheye

or panoramic sensors. To address this, we present CAM3R, a Camera-Agnostic, feed-forward Model for 3D Reconstruction capable of processing images from wide-angle camera models without prior calibration. Our framework consists of a two-view network which is bifurcated into a Ray Module (RM) to estimate per-pixel ray directions and a Cross-view Module (CVM) to infer radial distance with confidence maps, pointmaps, and relative poses. To unify these pairwise predictions into a consistent 3D scene, we introduce a Ray-Aware Global Alignment framework for pose refinement and scale optimization while strictly preserving the predicted local geometry. Extensive experiments on various camera model datasets, including panorama, fisheye and pinhole imagery, demonstrate that CAM3R establishes a new state-of-the-art in pose estimation and reconstruction.

1 Introduction

The reconstruction of 3D environments from 2D imagery remains a cornerstone task in computer vision. Recent foundation models [11, 14, 32–34] can achieve 3D reconstruction through a single feed-forward pass by learning a prior from large-scale 3D data. However, these models are grounded in the pinhole camera assumption because of its simple formulation and availability. This bias leads to significant geometric artifacts when these models are evaluated on wide-angle imagery captured via non-linear optics, such as fisheye or 360° panoramic lenses.

To mitigate this, a common solution involves undistorting wide-angle images before processing via downstream reconstruction models. This rectification process forces extreme stretching or discards peripheral regions with high curvature [28], and can also propagate errors; moreover, such a multi-stage pipeline defeats the purpose of end-to-end feed-forward models.

Recent deep learning models directly establish the relationship between a 3D imaging ray and its corresponding pixel, making minimal assumptions regarding the underlying camera model [16]. However, these methods have been confined to monocular contexts [21], and have not yet been extended to to multi-view settings.

In this paper, we propose **CAM3R**, a **C**amera-**A**gnostic feed-forward **M**odel for **3**D **R**econstruction from uncalibrated and unposed imagery across a variety of camera types. The core of our architecture is a two-view network capable of regressing a dense scene representation without requiring prior intrinsic and extrinsic metadata. Our approach simultaneously addresses two challenges: (a) handling non-linear distortions in wide-angle imagery and (b) extracting 3D contextual cues for pose estimation. For the former, we design a **Ray Module (RM)** which learns a ray representation for wide-angle lens geometries, even when the input pair originates from heterogeneous camera models. For the latter, we introduce a **Cross-view Module (CVM)**, which learns the relationship between image pairs to generate 3D representations: radial distance with confidence maps, pointmaps, and the relative pose between the two views. To integrate pairwise predictions into a globally consistent 3D reconstruction, we in-

troduce **Ray-Aware Global Alignment**. Traditional bundle adjustment often fails on distorted images due to the pinhole assumption, where pixel distances linearly correspond to 3D distances [29, 33]. Our Ray-Aware strategy lifts optimization into a 3D, ray-consistent space. Using constraints from the two-view network, we prune the scene graph to retain reliable edges and optimize camera poses and relative scales for global consistency.

In summary, our primary contributions are as follows:

- **Unified camera-agnostic framework.** We present **CAM3R**, a feed-forward pipeline that jointly performs calibration and 3D reconstruction from unposed imagery across multiple camera geometries.
- **Decoupled supervision strategy.** We introduce a supervision strategy for the Ray Module and Cross-view Module that jointly trains rays, dense pointmaps, and relative poses in a two-view setup.
- **Ray-Aware Global Alignment.** We propose a global alignment strategy that fuses local pairwise predictions into a globally consistent 3D coordinate system via ray-aware optimization, preserving structural fidelity across different lens types.

We demonstrate that our unified formulation achieves state-of-the-art results in two-view and multi-view reconstruction. Notably, CAM3R provides significant performance gains in challenging cross-modality scenarios, including large Field-of-View (FoV) and panoramic captures, while retaining state-of-the-art pinhole performance.

2 Related Work

2.1 Distorted Camera Intrinsics

Camera intrinsic calibration involves estimating the parameters necessary to establish a mapping between the two-dimensional image plane and the three-dimensional coordinate system. Classical approaches [4, 5] rely on fixed parametric forms, which prove effective for narrow-FoV sensors but struggle with extreme-FoV cameras. To support extreme-FoV geometries, many models [9, 13, 30] rely on higher-order parameters, but are prone to overfitting due to complex parameter dependencies. Popular deep-learning calibration frameworks [3,31] infer calibration parameters directly from a single image but often exhibit limited generalization and are susceptible to training-set biases. In most downstream tasks, particularly 3D reconstruction, the fundamental objective is to determine the direction and distance of the light ray that formed each pixel [16]. Building on this insight, recent works such as UniK3D [21] have shifted toward learning continuous representations that map each pixel to a 3D ray direction. However, these advancements remain largely confined to monocular settings.

2.2 Structure-from-Motion (SfM) and Multi-view Stereo (MVS)

SfM formulates 3D reconstruction by simultaneously solving for sparse geometry and the underlying camera trajectory from an unordered image collection.

Traditional frameworks, such as COLMAP [26], operate through a rigid sequential pipeline: (a) extracting and matching local features [6, 8, 18, 25], (b) solving two-view epipolar geometry, and (c) optimizing a multi-view sparse graph via global bundle adjustment [29]. To obtain dense surfaces, MVS [27] is employed to triangulate dense 3D geometry from SfM-calibrated viewpoints. However, such pipelines are highly susceptible to error propagation, where inaccuracies in the initial SfM pose estimates degrade the downstream dense reconstruction. Recently, a shift has emerged with geometric vision foundation models, such as DUSt3R [33] and VGGT [32], that bypass traditional multi-stage pipelines by directly regressing scene geometry (Cartesian pointmaps (X, Y, Z)) that implicitly couples with camera projection in a single feed-forward pass. Despite their versatility, these methods remain limited by their reliance on the pinhole camera model, which dominates the training data and can lead to degraded performance on images from wide-FoV lenses. CAM3R departs from this design along three axes. **(a) Two-view projection-invariant target.** DUSt3R encodes both the second view's camera *and* the relative pose into a single output in the first view's coordinate frame. CAM3R, instead, cleanly separates *what each camera sees* from *where each camera is*. We leverage the *radial distance* (Euclidean distance of the point in space from the camera's optical center), which is well-defined for any lens. Unlike UniK3D's monocular setting, we predict radial distance in a two-view cross-attention framework. **(b) Structurally necessary pose network.** A central challenge that DUSt3R does not address lies in integrating cross-camera radial distance in the same coordinate system for optimization. To this end, CAM3R introduces an explicit pose network to predict $(R_{2\to1}, \mathbf{t}_{2\to1})$. **(c) Confidence-weighted Global Alignment.** We replace DUSt3R's alignment with our Ray-Aware Global Alignment since camera-agnostic geometry calls for camera-agnostic optimization: the pinhole-based global alignment used by DUSt3R, in which pixel distances are assumed to vary linearly with 3D distances, is not appropriate for distorted imagery.

3 Method

In this section, we begin by providing a brief notation for a better understanding of CAM3R's architectural design. Following this, we describe our two-view feed-forward network comprising the Ray Module (RM) and the Cross-view Module (CVM), along with the multi-task supervised loss functions that govern their optimization. Finally, we introduce our Ray-Aware Global Alignment strategy to integrate the pairwise predictions from our two-view network into a globally consistent 3D reconstruction.

Notations.

- *Input Data*: Consider a pair of uncalibrated views $I_1(\mathbf{u}), I_2(\mathbf{u}) \in \mathbb{R}^{W\times H\times 3}$ captured from varying lens geometries (panorama, fisheye, pinhole), where $\mathbf{u} = (u, v)$ denotes the pixel coordinates of each image.

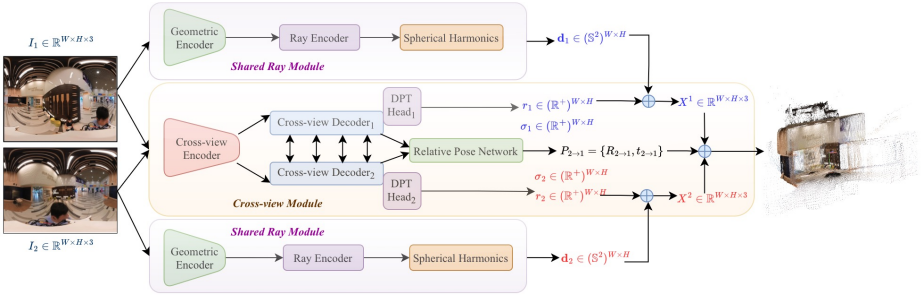

Fig. 2: CAM3R Overview. Given an input image pair (I_1, I_2), the framework operates through two parallel streams. The ***Shared Ray Module*** recovers the camera lens geometry by regressing SH coefficients to reconstruct continuous ray directional fields $\mathbf{d}_i$. The ***Cross-view Module*** extracts features and utilizes a dual transformer decoder for information exchange between the two views. Specialized DPT heads regress radial distances r_i with confidence maps σ_i; a Relative Pose Network estimates the rigid transformation $P_{2\to1}$. The local pointmaps $X^{i,i}$ are obtained by fusing rays $\mathbf{d}_i$ with radial distances r_i. The second view is transformed into the reference coordinate frame of the first view via $P_{2\to1}$ to produce the aligned 3D reconstruction.

- *Directional Rays:* Let $\mathbf{d}_i \in (\mathbb{S}^2)^{W\times H}$ denote the per-pixel rays for view i, where each unit vector $\mathbf{d}_i(u, v)$ gives the direction of a pixel. The mapping $\psi : \mathbb{R}^2 \to \mathbb{S}^2$ projects 2D pixel coordinates $\mathbf{u}$ onto the spherical domain (θ, ϕ). We expand the resulting directions in Spherical Harmonics (SH) Y_l^m of degree l and order m.
- *Radial Distance and Confidence:* Let $r_i \in (\mathbb{R}^+)^{W\times H}$ denote the per-pixel Euclidean distance from the camera's optical center along each ray $\mathbf{d}_i(\mathbf{u})$. To account for uncertainty, we predict a confidence map $\sigma_i \in (\mathbb{R}^+)^{W\times H}$, where higher values indicate greater reliability in the estimated radial distance.
- *Dense Pointmap Representation:* Let $X^{i,i} \in \mathbb{R}^{W\times H\times 3}$ be the pointmap for view i, given by the element-wise product of the ray directions and radial distances:

$$X^{i,i}(\mathbf{u}) = \mathbf{d}_i(\mathbf{u}) \cdot r_i(\mathbf{u}) \tag{1}$$

- *Coordinate Transformations:* We denote $X^{i,j}$ as the pointmap X^i expressed in the coordinate frame of camera j. Given camera poses $P_i, P_j \in SE(3)$, the transformation between frames is $X^{i,j}(\mathbf{u}) = P_j P_i^{-1} X^{i,i}(\mathbf{u})$.

3.1 CAM3R

We design CAM3R to address camera-agnostic 3D reconstruction via direct regression, drawing inspiration from UniK3D [21] and DUSt3R [33]. A detailed rationale for this network topology is provided in the supplementary material. As illustrated in Fig. 2, the network $\mathcal{F}$ consumes an image pair (I_1, I_2) and regresses per-pixel ray fields $\mathbf{d}_i$, radial distances r_i with confidence σ_i, and the relative rigid pose $P_{2\to1}$. By decoupling local ray directions from radial distance, CAM3R provides greater flexibility in handling heterogeneous camera geometries.

Ray Module (RM) serves as a backbone for estimating the internal camera geometry across pinhole, fisheye, and panoramic lens models. For each image I_i, a shared geometric encoder extracts class tokens T_i, which are then processed by a pair of Transformer Ray Encoder (T-Enc) layers. These layers output a compact set of SH expansion coefficients $\mathbf{c}_{l,m}^i$ over a finite basis Y_l^m up to a degree L. This bandwidth parameter L controls the capacity to capture high-frequency detail in the reconstructed ray directional field. A continuous ray directional field is reconstructed via an inverse SH transformation; for any pixel coordinate $\mathbf{u}$, the directional unit vector $\mathbf{d}_i(\mathbf{u})$ is given by:

$$\mathbf{d}_i(\mathbf{u}) = \frac{\sum_{l=1}^{L} \sum_{m=-l}^{l} \mathbf{c}_{l,m}^i \, Y_l^m(\psi(\mathbf{u}))}{\left\| \sum_{l=1}^{L} \sum_{m=-l}^{l} \mathbf{c}_{l,m}^i \, Y_l^m(\psi(\mathbf{u})) \right\|_2} \tag{2}$$

Cross-view Module (CVM) processes the images I_1, I_2 through a Siamese encoder to extract feature representations $F_1 = \text{Encoder}(I_1)$ and $F_2 = \text{Encoder}(I_2)$. These features are passed to a dual transformer decoder. Each decoder block facilitates information exchange via self-attention and cross-attention. The representations G for each branch are updated as:

$$G_1^b = \text{DecoderBlock}_1^b\big(G_1^{b-1}, G_2^{b-1}\big), \quad G_2^b = \text{DecoderBlock}_2^b\big(G_2^{b-1}, G_1^{b-1}\big) \tag{3}$$

A Dense Prediction Transformer (DPT) head [22] is integrated with the decoder blocks to regress the radial distance r_i and associated confidence map σ_i. We enforce positive values $r_i(\mathbf{u}) \in \mathbb{R}^+$ via an activation layer. Combining this radial distance with the per-pixel ray directions $\mathbf{d}_i(\mathbf{u})$ from the Ray Module, the 3D point coordinates $X^{i,i}(\mathbf{u})$ are obtained as in Eq. (1) in the local camera frame. To express the second view's geometry $X^{2,2}$ in the reference frame of the first view, we employ a relative pose network attached to the decoder blocks, which regresses a rigid transformation $P_{2\to1} = \{R_{2\to1}, \hat{\mathbf{t}}_{2\to1}\}$, where $R_{2\to1} \in SO(3)$ is the rotation matrix and $\hat{\mathbf{t}}_{2\to1} \in \mathbb{S}^2$ denotes a unit translation direction. The full translation vector is recovered by introducing a relative scale factor $s \in \mathbb{R}^+$ such that $\mathbf{t}_{2\to1} = s\,\hat{\mathbf{t}}_{2\to1}$. The transformation of the second pointmap $X^{2,2}$ into the coordinate system of the first view is then defined as

$$X^{2,1}(\mathbf{u}) = R_{2\to1}\, X^{2,2}(\mathbf{u}) + \mathbf{t}_{2\to1}. \tag{4}$$

3.2 Training Objectives

Asymmetric Angular Loss. To supervise the per-pixel directional rays $\mathbf{d}_i$ predicted by our model, we adopt an asymmetric angular loss inspired by [21]. Standard symmetric loss functions (*e.g.* L_2, L_1) often bias toward the most frequent modes in the training data, which are typically narrow-FoV pinhole images. To prevent an inward collapse, we employ a quantile regression framework that enforces a heavier penalty for angular underestimation than for overestimation. We decompose the predicted dense ray field $\mathbf{d}_i$ into its spherical coordinates

$(\hat{\theta}, \hat{\phi})$. Given the ground-truth angles θ^* and ϕ^*, the loss for a target quantile $\alpha \in [0, 1]$ is formulated as:

$$\mathcal{L}_{\text{AA}}^{\alpha}(\hat{\theta}, \theta^*) = \sum_{j\,:\,\hat{\theta}_j < \theta_j^*} \alpha \left|\hat{\theta}_j - \theta_j^*\right| + \sum_{j\,:\,\hat{\theta}_j \geq \theta_j^*} (1-\alpha) \left|\hat{\theta}_j - \theta_j^*\right| \tag{5}$$

The total angular objective $\mathcal{L}_A$ is defined as a weighted combination of the losses for both spherical coordinates:

$$\mathcal{L}_A = \beta\, \mathcal{L}_{\text{AA}}^{0.7}(\hat{\theta}, \theta^*) + (1-\beta)\, \mathcal{L}_{\text{AA}}^{0.5}(\hat{\phi}, \phi^*) \tag{6}$$

Local Regression Loss. To supervise the predicted 3D pointmaps $X^{i,i}$, we minimize their distance to the ground-truth geometry in their local coordinate systems. This objective ensures that the model recovers the 3D structure of each view, independent of the relative camera pose between the two views. Given the ground-truth pointmaps $\overline{X}^{i,i}$ for view $i \in \{1, 2\}$, we adopt the Mean Squared Error (MSE) formulation from [33]. Let Ω^i be the set of valid pixels with available ground truth. To resolve the scale ambiguity, we compute normalization factors for each view:

$$\eta_i = \text{mean}_{\mathbf{u}\in\Omega^i} \left\|X^{i,i}(\mathbf{u})\right\|_2, \quad \overline{\eta}_i = \text{mean}_{\mathbf{u}\in\Omega^i} \left\|\overline{X}^{i,i}(\mathbf{u})\right\|_2 \tag{7}$$

where each normalization factor is the average distance to the origin over all valid points in the respective frame. The point-wise regression loss $\mathcal{L}_{\text{regr}}$ is:

$$\mathcal{L}_{\text{regr}} = \sum_{i\in\{1,2\}} \sum_{\mathbf{u}\in\Omega^i} \left\| \frac{1}{\eta_i} X^{i,i}(\mathbf{u}) - \frac{1}{\overline{\eta}_i} \overline{X}^{i,i}(\mathbf{u}) \right\|_2^2 \tag{8}$$

Relative Pose Loss. To enforce multi-view consistency, we supervise the relative pose head by anchoring the translation magnitude to the scale of the reconstructed geometry. Let the ground-truth relative pose that aligns the second view's pointmaps to the first view's coordinate frame be $\overline{P}_{2\to1} = \{\overline{R}_{2\to1}, \overline{\mathbf{t}}_{2\to1}\}$. Because our model regresses geometry up to an unknown scale, we resolve the translation ambiguity by computing a scale factor $\tilde{s}$ as the ratio of the predicted pointmap magnitudes to the ground-truth magnitudes, detached from the gradient flow to serve as a static target. We then define the scale-adjusted ground-truth translation as $\mathbf{t}_{2\to1}^* = \tilde{s}\, \overline{\mathbf{t}}_{2\to1}$. The relative pose objective $\mathcal{L}_{\text{pose}}$ penalizes errors in both the rotation and the scale-anchored translation. For rotation, we use the geodesic distance on $SO(3)$, which provides a physically meaningful angular error. For translation, we use an MSE loss to jointly supervise the direction and the magnitude relative to the 3D scene:

$$\mathcal{L}_{\text{rot}} = \arccos\left(\frac{\text{Tr}\big(R_{2\to1}^{\top}\overline{R}_{2\to1}\big) - 1}{2}\right) \tag{9}$$

$$\mathcal{L}_{\text{trans}} = \|\mathbf{t}_{2\to1} - \mathbf{t}_{2\to1}^*\|_2^2 \tag{10}$$

The total pose objective is: $\mathcal{L}_{\text{pose}} = \mathcal{L}_{\text{rot}} + \mathcal{L}_{\text{trans}}$.

Overall Optimization Objective. The final training objective for CAM3R jointly optimizes three terms: the geometric Asymmetric Angular Loss ($\mathcal{L}_A$), the structural Local Regression Loss ($\mathcal{L}_{\text{regr}}$), and the positional Relative Pose Loss ($\mathcal{L}_{\text{pose}}$). The total loss $\mathcal{L}_{\text{total}}$ is a weighted summation of these components:

$$\mathcal{L}_{\text{total}} = \lambda_A \, \mathcal{L}_A + \lambda_{\text{regr}} \, \mathcal{L}_{\text{regr}} + \lambda_{\text{pose}} \, \mathcal{L}_{\text{pose}} \tag{11}$$

This formulation allows the network to be trained end-to-end, learning the complex interplay between lens geometry and the view consistency of the 3D scene structure.

3.3 Ray-Aware Global Alignment

To reconstruct the entire scene from an unordered collection of images $\{I_1, \ldots, I_N\}$, we propose a Ray-Aware Global Alignment framework. While conventional global alignment techniques such as DUSt3R [33] assume pinhole projection, our approach optimizes scene geometry within a purely ray-consistent 3D space through the multi-stage process described below.

Scene-Graph Pruning. We construct an exhaustive graph $G(V, E)$, where edges $e = (i, j) \in E$ denote directed image pairs. Each edge undergoes a two-view inference pass with CAM3R, followed by a two-stage pruning: (a) **Symmetric Pose Consistency**: We evaluate bidirectional consistency between reciprocal edges (i, j) and (j, i). An edge is pruned if the angular deviation between the predicted rotation $R_{j\to i}$ and the transposed $R_{i\to j}^{\top}$ exceeds a threshold τ_{rot}, or if the translation direction $\hat{\mathbf{t}}_{j\to i}$ fails to align. This retains only mathematically reversible camera trajectories for global optimization. (b) **Geometric Overlap Verification**: We compute strict Mutual Nearest Neighbor (MNN) correspondences in 3D space between $X^{i,i}$ and the transformed $X^{j,i}$. Edges with insufficient overlap (*e.g.* $< 20\%$ of the pixel count) are discarded to prevent erroneous links.

Global Consensus and Optimization. Following pruning, we aggregate pairwise predictions into per-image consensus fields. For each image I_i, we form the global ray field $\mathcal{D}_i$ as a confidence-weighted average of all normalized rays predicted across its valid incident edges. Given this ray field, we initialize the global radial distance field $\mathcal{R}_i$ in three steps: we align the pairwise radial distances along the consensus rays, resolve their relative scale via a median-based alignment, and fuse the aligned distances through confidence-weighted averaging. The resulting consensus ray field $\overline{\mathcal{D}_i}$ and radial field $\overline{\mathcal{R}_i}$ are frozen and define a per-pixel 3D point in the local coordinate system of camera i, $\mathbf{x}_i(\mathbf{u}) = \overline{\mathcal{R}_i}(\mathbf{u})\,\overline{\mathcal{D}_i}(\mathbf{u})$, which serves as a geometric prior. We then formulate global alignment as an optimization over the camera poses $\{P_i\}_{i=1}^N \in SE(3)$ and per-image scales $\{s_i\}_{i=1}^N \in \mathbb{R}^+$, where each scale s_i compensates for the unknown global scale of the predicted radial distances for image I_i. With the geometric priors fixed, let $\sigma_{i,j}(\mathbf{u})$ denote the predicted confidence for pixel $\mathbf{u}$ on edge (i, j), which down-weights unreliable correspondences. The objective minimizes the confidence-weighted 3D alignment

error over all connected pairs (i, j) in the pruned scene graph E:

$$\min_{P_i, s_i} \sum_{(i,j)\in E_{\text{pruned}}} \sum_{\mathbf{u}} \sigma_{i,j}(\mathbf{u}) \left\| P_i\big(s_i \mathbf{x}_i(\mathbf{u})\big) - P_j\big(s_j \mathbf{x}_j(\mathbf{u})\big) \right\|_2^2 \tag{12}$$

We employ a multi-stage alternating optimization scheme [19] to recover the global scene geometry. We first fix the scales and optimize the camera poses, then correct the scales while holding the poses constant. This alternating cycle is repeated for a few iterations to stabilize the camera trajectory, after which we jointly optimize poses and scales for global consistency across the scene graph. Further details on Ray-Aware Global Alignment are provided in the supplementary material.

4 Experiments

4.1 Training and Evaluation Datasets

We train and test our framework on a collection of datasets encompassing panoramic, fisheye, and perspective imagery across indoor and outdoor environments. To ensure cross-model generalization, we augment existing datasets to induce optical heterogeneity:

- 2D3DS [1] and 360Loc [10]: Panoramic collections from which we synthesize additional fisheye and perspective views.
- ADT [20]: We expand this fisheye dataset by generating synthetic perspective counterparts.
- MegaDepth [15]: Perspective dataset; no augmentation.
- CO3Dv2 [23]: This pinhole dataset is reserved for zero-shot evaluation and is not used during training. We expand it by generating synthetic fisheye counterparts.

Throughout this paper, references to these datasets denote the above augmented, multi-modal versions. Training and test splits largely follow the official protocols established by the dataset authors unless otherwise specified. For datasets where image pairs are not explicitly defined, we curate them using ground-truth camera poses, ensuring significant spatial overlap and wide-baseline configurations. Specific details regarding the synthetic fisheye and pinhole projections and dataset splits are provided in the supplementary material.

4.2 Training Configuration

We train our model using a two-phase curriculum [2]. First, we optimize our model exclusively on homogeneous pairs (*e.g.* panorama-panorama) to establish a primary version **CAM3R-homo** for learning the intra-model multi-view geometry. In the second phase, we introduce heterogeneous pairs (*e.g.* pinhole-panorama) to the training set to produce the final **CAM3R** model. We use the

AdamW optimizer [17] with an initial learning rate of 5×10^{-5}, a linear warmup, and cosine decay. Images are resized to 512 pixels on the long edge. Training is conducted on four NVIDIA H200 GPUs using balanced dataset sampling to ensure stability across modalities. The ViT backbones [7] of the Ray Module and Cross-view Module are initialized with the pre-trained weights of UniK3D [21] and DUSt3R [33], respectively. All other components in the network are initialized from scratch. Further specifications regarding the training curriculum are detailed in the supplementary material.

Runtime. On a single consumer-grade RTX A5000, a two-view forward pass at 384×512 takes ∼86 ms for the Ray Module and ∼91 ms for the Cross-view Module, requiring ∼14.5 GB of memory.

4.3 Two-view Pose Estimation

Relative pose estimation for a pair of images I_1 and I_2 involves recovering the rotation $R \in SO(3)$ and the translation $\hat{\mathbf{t}} \in \mathbb{S}^2$ that maps the coordinate system of the second view into the first. In our setting, this task is particularly challenging, as the images may originate from disparate camera models (*e.g.* a fisheye-panorama pair).

Datasets. We evaluate **CAM3R** on the held-out splits of the augmented versions of 2D3DS, MegaDepth, ADT, and 360Loc, and further assess zero-shot capability using the augmented CO3Dv2 dataset. Across all datasets, evaluation is performed on image pairs characterized by substantial spatial overlap and baseline configurations, resulting in a benchmark comprising homogeneous and heterogeneous camera models. The details regarding the pairwise sampling are provided in the supplementary material.

Baselines and metrics. We compare CAM3R with foundation models: DUSt3R [33], MASt3R [14], Pow3R [11], VGGT [32], and π^3 [34]. We report **Relative Rotation Accuracy** (RRA@15) and **Relative Translation Accuracy** (RTA@15), defined as the percentage of pairs with angular errors τ below 15°.

Results. As shown in Table 1, CAM3R achieves superior performance overall. While π^3 and VGGT benefit from explicit training on MegaDepth and VGGT additionally on ADT, CAM3R remains highly competitive on unseen MegaDepth and ADT scenes. On CO3Dv2, our model demonstrates strong zero-shot generalization, outperforming all baselines in translation accuracy (88.2% RTA@15). The largest performance gap appears in cross-model scenarios: while DUSt3R and MASt3R collapse on panoramic datasets (360Loc, 2D3DS), CAM3R maintains robust accuracy (*e.g.* 97.7% RRA@15 and 94.3% RTA@15 on 2D3DS). These quantitative gains are further evidenced by the qualitative pointcloud reconstructions in Fig. 3, which illustrate CAM3R's ability to preserve structural integrity under extreme radial distortion. In MegaDepth, CAM3R accurately recovers complex building architectures such as domes and tomb structures. Crucially, in fisheye sequences from ADT, CAM3R (marked green in Fig. 3) preserves wall planes and clocks, which appear severely distorted (*bent*) in baseline (marked red in Fig. 3) reconstructions.

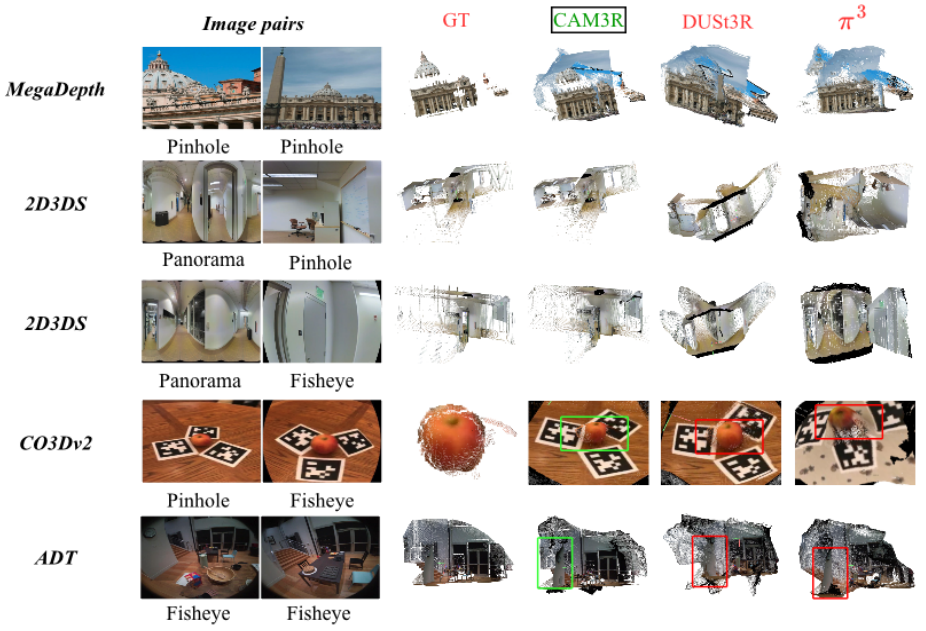

Fig. 3: Qualitative Two-view Reconstructions. Visualization of 3D pointclouds for image pairs across panorama, fisheye, pinhole. Despite extreme distortions, relative poses are accurately recovered and structural consistency is maintained. *Note that these are raw network outputs.*

4.4 Multi-view Pose Estimation

Multi-view relative pose estimation extends the two-view problem to a set of $N > 2$ images, where the goal is to recover a globally consistent set of camera poses $\{P_1, \ldots, P_N\}$ in a unified coordinate frame. We first perform pairwise inference with **CAM3R** across the exhaustive scene graph. We then apply our two-stage pruning (Sec. 3.3) to filter inconsistent edges, as illustrated in Fig. 4, and utilize Ray-Aware Global Alignment to recover the global scene.

Dataset. Similar to the two-view evaluation, we assess the multi-view performance of **CAM3R** on 2D3DS, MegaDepth, 360Loc, ADT, and CO3Dv2 (for

Table 1: Two-view relative pose estimation results. We report Accuracy at 15° (higher is better). Bold indicates best performance.

Model	2D3DS		MegaDepth		CO3Dv2 (Zero-shot)		360Loc		ADT	
	RRA@15	RTA@15	RRA@15	RTA@15	RRA@15	RTA@15	RRA@15	RTA@15	RRA@15	RTA@15
DUSt3R [33]	10.6	6.0	95.6	80.8	94.7	43.1	0.0	0.0	91.0	63.6
MASt3R [14]	18.3	9.3	69.7	56.4	**98.4**	33.4	39.8	5.3	96.6	63.5
Pow3R [11]	7.5	6.0	96.2	74.2	95.8	38.3	0.0	0.0	96.6	79.2
VGGT [32]	11.8	11.0	98.0	88.2	90.9	29.4	37.8	11.1	92.7	82.9
π^3 [34]	16.8	11.4	**99.8**	93.3	90.7	22.7	38.5	13.0	97.5	93.8
CAM3R-homo	65.4	56.8	97.2	92.6	96.1	66.5	58.3	54.7	98.2	93.4
CAM3R	**97.7**	**94.3**	96.8	**94.2**	97.5	**88.2**	**96.0**	**91.0**	**99.0**	**95.0**

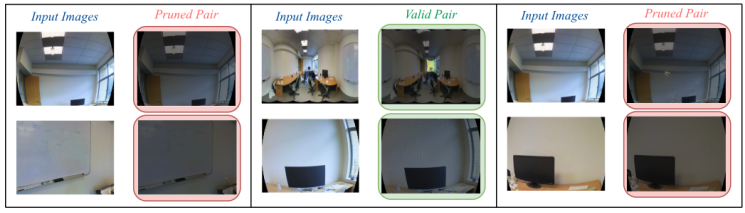

Fig. 4: Qualitative Pruning Analysis. From left to right: (1) Successful rejection of a non-overlapping pair; (2) A valid pair with dense 3D correspondences; (3) Rejection of a doppelganger case where visually similar computer monitors yield inconsistent relative geometry.

zero-shot evaluation). Unlike the two-view evaluation in Sec. 4.3, which relies on curated image pairs, the multi-view benchmark utilizes unstructured pools of images sampled from within a scene or subscene. Further technical specifics regarding the group-wise sampling protocols are detailed in the supplementary material.

Baselines and Metrics. Due to the performance degradation of DUSt3R, MASt3R, and Pow3R on heterogeneous optics in the two-view setting, we limit our multi-view comparative analysis to VGGT and π^3. We report **Accuracy (RRA@30/RTA@30)**, mean **Average Accuracy (mAA@30)**, and **Absolute Trajectory Error (ATE)**. ATE is computed via root-mean-square error (RMSE) after Umeyama alignment [12] to account for global scale and pose.

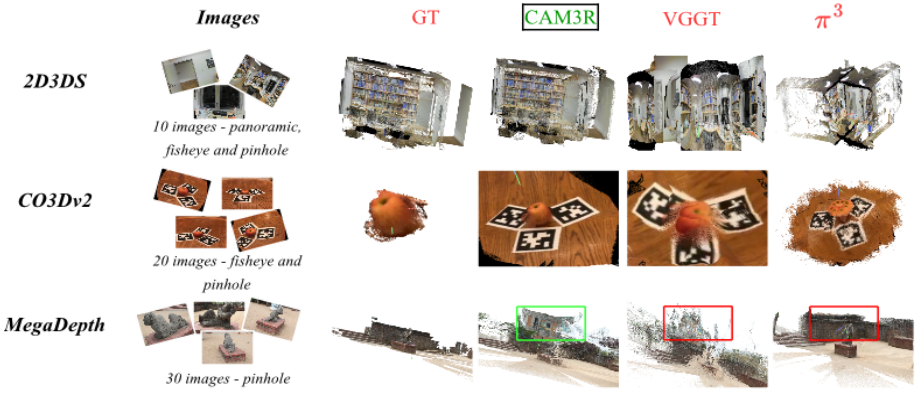

Fig. 5: Qualitative Multi-view Reconstructions. Despite high radial distortion and lack of scene-graph information, globally consistent poses and structural geometry are recovered through the Ray-Aware Global Alignment on CAM3R pairwise inference, effectively mitigating trajectory drift and scale ambiguity.

Results. As shown in Tab. 2, CAM3R achieves a better performance across the majority of datasets. VGGT and π^3 struggle with the 2D3DS and 360Loc datasets, whereas CAM3R demonstrates robustness to these wide-angle geometries, achieving 73.5% and 82.6% mAA, respectively. While π^3 shows competitive accuracy on standard perspective datasets such as MegaDepth, it benefits from

explicit training on that domain. On the CO3Dv2 benchmark, CAM3R achieves a robust 64.9% mAA in a zero-shot setting, while baseline models struggle to handle the mixture of fisheye and pinhole images. In Tab. 3, CAM3R reduces drift in wide-angle scenarios, reaching an ATE of 2.7 on 360Loc compared to the baselines. These quantitative gains are reflected in the qualitative multi-view reconstructions in Fig. 5, which illustrate our model's ability to recover globally consistent camera trajectories and dense scene geometry from unstructured, distorted image pools. In CO3Dv2, CAM3R (marked green) recovers well-defined object geometries (here, the apple structure), whereas baseline models (marked red) produce scattered pointclouds. Furthermore, in MegaDepth, our model achieves superior structural completion of complex buildings; while some background noise is present due to sky regions, the primary architectural integrity is preserved more effectively than in the baselines.

Table 2: Multi-view relative pose estimation results. We report Accuracy at 30° (RRA and RTA) and mAA@30 (higher is better). Bold indicates best performance.

Model	2D3DS			MegaDepth			CO3Dv2 (Zero-shot)			360Loc			ADT		
	RRA	RTA	mAA	RRA	RTA	mAA	RRA	RTA	mAA	RRA	RTA	mAA	RRA	RTA	mAA
VGGT	31.8	34.4	7.6	**100.0**	97.4	68.8	70.5	75.3	19.6	47.9	50.8	19.5	**100.0**	95.6	60.3
π^3	40.0	35.8	9.6	**100.0**	**98.4**	73.4	**89.5**	**91.6**	22.7	48.6	47.4	17.8	**100.0**	**100.0**	75.8
CAM3R + DUSt3R GA	72.3	55.1	38.8	86.5	72.3	68.5	69.7	56.3	46.1	66.7	59.2	55.5	78.3	71.1	70.8
CAM3R + Our GA	**94.0**	**91.5**	**73.5**	96.6	96.3	**87.4**	85.0	85.2	**64.9**	**98.7**	**91.2**	**82.6**	92.2	91.5	**77.3**

Table 3: Multi-view trajectory alignment error (ATE RMSE). Lower is better. Bold indicates best performance.

Model	2D3DS	MegaDepth	CO3Dv2 (Zero-shot)	360Loc	ADT
VGGT	3.8	0.7	1.3	6.3	0.5
π^3	2.9	**0.6**	**0.7**	5.8	**0.4**
CAM3R + DUSt3R GA	2.4	1.2	1.6	4.5	0.6
CAM3R + Our GA	**1.8**	0.8	1.1	**2.7**	**0.4**

4.5 3D Reconstruction Quality

The pose metrics above evaluate camera trajectories but not the recovered geometry. We report Pointcloud metrics (Accuracy, Completeness, Normal Consistency, and Chamfer Distance) and Radial Distance accuracy (Abs Rel, δ_1). We use *radial distance* rather than Z-depth, which is ill-defined for wide-FoV lenses. As shown in Tab. 4, CAM3R achieves a better pointcloud quality for most of the datasets. Notably, π^3 and VGGT can retain reasonable per-pixel radial accuracy (*e.g.* π^3 reaches δ_1=0.997 on CO3Dv2) yet still produce inferior pointclouds (CD 0.111 versus CAM3R's 0.068; see Tab. 4, with the qualitative comparison in Fig. 5). This is what CAM3R argues for: per-pixel distances appear plausible while 3D *structure* is bent, because Cartesian regression entangles lens projection with scene geometry. On pinhole-only MegaDepth, CAM3R matches π^3 in reconstruction quality (CD 0.026 vs. 0.024; δ_1=0.996). Within our pinhole evaluation, camera-agnosticism incurs no serious accuracy cost with wide-FoV capacity.

Table 4: 3D reconstruction quality: pointcloud and radial-distance metrics.

Dataset	Method	*Pointcloud Quality*				*Radial Dist.*	
		Acc↓	Comp↓	NC↑	CD↓	Abs Rel↓	δ_1 ↑
2D3DS (pano-pin-fish)	VGGT	0.520	1.101	0.443	1.022	0.276	0.625
	π^3	0.325	0.399	0.474	0.602	0.217	0.710
	CAM3R	**0.059**	**0.077**	**0.970**	**0.068**	**0.053**	**0.954**
360Loc (pano-pin-fish)	VGGT	0.890	3.665	0.547	2.270	0.436	0.385
	π^3	0.630	1.405	0.595	1.017	0.333	0.547
	CAM3R	**0.071**	**0.056**	**0.977**	**0.064**	**0.011**	**0.994**
MegaDepth (pinhole-pinhole)	VGGT	0.034	0.076	0.907	0.027	0.016	0.987
	π^3	**0.013**	0.058	**0.921**	**0.024**	**0.010**	**0.996**
	CAM3R	0.032	**0.022**	0.901	0.026	0.012	**0.996**
ADT (pinhole-fisheye)	VGGT	0.260	0.162	0.653	0.211	0.067	0.547
	π^3	0.166	0.084	0.763	0.125	0.040	0.737
	CAM3R	**0.016**	**0.013**	**0.982**	**0.014**	**0.018**	**0.966**
CO3Dv2 (pinhole-fisheye)	VGGT	0.093	0.221	0.613	0.157	0.035	0.985
	π^3	0.113	0.109	0.616	0.111	0.027	**0.997**
	CAM3R	**0.052**	**0.084**	**0.676**	**0.068**	**0.024**	0.986

4.6 Ablation Studies

Architecture vs. Data: Is Decoupling Necessary? A natural question is whether CAM3R's gains stem from its decoupled architecture or simply from exposure to distorted training data. Our design emerged through three stages (detailed in the supplementary material): (1) naively fine-tuning a DUSt3R-style model on distorted data, which forces the network to entangle non-linear projection with scene geometry and relative pose; (2) decoupling radial-distance prediction in each view's local frame using ground-truth rays; and (3) the final CAM3R, which learns rays via the Ray Module and aligns views via the Relative Pose Network, eliminating the ground-truth dependency. Fine-tuning DUSt3R on the *identical* training mixture proves insufficient, reaching only 17.8% RRA@15 on 2D3DS versus CAM3R's 97.7% (Tab. 5). The gains therefore arise from the projection-invariant decoupled representation, not from the training data or model capacity.

Table 5: Quantitative two-view validation of architectural design choices. We report two-view relative pose estimation metrics at 15° (↑). Comparing Vanilla DUSt3R, a naively fine-tuned DUSt3R (Stage 1), and our final decoupled CAM3R (Stage 3) reveals that simply exposing standard foundation models to distorted data is mathematically insufficient. Our model yields massive performance gains, particularly on extreme wide-angle datasets (*e.g.* 2D3DS [1] and 360Loc [10]).

Model Iteration	**2D3DS**		**MegaDepth**		**CO3Dv2 (Zero-shot)**		**360Loc**		**ADT**	
	RRA@15	RTA@15	RRA@15	RTA@15	RRA@15	RTA@15	RRA@15	RTA@15	RRA@15	RTA@15
Vanilla DUSt3R [33]	10.6	6.0	95.6	80.8	94.7	43.1	0.0	0.0	91.0	63.6
Fine-tuned DUSt3R (Stage 1)	17.8	10.9	94.2	72.5	94.3	52.7	13.0	9.2	87.5	65.4
CAM3R (Final - Stage 3)	**97.7**	**94.3**	**96.8**	**94.2**	**97.5**	**88.2**	**96.0**	**91.0**	**99.0**	**95.0**

Impact of Heterogeneous Training. We evaluate the necessity of the second training phase by comparing the full **CAM3R** against the phase-one baseline, **CAM3R-homo**. As shown in Tab. 1, the inclusion of heterogeneous

pairs (*e.g.* panorama-perspective) during the second stage of our training curriculum significantly improves two-view relative pose metrics.

Ray-Aware Global Alignment. We evaluate the efficacy of our Ray-Aware Global Alignment against the global alignment in DUSt3R. As shown in Tab. 2, the ray-based optimizer, applied to CAM3R's pairwise predictions, is superior across all datasets, where substituting DUSt3R's alignment lowers mAA by up to ∼35 points. Furthermore, the ATE RMSE results in Tab. 3 confirm that our approach suppresses cumulative drift, reducing trajectory error by up to 40% (*e.g.* from 4.5 to 2.7 on 360Loc).

5 Conclusion

In this paper, we address two challenges in 3D reconstruction: handling non-linear distortions in wide-angle imagery and extracting reliable cross-view cues for robust pose estimation. To this end, we introduce CAM3R, a regression-based framework designed to tackle both problems. The Ray Module learns pixel-wise ray representations and the Cross-view Module predicts pairwise pointmaps and relative poses. Building on these components, we extend the framework to multi-view reconstruction of unordered, distorted images through a Ray-Aware Global Alignment method. Extensive evaluations demonstrate that CAM3R achieves state-of-the-art performance on panoramic and fisheye imagery, while maintaining pinhole accuracy. Two limitations motivate future work: the separate ViT backbones increase memory, which parameter sharing or distillation into a unified encoder [24] could alleviate; and scene-graph construction requires $O(N^2)$ pairwise inferences, motivating an extension of the ray-based formulation to multi-view transformer architectures [32]. Overall, explicitly modeling camera rays offers a strong foundation for camera-agnostic 3D reconstruction.

Acknowledgements

We thank Prof. Anand Bhattad for helpful discussions and valuable feedback. This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 140D0423C0076. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

1. 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. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009)
3. Bogdan, O., Eckstein, V., Rameau, F., Bazin, J.C.: Deepcalib: A deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In: Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production. pp. 1–10 (2018)
4. Brown, D.C.: Close-range camera calibration. Photogrammetric engineering **37**(8), 855–866 (1971)
5. Conrady, A.: Lens-systems, decentered. Monthly Notices of the Royal Astronomical Society, Vol. 79, p. 384-390 **79**, 384–390 (1919)
6. 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)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), `https://arxiv.org/abs/2010.11929`
8. 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)
9. 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)
10. 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)
11. Jang, W., Weinzaepfel, P., Leroy, V., Agapito, L., Revaud, J.: Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1071–1081 (2025)
12. Kabsch, W.: A solution for the best rotation to relate two sets of vectors. Foundations of Crystallography **32**(5), 922–923 (1976)
13. 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)
14. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European conference on computer vision. pp. 71–91. Springer (2024)
15. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2041–2050 (2018)
16. Liao, K., Nie, L., Huang, S., Lin, C., Zhang, J., Zhao, Y., Gabbouj, M., Tao, D.: Deep learning for camera calibration and beyond: A survey (2025), `https://arxiv.org/abs/2303.10559`
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019), `https://arxiv.org/abs/1711.05101`

18. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
19. Neal, P., Eric, C., Borja, P., Jonathan, E.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning **3**(1), 1–122 (2011)
20. Pan, X., Charron, N., Yang, Y., Peters, S., Whelan, T., Kong, C., Parkhi, O., Newcombe, R., Ren, Y.C.: Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20133–20143 (2023)
21. 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)
22. 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)
23. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10901–10911 (2021)
24. Sarıyıldız, M.B., Weinzaepfel, P., Lucas, T., De Jorge, P., Larlus, D., Kalantidis, Y.: Dune: Distilling a universal encoder from heterogeneous 2d and 3d teachers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30084–30094 (2025)
25. 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)
26. Schönberger, J.L., Frahm, J.M.: Structure-from-Motion Revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
27. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise View Selection for Unstructured Multi-View Stereo. In: European Conference on Computer Vision (ECCV) (2016)
28. Sturm, P., Ramalingam, S., Tardif, J.P., Gasparini, S., Barreto, J.: Camera models and fundamental concepts used in geometric computer vision. Foundations and Trends in Computer Graphics and Vision **6**(1-2), 1–183 (2011)
29. 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)
30. Urban, S., Leitloff, J., Hinz, S.: Improved wide-angle, fisheye and omnidirectional camera calibration. ISPRS Journal of Photogrammetry and Remote Sensing **108**, 72–79 (2015)
31. Veicht, A., Sarlin, P.E., Lindenberger, P., Pollefeys, M.: Geocalib: Learning single-image calibration with geometric optimization. In: European Conference on Computer Vision. pp. 1–20. Springer (2024)
32. 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)
33. 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)

34. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-Equivariant Visual Geometry Learning. arXiv preprint arXiv:2507.13347 (2025)