Marginalized Bundle Adjustment: Multi-View Camera Pose from Monocular Depth Estimates

Shengjie Zhu[1,2] * Ahmed Abdelkader[1] Mark J. Matthews[1] Xiaoming Liu[2] Wen-Sheng Chu[1]
[1]Google [2]Michigan State University

Abstract

Structure-from-Motion (SfM) is a fundamental 3D vision task for recovering camera parameters and scene geometry from multi-view images. While recent deep learning advances enable accurate Monocular Depth Estimation (MDE) from single images without depending on camera motion, integrating MDE into SfM remains a challenge. Unlike conventional triangulated sparse point clouds, MDE produces dense depth maps with significantly higher error variance. Inspired by modern RANSAC estimators, we propose Marginalized Bundle Adjustment (MBA) to mitigate MDE error variance leveraging its density. With MBA, we show that MDE depth maps are sufficiently accurate to yield SoTA or competitive results in SfM and camera relocalization tasks. Through extensive evaluations, we demonstrate consistently robust performance across varying scales, ranging from few-frame setups to large multi-view systems with thousands of images. Our method highlights the significant potential of MDE in multi-view 3D vision. Code is available at https://marginalized-ba.github.io/.

1. Introduction

Structure-from-Motion (SfM) is a fundamental method in 3D vision for recovering 3D scene geometry (as point clouds) and camera parameters (intrinsics and extrinsics) from multi-view images. Its versatility has fueled a wide range of applications, including 3D reconstruction [23], robot navigation [26], camera re-localization [24], neural rendering [44], etc. Classical SfM pipelines operate by identifying sparse 2D correspondences from image pairs to jointly optimize 3D point positions and camera poses via Bundle Adjustment (BA). However, reliance on explicit feature matching makes these systems prone to failure in scenes with low texture or limited parallax, and thus compromises reconstruction accuracy.

Unlike classic SfM that relies on motion cues to infer geometry, deep learning advances now enable the inference of structure *independently* of motion via Monocular Depth Estimation (MDE) [5, 48]. Despite the availability of this rich structural prior, its integration into multi-view pipelines remains an open challenge. Crucially, dense MDE predictions remain underutilized. Existing works typically use them only to initialize sparse keypoints, discarding the dense data in favor of traditional BA refinement [7, 62, 74]. Meanwhile, alternative learning-based methods face other limitations. Scene coordinate regression methods [8–10] require expensive scene-specific fine-tuning. Methods formulating BA as network inference [72–74] suffer from high memory footprints that limit scalability. Finally, other approaches that train a depth model during BA [62] are too memory-intensive to leverage large SoTA foundation MDE models.

This raises a key question: *how can dense MDE predictions be leveraged for multi-view pose estimation?* The challenge is that monocular depth maps yield dense but high-variance point clouds, failing to meet the requirements of classical SfM for sparse, accurate features (Fig. 1). To bridge this gap, we propose a "Motion-from-Structure" approach that directly recovers camera motion from dense structural information provided by MDE. To faithfully showcase MDE for multi-view pose estimation, we avoid per-pixel refinement entirely, intervening only to resolve scale ambiguity through per-frame affine corrections.

To handle dense, high-variance inputs, we draw inspiration from RANSAC. Conventional RANSAC relies on discrete inlier counting. Despite being robust, it is non-differentiable and sensitive to the chosen error threshold. To overcome this limitation, we leverage the dense projective residuals induced by MDE's dense depth predictions. In this setting, we observe that the inlier count for a given threshold τ corresponds (in the limit) to the Cumulative Distribution Function (CDF), $F(\tau)$, for that empirical residual distribution. Based on this insight, we formulate a robust BA objective that maximizes the Area Under the Curve (AUC) of this empirical CDF (Fig. 3). This effectively integrates information across a range of thresholds to marginalize out the error threshold in BA, hence the name Marginalized Bundle Adjustment (MBA).

*Research conducted during an internship at Google.

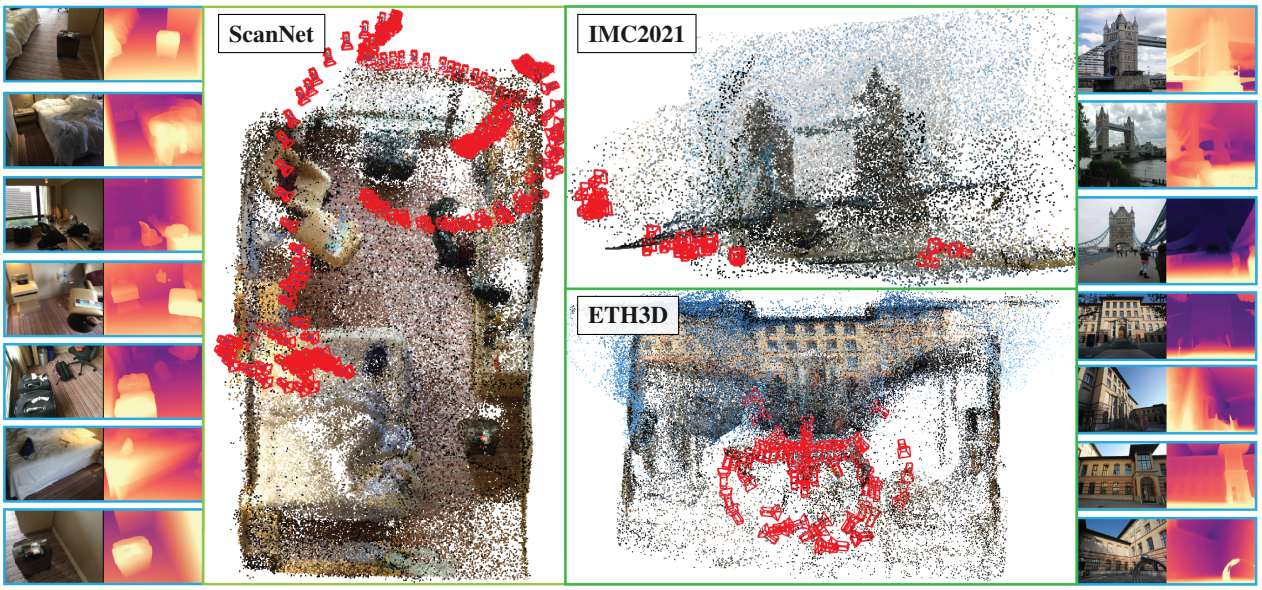


Figure 1. **Marginalized Bundle Adjustment (MBA).** Our method registers monocular depth maps into a consistent 3D coordinate system. Red camera icons indicate the viewpoints of registered depth maps. Monocular depth provides a strong structural prior, yet its predictions are inherently high-variance, as reflected by the noisy appearance of the reconstructed point cloud. This makes classical Bundle Adjustment designed for accurate sparse point cloud ill-suited. We therefore introduce a RANSAC-inspired Bundle Adjustment objective that leverages the depth maps' density to robustly accommodate their variance. Although depth-derived point clouds have lower visual fidelity, our experiments show that monocular depth already supports SoTA or competitive performance on diverse SfM benchmarks, as exemplified ScanNet [15], IMC2021 [6], and ETH3D [6]. It highlights significant potential of monocular depth models for multi-view vision tasks.

Since analytical AUC maximization is intractable, we derive a differentiable surrogate loss. Notably, this formulation generalizes MAGSAC [2], a special case of our loss if its assumed Chi-squared distribution is replaced with the empirical residual distribution. Our marginalized objective is general to serve as a robust scoring function in RANSAC, where it matches the performance of MAGSAC++ [3] on the two-view essential matrix estimation.

We integrate the MBA objective into a flexible, coarse-to-fine framework applicable to both SfM and camera re-localization. This pipeline is designed to accept various robust loss functions, where we opt to skip conventional components such as feature tracking and triangulation to highlight the benefits of MDE on its own. Through the power of monocular depth priors, our results show state-of-the-art or competitive results on multiple indoor and outdoor benchmarks. Additionally, the method proves highly scalable, capable of performing global BA over thousands of frames using a distributed cluster (Table 1), which validates the viability of MDE-based methods for large-scale reconstruction. We summarize our contributions as follows:

- First framework integrating general MDE models into SfM and re-localization tasks across varying scales.
- Novel and principled RANSAC-inspired objective function designed to handle dense, high-variance depth priors. This formulation is versatile and applicable to two-view RANSAC and multi-view Bundle Adjustment.
- SoTA or competitive performance on indoor & outdoor, small & large scale, SfM and re-localization benchmarks.

2. Related Work

Dense Depth Map and Point Cloud Estimation. Recent depth foundation models [31, 33, 51, 83] enable relatively accurate metric depth maps inference in-the-wild [5, 48, 49]. Methods [31, 64, 80] augment monocular depth with video. Next, monocular point cloud models including MoGe [76, 77] directly regress 3D point cloud instead of depth map, augmenting depth learning with intrinsic prior. Further, binocular point cloud models [75, 84] such as DUSt3R [79] and MASt3R [34] integrate motion prior via regressing spatially aligned point clouds. Yet, shown in Fig. 1, aforementioned models are inherently dense and high-variance. Our Marginalized BA function is designed to accommodate these characteristics. To our knowledge, we are also the first generalizable framework that extends depth and point cloud models to multi-view vision tasks.

RANSAC. RANSAC algorithms [2] robustly estimate low-DoF parameters under noise. Our work extends the RANSAC philosophy from low-DoF two-view tasks to the high-DoF multi-view problems. Namely, we handle noisy inputs from high-variance network regressions including dense depth maps. Several RANSAC variants [2, 67] have improved scoring function by moving from binary [22] to continuous formulations. MAGSAC [2] smooths binary

Dataset	ScanNet [14]	Tanks&Temples [30]	ETH3D [59]	IMC2021 [6]	7-Scenes [60]	Wayspots [9]
Images	391	1,106	76	25	8,000	1,157
Pairs	21,982	210,894	1,571	300	282,209	666,598

Table 1. **Maximum Size of Pose Graph** in each benchmarked dataset. The nodes and edges are its images and co-visible image pairs.

counting function by integrating residuals against a chi-squared prior distribution. We propose a generalized formulation that maximizes inlier counts aggregated at multiple residual thresholds. In essential-matrix estimation, our strategy matches MAGSAC++'s performance [2].

Deep Learning for Multi-view Pose Estimation. Traditional SfM pipelines like COLMAP [56] have been improved by learned image matching models, including PixSfM [40], DF-SfM [28], and Dense-SfM [32]. Subsequent works incorporates 3D priors. For example, FlowMap [62] formulates SfM as a self-supervised depth learning task, while MASt3R-SfM [18] introduces a dedicated point cloud aggregation strategy for models like DUSt3R [79] and MASt3R [34]. More recently, the entire SfM pipeline has been reformulated as an end-to-end learning problem [72–74]. However, high memory footprint typically limits these approaches to small-scale scenarios. In the related task of camera re-localization, methods often train a network to map images to world coordinates but usually require scene-specific fine-tuning [1, 10]. To our knowledge, MBA is the first framework to successfully apply general-purpose MDE models to both small and large-scale SfM problems. Notably, when using DUSt3R as its MDE model, MBA achieves higher accuracy than the specialized MASt3R-SfM pipeline (Tab. 3).

3. Method

We propose an MDE-based approach for multi-view pose estimation. As shown in Fig. 1, our pipeline keeps the depth maps fixed, and applies only affine corrections to account for the depth scale ambiguity. Our core contribution is a marginalized BA (MBA) objective that effectively leverages dense depth map predictions with high variance. In this section, we present our approach in the context of SfM, and defer the minor adaptations required for camera re-localization and two-view RANSAC towards the end.

3.1. System Overview

Problem Definition. Given as input an unordered collection of N RGB frames $\{\mathbf{I}_i\}_{i\in[N]}$, we optimize for camera intrinsics $\mathcal{K}=\{\mathbf{K}_i\}$ and extrinsics $\mathcal{P}=\{\mathbf{P}_i\}$. To do so, we precompute N depth maps $\mathcal{D} = \{\mathbf{D}_i = \mathcal{N}_{\mathbf{D}}(\mathbf{I}_i)\}$ and pairwise correspondence maps $\mathcal{C} = \{\mathbf{C}_{i,j} = \mathcal{N}_{\mathbf{C}}(\mathbf{I}_i, \mathbf{I}_j), i \neq j\}$, where $\mathcal{N}_{\mathbf{D}}$ and $\mathcal{N}_{\mathbf{C}}$ denote pre-trained depth and correspondence models, respectively (see examples in Fig. 4). We jointly optimize per-frame affine corrections $\mathcal{A} = \{\alpha_i, \beta_i \mid i \leq N\}$ for each depth map, producing scale-ambiguity corrected depth maps:

$$\mathbf{D}'_i = \alpha_i \cdot \mathbf{D}_i + \beta_i. \tag{1}$$

Optimization Objective. Denote $\mathcal{X} = \{\mathcal{P}, \mathcal{K}, \mathcal{A}\}$ as the set of all variables to optimize, and $\mathbf{X}_i = (\mathbf{P}_i, \mathbf{K}_i, \mathbf{A}_i)$, we define the objective as maximizing a scoring function $\mathcal{S}$:

$$\mathcal{X}^* = \underset{\mathcal{X}=(\mathcal{P},\mathcal{K},\mathcal{A})}{\arg\max}\ \mathcal{S}(\mathcal{X} \mid \mathcal{D}, \mathcal{C}). \tag{2}$$

We define $\mathcal{S}$ as the summation of a suitable quality function $\mathcal{Q}$ over frame pairs (I_i, I_j) over the pose graph $\mathcal{G}$:

$$\mathcal{S}(\mathcal{X} \mid \mathcal{D}, \mathcal{C}) = \sum_{(i,j)\in\mathcal{G}} \mathcal{Q}(\mathbf{X}_i, \mathbf{X}_j \mid \mathbf{D}_i, \mathbf{D}_j, \mathbf{C}_{i,j}). \tag{3}$$

The pose graph $\mathcal{G}$ (defined in Sec. 3.3) connects co-visibile frame pairs with enough high-confidence correspondences.

Algorithm Pipeline. In Fig. 2, we subsample dense depth and correspondence into a data matrix. Outlined in Sec. 3.3, with initialized intrinsics, extrinsics, and depth affine corrections, we sequentially execute coarse and fine stage SfM. Two stages are mostly same except for employing different pose graphs. Each stage applies gradient descent with fixed iterations. In the following Sec. 3.2, we first describe our core algorithm: the marginalized BA objective for dense, high-variance network regressions.

3.2. Marginalized Bundle Adjustment (MBA)

We begin with a naive yet robust binary quality function to realize the Eq. (3) quality function. Despite its robustness, the binarized function is non-differentiable and sensitive to the single threshold selected. Hence, we propose a smooth form integrating over multiple thresholds.

Subsample Dense Depth and Correspondence. For each co-visible frame pair of the pose graph $\mathcal{G}$, we sample a fixed number κ of paired pixels. Specifically, between frame i and j, we sample κ depth pixels on frame i and κ i-to-j correspondence pixels. We apply random sampling with replacement over the correspondences with a confidence score at least χ yielding a data matrix of size $|\mathcal{E}| \times \kappa \times 5$, where $\mathcal{E}$ denotes the edges of $\mathcal{G}$.

Projective Residuals. We define the residual $r_{i,j,k}$ as the 2D discrepancy in the kth sampled correspondence $c_{i,j,k} \in \mathbf{C}_{i,j}$. Denoting $c_{i,j,k}$ as $(p_{i,j,k}, q_{i,j,k}) \in I_i \times I_j$, we write

$$r_{i,j,k} = \left\| \pi_{i\to j}\left(\mathbf{D}'_i[p_{i,j,k}]\right) - q_{i,j,k} \right\|_2, \tag{4}$$

where the operator $\pi_{i\to j}$ projects the pixel $p_{i,j,k}$ in frame I_i to the frame I_j with its affine-aligned depth value in $\mathbf{D}'_i$

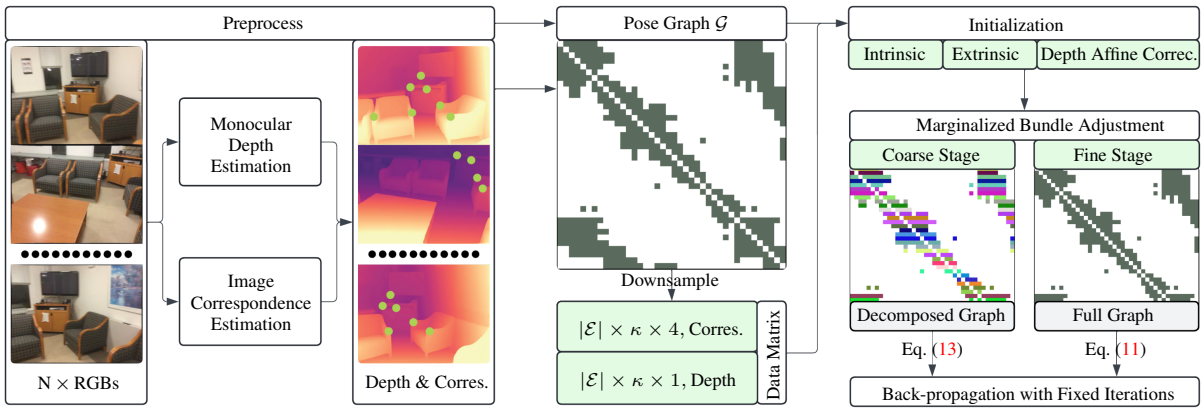


Figure 2. **System Overview.** With N RGBs, the system consumes dense depth maps and pairwise correspondence inferred by pretrained models. The system outputs intrinsics, extrinsics, and frame-wise depth affine corrections scalars. A sparse $N \times N$ pose graph is built from co-visible frames using correspondences. Dense inputs are subsampled into a data matrix of $|\mathcal{E}| \times \kappa \times 5$ (graph edges count) for scalable multi-GPU optimization. After initialization, the Bundle Adjustment proceeds from coarse to fine. In coarse stage, the BA objective is evaluated and summed over "star-shaped" subgraph $\mathcal{G}_i$ of each frame i. One subgraph includes itself plus its co-visible neighbors, marked as one colored row in coarse pose graph. Fine stage computes with full graph. The BA applies gradient descent for fixed iterations.

from Eq. (1). Pixel $q_{i,j,k}$ is the corresponding pixel of $p_{i,j,k}$. The projection is defined by the camera intrinsics $\mathbf{K}_i$, $\mathbf{K}_j$ and extrinsics $\mathbf{P}_i$, $\mathbf{P}_j$ [27]. Other robust norms may also be used in Eq. (4), *e.g.*, the Cauchy function used in [57].

Binary Quality with a Threshold. We start realizing Eq. (3) with a robust binary quality function:

$$\mathcal{S}^b(\mathcal{X} \mid \mathcal{D}, \mathcal{C}, \tau) = \sum_{(i,j)\in\mathcal{G}} \left(\sum_{k\in\kappa} \mathbb{1}[r_{i,j,k} < \tau] \right), \quad (5)$$

where the variable τ is the residual threshold and function $\mathbb{1}(\cdot)$ is the indicator function. Intuitively, a depth pixel is considered an inlier if its projective residual is below the threshold τ. The binarized scoring function in Eq. (5) is widely used in RANSAC algorithms [22] for its robustness on overly-sampled noisy inputs. RANSAC algorithm mostly addresses low Degree-of-Freedom (DoF) problem including essential matrix estimation [46]. In contrast, multi-view pose estimation problem solves multiple poses, possessing a significantly larger solution space. This necessitates a *continuous* scoring function to enable iterative optimization whereas $\mathcal{S}^b$ from Eq. (5) is discrete.

CDF as Smoothed Quality. The dense depth maps provide enough samples of projective residuals to utilize their distributional properties. Let R denote the set of all residuals. We model the residual r as a random variable following an empirical distribution $\mathcal{R}$ estimated with kernel density estimation (KDE) [61]. This gives residual r distribution $\mathcal{R}$:

$$r \sim \mathcal{R} = \mathrm{KDE}(R), \; R = \{r_{i,j,k} \mid (i,j) \in \mathcal{G}, [k] \in \kappa\}. \quad (6)$$

Please see Fig. 3 for an example distribution. Denote $p(r)$ and $F(\tau) = \Pr[r < \tau]$ as the probability distribution function (PDF) and cumulative distribution function (CDF) of $\mathcal{R}$, respectively. The binary scoring function at threshold τ in Eq. (5) is approximated as the CDF function at value τ.

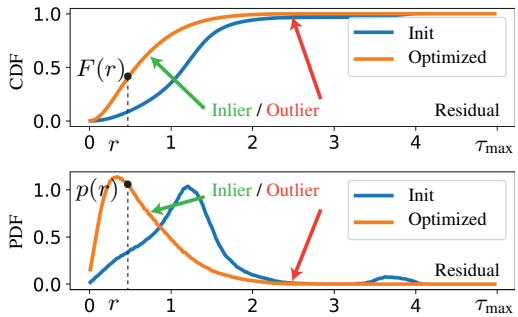


Figure 3. **CDF and PDF** of empirical residual distribution $\mathcal{R}$. In Eq. (8), our BA maximizes the area-under-the-curve of $\mathcal{R}$'s CDF curve up a maximum threshold. The BA automatically formulates a smoothed categorization of inliers versus outliers. The forward and backward computation (Eqn. 10 & 11) at the residual r is succinctly defined as indexing the curve at $F(r)$ and $p(r)$.

$$\mathcal{S}^b(\mathcal{X} \mid \mathcal{D}, \mathcal{C}, \tau) = \sum_{i,j,k} \mathbb{1}[r_{i,j,k} < \tau] \approx |R| \cdot F(\tau). \quad (7)$$

Marginalizing Thresholds. We generalize Eq. (5) by extending from single to multiple thresholds. We integrate Eq. (5) up to a maximum threshold $\tau_{\max}$, beyond which a residual is safely considered an outlier.

$$\mathcal{S}^m(\mathcal{X} \mid \mathcal{D}, \mathcal{C}) = \int_0^{\tau_{\max}} \mathcal{S}^b(\mathcal{X} \mid \mathcal{D}, \mathcal{C}, \tau)\, d\tau \approx |R| \int_0^{\tau_{\max}} F(r)\, d\tau. \quad (8)$$

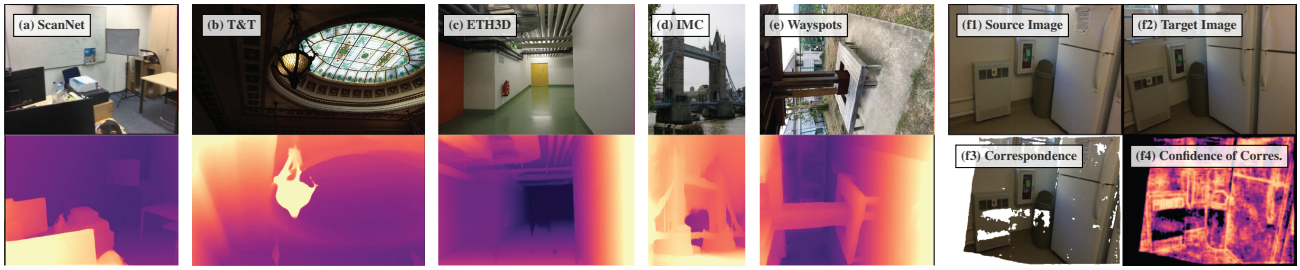


Figure 4. **Dense Depth and Correspondence** are inputs to our multi-view pose estimation system. We benchmark pose performance with depth map under diverse imaging conditions. Figures (a) to (e) include ScanNet [14] indoor images, T&T [30] and ETH3D [58] high resolution images, IMC2021 [6] internet-collected images, and Wayspots [9] flipped image. Fig. (f) visualizes dense correspondence.

Intuitively, depth pixels from dense depth maps vary at noise levels. In Eq. (5), a large threshold τ aids to register camera at approximately correct locations while a small τ improves accuracy but risks local minima. On the other hand, Eq. (8) leverages benefits from setting both small and large thresholds. Mathematically, it computes the area under a truncated CDF curve, as outlined in Fig. 3.

Marginalized Bundle Adjustment (MBA) Objective. Practically, we use a histogram-based KDE implementation. This renders Eq. (8) as a summation over a finite set of T thresholds with $\tau_i = \frac{i}{T} \cdot \tau_{\max}$.

$$\mathcal{S}^m(\mathcal{X} \mid \mathcal{D}, \mathcal{C}) = \sum_{0 \leq i \leq T} \mathcal{S}^b(\mathcal{X} \mid \mathcal{D}, \mathcal{C}, \tau_i) \approx |R| \cdot \frac{\tau_{\max}}{T} \cdot \sum_{i=0}^{T} F(\tau_i). \tag{9}$$

MBA Forward. In Fig. 3, to maximize Eq. (9), we propose a *surrogate loss* with pixel-wise forward and backward functions (plus a negative sign). Intuitively, Eq. (9) provides a smooth differentiable loss wrapped over each residual:

$$\mathcal{L}_{\text{MBA}} = \frac{1}{|R|} \sum_{i,j,k} -F(r_{i,j,k}) \cdot \mathbb{1}[r_{i,j,k} < \tau_{\max}]. \tag{10}$$

MBA Backward. Correspondingly, the backward is:

$$\frac{\partial \mathcal{L}_{\text{MBA}}}{\partial r_{i,j,k}} = -\frac{1}{|R|} p(r_{i,j,k}) \cdot \mathbb{1}[r_{i,j,k} < \tau_{\max}]. \tag{11}$$

Discussion. Despite employing Eq. (10) as loss, our optimization objective is Eq. (9). We prove its effectiveness via applying Eq. (9) as a scoring function to a two-view pose estimation with RANSAC. In Tab. 8, our scoring function performs comparable to SoTA method [2] with dense image correspondence inputs. Notably, MAGSAC [2] scoring function becomes a special case of Eq. (10) as replacing its assumed chi-squared distribution with the empirical distribution $\mathcal{R}$. See more details in Supp. Sec. 6.

Robustness. As a RANSAC inspired objective function, the proposed loss function Eq. (9) inherits its robustness. From the backward function Eq. (11), the gradient of extreme residual values, *i.e.*, those of low probability, is suppressed, as shown in Fig. 3. After convergence, the noise level of each depth pixel is implicitly captured by its residual probability, enabling BA to automatically distinguish inliers from outliers without a dedicated neural network.

Scalability. Shown in Fig. 2, our approach maintains an $|\mathcal{E}| \times \kappa \times 5$ matrix. It allows scaling up by parallelizing computation across multiple GPUs. In Tab. 1, we enable global BA over large-scale pose graph involving $8,000$ frames and $564,418$ co-visible frame pairs. This significantly surpasses recent methods including FlowMap [62], VGG-SfM [73], and PoseDiffusion [72], which run out of memory when processing more than 200 frames [18].

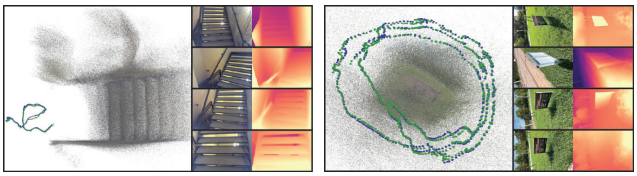
3.3. SfM Pipeline

The section outlines the sequentially executed, initialization, coarse-stage SfM, and fine-stage SfM processes. We also describe extension to camera re-localization.

Pose Graph Construction. As shown in Fig. 2, we construct an undirected pose graph $\mathcal{G}$ from a set of co-visible frame pairs (I_i, I_j). Co-visibility between two frames is measured as the percentage of pixels that are visible in both views, computed from correspondence maps $\mathbf{C}_{i,j} \in \mathcal{C}$. We include an edge $g_{i,j} \in \mathcal{G}$ if the co-visibility score exceeds a threshold ν, resulting a sparse pose graph.

Intrinsic Initialization. Each frame, we run DUSt3R [79] with two identical frames to extract a dense point cloud, then initialize per-frame intrinsics via RANSAC-based calibration [86]. If a shared intrinsic is assumed, we initialize it using the median focal length. See details at Supp. Sec. 7.6.

Camera Pose and Depth Adjustment Initialization. We construct a spanning tree from the pose graph $\mathcal{G}$ using a greedy strategy. The root node is chosen as the frame with the highest degree. At each step, we add a new node that maximizes the total degree of the tree. For each new frame i, we identify a registered, co-visible frame j. Using projection from frame j to i, we initialize the 6-DoF camera



(a) 7-Scenes [60], sequence Stairs (b) Wayspots [9], sequence Lawn

Figure 5. **Camera Re-localization** on 7-Scenes and Wayspots. Green and blue mark predicted and groundtruth odometry. We present challenging sequences of repetitive, textureless images. The images exhibit (1) scale changes, (2) flipping, and (3) a lack of distinguishable depth references. Consequently, the model predicts sub-optimal depth maps. Surprisingly, despite these significant challenges, depth maps still support accurate camera poses.

pose $\mathbf{P}_i$, with translation magnitude estimated following a Five-Point algorithm-based method. From the reverse projection i to j, we estimate depth scale adjustment s_i. The depth bias β_i is initialized to zero.

Coarse-Stage SfM. Suppose the frame i is poorly registered, its corresponding residual $r_{i,j,k}$ exhibits significantly large values. Due to the robustness property of Eq. (11), the residuals with larger values are automatically suppressed with smaller gradients, see outliers at Fig. 3. These characteristics cause poorly registered frames to become "stuck" in a local minimum. We propose a graph decomposition strategy to mitigate the occurrence of early local minima. For the graph $\mathcal{G}$, we decompose it into a N subgraphs $\mathcal{G}_i$:

$$\mathcal{G} = \sum_{i \in N} \mathcal{G}_i, \quad \mathcal{G}_i = \{\mathcal{V}_i,\ \mathcal{E}_i\}, \tag{12}$$

where $\mathcal{V}_i = \{\mathbf{I}_i\} \cup \mathcal{N}(\mathbf{I}_i)$ and $\mathcal{E}_i = \{(\mathbf{I}_i, \mathbf{I}_j) \mid \mathbf{I}_j \in \mathcal{N}(\mathbf{I}_i)\}$. Each subgraph $\mathcal{G}_i$ consists of the i-th frame $\mathbf{I}_i$ and its neighboring frames $\mathcal{N}(\mathbf{I}_i)$. In Fig. 2, we visualize each subgraph as a row of color in the sparse pose graph.

Beyond subgraph decomposition, we apply logarithmic operation to the L2 norm residual in Eq. (4) to enhance its robustness to large residuals from poorly registered frames. This renders the coarse-SfM BA forward function as:

$$\mathcal{L}_{\text{coarse}} = -\frac{1}{N} \sum_i \frac{1}{\|R_i\|} \sum_{(j,k) \in \mathcal{G}_i} \bar{F}_i(\bar{r}_{i,j,k}) \cdot \mathbb{1}[\bar{r}_{i,j,k} < \bar{\tau}_{\text{max}}], \tag{13}$$

where $\bar{r}_{i,j,k} = \log(1 + r_{i,j,k})$ and $\bar{F}_i$ is the CDF for the empirical distribution of logged residuals over the subgraph $\mathcal{G}_i$. Variable R_i is the set of valid pixels included in subgraph $\mathcal{G}_i$ up to a predefined maximum $\bar{\tau}_{\text{max}}$.

Fine-Stage SfM. As shown in Fig. 2, we use L2 norm BA objective, with forward and backward functions defined in Eq. (10) and Eq. (11), respectively. We evaluate over the entire graph. For both the coarse and fine stages, BA runs by applying gradient descent with a fixed number of iterations.

Extension to Camera Re-localization. In Fig. 5, camera re-localization estimates the poses of unregistered query images given a set of registered map frames. We extend to it with two changes. (1) Intrinsic initialization is skipped as provided. The poses and depth adjustments are initialized using the map frame with the highest visibility. (2) The pose graph includes both map and query frames. Gradients for map frames are disabled. We use groundtruth depth map of map frames if provided. Otherwise, we estimate depth maps using MDE, *e.g.*, the Wayspots map-free re-localization [9].

Extension to RANSAC. Using parallelized GPU implementation, we initialize 64 minimal solutions choosing the maximizer according to Eq. (9). Details in Supp.Sec. 7.1.

4. Experiments

We benchmark our method's pose accuracy on two fundamental 3D vision tasks: Structure-from-Motion and camera re-localization. For the SfM task we evaluate using small-scale ETH3D [59] IMC2021 [6], ScanNet [14], and T&T [30] datasets. For camera re-localization we evaluate using the 7-Scenes [60] and Wayspots [9] datasets. Unless otherwise specified, we use the SoTA model DUSt3R [79] as the default monocular depth estimator, while also supporting other MDEs (see Tab. 5 and Supp. Tab. 13). Following [79], we feed DUSt3R with two identical images and extract the z-channel of the resulting point cloud as the depth map. We use RoMa [20] for dense correspondence.

Implementation Details. We optimize coarse and fine stages using Adam [29] for total 50k iterations with a learning rate of 1e-3. We sample $\kappa = 200$ pixels at each frame pair. We set maximum residual value to $\bar{\tau}_{\text{max}} = 10$ for the coarse BA objective, Eq. (13), and $\tau_{\text{max}} = 20$ for the fine BA objective, Eq. (5). We include image pairs with $\nu \geq 15\%$ of their pixels co-visible. Correspondences with confidence $\chi > 0.2$ are considered valid. More in Supp.Sec. 7.6.

4.1. Structure-from-Motion Evaluations

Benchmark on ETH3D. ETH3D [59] contains unordered, high-resolution indoor and outdoor images with precisely calibrated groundtruth camera poses. In Table 2 we follow the protocol of MASt3R-SfM [18] in reporting the relative rotation accuracy (RRA) and relative translation accuracy (RTA) of poses at a fixed threshold. We achieve SoTA with substantial improvement in Tab. 2. We outperform the classic COLMAP [56], DF-SfM [28], and the scene-coordinate regression based method ACE-Zero [10]. We also outperform the learning-base VGG-SfM [73], and its counterparts FlowMap [62] and MASt3R-SfM [18], using depth maps and point clouds respectively. Notably, we outperform MASt3R-SfM using the same depth estimator of DUSt3R and a less-performant RoMa correspondence estimator. Our experiments indicate that the zero-shot monocular depth map already produces comparable or better performance than classic SfM algorithms.

Scene	COLMAP [56]		ACE-Zero [10]		FlowMap [62]		VGG-SfM [73]		DF-SfM [28]		MASt3R-SfM [18]		MBA (Ours)	
	RRA	RTA	RRA	RTA	RRA	RTA	RRA	RTA	RRA	RTA	RRA	RTA	RRA	RTA
courtyard	56.3	60.0	4.0	1.9	7.5	3.6	50.5	51.2	80.7	74.8	89.8	64.4	94.7	94.7
delivery area	34.0	28.1	27.4	1.9	29.4	23.8	22.0	19.6	82.5	82.0	83.1	81.8	83.1	83.0
electro	53.3	48.5	16.9	7.9	2.5	1.2	79.9	58.6	82.8	81.2	100.0	95.5	95.6	78.2
facade	92.2	90.0	74.5	64.1	15.7	16.8	57.5	48.7	80.9	82.6	74.3	75.3	100.0	99.2
kicker	87.3	86.2	26.2	16.8	1.5	1.5	100.0	97.8	93.5	91.0	100.0	100.0	100.0	98.9
meadow	0.9	0.9	3.8	0.9	3.8	2.9	100.0	96.2	56.2	58.1	58.1	58.1	100.0	58.1
office	36.9	32.3	0.9	0.0	0.9	1.5	64.9	42.1	71.1	54.5	100.0	98.5	100.0	86.2
pipes	30.8	28.6	9.9	1.1	6.6	12.1	100.0	97.8	72.5	61.5	100.0	100.0	100.0	96.7
playground	17.2	18.1	3.8	2.6	2.6	2.8	37.3	40.8	70.5	70.1	100.0	93.6	94.7	93.8
relief	16.8	16.8	16.8	17.0	6.9	7.7	59.6	57.9	32.9	32.9	34.2	40.2	100.0	98.9
relief 2	11.8	11.8	7.3	5.6	8.4	2.8	69.9	70.3	40.9	39.1	57.4	76.1	100.0	98.9
terrace	100.0	100.0	5.5	2.0	33.2	24.1	38.7	29.6	100.0	99.6	100.0	100.0	100.0	100.0
terrains	100.0	99.5	15.8	4.5	12.3	13.8	70.4	54.9	100.0	91.9	58.2	52.5	100.0	95.4
Average	49.0	47.8	16.4	9.7	10.1	8.8	65.4	58.9	74.2	70.7	81.2	79.7	**97.3**	**90.2**

Table 2. **Structure-from-Motion** benchmark on ETH3D dataset [59] in metrics RRA (@5°) and RTA (@5°) following MASt3R-SfM evaluation protocol. See performance with other MDE models at Supp.Tab. 13. See detailed evaluation protocol at Supp.Sec. 7.5

Type	Method	AUC@3°	AUC@5°	AUC@10°
Detector-Based	SIFT+NN + COLMAP [56] CVPR'16	24.87	34.47	45.94
	SIFT + NN + PixSfM [41] ICCV'21	26.45	35.73	47.24
	D2Net + NN + PixSfM [41] ICCV'21	10.27	13.12	17.25
	R2D2 + NN + PixSfM [41] ICCV'21	32.44	42.55	55.01
	SP + SG + PixSfM [41] ICCV'21	46.30	58.43	71.62
Detector-Free	LoFTR + PixSfM [41] ICCV'21	44.06	56.16	69.61
	LoFTR + DF-SfM [28] CVPR'24	46.55	58.74	72.19
	AspanTrans. + DF-SfM [28] CVPR'24	46.79	59.01	72.50
	MatchFormer + DF-SfM [28] CVPR'24	45.83	57.88	71.22
Dense Matching	DKM + Dense-SfM [32] CVPR'25	48.65	61.09	74.37
	RoMa + Dense-SfM [32] CVPR'25	48.48	60.79	73.90
Deep-based	VGG-SfM [73] CVPR'24	45.23	58.89	73.92
	VGGT [74] CVPR'25	39.23	52.74	71.26
	VGGT [74] + BA CVPR'25	**66.37**	**75.16**	**84.91**
Point-Based	Mast3r-SfM [18] 3DV'25	42.26	54.53	67.97
	Dense-SfM + Mast3r-SfM [18] 3DV'25	44.98	57.09	70.09
MDE	MBA (Ours)	47.29	60.37	74.59

Table 3. **Structure-from-Motion** on IMC2021 [6] dataset in AUC metric. We are comparable or better than Dense-SfM, DF-SfM, Mast3r-SfM, PixSfM, VGG-SfM and VGGT without BA.

Benchmark on IMC2021. The IMC2021 [6] dataset features internet images of tourist landmarks, organized into $1,020$ subsets each containing 5, 10, or 25 images. Fig. 1 shows example results of London Bridge. Despite including challenging elements including sky, river, tourists and trees, MDE depth maps still enable competitive pose performance. Table 3 presents quantitative results on IMC2021. Our method performs second to VGGT [74] + BA at AUC@10° and competitive results at other thresholds. We outperform the learning based methods VGG-SfM [73], and FlowMap [62] while consistently outperforming MASt3R-SfM [18]. Our method also ranks similarly with SoTA classic method Dense-SfM [32] while outperform others including COLMAP [56], PixSfM [41], and DF-SfM [28]. We observe that VGGT+BA significantly improves performance, suggesting BA remains a necessary component for foundation feed-forward SfM models. We contribute a new BA objective that exploits the density of per-pixel regressions. It potentially applies to recent feed-forward systems including VGG-SfM, VGGT , and Light3R-SfM [21] that already output dense depth maps or point clouds. Finally, we scale to large optimization problem (Tab. 1) while feed-forward baselines FlowMap and VGG-SfM run out of memory with more than 200 views [18].

Method	Inference	RRA@5 ↑	RTA@5 ↑	ATE ↓	Registration ↑
Cut3R [75] CVPR'25	FeedForward	18.8	25.8	0.017	100.0
Spann3R [71] 3DV'25		22.1	30.7	0.016	100.0
SLAM3R [42] CVPR'25		20.3	24.7	0.015	100.0
VGGT-SLAM [43] arXiv'25		57.3	67.9	**0.008**	100.0
Light3R-SfM [21] CVPR'25		52.0	52.8	0.011	100.0
SAIL [16] arXiv'25		70.4	74.7	**0.008**	100.0
GLOMAP [47] ECCV'24	Optimization	75.8	76.7	0.010	100.0
ACE0 [10] ECCV'24		56.9	57.9	0.015	100.0
DF-SfM [28] CVPR'24		69.6	69.3	0.014	76.2
FlowMap [62] 3DV'25		31.7	35.7	0.017	66.7
VGGSfM [73] CVPR'24		–	–	–	0.0
MASt3R-SfM [18] 3DV'25		49.2	54.0	0.011	100.0
DROID-SLAM [66] NeurIPS'21		31.3	40.3	0.021	100.0
SAIL-OPT [16] arXiv'25		71.5	**77.7**	**0.008**	100.0
MBA (ours)		**71.7**	77.0	0.009	100.0

Table 4. **Structure-from-Motion** on Tanks&Temple [30] dataset.

Method	Depth	Corres.	ACC@3°	ACC@5°	ACC@10°
COLMAP [56]	-	SuperPoint [17]	0.342	0.505	0.670
MBA (Ours)	ZoeDepth [5]	RoMa [20]	0.372	0.586	0.811
	DUSt3R [79]	RoMa [20]	0.403	**0.615**	0.820
	UniDepth [48]	RoMa [20]	**0.407**	0.612	**0.823**
	DUSt3R [79]	MASt3R [34]	0.384	0.596	0.811
	UniDepth [48]	MASt3R [34]	0.393	0.598	0.817

Table 5. **Structure-from-Motion** on the ScanNet dataset [14].

Benchmark on Tanks&Temples. T&T [30] is a large-scale SfM benchmark including 21 scenes, of around 300 images each, with pseudo-groundtruth from COLMAP. We follow concurrent work SAIL [16] in evaluation, where VGGSfM fails to converge on this benchmark. Our MDE-anchored method performs better or on-par with both feed-forward and optimization baselines; see Tab. 4.

Category	Method	Average (deg/cm)
FM	AS [55]PAMI'16	5.1/2.46
	HLoc [53]CVPR'19	**3.4/1.07**
E2E	SC-wLS [81]ECCV'22	6.6/1.45
	NeuMaps [65]CVPR'23	3.1/1.09
	PixLoc [54]CVPR'21	**2.9/0.98**
SCR	ACE [9]CVPR'23	2.8/0.93
	DSAC* [8]PAMI'22	2.7/1.41
	HSCNet [37]CVPR'20	2.7/0.90
	HSCNet++ [78]IJCV'24	**2.29/0.81**
APR	MAREPO [13]CVPR'24	**3.2/1.54**
MDE	MBA (Ours)	**2.7/0.92**

Table 6. **Camera relocalization** performance benchmark on the 7-Scenes dataset [60]. Per-scene results are in Supp.Tab. 15.

Benchmark on ScanNet. We further evaluate different combinations of MDE and correspondence models on ScanNet at Tab. 5. We compare with COLMAP only on frames it successfully registers, and our MDE-based SfM system still outperforms it on this large-scale SfM benchmark.

Benchmark on FastMap. We additionally compare to SoTA classic SfMs [36, 47] on large scenes in Supp.Tab. 12.

4.2. Camera Re-Localization Evaluations

Benchmark on 7-Scenes. For the task of camera relocalization on 7-Scenes, we follow DUSt3R and MAREPO [13] in categorizing baselines into four groups: feature matching (FM), end-to-end (E2E), scene coordinate regression (SCR), and absolute pose regression (APR). We follow DUSt3R in using DIR [25] retrieving the top 200 query-to-query and query-to-map images in evaluating our method. Table 6 presents relocalization metrics, ranking our method second to HSCNet++ [78], while outperforming ACE [9], DSAC* [8], MAREPO [13], and HLoc [53]. Note that our method is scene-agnostic, whereas HSCNet++ is a scene-specific approach. We note too that our method is a multi-view approach that assumes multiple test images are available, while baselines operate on each query independently.

Benchmark on Wayspots. Wayspots [9] is a map-free camera relocalization dataset comprised of handheld smartphones images. Unlike 7-Scenes, it does not provide groundtruth depth map for map images. We estimate depth maps for query and map images with DUSt3R and construct a pose graph via exhaustive image matching. Table 7 presents our results on Wayspots. Compared to baselines, we directly use rotated images without alignment (See example at Fig. 5), yet still achieve SoTA performance. This highlights our method's ability to leverage the generalization power of the MDE model. Baselines may face challenge as the scene-specific finetuning is nevertheless constrained to the texture and lighting conditions of the video.

Scene	MAREPO [13]	DSAC* [8]	ACE [9]	MBA (Ours)
Cubes	71.8%	83.8%	**97.0%**	75.1%
Bears	80.7%	82.6%	80.7%	**100%**
Winter Sign	0.0%	0.2%	1.0%	**9.3%**
Inscription	37.1%	54.1%	**49.0%**	28.3%
The Rock	99.8%	**100%**	**100%**	**100%**
Tendrils	29.3%	25.1%	34.9%	**51.5%**
Map	55.1%	56.7%	**56.5%**	45.1%
Square Bench	**70.7%**	69.5%	66.7%	58.6%
Statue	0.0%	0.0%	0.0%	0.0%
Lawn	34.2%	34.7%	35.8%	**85.0%**
Average-Accuracy	47.9%	50.7%	52.2%	**55.29%**

Table 7. **Camera relocalization** benchmark on the Wayspots [9].

Method ↓ AUC@ →	MegaDepth 5° ↑	MegaDepth 10° ↑	MegaDepth 20° ↑	ScanNet 5° ↑	ScanNet 10° ↑	ScanNet 20° ↑
LoFTR [63] CVPR'21	52.8	69.2	81.2	22.1	40.8	57.6
PDC-Net+ [68] TPAMI'23	51.5	67.2	78.5	20.3	39.4	57.1
DKM [19] CVPR'23	60.4	74.9	85.1	29.4	50.7	68.3
PMatch [85] CVPR'23	61.4	75.7	85.7	29.4	50.1	67.4
RoMa [20] CVPR'24	62.6	76.7	86.3	31.8	53.4	70.9
RoMa [20] + MAGSAC++ [3]	**68.0**	**79.8**	**88.0**	**32.9**	**54.6**	**71.3**
RoMa [20] + MBA (Ours)	66.9	79.3	87.6	32.4	54.1	71.0

Table 8. **RoMa benchmark** of two-view pose estimation with RANSAC on MegaDepth-1500 [38] and ScanNet-1500 [15].

4.3. Ablation Study

MDE Model. Our method accomodates different MDEs, and performance consistently improves with stronger ones. On ScanNet (Tab. 5), replacing ZoeDepth with UniDepth increases Acc@3° from 0.396 to 0.439. A similar improvement is observed on ETH3D; see Supp.Tab. 13.

Two-View RANSAC. Shown in the Tab. 8, the RANSAC-inspired scoring function Eq. (8) serves as an alternative scoring function in RANSAC, showing a comparable performance with the SoTA method MAGSAC++ [3].

Algorithmic Choices. In Supp.Tab. 10, we compare different optimization schemes and further demonstrate that increasing the sampling density yields marginal improvements in pose estimation accuracy. Finally, we compare the MBA loss to conventional robust loss functions [45, 57] where MBA loss consistently outperforms others.

5. Conclusion

We introduced "Marginalized Bundle Adjustment" (MBA), a multi-view pose estimation method that leverages monocular depth estimators (MDE). Our core contribution is a RANSAC-motivated BA objective enabled by dense network predictions. Our method scales to large datasets and generalizes across scenes, highlighting the potential of MDE in multi-view pose estimation task.

Limitations. Our use of a first-order optimizer leads to higher runtime compared to other SfM baselines. Beyond depth estimation models, it would be interesting to investigate a tight integration of our proposed MBA with feed-forward foundation models such as VGGT.

References

[1] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *ECCV*, 2022. 3

[2] Daniel Barath, Jiri Matas, and Jana Noskova. MAGSAC: Marginalizing Sample Consensus. In *CVPR*, 2019. 2, 3, 5, 1

[3] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. Magsac++, a fast, reliable and accurate robust estimator. In *CVPR*, 2020. 2, 8

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 1, 2

[5] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1, 2, 7, 3

[6] Xiaopeng Bi, Yu Chen, Xinyang Liu, Dehao Zhang, Ran Yan, Zheng Chai, Haotian Zhang, and Xiao Liu. Method towards cvpr 2021 image matching challenge. *arXiv preprint arXiv:2108.04453*, 2021. 2, 3, 5, 6, 7, 1

[7] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *CVPR*, 2023. 1

[8] Eric Brachmann and Carsten Rother. Visual Camera Re-Localization From RGB and RGB-D Images Using DSAC. *PAMI*, 2022. 1, 8, 4

[9] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated Coordinate Encoding: Learning to Relocalize in Minutes Using RGB and Poses. In *CVPR*, 2023. 3, 5, 6, 8, 1, 4

[10] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene Coordinate Reconstruction: Posing of Image Collections via Incremental Learning of a Relocalizer. In *ECCV*, 2024. 1, 3, 6, 7

[11] Shuai Chen, Zirui Wang, and Victor Prisacariu. Direct-PoseNet: Absolute Pose Regression with Photometric Consistency. In *3DV*, 2021. 4

[12] Shuai Chen, Xinghui Li, Zirui Wang, and Victor A Prisacariu. Dfnet: Enhance absolute pose regression with direct feature matching. In *ECCV*, 2022. 4

[13] Shuai Chen, Tommaso Cavallari, Victor Adrian Prisacariu, and Eric Brachmann. Map-Relative Pose Regression for Visual Re-Localization. In *CVPR*, 2024. 8, 4

[14] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *CVPR*, 2017. 3, 5, 6, 7, 1

[15] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2, 8, 1

[16] Junyuan Deng, Heng Li, Tao Xie, Weiqiang Ren, Qian Zhang, Ping Tan, and Xiaoyang Guo. Sail-recon: Large sfm by augmenting scene regression with localization. *arXiv preprint arXiv:2508.17972*, 2025. 7

[17] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, 2018. 7, 3

[18] Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. MASt3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *3DV*, 2025. 3, 5, 6, 7, 1, 2

[19] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense Kernelized Feature Matching for Geometry Estimation. In *CVPR*, 2023. 8

[20] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *CVPR*, 2024. 6, 7, 8, 1, 3

[21] Sven Elflein, Qunjie Zhou, and Laura Leal-Taixé. Light3r-sfm: Towards feed-forward structure-from-motion. In *CVPR*, 2025. 7

[22] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981. 2, 4

[23] Yasutaka Furukawa and Carlos Hernández. Multi-View Stereo: A Tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1

[24] Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time RGB-D camera relocalization. In *ISMAR*, 2013. 1

[25] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *ECCV*, 2016. 8

[26] Faiza Gul, Wan Rahiman, and Syed Sahal Nazli Alhady. A comprehensive study for robot navigation techniques. *Cogent Engineering*, 2019. 1

[27] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[28] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-Free Structure from Motion. In *CVPR*, 2024. 3, 6, 7, 1

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 6, 4

[30] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ToG*, 2017. 3, 5, 6, 7, 1, 2

[31] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *CVPR*, 2021. 2

[32] JongMin Lee and Sungjoo Yoo. Dense-sfm: Structure from motion with dense consistent matching. In *CVPR*, 2025. 3, 7, 1

[33] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 2

[34] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 2, 3, 7

[35] Hongdong Li and Richard Hartley. Five-point motion estimation made easy. In *ICPR*, 2006. 4

[36] Jiahao Li, Haochen Wang, Muhammad Zubair Irshad, Igor Vasiljevic, Matthew R. Walter, Vitor Campagnolo Guizilini, and Greg Shakhnarovich. Fastmap: Revisiting structure from motion through first-order optimization. *https://arxiv.org/abs/2505.04612*, 2025. 8, 1, 2, 3

[37] Xiaotian Li, Shuzhe Wang, Yi Zhao, Jakob Verbeek, and Juho Kannala. Hierarchical scene coordinate classification and regression for visual localization. In *CVPR*, 2020. 8, 4

[38] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 8, 1

[39] Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. Capturing, reconstructing, and simulating: the urbanscene3d dataset. In *ECCV*, 2022. 2

[40] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-Perfect Structure-From-Motion With Featuremetric Refinement. In *ICCV*, 2021. 3

[41] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. In *ICCV*, 2021. 7, 1

[42] Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In *CVPR*, 2025. 7

[43] Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv preprint arXiv:2505.12549*, 2025. 7

[44] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021. 1

[45] Thamsanqa Mlotshwa, Heinrich van Deventer, and Anna Sergeevna Bosman. Cauchy loss function: Robustness under gaussian and cauchy noise. In *Southern African Conference for Artificial Intelligence Research*, 2022. 8

[46] David Nistér. An efficient solution to the five-point relative pose problem. *PAMI*, 2004. 4

[47] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *ECCV*, 2024. 7, 8, 2

[48] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *CVPR*, 2024. 1, 2, 7, 3

[49] Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unik3d: Universal camera monocular 3d estimation. *arXiv preprint arXiv:2503.16591*, 2025. 2

[50] Nicholas Pilkington. Dronedeploy nerf dataset. https://github.com/nickponline/dd-nerf-dataset, 2022. 2

[51] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *PAMI*, 2022. 2

[52] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *ECCV*, 2022. 1, 2

[53] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *CVPR*, 2019. 8, 4

[54] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. Back to the Feature: Learning Robust Camera Localization From Pixels To Pose. In *CVPR*, 2021. 8, 4

[55] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *PAMI*, 2016. 8, 4

[56] Johannes L. Schonberger and Jan-Michael Frahm. Structure-From-Motion Revisited. In *CVPR*, 2016. 3, 6, 7, 1, 2

[57] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 4, 8

[58] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In *CVPR*, 2017. 5, 1, 3

[59] Thomas Schöps, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle Adjusted Direct RGB-D SLAM. In *CVPR*, 2019. 3, 6, 7, 1

[60] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *CVPR*, 2013. 3, 6, 8, 1, 4

[61] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CR, 1986. 4

[62] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. In *3DV*, 2025. 1, 3, 5, 6, 7

[63] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *CVPR*, 2021. 8

[64] Chengzhou Tang and Ping Tan. BA-net: Dense bundle adjustment networks. In *ICLR*, 2019. 2

[65] Shitao Tang, Sicong Tang, Andrea Tagliasacchi, Ping Tan, and Yasutaka Furukawa. NeuMap: Neural Coordinate Mapping by Auto-Transdecoder for Camera Localization. In *CVPR*, 2023. 8, 4

[66] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Advances in neural information processing systems*, 2021. 7

[67] Philip HS Torr and Andrew Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer vision and image understanding*, 2000. 2

[68] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. PDC-Net+: Enhanced Probabilistic Dense Correspondence Network. *PAMI*, 2023. 8

[69] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *CVPR*, 2023. 4

[70] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *CVPR*, 2022. 2

[71] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *3DV*, 2025. 7

[72] Jianyuan Wang, Christian Rupprecht, and David Novotny. PoseDiffusion: Solving Pose Estimation via Diffusion-aided Bundle Adjustment. In *ICCV*, 2023. 1, 3, 5

[73] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. VGGSfM: Visual Geometry Grounded Deep Structure From Motion. In *CVPR*, 2024. 5, 6, 7, 1, 3

[74] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 1, 3, 7

[75] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025. 2, 7

[76] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *CVPR*, 2025. 2

[77] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 2

[78] Shuzhe Wang, Zakaria Laskar, Iaroslav Melekhov, Xiaotian Li, Yi Zhao, Giorgos Tolias, and Juho Kannala. Hscnet++: Hierarchical scene coordinate classification and regression for visual localization with transformer. *IJCV*, 2024. 8, 4

[79] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2, 3, 5, 6, 7, 4

[80] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. Deepsfm: Structure from motion via deep bundle adjustment. In *ECCV*, 2020. 2

[81] Xin Wu, Hao Zhao, Shunkai Li, Yingdian Cao, and Hongbin Zha. Sc-wls: Towards interpretable feed-forward camera relocalization. In *ECCV*, 2022. 8, 4

[82] Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič, Dahua Lin, Michael Zollhöfer, and Christian Richardt. VR-NeRF: High-fidelity virtualized walkable spaces. In *SIGGRAPH Asia*, 2023. 2

[83] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 2

[84] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3r: A simple approach for estimating geometry in the presence of motion. In *ICLR*, 2025. 2

[85] Shengjie Zhu and Xiaoming Liu. Pmatch: Paired masked image modeling for dense geometric matching. In *CVPR*, 2023. 8

[86] Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: In-the-wild monocular camera calibration. In *NeurIPS*, 2023. 5, 4