

Global Structure-from-Motion Revisited

Linfei Pan¹, Dániel Baráth¹, Marc Pollefeys^{1,2}, and Johannes L. Schönberger²
¹ ETH Zurich ² Microsoft

Abstract. Recovering 3D structure and camera motion from images has been a long-standing focus of computer vision research and is known as Structure-from-Motion (SfM). Solutions to this problem are categorized into incremental and global approaches. Until now, the most popular systems follow the incremental paradigm due to its superior accuracy and robustness, while global approaches are drastically more scalable and efficient. With this work, we revisit the problem of global SfM and propose GLOMAP as a new general-purpose system that outperforms the state of the art in global SfM. In terms of accuracy and robustness, we achieve results on-par or superior to COLMAP, the most widely used incremental SfM, while being orders of magnitude faster. We share our system as an open-source implementation at <https://github.com/colmap/glomap>.

1 Introduction

Recovering 3D structure and camera motion from a collection of images remains a fundamental problem in computer vision that is highly relevant for a variety of downstream tasks, such as novel-view-synthesis [38, 50] or cloud-based mapping and localization [39, 58]. The literature commonly refers to this problem as Structure-from-Motion (SfM) [75] and, over the years, two main paradigms for solving it have emerged: *incremental* and *global* approaches. Both of them start with image-based feature extraction and matching followed by two-view geometry estimation to construct the initial view graph of the input images. Incremental methods then seed the reconstruction from two views and sequentially expand it by registering additional camera images and associated 3D structure. This sequential process interleaves absolute camera pose estimation, triangulation, and bundle adjustment, which, despite achieving high accuracy and robustness, limits scalability due to the costly repeated bundle adjustments. In contrast, global methods recover the camera geometry for all input images at once in separate rotation and translation averaging steps by jointly considering all two-view geometries in the view graph. Typically, the globally estimated camera geometry is then used as an initialization for triangulation of the 3D structure before a final global bundle adjustment step. While state-of-the-art incremental approaches are considered more accurate and robust, the reconstruction process of global approaches is more scalable and, in practice, orders of magnitude faster. In this paper, we revisit the problem of global SfM and propose a comprehensive system achieving a similar level of accuracy and robustness as state-of-the-art incremental SfM (*e.g.* Fig. 1a) while maintaining the efficiency and scalability of global approaches.

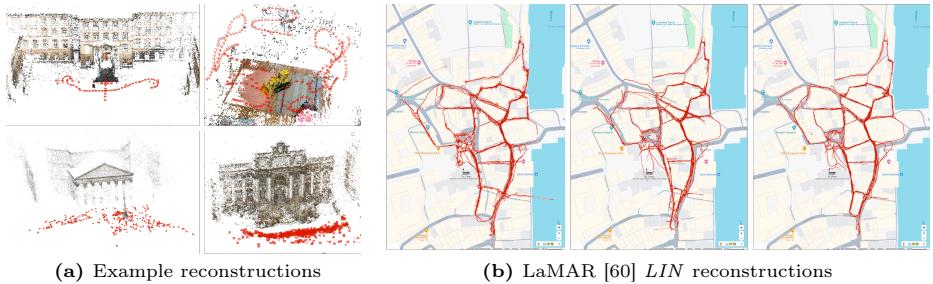


Fig. 1: Proposed GLOMAP produces satisfying reconstructions on various datasets. For (b), from left to right are estimated by Theia [71], COLMAP [62], GLOMAP. While baseline models fail to produce reliable estimations, GLOMAP achieves high accuracy.

The main reason for the accuracy and robustness gap between incremental and global SfM lies in the global translation averaging step. Translation averaging describes the problem of estimating global camera positions from the set of relative poses in the view graph with the camera orientations recovered before by rotation averaging. This process faces three major challenges in practice. The first being scale ambiguity: relative translation from estimated two-view geometry can only be determined up to scale [27]. As such, to accurately estimate global camera positions, triplets of relative directions are required. However, when these triplets form skewed triangles, the estimated scales are especially prone to noise in the observations [47]. Second, accurately decomposing relative two-view geometry into rotation and translation components requires prior knowledge of accurate camera intrinsics. Without this information, the estimated translation direction is often subject to large errors. The third challenge arises for nearly co-linear motion that leads to a degenerate reconstruction problem. Such motion patterns are common, especially in sequential datasets. These issues collectively contribute to the instability of camera position estimation, severely affecting the overall accuracy and robustness of existing global SfM systems. Motivated by the difficulties in translation averaging, significant research efforts have been dedicated to this problem. Many of the recent approaches [5, 10, 17–19, 32, 79] share a common characteristic with incremental SfM as they incorporate image points into the problem formulation. Building on this insight, we propose a global SfM system that directly combines the estimation of camera positions and 3D structure in a single global positioning step.

The main contribution of this work is the introduction of a general-purpose global SfM system, termed GLOMAP. The core difference to previous global SfM systems lies in the step of global positioning. Instead of first performing ill-posed translation averaging followed by global triangulation, our proposed method performs joint camera and point position estimation. GLOMAP achieves a similar level of robustness and accuracy as state-of-the-art incremental SfM systems [62] while maintaining the efficiency of global SfM pipelines. Unlike most previous global SfM systems, ours can deal with unknown camera intrinsics (*e.g.*, as found in internet photos) and robustly handles sequential image data

(e.g., handheld videos or self-driving car scenarios). We share our system as an open-source implementation at <https://github.com/colmap/globmap>.

2 Review of Global Structure-from-Motion

Global SfM pipelines generally consist of three main steps: correspondence search, camera pose estimation, and joint camera and structure refinement. The next sections provide a detailed review of state-of-the-art algorithms and frameworks.

2.1 Correspondence Search

Both incremental and global SfM begin with salient image feature extraction from the input images $\mathcal{I} = \{I_1, \dots, I_N\}$. Traditionally, feature points [21, 44] are detected and then described with compact signatures derived from the local context around the detection. This is followed by the search for feature correspondences between pairs of images (I_i, I_j) , which starts by efficiently identifying subsets of images [4] with overlapping fields of view and subsequently matching them in a more costly procedure [44, 59]. The matching is usually first done purely based on compact visual signatures producing a relatively large fraction of outliers initially. These are then verified by robustly [8] recovering the two-view geometry for overlapping pairs. Based on the geometric configuration of the cameras, this yields either a homography \mathbf{H}_{ij} for planar scenes with general motion and pure camera rotation with general scenes, or a fundamental matrix \mathbf{F}_{ij} (uncalibrated) and essential matrix \mathbf{E}_{ij} (calibrated) for general scenes and general motion. When the camera intrinsics are approximately known, these can be decomposed [27] into relative rotation $\mathbf{R}_{ij} \in \text{SO}(3)$ and translation $\mathbf{t}_{ij} \in \mathbb{R}^3$.

The computed two-view geometries with associated inlier correspondences define the view graph \mathcal{G} that serves as the input to the global reconstruction steps. In our pipeline, we rely on COLMAP’s [62] correspondence search implementation [61] with RootSIFT features and scalable bag-of-words image retrieval [63] to find candidate overlapping pairs for brute-force feature matching.

2.2 Global Camera Pose Estimation

Global camera pose estimation is the key step distinguishing global from incremental SfM. Instead of sequentially registering cameras with repeated triangulation and bundle adjustment, global SfM seeks to estimate all the camera poses $\mathbf{P}_i = (\mathbf{R}_i, \mathbf{c}_i) \in \text{SE}(3)$ at once using the view graph \mathcal{G} as input. To make the problem tractable, it is typically decomposed into separate rotation and translation averaging steps [53, 71] with some works also refining the view graph before [72], or directly estimating camera poses from the view-graph of two-view geometries [35, 36]. The main challenge lies in dealing with noise and outliers in the view graph by careful modeling and solving of the optimization problems.

Rotation Averaging, sometimes also referred to as rotation synchronization, has been studied for several decades [26, 49] and is related to pose graph optimization (PGO) algorithms [12, 13]. It is typically formulated as a non-linear

optimization, penalizing the deviation of the global rotation from estimated relative poses. Specifically, absolute rotations \mathbf{R}_i and relative rotations \mathbf{R}_{ij} should ideally satisfy the constraint $\mathbf{R}_{ij} = \mathbf{R}_j \mathbf{R}_i^\top$. However, in practice, this does not hold exactly due to noise and outliers. Thus, the problem is generally modeled with a robust least-metric objective and optimized as:

$$\arg \min_{\mathbf{R}} \sum_{i,j} \rho(d(\mathbf{R}_j^\top \mathbf{R}_{ij} \mathbf{R}_i, \mathbf{I})^p). \quad (1)$$

Hartley *et al.* [26] provide a comprehensive overview of various choices of robustifiers ρ (*e.g.*, Huber), rotation parameterizations (*e.g.*, quaternion or axis-angle), and distance metrics d (*e.g.*, chordal distance or geodesic distance).

Based on these principles, a multitude of methods have been proposed. Govindu [24] linearizes the problem via quaternions, while Martinec and Papadilla [49] relax the problem by omitting certain constraints on rotation matrices. Eriksson *et al.* [22] leverage strong duality. The tractability condition of the problem is examined by Wilson *et al.* [78]. Approaches utilizing semidefinite programming-based (SDP) relaxations [5, 23] ensure optimality guarantees by minimizing chordal distances [26]. Dellaert *et al.* [20] sequentially elevate the problem into higher-dimensional rotations within $\text{SO}(n)$ to circumvent local minima where standard numerical optimization techniques might fail [40, 48]. Various robust loss functions have been explored to handle outliers [14, 15, 25, 68, 82]. Recently, learning-based methodologies have emerged. NeuRoRa [57], MSP [81], and PoGO-Net [41] leverage Graph Neural Networks to eliminate outliers and to estimate absolute camera poses. DMF-synch [73] relies on matrix factorization techniques for pose extraction. In this work, we use our own implementation of Chatterjee *et al.* [14] as a scalable approach that provides accurate results in presence of noisy as well as outlier-contaminated input rotations.

Translation Averaging. After rotation averaging, the rotations \mathbf{R}_i can be factored out from the camera poses. What remains to be determined are the camera positions \mathbf{c}_i . Translation averaging describes the problem of estimating global camera positions that are maximally consistent with the pairwise relative translations \mathbf{t}_{ij} based on the constraint $\mathbf{t}_{ij} = \frac{\mathbf{c}_j - \mathbf{c}_i}{\|\mathbf{c}_j - \mathbf{c}_i\|}$. However, due to noise and outliers as well as the unknown scale of the relative translations, the task is especially challenging. In principle, the camera pose can be uniquely determined if the view graph has the property of *parallel rigidity*. Parallel rigidity, also known as bearing rigidity, has been researched in different fields of computer vision [7, 56], robotics [37], decision and control [84] as well as computer-aided design [67]. Arrigoni *et al.* [6] offers a unified review of this topic.

Different translation averaging methods have been proposed over the past years. The pioneering work by Govindu [24] minimizes the cross-product between the relative camera locations and the observed directions. Jiang *et al.* [34] linearizes the problem in units of triplets. Wilson *et al.* [79] optimizes the difference of directions directly and designs a dedicated outlier filtering mechanism. Ozyesil *et al.* [55] proposes a convex relaxation to the original problem and solves the Least Unsquared Deviations (LUD) problem with an L_1 loss for ro-

bustness. Zhuang *et al.* [85] realizes the sensitivity of the LUD method in terms of camera baseline and proposes the Bilinear Angle-based Translation Averaging (BATA) error for optimization. While significant improvements have been made in these works, translation averaging generally only works reliably when the view graph is well connected. The problem is also inherently ill-posed and sensitive to noisy measurements when cameras are subject to or close to co-linear motion. Furthermore, extraction of relative translation from two-view geometry is only possible with known camera intrinsics. When such information is inaccurate, the extracted translations are not reliable. Inspired by the observation that point tracks generally help in translation averaging [5, 17–19, 32, 79], in our proposed system, we *skip* the step of translation averaging. Instead, we directly perform a joint estimation of the camera and point positions. We refer to this step as global positioning with details introduced in Section 3.2.

Structure for Camera Pose Estimation. Several works have explored incorporating 3D structure into camera position estimation. Ariel *et al.* [5] directly use the correspondences in two-view geometry for estimating global translation. Wilson *et al.* [79] discovered that 3D points can be treated in a similar manner as the camera centers and, thus, can be easily incorporated into the optimization. Cui *et al.* [18] extends [34] by including point tracks into the optimization problem with linear relations. To reduce scale drifting, Holynski *et al.* [32] integrates line and plane features into the optimization problem. Manam *et al.* [46] incorporates the correspondences by reweighing the relative translation in the optimization. LiGT [10] proposes a “pose only” method for solving the camera positions with a linear global translation constraint imposed by points. The common theme of these works is that incorporating constraints on the 3D scene structure aids the robustness and accuracy of camera position estimation, which we take as an inspiration for our work.

2.3 Global Structure and Pose Refinement

After recovering the cameras, the global 3D structure can be obtained via triangulation. Together with the camera extrinsics and intrinsics, the 3D structure is then typically refined using global bundle adjustment.

Global Triangulation. Given two-view matches, transitive correspondences can be leveraged for boosting completeness and accuracy [62]. Moulon *et al.* [52] presents an efficient way for concatenating tracks. Triangulating multi-view points has a long research history [29, 45]. Common practices for such a task are the direct linear transformation (DLT) and midpoint methods [2, 27, 29]. Recently, LOST [31] was proposed as an uncertainty-based triangulation. Yet, the above triangulation mechanisms often break in the presence of arbitrary levels of outliers. In this regard, Schönberger *et al.* [62] proposes a RANSAC-based triangulation scheme, seeking to establish multiple point tracks in the presence of mismatches. Instead, our approach directly estimates 3D points by a single, joint global optimization method together with the camera positions (see Section 3.2).

Global Bundle Adjustment is essential in obtaining accurate final 3D structure $\mathbf{X}_k \in \mathbb{R}^3$, camera extrinsics \mathbf{P}_i and camera intrinsics π_i . It is formulated as

a joint robust optimization by minimizing reprojection errors as

$$\arg \min_{\pi, \mathbf{P}, \mathbf{X}} \sum_{i,k} \rho(\|\pi_i(\mathbf{P}_i, \mathbf{X}_k) - x_{ik}\|_2). \quad (2)$$

Please refer to Triggs *et al.* [74] for a comprehensive review of bundle adjustment.

2.4 Hybrid Structure-from-Motion

To combine the robustness of incremental and efficiency of global SfM, previous works have formulated hybrid systems. HSfM [17] proposes to incrementally estimate camera positions with rotations. Liu *et al.* [43] proposes a graph partitioning method by first dividing the whole set of images into overlapping clusters. Within each cluster, camera poses are estimated via a global SfM method. However, such methods are still not applicable when camera intrinsics are inaccurate according to their formulation. Our method overcomes this limitation by different modeling of the objective in the global positioning step.

2.5 Frameworks for Structure-from-Motion

Multiple open-source SfM pipelines are available. The incremental SfM paradigm is currently the most widely used due to its robustness and accuracy in real-world scenarios. Bundler [69] and VisualSfM [80] are systems dating back a decade ago. Building upon these, Schönberger *et al.* [62] developed COLMAP, a general-purpose SfM and multi-view stereo [64] system. COLMAP is versatile and has demonstrated robust performance across many datasets, making it the standard tool for image-based 3D reconstruction in recent years.

Several open-source pipelines are available for global SfM as well. OpenMVG [53] stands out as a prominent framework in this category. Starting with geometrically verified matches, it estimates the relative pose using a contrario RANSAC [51]. Following this, OpenMVG assesses rotation consistency through adjusted cycle length weighting to eliminate outlier edges and solves for global rotation using the remaining edges with a sparse eigenvalue solver. Global translations are refined through the trifocal tensor and then subjected to translation averaging using the L_∞ method. Finally, OpenMVG performs global triangulation via per-point optimization and a global bundle adjustment.

Theia [71] is another well-established global SfM pipeline. It adopts a similar approach to OpenMVG by initially estimating global rotations via averaging and then estimating camera positions through translation averaging. For rotation averaging, Theia employs a robust scheme from [14]. For translation averaging, it defaults to using the LUD method [55]. The pipeline concludes with global triangulation and bundle adjustment, similar to OpenMVG.

Several learning-based pipelines are available. PixSfM [42] proposes a joint refinement mechanism over features and structure to achieve sub-pixel accurate reconstruction and can be combined with our system. VGGsfM [76] proposes an end-to-end learning framework for the SfM task, and Zhuang *et al.* [83] proposes

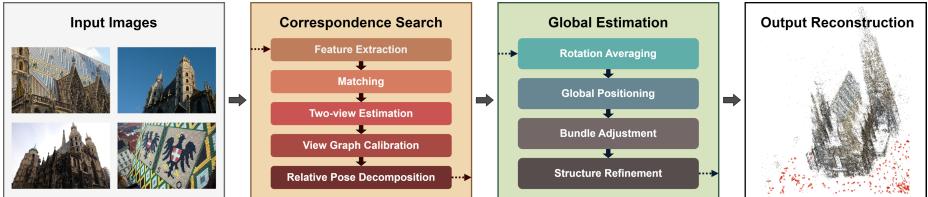


Fig. 2: Pipeline of proposed GLOMAP system, a global Structure-from-Motion framework, that distinguishes itself from other global methods by merging the translation averaging and triangulation phase into a single global positioning step.

to operate on pixel-wise correspondences to regress camera position directly. However, these two methods are limited to handling tens of images.

In this paper, we propose a new end-to-end global SfM pipeline and release it to the community as an open-source contribution to facilitate downstream applications and further research.

3 Technical Contributions

This section presents our key technical contributions to improve upon the state of the art in global SfM (*cf.* Fig. 2) and close the gap to incremental SfM in terms of robustness and accuracy.

3.1 Feature Track Construction

Feature tracks must be carefully constructed as to achieve accurate reconstruction. We start by only considering inlier feature correspondences produced by two-view geometry verification. In this step, we distinguish between the initial classification of the two-view-geometry [62]: if homography \mathbf{H} best describes the two-view geometry, we use \mathbf{H} for the verification of inliers. The same principle is applied to essential matrix \mathbf{E} and fundamental matrix \mathbf{F} . We further filter outliers by performing a cheirality test [28, 77]. Matches that are close to any of the epipoles or have small triangulation angles are also removed to avoid singularities due to large uncertainties. After pairwise filtering of all view graph edges, we form feature tracks by concatenating all remaining matches.

3.2 Global Positioning of Cameras and Points

This step aims to jointly recover point and camera positions (see Fig. 3). Instead of performing translation averaging followed by global triangulation, we directly perform a joint global triangulation and camera position estimation. Different from most previous works, our objective function is initialization-free and consistently converges to a good solution in practice. In standard incremental and global SfM systems, feature tracks are verified and optimized by reprojection errors to ensure reliable and accurate triangulations. However, the reprojection error across multiple views is highly non-convex, thus requiring careful

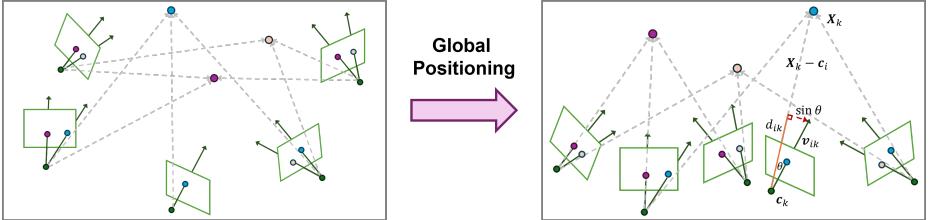


Fig. 3: Global Positioning. The left figure visualizes the initial configuration, depicting randomly initialized cameras and points. Black arrows, traversing through colored circles on the image planes, denote the measurements. Dashed lines represent the actual image rays, which are subject to optimization by adjusting the positions of the points and the cameras while their orientations remain constant. The right figure displays the outcome following the minimization of angles between the measurements (solid lines) and the image rays from the 3D points (dashed lines).

initialization. Moreover, the error is unbounded, so it is not robust to outliers. To overcome these challenges, we build upon the objective function proposed by [85] and use normalized direction differences as an error measure. The original formulation was proposed in terms of the relative translations, whereas, in our formulation, we discard the relative translation constraints and only include camera ray constraints. Concretely, our problem is modeled and optimized as:

$$\arg \min_{\mathbf{X}, \mathbf{c}, d} \sum_{i,k} \rho(\|\mathbf{v}_{ik} - d_{ik}(\mathbf{X}_k - \mathbf{c}_i)\|_2), \quad \text{subject to } d_{ik} \geq 0, \quad (3)$$

where the \mathbf{v}_{ik} is the globally rotated camera ray observing point \mathbf{X}_k from camera \mathbf{c}_i , while d_{ik} is a normalizing factor. We use Huber [33] as a robustifier ρ and Levenberg–Marquardt [40] from Ceres [3] as the optimizer. All point and camera variables are initialized by a uniform random distribution in the range $[-1, 1]$ while the normalization factors are initialized as $d_{ik} = 1$. We down-weight terms involving cameras with unknown intrinsics by factor 2 to reduce their influence.

Compared to reprojection errors, this has several advantages. The first is robustness. While reprojection errors are unbounded, the above is equivalent to

$$\begin{cases} \sin \theta & \text{if } \theta \in [0, \pi/2], \\ 1 & \text{if } \theta \in [\pi/2, \pi], \end{cases} \quad (4)$$

where θ is the angle between \mathbf{v}_{ik} and $\mathbf{X}_k - \mathbf{c}_i$ for optimal d_{ik} [85]. Thus, the error is strictly bounded to the range $[0, 1]$. As such, outliers do not heavily bias the result. Secondly, the objective function, as we experimentally show, converges reliably with random initialization due to its bilinear form [85].

Compared with classical translation averaging, discarding the relative translation terms in the optimization has two key advantages. First, the applicability of our method on datasets with inaccurate or unknown camera intrinsics as well as degenerate cameras not following the expected pinhole model (*e.g.*, when dealing with arbitrary internet photos). This is because the knowledge of accurate intrinsics is required to solve for relative translation. When they deviate

from the expected value, the estimated two-view translations suffer from large errors. Since translation averaging is inherently ill-posed due to unknown scale, recovering camera positions from noisy and outlier-contaminated observations is challenging, especially as relative translation errors exacerbate with longer baselines. Our proposed pipeline, instead, relies on careful filtering of two-view geometry and the error is defined w.r.t. the camera rays. Hence, poor camera intrinsics only bias the estimation of individual cameras instead of also biasing other overlapping cameras. Second, the applicability of global SfM in co-linear motion scenarios, which is a known degenerate case for translation averaging. Compared to pairwise relative translations, feature tracks constrain multiple overlapping cameras. As such, our proposed pipeline can deal more reliably in common forward or sideward motion scenarios (see Section 4.3).

3.3 Global Bundle Adjustment

The global positioning step provides a robust estimation for cameras and points. However, the accuracy is limited, especially when camera intrinsics are not known in advance. As a further refinement, we perform several rounds of global bundle adjustment using Levenberg-Marquardt and the Huber loss as a robustifier. Within each round, camera rotations are first fixed, then jointly optimized with intrinsics and points. Such a design is particularly important for reconstructing sequential data. Before constructing the first bundle adjustment problem, we apply a pre-filtering of 3D point observations based on the angular error while allowing a larger error for uncalibrated cameras. Afterward, we filter tracks based on reprojection errors in image space. Iterations are halted when the ratio of filtered tracks falls below 0.1%.

3.4 Camera Clustering

For images collected from the internet, non-overlapping images can be wrongly matched together. Consequently, different reconstructions collapse into a single one. To overcome this issue, we post-process the reconstruction by performing clustering of cameras. First, the covisibility graph \mathcal{G} is constructed by counting the number of visible points for each image pair. Pairs with fewer than 5 counts are discarded as the relative pose cannot be reliably determined below this number, and the median of the remaining pairs is used to set inlier threshold τ . Then, we find well-constrained clusters of cameras by finding strongly connected components in \mathcal{G} . Such components are defined by only connecting pairs with more than τ counts. Afterwards, we carefully attempt to merge two strong components, if there are at least two edges with more than 0.75τ counts. We recursively repeat this procedure until no more clusters can be merged. Each connected component is output as a separate reconstruction.

3.5 Proposed Pipeline

The pipeline of the proposed method is summarized in Fig. 2. It consists of two major components: *correspondence search* and *global estimation*. For correspon-

Table 1: Result on ETH3D SLAM [65] dataset. Each row represents the average results on scenes with the same prefix. The proposed system outperforms global SfM baselines by a large margin, and also achieves better results than COLMAP [62].

	Recall @ 0.1m				AUC @ 0.1m				AUC @ 0.5m				Time (s)												
	OpenMVG		Theia		GLOMAP		COLMAP		OpenMVG		Theia		GLOMAP		COLMAP		OpenMVG		Theia		GLOMAP		COLMAP		
cables	76.8	88.0	88.0	68.0	59.1	60.2	76.2	56.5	77.4	82.4	85.6	70.4	103.1	339.5	195.6	2553.4	1.5	5.3	10.0	196.2	1.5	5.3	10.0	196.2	
camera	2.2	26.6	32.6	34.8	2.0	12.9	21.9	24.7	2.2	26.2	30.8	33.2	78.3	52.6	111.0	1057.7	78.3	52.6	111.0	1057.7	78.3	52.6	111.0	1057.7	
ceiling	6.4	22.8	28.7	28.6	2.9	17.0	22.3	15.3	8.2	21.7	27.4	26.3	376.2	195.3	150.0	1115.1	376.2	195.3	150.0	1115.1	376.2	195.3	150.0	1115.1	
desk	28.0	32.2	32.3	24.4	16.0	29.2	28.5	21.1	28.6	31.6	31.6	23.7	150.3	70.5	142.1	1230.8	150.3	70.5	142.1	1230.8	150.3	70.5	142.1	1230.8	
einstein	32.9	47.8	48.5	33.6	22.7	32.1	36.5	25.4	34.2	46.7	47.9	36.7	114.4	356.7	144.3	731.2	114.4	356.7	144.3	731.2	114.4	356.7	144.3	731.2	
kidnap	73.1	73.3	73.3	73.3	63.4	62.3	70.3	68.5	71.2	71.1	72.7	72.3	91.9	60.2	77.6	983.8	91.9	60.2	77.6	983.8	91.9	60.2	77.6	983.8	
large	35.4	48.6	49.0	44.5	18.1	37.8	45.8	20.7	33.7	46.6	48.4	43.4	33.5	29.7	44.3	301.2	33.5	29.7	44.3	301.2	33.5	29.7	44.3	301.2	
mannequin	49.1	62.2	67.4	59.3	36.7	46.5	61.4	52.8	49.1	59.7	66.4	58.4	859.7	109.0	788.9	9995.1	859.7	109.0	788.9	9995.1	859.7	109.0	788.9	9995.1	
motion	18.8	16.9	39.8	17.7	11.0	11.9	22.5	12.9	23.3	19.2	45.9	19.7	313.8	167.5	533.3	2349.7	313.8	167.5	533.3	2349.7	313.8	167.5	533.3	2349.7	
planar	30.6	100.0	100.0	100.0	12.5	97.8	98.7	98.3	38.0	99.6	99.7	99.7	21.7	35.7	28.7	202.7	21.7	35.7	28.7	202.7	21.7	35.7	28.7	202.7	
plant	77.0	89.1	93.3	92.8	62.9	75.3	82.0	82.3	77.1	88.2	93.4	92.5	721.3	118.3	434.4	6573.9	721.3	118.3	434.4	6573.9	721.3	118.3	434.4	6573.9	
reflective	12.6	16.1	22.0	26.2	6.7	9.0	12.1	9.2	16.5	23.0	31.3	33.5	63.2	136.8	74.5	561.1	63.2	136.8	74.5	561.1	63.2	136.8	74.5	561.1	
repetitive	26.3	28.5	32.7	28.5	23.9	15.2	29.2	27.2	25.8	27.0	32.0	28.3	91.7	170.5	239.7	469.6	91.7	170.5	239.7	469.6	91.7	170.5	239.7	469.6	
sfn	83.2	94.1	97.0	55.9	55.9	57.9	53.7	79.6	35.7	80.0	88.7	95.2	55.8	182.0	97.8	221.5	2777.7	182.0	97.8	221.5	2777.7	182.0	97.8	221.5	2777.7
sofa	11.2	22.0	23.9	32.2	5.5	13.1	22.1	28.6	10.3	21.3	23.5	31.6	9.1	8.9	10.1	157.3	9.1	8.9	10.1	157.3	9.1	8.9	10.1	157.3	
table	79.1	93.7	94.3	99.9	68.2	84.3	95.6	76.9	89.2	92.3	99.1	76.9	180.0	97.8	221.5	2777.7	180.0	97.8	221.5	2777.7	180.0	97.8	221.5	2777.7	
vicon	64.6	84.2	97.0	81.1	20.4	57.1	80.5	38.9	71.7	87.6	93.7	75.0	50.9	88.2	46.2	474.8	50.9	88.2	46.2	474.8	50.9	88.2	46.2	474.8	
Average	48.2	62.8	66.4	57.9	34.9	46.0	57.0	47.6	48.6	61.1	65.7	57.9	120.8	91.8	133.5	1115.4	120.8	91.8	133.5	1115.4	120.8	91.8	133.5	1115.4	

dence search, it starts with feature extractions and matching. Two-view geometry, including fundamental matrix, essential matrix, and homography, are estimated from the matches. Geometrically infeasible matches are excluded. Then, view graph calibration is performed similar to Sweeney *et al.* [72] on geometrically verified image pairs. With the updated camera intrinsics, relative camera poses are estimated. As for global estimation, global rotations are estimated via averaging [14] and inconsistent relative poses are filtered by thresholding the angular distance between \mathbf{R}_{ij} and $\mathbf{R}_j \mathbf{R}_i^\top$. Then, the positions of cameras and points are jointly estimated via global positioning, followed by global bundle adjustment. Optionally, the accuracy of the reconstruction can be further boosted with structure refinement. Within this step, points are retriangulated with the estimated camera pose, and rounds of global bundle adjustment are performed. Camera clustering can also be applied to achieve coherent reconstructions.

4 Experiments

To demonstrate the performance of our proposed *GLOMAP* system, we conduct extensive experiments on various datasets, ranging from calibrated to uncalibrated and from unordered to sequential scenarios. More specifically, we compare against the state-of-the-art frameworks (OpenMVG [53], Theia [71], COLMAP [62]) on the ETH3D [65, 66], LaMAR [60], Image Matching Challenge 2023 (IMC 2023) [16], and MIP360 [9] datasets. Furthermore, we present ablations to study the behavior of different components of our proposed system. **Metrics.** For all evaluations, we adopt two standard metrics. For unordered image data, we report the AUC (Area Under the recall Curve) scores calculated from the maximum of relative rotation and translation error between every image pair, similar to [16, 30, 65]. Such an error formulation considers the deviation between every possible camera pair. For sequential image data, we report the AUC scores calculated from the camera position error after globally aligning the

Table 2: Results on ETH3D MVS (rig) [66]. Proposed GLOMAP largely outperforms other SfM systems by a large margin while maintaining the efficiency of global SfM.

	AUC @ 1°			AUC @ 3°			AUC @ 5°			Time (s)						
	OpenMVG	Theia	GLOMAP	COLMAP	OpenMVG	Theia	GLOMAP	COLMAP	OpenMVG	Theia	GLOMAP	COLMAP	OpenMVG	Theia	GLOMAP	COLMAP
delivery_area	0.0	47.8	75.9	66.6	0.1	81.0	91.2	87.5	0.3	88.2	94.6	92.2	235.7	259.5	519.9	1745.9
electro	0.2	25.9	47.5	38.4	1.8	61.6	72.9	65.2	2.9	73.5	81.7	75.2	99.9	151.8	429.2	3283.1
forest	0.0	65.6	74.1	74.7	0.0	87.1	90.0	90.3	0.1	91.8	93.5	93.7	306.6	563.2	1658.4	7571.6
playground	0.0	23.1	40.0	0.0	0.1	62.1	72.5	0.0	0.2	74.1	81.0	0.0	121.5	598.6	1008.3	350.2
terrains	0.7	39.6	50.2	48.0	2.6	71.2	78.1	76.8	3.1	79.6	85.3	84.3	62.6	179.4	353.0	1333.9
Average	0.2	40.4	57.6	45.5	0.9	72.6	80.9	64.0	1.3	81.4	87.2	69.1	165.3	350.5	793.8	2857.0

Table 3: Results on ETH3D MVS (DSLR) [66]. On this dataset, the proposed method outperforms other global SfM baselines and is comparable to COLMAP [62].

	AUC @ 1°			AUC @ 3°			AUC @ 5°			Time (s)							
	OpenMVG	Theia	GLOMAP	COLMAP	OpenMVG	Theia	GLOMAP	COLMAP	OpenMVG	Theia	GLOMAP	COLMAP	OpenMVG	Theia	GLOMAP	COLMAP	
courtyard	68.2	91.2	87.8	87.3	81.7	97.0	95.9	95.8	85.9	98.2	97.6	97.5	11.2	11.9	24.5	39.2	
delivery_area	91.4	92.0	92.8	92.3	97.1	97.3	97.6	97.4	98.3	98.4	98.6	98.5	235.7	3.8	9.2	26.4	
electro	72.6	49.8	79.3	70.0	81.8	54.9	86.7	75.7	83.9	57.9	88.5	77.0	99.9	2.8	7.5	23.9	
façade	89.4	88.1	91.1	90.3	96.4	94.3	97.0	96.7	97.8	95.5	98.2	98.0	35.1	47.4	91.4	113.5	
kicker	82.0	71.0	86.3	86.6	89.6	77.6	94.3	91.3	91.2	79.2	96.5	92.2	1.7	3.5	6.4	16.2	
meadow	10.5	17.3	74.7	61.6	13.8	21.5	90.4	77.8	14.7	23.4	94.2	81.5	0.3	1.8	2.1	5.1	
office	19.3	27.6	59.6	45.2	24.1	32.8	81.8	56.5	25.6	35.5	88.6	59.9	0.4	0.7	1.3	19.1	
pipes	29.4	41.0	89.8	86.3	36.3	53.5	96.6	95.4	41.1	56.7	98.0	97.3	0.3	0.6	0.8	3.4	
playground	49.4	72.3	91.2	90.6	55.6	77.8	97.0	96.8	56.9	78.9	98.2	98.1	121.5	3.0	7.7	29.8	
relief	89.8	66.9	93.7	93.3	96.6	73.0	97.9	97.8	97.9	76.4	98.7	98.7	3.7	4.7	12.7	34.1	
relief_2	11.9	94.1	95.0	94.9	12.4	98.0	98.3	98.3	98.2	12.5	98.8	99.0	98.9	0.7	3.2	5.4	24.1
terrace	91.9	94.0	92.8	92.3	97.3	98.0	97.6	97.4	98.4	98.8	98.6	98.5	1.1	1.6	4.0	10.9	
terrains	70.9	86.4	82.2	81.9	89.3	95.3	93.7	93.6	93.4	97.1	96.2	96.1	62.6	2.1	6.8	21.7	
botanical_garden	26.0	51.0	87.2	5.1	30.1	74.5	95.6	5.3	30.9	82.4	97.3	5.4	1.0	1.0	4.2	9.0	
boilers	89.6	89.8	91.0	90.6	96.4	96.5	97.0	96.9	97.9	97.9	98.2	98.1	4.1	2.6	8.3	18.5	
bridge	44.9	89.8	91.8	91.6	47.9	95.3	97.2	97.2	48.5	96.4	98.3	98.3	35.1	31.3	87.2	91.9	
door	93.8	93.8	95.1	96.9	97.9	97.9	98.4	99.0	98.8	98.8	99.0	99.4	1.2	1.0	1.8	4.2	
exhibition_hall	11.0	84.4	25.5	85.4	15.2	93.5	29.7	94.3	16.3	96.0	30.7	96.5	20.6	49.8	68.3	72.0	
lecture_room	80.2	69.4	83.3	82.9	92.7	79.6	94.1	94.1	95.6	82.6	96.5	96.4	3.3	2.4	7.4	10.9	
living_room	88.0	84.8	88.0	88.3	95.8	92.6	95.8	95.8	97.4	94.3	97.4	97.5	9.4	10.5	22.5	49.1	
lounge	34.1	34.1	34.0	33.9	35.4	35.4	35.3	35.3	35.6	35.6	35.6	35.6	0.4	1.0	1.2	1.8	
observatory	58.4	65.8	65.4	64.4	83.7	87.1	87.0	86.5	89.9	92.2	92.1	91.8	5.8	4.8	13.2	24.4	
old_computer	23.3	49.8	53.6	58.7	41.5	59.8	61.4	90.2	48.4	62.0	63.0	92.7	5.9	4.1	8.9	19.6	
statue	96.4	98.8	98.8	98.7	98.8	99.6	99.6	99.6	99.8	99.8	99.8	99.7	0.9	1.6	5.7	7.4	
terrace_2	87.9	90.8	91.0	90.7	95.7	96.9	96.9	96.8	97.4	98.1	98.2	98.1	1.0	1.5	3.3	10.2	
Average	60.4	71.8	80.8	79.2	68.1	79.2	88.5	86.5	70.1	81.2	90.3	88.1	26.5	7.9	16.5	27.5	

reconstruction to the ground truth using a robust RANSAC scheme [62]. When images are taken in sequences, especially in the case when cameras are nearly co-linear, the relative error does not capture the scale drift well. Thus, we directly focus on the camera positions. For a fair comparison, we use the same feature matches as input to all methods and thus also exclude correspondence search from the reported runtimes. We also tried using OpenMVG’s and Theia’s correspondence search implementations but consistently obtained better results using COLMAP. We employ Kapture [1] for importing verified matches to OpenMVG. We use fixed settings for GLOMAP and the default recommended settings for OpenMVG and Theia.

4.1 Calibrated Image Collections

ETH3D SLAM [65] is a challenging dataset containing sequential data with sparse features, dynamic objects, and drastic illumination changes. We evaluated our method on the training sequences that come with millimeter-accurate ground truth. Ground truth is not available for test sequences and some frames, so we do not consider them. The results are presented in Table 1. Each row

Table 4: Results on LaMAR [60] datasets. The proposed method largely outperforms other baselines as well as COLMAP [62]. For LIN, structure refinement is not performed for GLOMAP due to memory limitation (marked as *).

	Recall @ 1m				AUC @ 1m				AUC @ 5m				Time (s)											
	OpenMVG		Theia		GLOMAP		COLMAP		OpenMVG		Theia		GLOMAP		COLMAP		OpenMVG		Theia		GLOMAP		COLMAP	
CAB	-	6.0	11.6	13.0	-	3.2	4.7	5.8	-	9.3	16.9	19.2	-	1345.6	6162.2	1194033.6	-	1345.6	6162.2	1194033.6	-	1345.6	6162.2	1194033.6
HGE	-	8.3	48.4	38.9	-	2.9	22.2	18.0	-	9.4	50.3	46.9	-	1182.4	12587.2	249771.1	-	1182.4	12587.2	249771.1	-	1182.4	12587.2	249771.1
LIN*	-	18.8	87.3	44.2	-	7.0	46.7	17.7	-	38.5	85.6	52.3	-	2097.9	18466.6	620176.4	-	2097.9	18466.6	620176.4	-	2097.9	18466.6	620176.4
Average	-	11.0	49.1	32.0	-	4.4	24.5	13.8	-	19.1	50.9	39.4	-	1542.0	12405.3	354660.4	-	1542.0	12405.3	354660.4	-	1542.0	12405.3	354660.4

in the table averages the results across sequences sharing the same prefix with full results in the suppl. material. The results demonstrate that our proposed GLOMAP system achieves approximately 8% higher recall and scores 9 and 8 additional points in AUC at the 0.1m and 0.5m thresholds, respectively, compared to COLMAP, which is also one order of magnitude slower. Against other global SfM pipelines, GLOMAP shows a 18% and 4% improvement in recall and around 11 points higher AUC at 0.1m, confirming its robustness.

ETH3D MVS (rig) [66] contains, per scene, about 1000 multi-rig exposures with each 4 images. The dataset contains both outdoor and indoor scenes with millimeter-accurate ground truth for 5 training sequences. We do not fix the pose of the rig for any of the methods. Results on this dataset can be found in Table 2. Ours successfully reconstructs all scenes. In contrast, OpenMVG performs poorly on all scenes while COLMAP fails for one, and Theia performs consistently worse than ours. On the sequences where COLMAP succeeds, ours achieves similar or higher accuracy. Our runtime is a little slower than global SfM baselines and about 3.5 times faster than COLMAP.

ETH3D MVS (DSLR) [66] features an unordered collection of high-resolution images of outdoor and indoor scenes with millimeter-accurate ground truth for both training and testing sequences and results reported in Table 3. Consistent with other ETH3D datasets, ours outperforms OpenMVG and Theia while achieving similar accuracy as COLMAP. For *exhibition_hall*, GLOMAP performs inaccurately because of rotational symmetry of the scene, causing rotation averaging to collapse. Due to the small scale of the scenes, all methods achieve comparable runtimes.

LaMAR [60] is a large-scale indoor and outdoor benchmark with each scene containing several tens of thousands of images captured by a variety of AR devices and smartphones. For this dataset, we use the retrieval pipeline from the benchmark [60] to establish matches. The results on this dataset can be found in Table. 4, and the qualitative result can be found in Figure 1b. GLOMAP achieves significantly more accurate reconstruction on *HGE* and *LIN* compared to all other baselines, including COLMAP [62] while being orders of magnitude faster than COLMAP. On *CAB*, all methods, including COLMAP, perform poorly, especially upon visual inspection, on this extremely challenging benchmark due to many forward motion trajectories, drastic day-night illumination changes, and many symmetries across floors/rooms and repetitive facades.

Table 5: Results on IMC 2023 [16]. Our GLOMAP method comes close to COLMAP generated ground truth while outperforming global SfM by a large margin.

	AUC @ 3°				AUC @ 5°				AUC @ 10°				Time (s)							
	OpenMVG		Theia	GLOMAP	COLMAP	OpenMVG		Theia	GLOMAP	COLMAP	OpenMVG		Theia	GLOMAP	COLMAP	OpenMVG		Theia	GLOMAP	COLMAP
	OpenMVG	Theia	GLOMAP	COLMAP	OpenMVG	Theia	GLOMAP	COLMAP	OpenMVG	Theia	GLOMAP	COLMAP	OpenMVG	Theia	GLOMAP	COLMAP	OpenMVG	Theia	GLOMAP	COLMAP
bike	-	0.0	35.0	77.9	-	0.0	38.9	86.7	-	0.0	41.9	93.4	-	1.4	1.5	1.1	-	1.7	1.4	0.6
chairs	-	0.0	82.6	78.8	-	0.0	89.6	0.8	-	0.0	94.8	0.8	-	3.4	3.4	5.9	-	3.4	3.4	5.9
fountain	22.1	57.1	91.2	91.3	24.2	61.6	94.7	94.8	25.7	64.9	97.4	97.4	0.8	1.4	1.4	1.1	0.8	1.4	1.4	0.6
cypress	1.6	17.3	67.1	45.7	1.7	21.8	73.8	48.9	1.7	28.0	80.5	51.4	0.3	1.3	2.6	1.1	0.3	1.3	2.6	1.1
discuri	0.4	1.7	59.4	58.7	0.4	2.5	61.9	61.4	0.5	4.5	64.4	63.9	41.9	24.7	115.6	156.3	41.9	24.7	115.6	156.3
wall	57.1	84.6	95.3	88.6	73.9	88.9	97.2	93.2	87.0	92.1	98.6	96.6	23.8	28.1	77.8	63.5	23.8	28.1	77.8	63.5
kyiv-puppet-theater	0.7	1.0	10.0	0.3	0.7	1.1	12.0	0.3	0.9	1.5	19.5	0.3	0.3	2.5	2.6	0.3	0.3	2.5	2.6	0.3
brandenburg_gate	21.4	42.3	68.7	70.1	35.0	52.8	75.2	77.2	53.6	65.2	81.5	83.9	1171.8	199.9	368.4	1472.1	1171.8	199.9	368.4	1472.1
british_museum	18.5	34.9	62.0	61.6	32.0	47.4	72.7	72.7	51.3	63.8	83.5	83.9	78.1	84.6	117.6	318.0	78.1	84.6	117.6	318.0
bucksingham_palace	4.1	26.0	85.9	80.5	12.5	37.2	89.1	86.2	34.0	53.0	92.1	91.1	429.7	173.3	484.1	4948.2	429.7	173.3	484.1	4948.2
colosseum_exterior	37.3	69.0	80.5	80.7	52.8	77.1	85.8	86.1	69.2	84.7	90.5	90.8	542.8	489.2	767.9	13561.7	542.8	489.2	767.9	13561.7
grand_place_brussels	18.3	34.7	71.8	68.7	32.4	50.8	78.5	76.6	50.5	67.6	84.6	84.2	124.2	87.6	220.7	520.3	124.2	87.6	220.7	520.3
lincoln_statue	1.0	30.7	68.4	67.2	4.4	36.3	72.5	71.5	23.8	41.7	76.6	75.3	99.9	46.1	103.9	362.3	99.9	46.1	103.9	362.3
notre_dame_facade	32.5	52.3	67.2	69.1	43.0	59.1	70.8	72.5	55.2	65.5	74.3	75.5	2592.1	2393.9	4146.9	152135.4	2592.1	2393.9	4146.9	152135.4
pantheon_exterior	49.6	68.3	77.0	79.4	62.3	74.9	81.8	83.5	74.4	81.2	86.2	87.2	285.7	169.3	488.8	1454.6	285.7	169.3	488.8	1454.6
piazza_san_marco	2.6	48.6	72.7	58.0	6.4	62.0	82.3	71.0	23.7	75.9	90.6	83.7	16.7	12.6	43.4	74.2	16.7	12.6	43.4	74.2
sacre_coeur	37.1	73.7	78.8	78.5	50.8	77.5	81.1	80.9	64.8	80.9	83.0	82.8	196.1	130.4	266.9	682.2	196.1	130.4	266.9	682.2
sagrada_familia	30.3	50.7	53.8	54.3	39.2	55.5	58.7	58.9	48.9	60.0	62.9	62.9	42.6	48.6	140.9	237.0	42.6	48.6	140.9	237.0
st_pauls_cathedral	1.9	60.5	72.4	72.7	6.5	70.4	80.2	78.9	25.8	79.7	85.8	85.0	61.9	45.0	101.0	241.4	61.9	45.0	101.0	241.4
st_peters_square	31.1	56.3	79.5	83.6	46.0	66.8	84.2	87.8	64.0	77.5	88.8	91.6	961.0	621.1	1177.9	6051.9	961.0	621.1	1177.9	6051.9
taj_mahal	38.6	58.6	72.1	68.9	51.0	65.7	77.3	75.5	64.9	73.3	82.4	81.7	380.1	379.0	630.7	14528.9	380.1	379.0	630.7	14528.9
trevi_fountain	43.4	64.6	79.0	80.5	56.2	72.7	82.8	84.4	69.3	80.3	86.4	87.8	1202.9	669.6	1676.0	12294.4	1202.9	669.6	1676.0	12294.4
Average	20.4	42.4	69.6	65.3	28.7	49.2	74.6	70.4	40.4	56.4	79.4	75.1	412.6	255.1	497.3	4050.1	412.6	255.1	497.3	4050.1

4.2 Uncalibrated Images Collections

IMC 2023 [16] contains unordered image collections over complex scenes. Images are collected from various sources and often lack prior camera intrinsics. The ground truth of the dataset is built by COLMAP [62] with held out imagery. As the accuracy of this dataset is not very high, we follow the same scheme as He *et al.* [30] to report the AUC scores at 3°, 5°, 10°. The results on training sets can be found in Table 5. On this dataset, the average AUC scores of the proposed method at 3°, 5° and 10° is several times higher than other global SfM baselines. The runtime is similar to other global SfM pipelines. Compared to COLMAP [62], the proposed method is about 4 points higher in AUC scores at 3°, 5°, and 10°, and is about 8 times faster.

MIP360 [9] contains 7 object-centric scenes with high-resolution images taken by the same camera. The provided COLMAP model is considered as (pseudo) ground truth for this dataset. Similarly, as the accuracy of ground truth is limited, the AUC scores at 3°, 5°, and 10° are reported. COLMAP reconstructions are re-estimated with the same matches as other methods. Results are summarized in Table 6 and our method is significantly closer to the reference model compared with other global SfM methods while rerunning COLMAP produces similar results as ours. Ours is more than 1.5 times faster than COLMAP.

4.3 Ablation

To demonstrate the effectiveness of the global position strategy, we conduct experiments by replacing the component by 1) adding only relative translation constraints, denoted as (BATA, cam), and 2) adding both points as well as translation constraints (BATA, cam+pt). For the (BATA, cam+pt) experiment, we use a similar weighting strategy for two types of constraints as implemented in Theia [71]. We also compare the result of replacing the global positioning by

Table 6: Results on MIP360 [9] datasets. The proposed method largely outperforms other baselines while obtaining similar results as COLMAP [62].

	AUC @ 3°				AUC @ 5°				AUC @ 10°				Time (s)			
	OpenMVG	Theia	GLOMAP	COLMAP	OpenMVG	Theia	GLOMAP	COLMAP	OpenMVG	Theia	GLOMAP	COLMAP	OpenMVG	Theia	GLOMAP	COLMAP
bicycle	89.2	12.0	95.8	95.8	92.7	14.7	97.5	97.5	95.3	17.9	98.7	98.7	123.1	28.6	66.9	120.6
bonsai	9.4	87.0	98.5	92.6	27.9	91.8	99.1	95.6	61.2	95.6	99.5	97.8	176.0	194.5	467.5	662.5
counter	96.9	98.9	99.3	99.2	98.1	99.4	99.6	99.5	99.1	99.7	99.8	99.8	46.6	71.0	203.9	270.5
garden	95.0	36.8	97.3	97.3	97.0	37.9	98.4	98.4	98.5	38.9	99.2	99.2	40.2	39.6	128.9	291.8
kitchen	88.5	93.5	94.8	94.9	93.1	95.8	96.9	97.0	96.5	97.6	98.4	98.5	127.0	187.3	426.9	619.1
room	39.6	26.0	97.7	96.2	42.2	26.8	98.6	97.7	44.9	27.6	99.3	98.9	96.3	85.6	216.3	371.6
stump	95.3	7.1	99.1	99.1	97.1	7.5	99.5	99.5	98.5	8.0	99.7	99.7	38.7	10.9	36.6	83.9
Average	73.4	51.6	97.5	96.5	78.3	53.4	98.5	97.9	84.9	55.0	99.2	98.9	92.6	88.2	221.0	345.7

Table 7: Ablation on global positioning constraints shows points alone perform best.

	ETH3D DSLR				IMC 2023			
	AUC@1°	AUC@3°	AUC@5°	Time (s)	AUC@3°	AUC@5°	AUC@10°	Time (s)
LUD	77.2	85.7	87.9	37.1	64.3	69.5	74.6	987.2
cam	73.5	80.5	82.4	21.2	62.1	67.0	71.6	886.9
BATA pt+cam	80.1	87.7	89.5	17.1	68.6	73.6	78.3	541.2
pt	80.8	88.5	90.3	16.5	69.6	74.6	79.4	497.3

Theia’s LUD [55]. For the experiments (BATA, cam) and LUD, we perform extra global positioning with fixed cameras to obtain point positions for subsequent bundle adjustment. We tested on both ETH3D MVS (DSLR) [66] and IMC 2023 [9]. Results are summarized in Table 7. We see that relative translation constraints deteriorate convergence and overall performance.

4.4 Limitations

Though generally achieving satisfying performance, there still remain some failure cases. The major cause is a failure of rotation averaging, *e.g.*, due to symmetric structures (see *Exhibition_Hall* in Table 3). In such a case, our method could be combined with existing approaches like Doppelganger [11]. Also, since we rely on traditional correspondence search, incorrectly estimated two-view geometries or the inability to match image pairs altogether (*e.g.*, due to drastic appearance or viewpoint changes) will lead to degraded results or, in the worst case, catastrophic failures.

5 Conclusion

In summary, we proposed GLOMAP as a new global SfM pipeline. Previous systems within this category have been considered more efficient but less robust than incremental approaches. We revisited the problem and concluded that the key lies in the use of points in the optimization. Instead of estimating camera positions via ill-posed translation averaging and separately obtaining 3D structure from point triangulation, we merge them into a single global positioning step. Extensive experiments on various datasets show that the proposed system achieves comparable or superior results to incremental methods in terms of accuracy and robustness while being orders of magnitude faster. The code is made available as open-source under a commercially friendly license.

Acknowledgment

The authors thank Philipp Lindenberger for the thoughtful discussions and comments on the text. This work was partially funded by the Hasler Stiftung Research Grant via the ETH Zurich Foundation and the ETH Zurich Career Seed Award. Linfei Pan was supported by gift funding from Microsoft.

References

1. Kapture toolbox. <https://github.com/naver/kapture>
2. Abdel-Aziz, Y.I., Karara, H.M., Hauck, M.: Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. *Photogrammetric engineering & remote sensing* **81**(2), 103–107 (2015)
3. Agarwal, S., Mierle, K., Team, T.C.S.: Ceres Solver (3 2022)
4. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307 (2016)
5. Arie-Nachimson, M., Kovalsky, S.Z., Kemelmacher-Shlizerman, I., Singer, A., Basri, R.: Global motion estimation from point matches. In: 2012 Second international conference on 3D imaging, modeling, processing, visualization & transmission. pp. 81–88. IEEE (2012)
6. Arrigoni, F., Fusiello, A.: Bearing-based network localizability: A unifying view. *IEEE transactions on pattern analysis and machine intelligence* **41**(9), 2049–2069 (2018)
7. Arrigoni, F., Fusiello, A., Rossi, B.: On computing the translations norm in the epipolar graph. In: 2015 International Conference on 3D Vision. pp. 300–308. IEEE (2015)
8. Barath, D., Noskova, J., Ivashechkin, M., Matas, J.: Magsac++, a fast, reliable and accurate robust estimator. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1304–1312 (2020)
9. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mipnerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5470–5479 (2022)
10. Cai, Q., Zhang, L., Wu, Y., Yu, W., Hu, D.: A pose-only solution to visual reconstruction and navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(1), 73–86 (2021)
11. Cai, R., Tung, J., Wang, Q., Averbuch-Elor, H., Hariharan, B., Snavely, N.: Dopelgangers: Learning to disambiguate images of similar structures. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 34–44 (2023)
12. Carlone, L., Aragues, R., Castellanos, J.A., Bona, B.: A linear approximation for graph-based simultaneous localization and mapping. In: *Robotics: Science and Systems*. vol. 7, pp. 41–48. MIT Press Cambridge, MA, USA (2012)
13. Carlone, L., Calafiore, G.C.: Convex relaxations for pose graph optimization with outliers. *IEEE Robotics and Automation Letters* **3**(2), 1160–1167 (2018)
14. Chatterjee, A., Govindu, V.M.: Efficient and robust large-scale rotation averaging. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 521–528 (2013)

15. Chatterjee, A., Govindu, V.M.: Robust relative rotation averaging. *IEEE transactions on pattern analysis and machine intelligence* **40**(4), 958–972 (2017)
16. Chow, A., Trulls, E., HCL-Jevster, Yi, K.M., lcmrll, old ufo, Dane, S., tanjigou, WastedCode, Sun, W.: Image matching challenge 2023 (2023), <https://kaggle.com/competitions/image-matching-challenge-2023>
17. Cui, H., Gao, X., Shen, S., Hu, Z.: Hsfn: Hybrid structure-from-motion. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1212–1221 (2017)
18. Cui, Z., Jiang, N., Tang, C., Tan, P.: Linear global translation estimation with feature tracks. arXiv preprint arXiv:1503.01832 (2015)
19. Cui, Z., Tan, P.: Global structure-from-motion by similarity averaging. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2015)
20. Dellaert, F., Rosen, D.M., Wu, J., Mahony, R., Carlone, L.: Shonan rotation averaging: global optimality by surfing $SO(p)^n$. In: European Conference on Computer Vision. pp. 292–308. Springer (2020)
21. 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)
22. Eriksson, A., Olsson, C., Kahl, F., Chin, T.J.: Rotation averaging and strong duality. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 127–135 (2018)
23. Fredriksson, J., Olsson, C.: Simultaneous multiple rotation averaging using lagrangian duality. In: Asian Conference on Computer Vision. pp. 245–258. Springer (2012)
24. Govindu, V.M.: Combining two-view constraints for motion estimation. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. vol. 2, pp. II–II. IEEE (2001)
25. Hartley, R., Aftab, K., Trumpf, J.: L1 rotation averaging using the weiszfeld algorithm. In: CVPR 2011. pp. 3041–3048. IEEE (2011)
26. Hartley, R., Trumpf, J., Dai, Y., Li, H.: Rotation averaging. *International journal of computer vision* **103**, 267–305 (2013)
27. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
28. Hartley, R.I.: Cheirality invariants. In: Proc. DARPA Image Understanding Workshop. vol. 3. Citeseer (1993)
29. Hartley, R.I., Sturm, P.: Triangulation. *Computer vision and image understanding* **68**(2), 146–157 (1997)
30. He, X., Sun, J., Wang, Y., Peng, S., Huang, Q., Bao, H., Zhou, X.: Detector-free structure from motion. arXiv preprint arXiv:2306.15669 (2023)
31. Henry, S., Christian, J.A.: Absolute triangulation algorithms for space exploration. *Journal of Guidance, Control, and Dynamics* **46**(1), 21–46 (2023)
32. Holynski, A., Geraghty, D., Frahm, J.M., Sweeney, C., Szeliski, R.: Reducing drift in structure from motion using extended features. In: 2020 International Conference on 3D Vision (3DV). pp. 51–60. IEEE (2020)
33. Huber, P.J.: Robust estimation of a location parameter. In: Breakthroughs in statistics: Methodology and distribution, pp. 492–518. Springer (1992)
34. Jiang, N., Cui, Z., Tan, P.: A global linear method for camera pose registration. In: Proceedings of the IEEE international conference on computer vision. pp. 481–488 (2013)

35. Kasten, Y., Geifman, A., Galun, M., Basri, R.: Algebraic characterization of essential matrices and their averaging in multiview settings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5895–5903 (2019)
36. Kasten, Y., Geifman, A., Galun, M., Basri, R.: Gpsfm: Global projective sfm using algebraic constraints on multi-view fundamental matrices. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3264–3272 (2019)
37. Kennedy, R., Daniilidis, K., Naroditsky, O., Taylor, C.J.: Identifying maximal rigid components in bearing-based localization. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 194–201. IEEE (2012)
38. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
39. Kipman, A.: Azure Spatial Anchors approach to privacy and ethical design. <https://www.linkedin.com/pulse/azure-spatial-anchors-approach-privacy-ethical-design-alex-kipman> (2019)
40. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. Quarterly of applied mathematics **2**(2), 164–168 (1944)
41. Li, X., Ling, H.: Pogo-net: pose graph optimization with graph neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5895–5905 (2021)
42. Lindenberger, P., Sarlin, P.E., Larsson, V., Pollefeys, M.: Pixel-perfect structure-from-motion with featuremetric refinement. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5987–5997 (2021)
43. Liu, Z., Qv, W., Cai, H., Guan, H., Zhang, S.: An efficient and robust hybrid sfm method for large-scale scenes. Remote Sensing **15**(3), 769 (2023)
44. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**, 91–110 (2004)
45. Lu, F., Hartley, R.: A fast optimal algorithm for l 2 triangulation. In: Computer Vision–ACCV 2007: 8th Asian Conference on Computer Vision, Tokyo, Japan, November 18–22, 2007, Proceedings, Part II 8. pp. 279–288. Springer (2007)
46. Manam, L., Govindu, V.M.: Correspondence reweighted translation averaging. In: European Conference on Computer Vision. pp. 56–72. Springer (2022)
47. Manam, L., Govindu, V.M.: Sensitivity in translation averaging. Advances in Neural Information Processing Systems **36** (2024)
48. Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. Journal of the society for Industrial and Applied Mathematics **11**(2), 431–441 (1963)
49. Martinec, D., Pajdla, T.: Robust rotation and translation estimation in multiview reconstruction. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2007)
50. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
51. Moisan, L., Moulon, P., Monasse, P.: Automatic homographic registration of a pair of images, with a contrario elimination of outliers. Image Processing On Line **2**, 56–73 (2012)
52. Moulon, P., Monasse, P.: Unordered feature tracking made fast and easy. In: CVMP 2012. p. 1 (2012)
53. Moulon, P., Monasse, P., Perrot, R., Marlet, R.: OpenMVG: Open multiple view geometry. In: International Workshop on Reproducible Research in Pattern Recognition (2016)

54. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)* **41**(4), 1–15 (2022)
55. Ozyesil, O., Singer, A.: Robust camera location estimation by convex programming. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2674–2683 (2015)
56. Ozyesil, O., Singer, A., Basri, R.: Stable camera motion estimation using convex programming. *SIAM Journal on Imaging Sciences* **8**(2), 1220–1262 (2015)
57. Purkait, P., Chin, T.J., Reid, I.: Neurora: Neural robust rotation averaging. In: *European Conference on Computer Vision*. pp. 137–154. Springer (2020)
58. Reinhardt, T.: Google Visual Positioning Service. <https://ai.googleblog.com/2019/02/using-global-localization-to-improve.html> (2019)
59. 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)
60. Sarlin, P.E., Dusmanu, M., Schönberger, J.L., Speciale, P., Gruber, L., Larsson, V., Miksik, O., Pollefeys, M.: Lamar: Benchmarking localization and mapping for augmented reality. In: *European Conference on Computer Vision*. pp. 686–704. Springer (2022)
61. Schönberger, J.L.: Robust Methods for Accurate and Efficient 3D Modeling from Unstructured Imagery. Ph.D. thesis, ETH Zürich (2018)
62. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2016)
63. Schönberger, J.L., Price, T., Sattler, T., Frahm, J.M., Pollefeys, M.: A vote-and-verify strategy for fast spatial verification in image retrieval. In: *Asian Conference on Computer Vision (ACCV)* (2016)
64. 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)
65. Schöps, T., Sattler, T., Pollefeys, M.: BAD SLAM: Bundle adjusted direct RGB-D SLAM. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2019)
66. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2017)
67. Servatius, B., Whiteley, W.: Constraining plane configurations in computer-aided design: Combinatorics of directions and lengths. *SIAM Journal on Discrete Mathematics* **12**(1), 136–153 (1999)
68. Sidhartha, C., Govindu, V.M.: It is all in the weights: Robust rotation averaging revisited. In: *2021 International Conference on 3D Vision (3DV)*. pp. 1134–1143. IEEE (2021)
69. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. In: *ACM siggraph 2006 papers*, pp. 835–846 (2006)
70. Strecha, C., Von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: *2008 IEEE conference on computer vision and pattern recognition*. pp. 1–8. Ieee (2008)
71. Sweeney, C.: Theia multiview geometry library: Tutorial & reference. <http://theia-sfm.org>

72. Sweeney, C., Sattler, T., Hollerer, T., Turk, M., Pollefeys, M.: Optimizing the viewing graph for structure-from-motion. In: Proceedings of the IEEE international conference on computer vision. pp. 801–809 (2015)
73. Tejas, G., Zara, G., Rota, P., Fusiello, A., Ricci, E., Arrigoni, F.: Rotation synchronization via deep matrix factorization. arXiv preprint arXiv:2305.05268 (2023)
74. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment — a modern synthesis. In: Triggs, B., Zisserman, A., Szeliski, R. (eds.) Vision Algorithms: Theory and Practice. pp. 298–372. Springer Berlin Heidelberg, Berlin, Heidelberg (2000)
75. Ullman, S.: The interpretation of structure from motion. Proceedings of the Royal Society of London. Series B. Biological Sciences **203**(1153), 405–426 (1979)
76. Wang, J., Karaev, N., Rupprecht, C., Novotny, D.: Visual geometry grounded deep structure from motion (2023)
77. Werner, T., Pajdla, T.: Chirality in epipolar geometry. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. vol. 1, pp. 548–553. IEEE (2001)
78. Wilson, K., Bindel, D., Snavely, N.: When is rotations averaging hard? In: ECCV (2016)
79. Wilson, K., Snavely, N.: Robust global translations with 1dsfm. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part III 13. pp. 61–75. Springer (2014)
80. Wu, C.: Towards linear-time incremental structure from motion. In: 2013 International Conference on 3D Vision-3DV 2013. pp. 127–134. IEEE (2013)
81. Yang, L., Li, H., Rahim, J.A., Cui, Z., Tan, P.: End-to-end rotation averaging with multi-source propagation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11774–11783 (2021)
82. Zhang, G., Larsson, V., Barath, D.: Revisiting rotation averaging: Uncertainties and robust losses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17215–17224 (2023)
83. Zhang, J.Y., Lin, A., Kumar, M., Yang, T.H., Ramanan, D., Tulsiani, S.: Cameras as rays: Pose estimation via ray diffusion. arXiv preprint arXiv:2402.14817 (2024)
84. Zhao, S., Zelazo, D.: Localizability and distributed protocols for bearing-based network localization in arbitrary dimensions. Automatica **69**, 334–341 (2016)
85. Zhuang, B., Cheong, L.F., Lee, G.H.: Baseline desensitizing in translation averaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4539–4547 (2018)

Supplemental Materials

S1 Additional Comparisons

In this section, we present extra experiments in comparison with HSfM [17] and LiGT [10], which are two more camera pose estimation pipelines.

HSfM [17] is a hybrid Structure-from-Motion (SfM) pipeline that estimates camera rotations with global rotation averaging and estimates camera translation with incremental SfM. In this experiment, we compare with the algorithm implemented in the Theia [71] library. By default, the Theia implementation does not run a full bundle adjustment (BA) at the end of the HSfM reconstruction. For a fair comparison, we added these extra steps in the pipeline. Otherwise, we use the default parameter setting.

LiGT [10] is a camera pose estimation algorithm based on linear constraints posed by points. We use the implementation in OpenMVG [53], provided by the authors¹, and experiment with the same pipeline as OpenMVG experiments.

The results on the datasets providing camera intrinsics are reported in Table S1. On the ETH3D MVS rig and DSLR [66] datasets, HSfM achieves comparable results to the proposed GLOMAP. However, on the ETH3D SLAM [65] and LaMAR [60] datasets, HSfM fails, leading to significantly less accurate reconstructions than all other pipelines. We attribute this to the sparsity and the sequential nature of these datasets.

Results on the datasets lacking camera calibrations are summarized in Table S2. On the IMC 2023 and MIP360 datasets, HSfM substantially falls behind all tested methods in terms of accuracy. This is expected as HSfM assumes known intrinsics which are then kept fixed until the last BA step. On these datasets, we are only given coarse intrinsics prior, which is not sufficient for reconstruction without additional refinement steps.

Table S1: Results on datasets with known camera intrinsics.

	ETH3D SLAM				ETH3D MVS (rig)				ETH3D MVS (DSLR)				LaMAR			
	R@0.1m	AUC@0.1m	AUC@0.5m	Time (s)	AUC@1°	AUC@3°	AUC@5°	Time (s)	AUC@1°	AUC@3°	AUC@5°	Time (s)	R@0.1m	AUC@0.1m	AUC@5m	Time (s)
OpenMVG	48.2	34.9	48.6	120.8	0.2	0.9	1.3	165.3	60.4	68.1	70.1	26.5	-	-	-	-
Theia	62.8	46.0	61.1	91.8	40.4	72.6	81.4	350.5	71.8	79.2	81.2	7.9	11.1	4.4	19.1	1542.0
HSfM	42.5	36.7	41.7	117.2	40.5	72.6	81.3	745.6	65.0	70.5	71.7	17.2	10.4	6.9	10.9	8670.0
LiGT	34.0	27.8	33.4	210.3	0.0	0.0	0.1	127.1	50.0	56.2	57.8	18.9	-	-	-	-
GLOMAP	66.4	57.0	65.7	133.5	57.6	80.9	87.2	793.8	80.8	88.5	90.3	16.5	49.1	24.5	50.9	12405.3
COLMAP	57.9	47.6	57.9	1115.4	45.5	64.0	69.1	2857.0	79.2	86.5	88.1	27.5	32.0	13.8	39.4	354660.4

For more direct comparison with LiGT [10], additional results on Strecha [70] dataset are summarized in Table S3.

¹ <https://github.com/openMVG/openMVG/pull/2065>

Table S2: Results on datasets with missing camera intrinsics.

	IMC 2023				MIP360			
	AUC@3°	AUC@5°	AUC@10°	Time (s)	AUC@3°	AUC@5°	AUC@10°	Time (s)
OpenMVG	20.4	28.7	40.4	412.6	73.4	78.3	84.9	92.6
Theia	42.5	49.2	56.5	149.6	51.6	53.4	55.0	88.2
HSM	29.7	36.5	44.2	325.2	37.1	40.7	44.5	248.9
LiGT	8.9	14.0	22.9	438.7	43.6	49.2	57.1	107.1
GLOMAP	69.6	74.6	79.4	497.3	97.5	98.5	99.2	221.0
COLMAP	65.3	70.4	75.1	4051.0	96.5	97.9	98.9	345.7

Table S3: Average camera position errors (in mm) for Strecha dataset [70]. Rows with * are taken from the paper [10].

	Herz-Jesus-P8	Herz-Jesus-P25	Fountain-P11	Entry-P10	Castle-P19	Castle-P30
LiGT*	5.01	6.86	3.17	5.50	41.88	49.84
LiGT	3.54	5.29	2.81	9.08	24.72	36.36
GLOMAP	4.13	5.40	2.79	6.32	24.95	22.36

S2 Additional Reconstruction Results

More reconstruction results can be found in Fig. S1. From the figure, one can see that the proposed GLOMAP reconstructs the scenes accurately, robustly obtaining the general structure and, also the fine details.

Table S4: Synthesizing result on MIP360 [9] datasets. The proposed method largely outperforms other baselines while obtaining similar results as COLMAP [62]. For scenes where testing images not all registered (marked *italic*), reference camera pose provided by the dataset is used. The differences on *bicycle*, *bonsai*, *garden*, *room* and *stump* are evident. See Fig. S2 for details.

	PSNR				SSIM			
	OpenMVG	Theia	GLOMAP	COLMAP	OpenMVG	Theia	GLOMAP	COLMAP
bicycle	23.01	<i>17.75</i>	23.13	<i>23.15</i>	0.526	0.352	0.531	<i>0.532</i>
bonsai	23.88	28.54	<i>30.36</i>	29.66	0.767	0.872	<i>0.904</i>	0.896
counter	26.76	<i>26.78</i>	26.72	<i>26.81</i>	0.835	<i>0.836</i>	0.835	<i>0.837</i>
garden	<i>24.97</i>	<i>20.19</i>	<i>24.97</i>	<i>24.98</i>	<i>0.653</i>	<i>0.456</i>	<i>0.655</i>	<i>0.653</i>
kitchen	<i>29.32</i>	29.02	<i>29.35</i>	<i>29.23</i>	<i>0.853</i>	0.841	0.855	<i>0.851</i>
room	19.11	<i>17.07</i>	<i>29.41</i>	<i>29.14</i>	0.691	<i>0.643</i>	<i>0.876</i>	<i>0.871</i>
stump	23.56	<i>19.43</i>	23.81	<i>23.98</i>	0.584	<i>0.408</i>	0.595	<i>0.602</i>
Average	24.37	22.68	<i>26.82</i>	<i>26.71</i>	0.701	0.630	<i>0.750</i>	0.749

S3 Novel View Synthesis

To examine the impact of reconstruction quality on a downstream task, this section presents results on novel view synthesis. We conduct experiments with Instant-NGP [54], a popular method for synthesizing images. The tested dataset is MIP360 [9], which was originally proposed for this particular application.

Quantitative results can be found in Table S4 and qualitative results in Figure S2. We adopt the standard metric, PSNR (peak signal-to-noise ratio) and

Table S5: Ablation on robustness of global positioning to different noise levels of points.

Noise level	0px	1px	2px	4px	8px	16px	32px	64px
electro	100.0	99.3	98.1	97.2	94.2	91.3	82.1	63.0
facade	98.1	97.7	97.9	98.1	97.3	95.4	94.2	90.8
kicker	100.0	99.5	98.1	98.5	90.5	89.5	86.0	78.0
meadow	100.0	99.3	98.0	95.8	92.6	89.2	75.3	63.8
office	100.0	98.6	96.6	94.2	87.3	78.5	45.5	42.5
pipes	100.0	98.7	97.7	97.5	94.3	88.4	82.4	56.4
playground	99.7	99.5	96.4	96.6	96.0	90.4	85.8	72.2
relief	100.0	99.8	99.5	99.3	98.9	95.2	93.9	87.4
relief_2	100.0	99.8	99.5	99.1	98.4	96.7	92.5	88.3
terrace	100.0	99.7	99.5	99.1	98.5	94.0	91.7	80.5
terrains	100.0	99.8	99.4	99.0	97.9	95.8	89.1	84.0

SSIM (Structural similarity index measure). From the table, it can be observed that synthesis results with GLOMAP and COLMAP [62] reconstruction achieve similar scores both in PSNR and SSIM. For OpenMVG [53] and Theia [71], though they achieve similar scores in some scenes as our reconstruction, they *fail* on several scenes. Qualitatively, the synthesized results for Theia and OpenMVG are more blurred for several scenes, indicating the poor quality of the camera pose.

S4 Effect of Camera Clustering

For unordered internet image collection, obtaining clean and coherent is not trivial and we propose camera clustering technique for this purpose. Qualitative results of the mechanism can be found in Figure S3. The comparison shows the effectiveness of the proposed mechanism in pruning the floating structures.

S5 Robustness of Global Positioning.

This experiment validates the robustness in the presence of image noise. To construct the experiment, we synthesize perfect image observations by projecting COLMAP triangulations to the ground truth cameras, and then, for each observation, we add random Gaussian noise to the reprojections. We do not run global bundle adjustment in this experiment to isolate the performance of global positioning. Results can be found in Table S5. For perfect image observations with 0px noise, we reliably converge to the ground truth, underlining the effectiveness of random initialization. As the noise level increases to extreme values, the AUC scores degrade gradually, indicating a high level of robustness of our proposed global positioning.

S6 Detailed Results of ETH3D SLAM

Per-sequence result for ETH3D SLAM can be found in Table S6. From the table, one can see that the relative performance is consistent with the averaged results across sequences sharing the same prefix.

Table S6: Per sequence result for ETH3D SLAM

	Recall @ 0.1				AUC @ 0.1				AUC @ 0.5				Time (s)											
	OpenMVG		Theia		GLOMAP		COLMAP		OpenMVG		Theia		GLOMAP		COLMAP		OpenMVG		Theia		GLOMAP		COLMAP	
cables_1	66.5	100.0	100.0	100.0	32.5	51.9	80.5	93.8	72.3	90.4	96.1	98.8	289.0	1001.2	549.8	17501.8								
cables_2	100.0	100.0	100.0	40.2	90.9	74.6	88.0	16.7	98.2	94.9	97.6	49.4	5.2	3.8	11.6	1	96.8							
cables_3	63.9	63.9	63.9	63.9	53.9	54.1	60.1	58.9	61.9	62.0	63.2	62.9	15.0	13.5	25.5	1	61.5							
camera_shake_1	6.6	36.2	47.7	46.5	6.1	18.2	17.4	22.7	6.5	36.4	40.1	43.0	1.5	3.9	6.7	1	74.5							
camera_shake_2	-	20.9	24.3	51.1	-	8.6	22.1	45.0	-	19.0	24.0	49.9	-	-	9.8	19.3	511.3							
camera_shake_3	-	22.8	28.8	6.7	-	12.0	26.2	6.4	-	23.2	28.2	6.6	-	-	2.1	3.9	2.9							
ceiling_1	-	9.3	9.3	9.3	-	8.4	8.8	8.7	-	9.1	9.2	9.2	-	-	17.2	40.5	116.9							
ceiling_2	12.7	36.4	48.0	47.9	5.9	25.6	35.8	21.9	16.3	34.4	45.7	43.3	78.3	87.9	181.4	1998.6								
desk_3	37.9	46.3	46.3	28.2	16.7	42.0	39.8	27.6	39.8	45.4	45.0	128.1	126.5	255.1	124.3	767.5								
desk_changing_1	18.1	18.2	18.4	20.6	15.2	16.4	17.2	14.7	17.5	17.8	18.1	19.4	625.9	135.4	175.6	1462.6								
einstein_1	90.4	85.8	77.6	45.4	46.9	35.1	44.4	18.0	88.5	85.8	88.0	56.8	122.6	120.2	139.0	1920.6								
einstein_2	39.8	77.5	78.4	23.7	18.8	51.0	49.0	11.0	52.4	75.0	74.7	34.3	263.4	180.9	472.3	4765.7								
einstein_dark	-	5.6	7.0	5.0	-	3.2	2.8	2.0	-	6.3	7.4	6.1	-	-	20.4	31.2	267.8							
einstein_flashlight	-	17.4	19.1	17.3	-	11.2	16.3	9.4	-	16.2	18.5	17.2	-	-	36.4	84.8	739.1							
einstein_change_1	-	17.8	17.8	11.6	-	16.6	17.1	11.2	-	17.6	17.7	11.5	-	-	8.3	19.4	21.0							
einstein_change_2	[100.0]	100.0	100.0	100.0	93.2	89.3	96.9	96.0	98.6	97.9	99.4	99.2	64.9	27.6	86.3	388.1								
einstein_change_3	-	30.3	30.0	32.0	-	18.6	29.2	30.5	-	28.0	29.9	31.7	-	-	100.1	161.5	513.7							
kidnap_1	73.1	73.3	73.3	73.3	63.4	62.3	70.3	68.5	71.2	71.1	72.7	72.3	114.4	356.7	144.3	731.2								
large_loop_1	35.4	48.6	49.0	44.5	18.1	37.8	45.8	20.7	33.7	46.6	48.4	43.4	91.9	60.2	77.6	98.3								
mannequin_1	26.8	43.1	53.2	18.8	17.5	16.3	45.8	17.7	31.2	38.0	51.7	18.6	18.6	30.0	17.7	45.4								
mannequin_3	23.5	34.2	35.0	41.4	21.2	13.3	32.5	38.7	23.0	37.0	34.5	40.8	10.6	13.9	16.8	72.3								
mannequin_4	75.3	96.3	96.7	96.9	49.7	68.6	89.8	85.6	70.2	90.9	95.3	94.6	40.2	30.0	50.2	52.0								
mannequin_5	43.4	68.9	77.3	80.2	15.9	34.1	63.9	59.8	52.5	66.4	74.6	78.0	88.0	82.4	94.2	1261.3								
mannequin_7	13.9	15.2	19.6	18.3	9.0	14.6	18.0	17.8	13.8	15.1	19.3	18.2	6.8	6.2	12.5	19.7								
mannequin_face_1	100.0	100.0	100.0	100.0	92.6	97.2	98.6	98.1	98.5	99.4	99.7	99.6	99.8	99.8	99.8	53.2	17.6	76.3	244.4					
mannequin_face_2	100.0	100.0	100.0	100.0	98.1	98.1	99.0	99.0	99.6	99.6	99.8	99.8	16.8	21.7	52.4	285.6								
mannequin_head	41.9	56.8	55.7	14.0	11.2	43.5	52.6	13.1	36.1	54.1	56.8	13.8	26.8	47.8	29.9	10.2								
motion_1	18.8	16.9	39.8	17.7	11.0	11.9	22.5	12.9	23.3	19.2	45.9	19.7	859.7	109.0	788.9	9995.1								
planar_2	24.0	100.0	100.0	100.0	9.2	99.0	99.1	99.1	31.8	99.8	99.8	99.8	330.5	149.9	540.3	1220.0								
planar_3	37.2	100.0	100.0	100.0	15.7	96.7	98.3	97.5	44.2	99.3	99.7	99.5	297.1	185.1	526.3	3478.5								
plant_1	100.0	100.0	100.0	100.0	90.1	97.8	98.0	98.5	98.0	99.6	99.7	99.7	3.9	3.9	1.9	3.9	10.7							
plant_2	100.0	100.0	100.0	100.0	98.1	98.5	98.6	99.6	99.7	99.8	99.7	99.7	7.4	7.6	20.4	45.3								
plant_3	100.0	100.0	100.0	100.0	54.4	96.2	93.2	93.8	90.9	99.2	98.6	98.8	15.0	7.7	14.6	45.9								
plant_4	100.0	100.0	100.0	100.0	97.9	98.8	98.7	98.9	99.6	99.8	99.7	99.8	3.8	3.6	16.0	19.1								
plant_5	100.0	100.0	100.0	100.0	95.7	96.3	98.3	97.0	99.1	99.3	99.7	99.4	7.3	6.9	18.5	36.3								
plant_scene_1	43.9	77.8	77.8	98.5	36.1	52.2	71.8	85.0	42.3	72.7	76.6	95.8	35.4	25.0	39.6	441.4								
plant_scene_2	41.6	84.4	99.0	76.1	15.8	27.8	58.3	35.5	59.2	83.4	91.3	82.3	80.4	54.2	64.8	64.6								
plant_scene_3	30.4	50.3	69.6	67.8	15.1	34.6	38.6	50.8	27.7	52.3	81.9	64.4	20.2	179.1	51.5	378.5								
reflective_1	12.6	16.1	22.0	26.2	6.7	9.0	12.1	9.2	16.5	23.0	31.3	33.5	721.3	118.3	434.4	657.93								
repetitive	26.3	28.5	32.7	28.5	23.9	15.2	29.2	27.2	25.8	27.0	32.0	28.3	63.2	136.8	74.5	561.1								
sfm_bench	72.0	96.6	98.3	100.0	61.0	88.6	92.7	94.1	69.8	97.2	98.8	73.9	41.3	103.6	46.1									
sfm_garden	76.4	80.0	87.8	84.7	32.5	31.1	57.2	29.4	76.6	79.7	90.3	82.3	272.8	698.3	855.2	798.1								
sfm_house_loop	70.5	100.0	100.0	42.4	53.0	76.6	86.3	31.2	67.0	95.3	97.3	1	40.8	95.6	96.5	222.8	1030.1							
sfm_lab_room_1	99.6	99.6	99.6	20.9	75.7	44.0	77.9	11.7	94.8	88.5	95.3	23.5	12.8	14.3	32.1	37.9								
sfm_lab_room_2	97.6	94.4	99.2	31.2	67.2	28.2	84.0	12.3	91.5	84.9	96.2	33.5	3.3	2.3	5.0	20.7								
sofa_1	10.6	21.3	25.9	27.0	5.7	11.7	22.9	24.1	9.6	22.1	25.3	16.4	13.3	21.1	14.1	252.8								
sofa_2	18.6	21.8	21.8	43.1	8.3	11.3	20.4	38.7	16.6	19.7	21.5	42.4	9.2	6.1	9.8	137.2								
sofa_3	15.5	19.5	22.5	28.9	8.1	15.4	21.1	25.0	14.9	20.4	22.2	28.1	4.9	2.9	5.1	50.4								
sofa_4	-	25.3	25.3	29.9	-	14.0	23.8	26.8	-	23.2	25.1	29.3	-	-	5.5	11.5	188.8							
table_3	100.0	100.0	100.0	100.0	85.9	90.3	97.7	96.8	97.2	98.1	99.5	99.4	299.0	83.0	281.2	2116.8								
table_4	100.0	100.0	100.0	100.0	89.1	88.2	95.9	96.0	97.8	97.6	99.2	99.2	142.0	68.8	182.8	2817.7								
table_7	37.2	81.0	83.0	99.7	29.6	29.0	59.2	94.1	35.7	72.0	78.2	98.6	104.9	141.7	200.5	13987.7								
vicon_light_1	47.2	68.5	94.1	97.8	14.4	25.6	65.8	55.4	60.8	77.5	88.4	89.9	67.9	128.7	53.3	345.1								
vicon_light_2	82.0	100.0	100.0	64.3	26.4	88.6	95.3	22.3	82.7	97.7	99.1	60.1	33.8	47.8	39.0	604.6								
Average	48.2	62.8	66.4	57.9	34.9	46.0	57.0	47.6	48.6	61.1	65.7	57.9	120.8	91.8	133.5	1115.4								



Fig. S1: Example reconstructions from the proposed GLOMAP on various datasets.

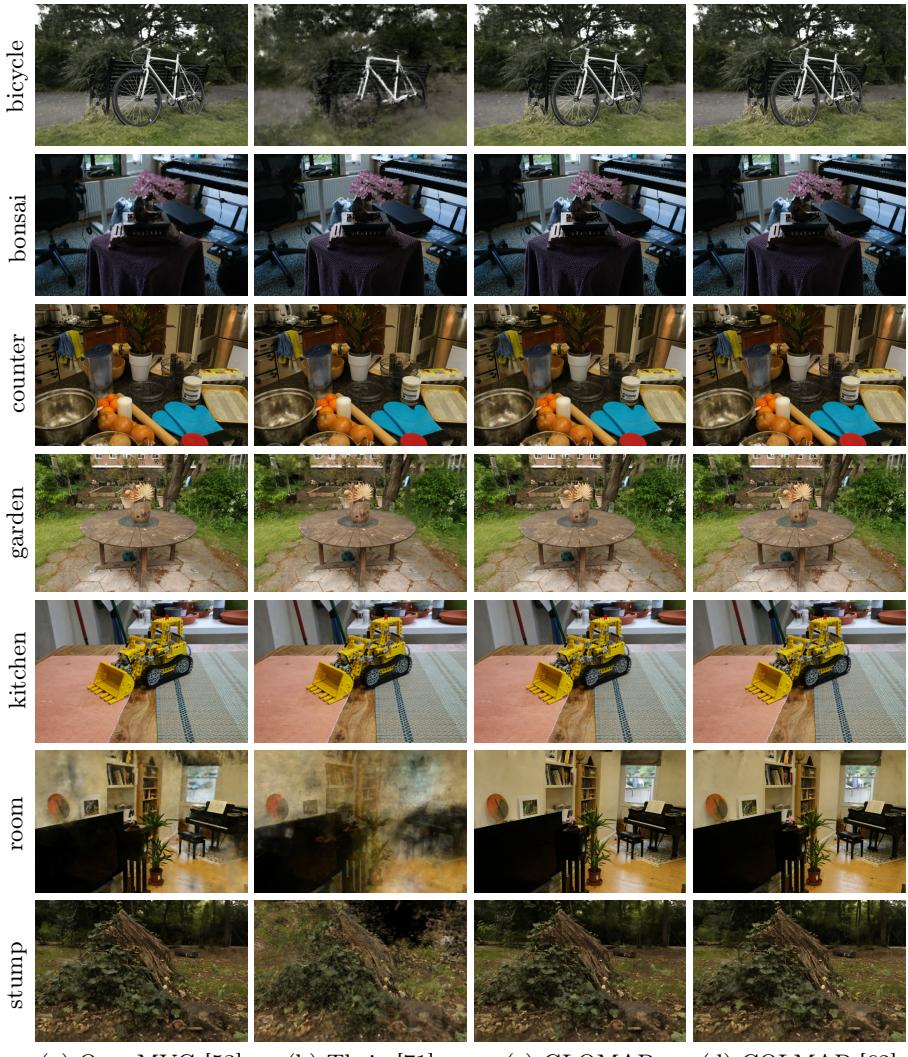
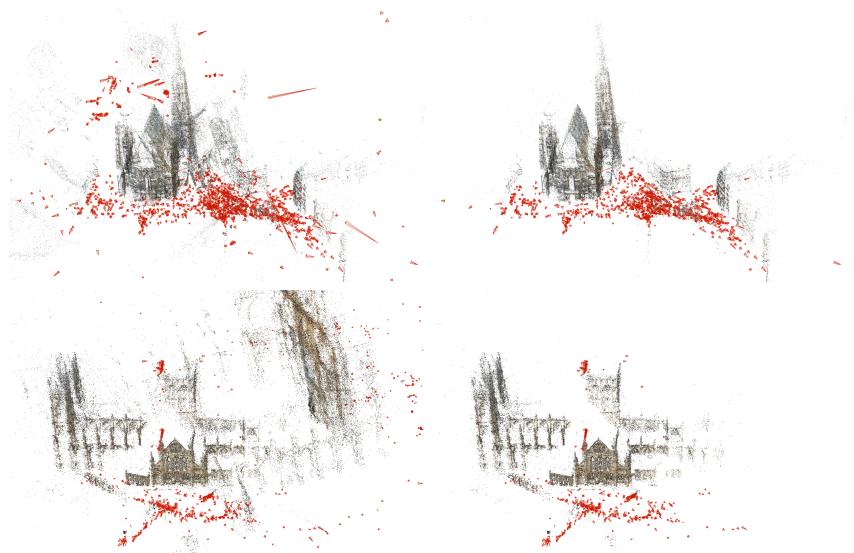


Fig. S2: Qualitative results for novel view synthesis with Instant-NGP [54]. The differences on *bicycle*, *bonsai*, *garden*, *room* and *stump* are visually evident.

Vienna Cathedral



Yorkminster



Fig. S3: Qualitative results of camera clustering on 1DSfM [79] datasets.