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

MGSfM: Multi-Camera Geometry Driven Global Structure-from-Motion

Peilin Tao[1,2*] Hainan Cui[1,2*†] Diantao Tu[1,2] Shuhan Shen[1,2†]
[1] Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
taopeilin2023@ia.ac.cn, hncui@nlpr.ia.ac.cn, tudiantao2020@ia.ac.cn, shshen@nlpr.ia.ac.cn

Abstract

Multi-camera systems are increasingly vital in the environmental perception of autonomous vehicles and robotics. Their physical configuration offers inherent fixed relative pose constraints that benefit Structure-from-Motion (SfM). However, traditional global SfM systems struggle with robustness due to their optimization framework. We propose a novel global motion averaging framework for multi-camera systems, featuring two core components: a decoupled rotation averaging module and a hybrid translation averaging module. Our rotation averaging employs a hierarchical strategy by first estimating relative rotations within rigid camera units and then computing global rigid unit rotations. To enhance the robustness of translation averaging, we incorporate both camera-to-camera and camera-to-point constraints to initialize camera positions and 3D points with a convex distance-based objective function and refine them with an unbiased non-bilinear angle-based objective function. Experiments on large-scale datasets show that our system matches or exceeds incremental SfM accuracy while significantly improving efficiency. Our framework outperforms existing global SfM methods, establishing itself as a robust solution for real-world multi-camera SfM applications. The code is available at https://github.com/3dv-casia/MGSfM/.

1. Introduction

Multi-camera systems have gained significant popularity in autonomous vehicles [28, 32, 49] and mobile robotics [9, 28, 43]. By rigidly mounting multiple cameras on a vehicle, these systems provide more comprehensive visual observations of the environment, enabling applications such as autonomous driving [32, 49], visual localization and mapping [21, 45, 52], and 3D object detection [23, 38]. To reconstruct 3D structure and recover camera motions, structure-from-motion (SfM) is a fundamental technique. It begins by calculating feature correspondences between overlapping images to construct a view graph [2, 3]. Then, cameras are registered, and 3D points are triangulated. Finally, camera parameters and 3D points are jointly refined through bundle adjustment (BA) [42, 47]. Depending on the camera registration manner, existing SfM pipelines can be classified into two categories: incremental and global methods.

Incremental SfM [44, 53] starts with an initial pair of images to form the initial scene and iteratively register additional overlapping images via the RANSAC [10] based Perspective-n-Point (PnP) method [17, 19]. While incremental SfM is generally robust against outlier feature matches, its primary drawbacks include strong sensitivity to image registration order, time-consuming repeated bundle adjustments, and the gradual accumulation of errors as the scene expands, making it unsuitable for large-scale reconstruction tasks. In contrast, global SfM methods [5, 41] first perform rotation averaging [7, 26] to estimate global camera rotations, followed by translation averaging [40, 55] to estimate global camera translations. By registering all cameras simultaneously, global approaches achieve uniform error distribution and higher efficiency. For rotation averaging, existing methods based on Lie algebra structures [7, 22] have been extensively studied. However, translation averaging methods relying solely on relative translations encounter degeneration when the camera motion trajectory is approximately collinear [29, 36]. To address these issues, recent works [41, 46] jointly estimate cameras and 3D points by incorporating feature tracks into translation averaging.

Nevertheless, existing SfM solutions may still fail in large-scale reconstruction tasks, especially when confronted with numerous mismatched feature tracks or a lack of loop closures. Given that multi-camera systems are increasingly employed in image acquisition platforms, a promising approach is to exploit the inherent multi-camera constraints, whereby each rigidly mounted camera maintains a fixed internal pose relative to the other cameras within the same rigid unit. Several works [12, 13] leverage this property by designating one camera as the reference camera, with the non-reference camera poses derived from the reference pose and the fixed

*These authors contributed equally to this work.
†Corresponding author.

internal relative poses. Thus, they reduce the SfM problem to estimating only the reference poses and the internal relative poses, leading to improved efficiency and robustness. The incremental method [13] first estimates the internal relative poses through incremental reconstruction and then incrementally registers the rigid unit to reconstruct the entire scene with fixed internal relative poses. While this approach exhibits strong robustness and mitigates scale drift with fixed internal poses, incremental registration still risks the accumulation of drift over large-scale scenes and exhibits low efficiency. The global method [12] demonstrates extremely high efficiency by jointly estimating the reference camera poses and internal relative poses. However, jointly estimating reference camera rotations and internal relative rotations is highly non-convex and sensitive to outlier relative rotations. The internal relative rotations can be estimated in advance through relative rotation averaging only when there is sufficient overlap in the field of view between two cameras in a rigid unit [12], which limits its broad applicability. Moreover, only camera-to-camera constraints are employed to minimize a biased distance-based objective function during translation averaging, making it less accurate to degenerate cases, such as rigid units moving along a straight line without overlapping fields of view between internal cameras.

To tackle these challenges, we introduce a novel multi-camera-geometry driven global SfM framework, abbreviated as MGSfM. By fusing multi-camera constraints, we transform the problem of estimating camera poses into the problem of estimating the rigid unit poses and internal camera poses. Since the rotation averaging problem is well-studied using Lie algebra structures, we decouple the multi-camera rotation averaging problem by first estimating the internal camera rotations, followed by computing the rigid unit rotations with fixed internal camera rotations. We begin by disregarding multi-camera constraints and applying conventional rotation averaging to estimate the internal camera rotations, thereby relaxing the overlap requirements of traditional systems. Subsequently, given the estimated internal camera rotations, the rigid unit rotation estimation employs the Baker–Campbell–Hausdorff formula [48] for linear optimization, enhancing robustness. For the translation averaging problem, inspired by recent advancements integrating feature tracks with angle-based refinement [41, 46], we propose a hybrid multi-camera translation averaging method. Our approach combines distance-based and angle-based objective functions while utilizing relative translations and feature tracks as hybrid inputs. This design mitigates degenerate cases and improves overall accuracy. Camera positions are first initialized using a convex distance-based objective function under the L_1 norm, followed by angle-based refinement under the L_2 norm. Next, 3D points are triangulated using a convex cross-product-form objective function under the L_1 norm. Finally, with a robust initialization, a non-bilinear angle-based refinement is proposed to refine camera and 3D point positions, exhibiting improved robustness and efficiency compared to the bilinear refinement [41, 55].

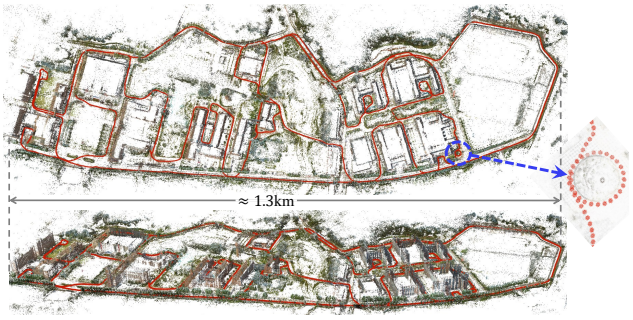

Figure 1. Our method successfully reconstructs a self-collected dataset named CAMPUS. Captured using a six-camera system, this dataset comprises over 29,000 images. Notably, many state-of-the-art methods—such as COLMAP [44], MCSfM [13], GLOMAP [41], and MMA [12]—fail to produce satisfactory results, as demonstrated in supplemental material.

In summary, this work makes two main **contributions**. First, we propose a decoupled multi-camera rotation averaging framework that transforms the highly non-convex multi-camera rotation averaging problem into a two-step standard single-camera problems. Second, we introduce a hybrid multi-camera translation averaging algorithm that integrates both camera-to-camera and camera-to-point constraints and incorporates a hybrid optimization framework comprising a convex distance-based initialization and an unbiased angle-based refinement. Various experiments show the superiority of our method compared to the state-of-the-art methods in terms of robustness, accuracy, and efficiency. Figure 1 illustrates our reconstruction result on a challenging large-scale dataset captured by a six-camera Insta360 Pro2 system. While many existing methods fail to deliver satisfactory reconstructions, our approach achieves accurate results in just one hour, demonstrating its practical effectiveness in large-scale, complex real-world scenarios.

2. Related Work

Rotation Averaging. Rotation averaging is a key step in global SfM, aimed at estimating global camera rotations from pairwise relative rotations. Early methods linearized the problem on the Lie group of rotations [22]. Crandall *et al*. [11] initialize the camera rotations with the discrete Markov Random Field formulations and refine them with Levenberg-Marquardt optimization. Wilson *et al*. [51] investigate the view-graph's density and consistency to understand when rotation averaging becomes hard. Chatterjee and Govindu [7] propose to initialize camera rotations via a maximum spanning tree and then refine them using an iterative re-weighted least squares (IRLS) formulation. Dellaert *et al*. [16] search a sequence of higher-dimensional lifts of the rotation averaging problem to find a globally optimal solu-

tion. Zhang *et al*. [54] propose to better model the underlying noise distributions by directly propagating the uncertainty from the point correspondences into the rotation averaging. Although some incremental rotation averaging methods [18] are proposed, global rotation averaging methods ensure uniform error distribution for large-scale scenes.

Translation Averaging. Traditional translation averaging methods aim to derive global camera positions from pairwise relative translations. Ozyesil *et al*. [40] propose a convex least unsquared deviations formulation to enhance the robustness. Zhuang *et al*. [55] introduce an unbiased, angle-based, bilinear formulation to mitigate the impact of varying camera baseline scales. Manam *et al*. [35] filter outlier feature matches and refine relative translations through an iterative averaging scheme. These methods, which rely solely on relative translations, encounter degeneration when the camera motion is collinear [29]. To address this problem, feature tracks are incorporated during translation averaging. The scales of camera baselines are estimated based on the depth consistency of feature rays in [14]. Additional constraints are formulated among cameras observing the same 3D point to eliminate the ambiguity in baseline scales in [5, 15, 33]. To mitigate the impact of inaccuracies from implicitly represented 3D points, camera positions and 3D points are jointly estimated in [41, 46, 50]. With random initialization, a chordal distance metric angle-based objective is employed in [50], and a bilinear angle-based objective is employed in [41]. In contrast, camera positions and 3D points are jointly initialized using a convex objective in [46].

Muti-Camera Reconstruction. Multi-camera-based simultaneous localization and mapping (SLAM) systems, such as ORB-SLAM2 [37] and ORB-SLAM3 [6], leverage fixed relative poses to achieve efficient mapping. However, these systems require pre-calibrated internal camera poses and are unable to process multiple video sequences collectively, such as those obtained through crowdsourcing. In comparison, some multi-camera-based SfM systems leverage the fixed internal relative pose constraints to estimate the relative poses automatically. This allows one camera to be designated the "reference camera", reducing the parameter space to its global poses plus the constant internal relative poses. In incremental settings, MCSfM [13] formulates an incremental pipeline in which rigid units are successively registered, minimizing scale drift and boosting robustness thanks to the fused observations from multiple cameras. However, incremental solutions remain susceptible to gradual error accumulation and heavy repeated bundle adjustments. On the other hand, MMA [12] introduces a global multi-camera motion averaging framework by extending rotation and translation averaging to consider multi-camera constraints. It simultaneously solves for the reference camera poses and internal relative poses, thereby improving efficiency over incremental systems. Despite these benefits, it may fail in large-scale scenes where a large fraction of relative poses are outliers, and it cannot solve degenerate collinear configurations. Internal camera poses in a rig are typically estimated based on a pre-built map for visual localization or calibration [24, 27]. In contrast, our method jointly estimates both camera poses and the 3D structure in a unified optimization framework.

3. Multi-Camera Driven Global SfM

In this section, we present the details of our multi-camera driven global SfM system. We first construct the view graph by computing the feature correspondences and relative poses between image pairs. Next, we define the multi-camera model by transforming the problem of estimating camera poses into the problem of estimating the rigid unit poses and the internal camera poses. Based on the multi-camera model, we first introduce a decoupled multi-camera rotation averaging method by estimating the internal camera rotations first, then computing the rigid unit rotations with fixed internal rotations. Then, we propose a hybrid, multi-camera translation averaging method that uses a hybrid distance-based and angle-based objective function with hybrid relative translations and feature tracks as input. Finally, we introduce the multi-camera bundle adjustment. For clarity, we use calligraphic letters (e.g.,$\boldsymbol{\mathcal{C}}$,$\boldsymbol{\mathcal{R}}$,$\boldsymbol{\mathcal{P}}$) to denote the corresponding sets of variables (e.g., camera centers $\boldsymbol{c}_i$, rotations $\boldsymbol{R}_i$, 3D points $\boldsymbol{p}_k$) being optimized in the objective functions.

3.1. View Graph Construction

Given images collected by the multi-camera system, scale-invariant image features [34] are first detected for all images. Since the multi-camera system is typically mounted on a continuously moving platform and the images are captured sequentially, we begin by sequentially matching each image with its adjacent images captured by the same camera. Next, loop detection is performed, where each image is matched with the top-K similar images identified by the image retrieval method [44].

Given initial feature matches and camera intrinsics for each image pair, we validate the feature matches using the RANSAC-based [10] 5-point algorithm [39]. If a sufficient number of inlier feature matches are detected between image pairs, the images are considered matched. The relative rotations $\boldsymbol{R}_{ij}$ and the normalized relative translations $\boldsymbol{t}_{ij}$ are then decomposed from the corresponding essential matrices. After obtaining matched image pairs, a view graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ is constructed, where each node in $\mathcal{V}$ represents an image, and each edge $ij \in \mathcal{E}$ represents the relative pose $(\boldsymbol{R}_{ij}, \boldsymbol{t}_{ij})$ and corresponding feature matches between two images $i, j \in \mathcal{V}$. The global camera rotations and positions $(\boldsymbol{R}_i, \boldsymbol{c}_i)$, and the relative camera rotations and translations $(\boldsymbol{R}_{ij}, \boldsymbol{t}_{ij})$ satisfy the following equations:

$$\boldsymbol{R}_{ij} = \boldsymbol{R}_j \boldsymbol{R}_i{}^{\top}, \quad \boldsymbol{R}_j{}^{\top} \boldsymbol{t}_{ij} = \frac{\boldsymbol{c}_i - \boldsymbol{c}_j}{||\boldsymbol{c}_i - \boldsymbol{c}_j||_2}. \tag{1}$$

3.2. Multi-Camera Model Definition

Since the cameras in a multi-camera system are triggered approximately synchronously, the images captured simultaneously are regarded as forming a rigid unit in which the internal camera poses remain fixed. Let $(\boldsymbol{R}_i^r, \boldsymbol{t}_i^r)$ denote the internal camera rotation and translation of image i in the corresponding rigid unit coordinate system, and let $(\boldsymbol{R}_i^g, \boldsymbol{c}_i^g)$ denote the rotation and position of the rigid unit corresponding to image i in the global coordinate system. To remove rotational and positional ambiguity, we set the pose of an arbitrary internal camera and the pose of an arbitrary rigid unit to a fixed pose like $(\boldsymbol{I}, \boldsymbol{0})$. The global camera rotation and position of image i can then be expressed in terms of the pose of its corresponding rigid unit and internal camera as:

$$\boldsymbol{R}_i = \boldsymbol{R}_i^r \boldsymbol{R}_i^g, \quad \boldsymbol{c}_i = \boldsymbol{c}_i^g - \boldsymbol{R}_i{}^\top \boldsymbol{t}_i^r. \tag{2}$$

Thus, by incorporating multi-camera constraints, robustness and efficiency are enhanced through increased camera connectivity and a reduced number of unknown variables.

3.3. Decoupled Multi-Camera Rotation Averaging

Based on Eqs. (1) and (2), relative camera rotation $\boldsymbol{R}_{ij}$ can be represented in terms of the internal camera rotations and global rigid unit rotations as:

$$\boldsymbol{R}_{ij} = \boldsymbol{R}_j^r \boldsymbol{R}_j^g \boldsymbol{R}_i^{g\top} \boldsymbol{R}_i^{r\top}. \tag{3}$$

For any two images i and j within the same rigid unit, we have $\boldsymbol{R}_j^g = \boldsymbol{R}_i^g$. Hence, for the relative rotations of the image pair ij in the same rigid unit, Eq. (3) can be simplified as:

$$\boldsymbol{R}_{ij}^r = \boldsymbol{R}_j^r \boldsymbol{R}_j^g \boldsymbol{R}_i^{g\top} \boldsymbol{R}_i^{r\top} = \boldsymbol{R}_j^r \boldsymbol{R}_i^{r\top} \tag{4}$$

When there is an insufficient overlapping field of view between images in the rigid units, the traditional multi-camera rotation averaging method, such as MRA [12], jointly estimates the global rigid unit rotations and internal camera rotations. However, the multiplication of four unknown rotation matrices cannot use the Baker–Campbell–Hausdorff formula [48] to generate linear optimization, which renders the optimization challenging and sensitive to outlier relative rotations. To address this, we decouple the estimation by first estimating the internal camera rotations, followed by estimating the rigid unit rotations using the known internal camera rotations. Through decoupled estimation of internal camera rotations and global rigid unit rotations, each step can be computed effectively on Lie algebra structures [7] and use the Baker–Campbell–Hausdorff formula [48].

3.3.1. Internal Rotation Estimation

Let $\rho(\cdot)$ represent the robust estimator function, and $d(\cdot)$ represents the geodesic distance metric [26]. We first independently estimate the global camera rotations via conventional rotation averaging by averaging all relative rotations:

$$\tilde{\boldsymbol{\mathcal{R}}} = \arg\min_{\boldsymbol{\mathcal{R}}} \sum_{i,j} \rho(d(\boldsymbol{R}_{ij}, \boldsymbol{R}_j \boldsymbol{R}_i^\top)). \tag{5}$$

We employ the method of Chatterjee *et al*. [7] to solve Eq. (5). Specifically, camera rotations are first initialized using image pairs from the maximum spanning tree of the view graph. Then, they are refined under the L_1 norm for a few iterations, followed by further refinement using the IRLS method.

Given the known initial global camera rotations $\tilde{\boldsymbol{\mathcal{R}}}$ from Eq. (5), the relative camera rotations of image pairs in the same rigid unit from Eq. (4) are calculated by:

$$\boldsymbol{R}_j^r \boldsymbol{R}_i^{r\top} = \boldsymbol{R}_{ij}^r = \tilde{\boldsymbol{R}}_j \tilde{\boldsymbol{R}}_i^\top. \tag{6}$$

To remove rotational ambiguity, we assume that the internal rotation $\boldsymbol{R}_i^r$ in each rigid unit is set to the identity matrix. From Eq. (6), the remaining internal rotations for each image like j in the same rigid unit are computed as:

$$\hat{\boldsymbol{R}}_j^r = \tilde{\boldsymbol{R}}_j \tilde{\boldsymbol{R}}_i^\top \boldsymbol{R}_i^r = \tilde{\boldsymbol{R}}_j \tilde{\boldsymbol{R}}_i^\top. \tag{7}$$

Since each rigid unit provides a solution for each remaining internal camera rotation, we obtain a set of possible solutions for each remaining internal camera rotation. To enhance robustness against outliers, we employ the method in [25] to estimate each internal camera rotation $\boldsymbol{R}^r$ by computing the geodesic median of all potential estimates $\hat{\boldsymbol{\mathcal{R}}}^r$ in $\mathrm{SO}(3)$.

3.3.2. Rigid Unit Rotation Estimation

With the known internal camera rotations and given all relative camera rotations in the view graph, the relative rigid unit rotations are computed by transforming Eq. (3):

$$\boldsymbol{R}_{ij}^g = \boldsymbol{R}_j^g \boldsymbol{R}_i^{g\top} = \boldsymbol{R}_j^{r\top} \boldsymbol{R}_{ij} \boldsymbol{R}_i^r. \tag{8}$$

Then, global rigid unit rotations are estimated through rotation averaging on Lie algebra structures. Similar to Eq. (5), the objective function is formulated by:

$$\min_{\boldsymbol{\mathcal{R}}^g} \sum_{i,j} \rho(d(\boldsymbol{R}_{ij}^g, \boldsymbol{R}_j^g \boldsymbol{R}_i^{g\top})), \tag{9}$$

which can also be effectively solved with the method of Chatterjee *et al*. [7]. To leverage the larger field of view provided by multiple cameras for enhanced robustness, we construct a maximum spanning tree based on the number of inlier feature matches from all overlapping image pairs between rigid units, and use it to estimate the initial global rotations of the rigid units. Finally, given the estimated internal camera rotations $\boldsymbol{R}_j^r$ and global rigid unit rotations $\boldsymbol{R}_j^g$, we obtain the global camera rotations using Eq. (2).

In conclusion, our approach decouples the estimation of internal camera rotations from global rigid unit rotations without requiring cameras in a multi-camera system to have overlapping fields of view. Leveraging the effectiveness of existing methods based on Lie algebra structures, our method demonstrates strong robustness and efficiency.

3.4. Hybrid Multi-Camera Translation Averaging

From Eqs. (1) and (2), relative translations $\boldsymbol{t}_{ij}$ are expressed in terms of global rigid unit positions and internal camera translations as follows:

$$\boldsymbol{R}_j{}^\top \boldsymbol{t}_{ij} = \frac{\mathbb{C}_{ij}}{||\mathbb{C}_{ij}||_2}, \tag{10}$$

where $\mathbb{C}_{ij} \triangleq \boldsymbol{c}_i - \boldsymbol{c}_j = \boldsymbol{c}_i^g - \boldsymbol{c}_j^g + \boldsymbol{R}_j{}^\top \boldsymbol{t}_j^r - \boldsymbol{R}_i{}^\top \boldsymbol{t}_i^r$.

Existing work like MMA [12] performs translation averaging by employing only relative translations to formulate a convex distance-based objective function by:

$$\min_{\boldsymbol{\mathcal{C}}^g, \mathcal{S}, \boldsymbol{\mathcal{T}}^r} \sum_{i,j} ||s_{ij} \boldsymbol{R}_j{}^\top \boldsymbol{t}_{ij} - \mathbb{C}_{ij}||_1, \text{ s.t. } s_{ij} \geq 1. \tag{11}$$

Here s_{ij} denotes the scale between camera i and camera j. Although Eq. (11) is a convex problem, the biased distance-based objective function compromises accuracy [55]. Two types of angle-based objectives are usually used to address this. A non-bilinear objective function is formulated as:

$$\min_{\boldsymbol{\mathcal{C}}} \sum_{i,j} \rho(||\boldsymbol{R}_j{}^\top \boldsymbol{t}_{ij} - \frac{\boldsymbol{c}_i - \boldsymbol{c}_j}{||\boldsymbol{c}_i - \boldsymbol{c}_j||_2}||_2). \tag{12}$$

While efficient, such a formulation with random initialization is sensitive to outliers [50]. A bilinear objective function is formulated by incorporating normalizing variables d_{ij} as:

$$\min_{\boldsymbol{\mathcal{C}}, \mathcal{D}} \sum_{i,j} \rho(||\boldsymbol{R}_j{}^\top \boldsymbol{t}_{ij} - d_{ij}(\boldsymbol{c}_i - \boldsymbol{c}_j)||_2), \text{ s.t. } d_{ij} \geq 0. \tag{13}$$

While incorporating additional variables improves convergence, it compromises efficiency. Furthermore, the exclusive use of camera-to-camera constraints derived from relative translations renders these methods susceptible to degenerate cases [29, 46]. This issue can be addressed by incorporating camera-to-point constraints. The mathematical expression of the camera-to-point constraint is given by:

$$\boldsymbol{R}_i{}^\top \boldsymbol{f}_{ik} = \frac{\boldsymbol{p}_k - \boldsymbol{c}_i}{||\boldsymbol{p}_k - \boldsymbol{c}_i||_2}, \quad \boldsymbol{f}_{ik} = \frac{\pi_i^{-1}(\boldsymbol{x}_{ik})}{||\pi_i^{-1}(\boldsymbol{x}_{ik})||_2}, \tag{14}$$

where $\boldsymbol{p}_k$ is the position of point k, $\boldsymbol{x}_{ik}$ is the feature point in image i corresponding to point k, π_i is the intrinsics of camera i and $\boldsymbol{f}_{ik}$ is the normalized feature ray from image i to point k in the camera coordinate system. The method GLOMAP [41] exclusively uses feature tracks to formulate a bilinear angle-based objective function as follows:

$$\min_{\boldsymbol{\mathcal{C}}, \mathcal{D}, \boldsymbol{\mathcal{P}}} \sum_{i,k} \rho(||\boldsymbol{R}_i{}^\top \boldsymbol{f}_{ik} - d_{ik}(\boldsymbol{p}_k - \boldsymbol{c}_i)||_2), \text{ s.t. } d_{ik} \geq 0. \tag{15}$$

Although this bilinear formulation converges reliably [41], the random initialization strategy used to solve this non-convex problem compromises robustness, and inclusion of redundant variables increase computational overhead.

In summary, a single optimization framework usually cannot fulfill the large-scale translation averaging task. Hence, we propose to handle this in a hybrid way, where the distance-based method is used for robust initialization, and the non-bilinear angle-based method is used for further refinement.

3.4.1. Camera Position Initialization

Given global camera rotations and relative translations from the view graph, camera positions are first initialized using Eq. (11) and subsequently refined using a non-bilinear objective function, as formulated below:

$$\min_{\boldsymbol{\mathcal{C}}^g, \boldsymbol{\mathcal{T}}^r} \sum_{i,j} \rho(||\boldsymbol{R}_j{}^\top \boldsymbol{t}_{ij} - \frac{\mathbb{C}_{ij}}{||\mathbb{C}_{ij}||_2}||_2). \tag{16}$$

3.4.2. Robust 3D Point Triangulation

With estimated global camera poses, we utilize the convex distance-based cross-product-form objective function to robustly triangulate the initial 3D points under L_1 norm as:

$$\min_{\boldsymbol{\mathcal{P}}} \sum_{i,k} ||\boldsymbol{R}_i{}^\top \boldsymbol{f}_{ik} \times (\boldsymbol{p}_k - \boldsymbol{c}_i)||_1. \tag{17}$$

3.4.3. Joint Angle-based Refinement

Finally, based on Eqs. (2) and (14), we derive the camera-to-point constraints from feature tracks to jointly refine all camera positions and 3D points through the multi-camera-based non-bilinear angle-based objective function as:

$$\min_{\boldsymbol{\mathcal{C}}^g, \boldsymbol{\mathcal{P}}, \boldsymbol{\mathcal{T}}^r} \sum_{i,k} \rho(||\boldsymbol{R}_i{}^\top \boldsymbol{f}_{ik} - \frac{\boldsymbol{p}_k - \boldsymbol{c}_i^g + \boldsymbol{R}_i{}^\top \boldsymbol{t}_i^r}{||\boldsymbol{p}_k - \boldsymbol{c}_i^g + \boldsymbol{R}_i{}^\top \boldsymbol{t}_i^r||_2}||_2). \tag{18}$$

In conclusion, relative translations and feature tracks within a robust distance-based objective function primarily serve as initialization, while feature tracks within a non-bilinear objective function are mainly utilized for refinement.

3.5. Multi-Camera Bundle Adjustment

Given camera parameters and 3D points, conventional bundle adjustment refines these parameters simultaneously by minimizing the reprojection error, which is formulated as:

$$\min_{\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{P}}, \boldsymbol{\mathcal{R}}, \Pi} \sum_{i,k} \rho(||\pi_i(\boldsymbol{R}_i(\boldsymbol{p}_k - \boldsymbol{c}_i)) - \boldsymbol{x}_{ik}||_2). \tag{19}$$

By integrating multi-camera constraints from Eq. (2), the multi-camera bundle adjustment is formulated as:

$$\min_{\substack{\boldsymbol{\mathcal{C}}^g, \boldsymbol{\mathcal{P}}, \boldsymbol{\mathcal{R}}^g, \\ \boldsymbol{\mathcal{R}}^r, \boldsymbol{\mathcal{T}}^r, \Pi}} \sum_{i,k} \rho(||\pi_i(\boldsymbol{R}_i^r \boldsymbol{R}_i^g (\boldsymbol{p}_k - \boldsymbol{c}_i^g) + \boldsymbol{t}_i^r) - \boldsymbol{x}_{ik}||_2). \tag{20}$$

In our framework, the rotations of internal cameras and rigid units remain fixed during the initial round of multi-camera bundle adjustment and are subsequently jointly optimized alongside 3D points and translations.

Data		COLMAP[44]				GLOMAP[41]				MMA[12]				MCSfM[13]				DMRA+MGP				MGSfM			
Name	N	$\tilde{e}_r$	$\bar{e}_r$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_r$	$\bar{e}_r$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_r$	$\bar{e}_r$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_r$	$\bar{e}_r$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_r$	$\bar{e}_r$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_r$	$\bar{e}_r$	$\tilde{e}_t$	$\bar{e}_t$
00	9082	0.4	0.5	0.9	3.8	0.4	**0.4**	0.8	1.3	0.7	0.8	1.5	1.9	**0.3**	**0.4**	0.6	0.8	0.4	**0.4**	**0.5**	**0.7**	0.4	**0.4**	**0.5**	**0.7**
01	2202	**0.2**	0.5	1.5	3.0	0.5	1.4	4.5	1e2	0.7	1.1	16.3	23.2	**0.2**	0.7	1.2	2.7	0.4	1.4	0.8	31.7	**0.2**	**0.3**	**0.6**	**1.1**
02	9322	0.5	0.7	4.0	12.3	0.4	0.5	5.4	1e3	1.8	2.3	7.0	10.1	**0.3**	**0.3**	1.0	**1.4**	0.4	0.4	1.3	1.8	0.4	0.4	**0.9**	**1.4**
03	1602	0.1	0.2	0.2	0.8	0.2	0.2	0.4	0.9	0.7	0.7	0.4	1.9	**0.2**	**0.2**	**0.2**	**0.4**	**0.2**	**0.2**	**0.2**	**0.4**	**0.2**	**0.2**	**0.2**	**0.4**
04	542	**0.1**	**0.1**	**0.1**	**0.2**	**0.1**	**0.1**	0.2	0.5	0.3	0.3	2.4	13.3	**0.1**	0.2	**0.1**	**0.2**	**0.1**	**0.1**	**0.1**	**0.2**	**0.1**	**0.1**	**0.1**	**0.2**
05	5522	0.6	0.6	0.8	2.2	**0.2**	**0.3**	0.3	0.6	0.6	1.3	0.6	1.2	0.3	**0.3**	**0.2**	0.5	**0.2**	**0.3**	**0.2**	**0.3**	**0.2**	**0.3**	**0.2**	**0.3**
06	2202	0.2	0.4	0.2	1.6	**0.1**	**0.2**	0.3	0.5	0.5	0.5	0.3	1.0	0.2	0.4	**0.1**	0.8	**0.1**	**0.2**	**0.1**	**0.1**	**0.1**	**0.2**	**0.1**	**0.1**
07	2202	1.1	1.1	0.5	2.2	0.3	**0.3**	0.3	**0.3**	0.6	0.6	0.9	2.2	0.5	0.7	0.7	3.1	**0.2**	**0.3**	**0.2**	**0.3**	**0.2**	**0.3**	**0.2**	**0.3**
08	8142	0.7	0.9	6.1	11.5	0.7	0.9	3.1	4.7	0.8	0.8	2.6	3.1	**0.4**	0.6	1.2	2.6	**0.4**	0.5	1.0	1.5	**0.4**	**0.4**	**0.9**	**1.4**
09	3182	0.6	0.6	1.1	3.4	**0.3**	**0.3**	1.7	2.8	0.6	0.6	1.7	3.5	**0.3**	**0.3**	**0.5**	**1.0**	**0.3**	**0.3**	**0.5**	**1.0**	**0.3**	**0.3**	**0.5**	**1.0**
10	2402	0.8	1.4	2.5	4.3	**0.3**	**0.3**	0.7	1.3	0.7	0.8	0.6	1.2	0.4	0.5	**0.4**	**1.2**	0.4	0.6	0.6	1.4	0.4	0.5	0.6	1.3

Table 1. Camera pose accuracy on the KITTI Odometry dataset. N shows the number of images; $\tilde{e}_r$ and $\bar{e}_r$ respectively denote the median and mean rotation error in degrees; $\tilde{e}_t$ and $\bar{e}_t$ respectively denote the median and mean position error in meters. The best results are **bolded**.

4. Experiment

We conduct experiments on real images collected by various multi-camera systems, including a stereo camera from the KITTI Odometry dataset [20], a four-camera system from the KITTI-360 dataset [31], two types of multi-camera systems mounted on street-view cars, and the Insta360 Pro2 system equipped with six fisheye cameras. Notably, only the two KITTI datasets provide ground-truth camera poses.

We compare our system MGSfM, with four state-of-the-art SfM techniques: the incremental method COLMAP [44], the global method GLOMAP [41], the multi-camera-based incremental method MCSfM [13], and the multi-camera-based global method MMA [12]. Additionally, we propose a method named “DMRA+MGP”, which employs our decoupled multi-camera rotation averaging (DMRA) and multi-camera bundle adjustment techniques, but modifies the hybrid multi-camera translation averaging step by integrating multi-camera constraints from Eq. (2) into the global positioning (MGP) objective function of GLOMAP [41] (see Eq. (15)). For the DMRA step, we utilize the method suggested in [7] to solve Eqs. (5) and (9). We employ the alternating direction method of multipliers (ADMM) optimizer [4] to solve the L_1 norm optimization problem in Eqs. (11) and (17). We use the Levenberg-Marquardt algorithm [30] from Ceres [1] as the optimizer to solve both the angle-based refinement problem and the bundle adjustment. In this framework, the Cauchy kernel is employed in Eqs. (16) and (18), while the Huber kernel is applied in Eq. (20).

All methods share the same view graph as input, which contains relative poses and the corresponding feature matches. For accuracy evaluation, we compare the median and mean errors of both camera rotations and positions after bundle adjustment on two kinds of KITTI datasets. We compare the reconstruction runtime on the larger KITTI-360 dataset for efficiency evaluation, excluding the view graph construction step since it is identical across all methods.

02 08

Figure 2. Our reconstruction results for data 02 and data 08 of the KITTI Odometry datasets.

4.1. Evaluation of KITTI Odometry Datasets

This dataset was collected using a stereo camera mounted on a driving car, where most camera motion trajectories are approximately collinear. Since there are moving cars and pedestrians on the street, we first use an off-the-shelf semantic segmentation algorithm [8] to detect these objects and avoid extracting feature points from these regions.

The accuracy comparison of the camera poses estimated by different methods is presented in Tab. 1, with MGSfM generally achieving the highest accuracy among most datasets. For scenes without loop closure, such as data 02 and 08, classical SfM methods yield camera poses with low accuracy due to scale drift. Comparing the COLMAP and MCSfM, the incorporation of multi-camera constraints improves the accuracy of the estimated camera poses, but the accumulated error still causes scene drift. In comparison, our global multi-camera system is more accurate. Since there is an overlap between the stereo cameras, global rotations of reference cameras in MMA [12] are estimated with known internal camera rotations. However, it still exhibits large position errors because only a biased distance-based objective function is used. For data 01, which contains numerous outlier feature matches, MGSfM also yields the best camera poses, while DMRA+MGP fails, demonstrating the strong robustness of our hybrid translation averaging. Fig. 2 presents two sample reconstruction results on the large-scale data 02 and 08, which are produced by our system MGSfM.

Data		COLMAP[44]				GLOMAP[41]				MMA[12]				MCSfM[13]				DMRA+MGP				MGSfM			
Name	N	$\tilde{e}_r$	$\bar{e}_r$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_r$	$\bar{e}_r$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_r$	$\bar{e}_r$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_r$	$\bar{e}_r$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_r$	$\bar{e}_r$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_r$	$\bar{e}_r$	$\tilde{e}_t$	$\bar{e}_t$
0000	9216	1.1	1.5	17.7	26.1	**0.4**	**0.7**	1.5	1.9	0.7	1.0	4.7	5.6	**0.4**	**0.7**	1.0	1.1	**0.4**	**0.7**	0.9	1.0	**0.4**	**0.7**	**0.8**	**0.9**
0002	11508	0.8	1.2	4.0	5.8	0.9	1.2	5.3	13.9	10.7	13.8	30.5	62.2	0.8	**1.1**	**1.1**	1.5	**0.7**	**1.1**	1.2	**1.4**	**0.7**	**1.1**	1.2	**1.4**
0003	828	-	-	-	-	0.4	0.7	3.5	11.2	3.5	6.3	4.7	17.8	2.6	3.9	1.4	13.9	0.4	0.6	**0.8**	1.9	**0.3**	**0.5**	**0.8**	**1.1**
0004	9272	0.7	1.3	7.3	13.2	0.7	1.2	8.4	29.8	4.6	11.7	17.4	54.1	0.6	1.0	1.3	1.8	0.6	1.0	1.3	18.3	**0.5**	**0.9**	**1.1**	**1.4**
0005	5396	1.6	1.8	6.2	8.3	0.8	1.2	5.5	31.9	5.6	19.7	31.5	1e2	0.7	1.1	0.9	**1.0**	0.5	**1.0**	**0.8**	1.1	**0.4**	**1.0**	**0.8**	1.2
0006	7760	0.8	1.1	17.0	36.0	1.2	1.4	19.4	2e2	11.4	10.9	39.7	52.2	1.7	1.9	4.0	5.9	**0.7**	**0.9**	**1.5**	61.7	**0.7**	1.0	**1.5**	**1.9**
0007	2720	20.7	26.5	88.4	2e3	1.1	1.2	1e2	3e2	16.2	13.9	1e2	4e2	**0.3**	**0.6**	**1.4**	6.0	1.4	1.6	7e2	1e3	0.6	0.8	2.0	**5.5**
0009	11248	6.3	7.7	61.4	78.5	0.5	**0.8**	1.1	1.5	2.7	5.1	5.9	19.8	**0.4**	**0.8**	**0.9**	**1.0**	**0.4**	**0.8**	**0.9**	**1.0**	**0.4**	**0.8**	**0.9**	**1.0**
0010	7672	16.0	21.8	4e2	6e2	1.8	1.6	2e2	4e2	14.5	12.1	68.0	2e2	**1.0**	**1.1**	2.7	8.9	**1.0**	**1.1**	2.1	3e2	1.1	1.2	**1.8**	**3.4**

Table 2. Camera pose accuracy on the KITTI-360 dataset. N shows the number of images. $\tilde{e}_r$ and $\bar{e}_r$ respectively denote the median and mean rotation error in degrees; $\tilde{e}_t$ and $\bar{e}_t$ respectively denote the median and mean position error in meters. The best results are **bolded**.

Figure 3. This figure shows the multi-camera system setup and the reconstruction results estimated by MGSfM on the KITTI-360 dataset.

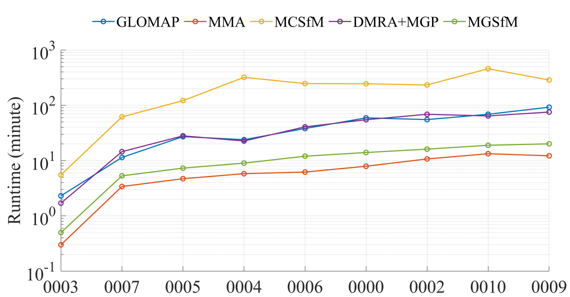

Figure 4. Runtime comparison (log scale) of different methods on the KITTI-360 dataset. Scenes are sorted by the ascending runtime of MGSfM to facilitate visualization.

4.2. Evaluation of KITTI-360 Datasets

The KITTI-360 dataset was captured using a station wagon equipped with a 180° fisheye camera on each side and a 90° perspective stereo camera at the front (see the top-left corner of Fig. 3). To mitigate the impact of dynamic objects, we also employ the semantic segmentation algorithm from [8]. Compared to the KITTI Odometry benchmark, KITTI-360 is more challenging due to the wider spatial distribution of the cameras, which results in almost no overlap between the fisheye and stereo cameras. The accuracy of estimated camera poses is summarized in Tab. 2, and the reconstruction results of MGSfM are shown in Fig. 3.

Since COLMAP does not incorporate multi-camera constraints and suffers from error accumulation, its estimated camera poses exhibit significant errors. GLOMAP achieves accuracy comparable to that of MGSfM only in data 0009, where there are few outlier feature matches and sufficient loop closures. The camera rotations estimated by MMA exhibit large errors because it performs rotation averaging by jointly estimating internal camera rotations and global reference rotations, rendering it sensitive to outlier relative rotations. Although MCSfM mitigates the error accumulation by fusing multi-camera constraints, it only outperforms MGSfM in the small-scale scene 0007. By initializing camera positions and 3D points before the angle-based refinement stage using camera-to-point constraints, MGSfM demonstrates higher robustness and accuracy than DMRA+MGP in challenging scenes such as data 0007 and data 0010. Although DMRA+MGP and MGSfM use the same camera rotations, MGSfM achieves better results in most datasets.

The reconstruction runtimes are shown in Fig. 4. Except for MMA, which only uses relative translations for translation averaging, MGSfM is faster than all other methods in comparison and more than $10\times$ faster than MCSfM. Although the method DMRA+MGP fuses multi-camera constraints, its efficiency is not significantly improved as the

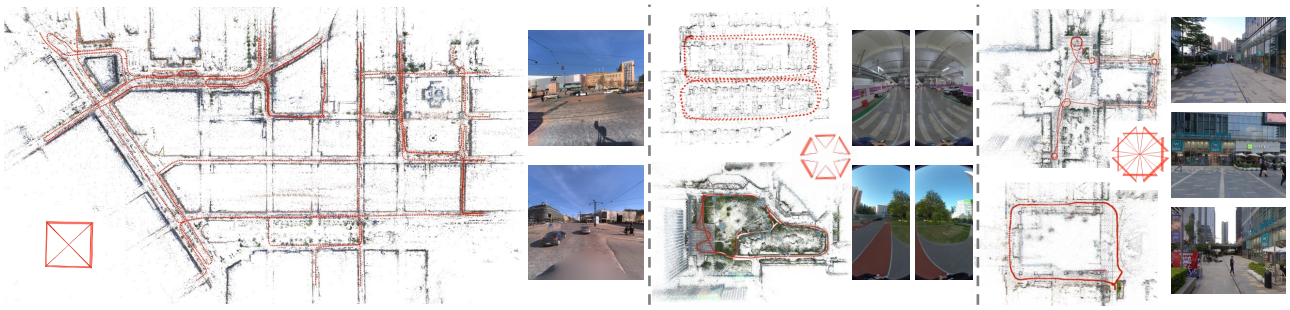

Figure 5. This figure shows the reconstruction results of five self-collected datasets. The left dataset contains 12k+ images. The middle two datasets contain 2k+ and 5k+ images, respectively. The right two datasets contain 6k+ and 2k+ images, respectively.

bilinear objective requires additional optimization of numerous redundant scale variables for all feature rays.

4.3. Evaluation of Self-Collected Datasets

To show the scalability of our system, we conduct experiments on five self-collected datasets captured by three types of multi-camera systems. The datasets and the corresponding reconstruction results from MGSfM are shown in Fig. 5. Our system successfully reconstructs all scenes across various scales, demonstrating its effectiveness across different multi-camera systems. Especially for the four-camera system dataset named STREET, our system only needs half an hour to perform the motion averaging on more than 12,000 images. More results are shown in our supplemental material.

4.4. Ablation Study

We use the KITTI-360 dataset [31] to analyze the impact of different input image observations and different angle-based objective functions on the performances of multi-camera translation averaging. As shown in Tab. 3, using the same known global rotations and view graph as input, we compare the accuracy of camera positions from six multi-camera translation averaging methods that respectively use only relative translations (denoted as "Only trans"), only feature tracks (denoted as "Only tracks"), or a hybrid of both (denoted as "Hybrid"), while employing either bilinear or non-bilinear angle-based objective functions.

For methods exclusively using relative translations, we first employ Eq. (11) to initialize camera positions, followed by angle-based refinement. As shown in the "Only trans" column of Tab. 3, bilinear formulations sometimes yield significantly larger mean distance errors by disrupting the continuity of the camera trajectory, whereas non-bilinear formulations are more robust to outlier relative translations. For methods exclusively using feature tracks, camera positions and 3D points are jointly optimized with random initialization. As shown in the "Only tracks" column, bilinear formulations (corresponding to DMRA+MGP in Tab. 2) exhibit better accuracy than their non-bilinear counterparts, indicating that the non-bilinear function is not good at handling random initializations. For methods using hybrid input, camera positions and 3D points are jointly refined with reasonable initial values. As shown in the "Hybrid" column, non-bilinear formulations (corresponding to MGSfM in Tab. 2) show better robustness than bilinear formulations and achieve higher accuracy than methods relying solely on translations, particularly in the challenging data 0010.

Input	Only trans				Only tracks				Hybrid			
Bilinear	yes		no		yes		no		yes		no	
data	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_t$	$\bar{e}_t$
0000	0.9	1.0	0.9	1.0	0.9	1.0	1.1	2e1	**0.8**	**0.9**	**0.8**	**0.9**
0002	1.3	2.5	1.4	1.6	**1.2**	**1.4**	1.5	1e2	1.3	3.5	**1.2**	**1.4**
0003	1.8	5e1	**0.7**	1.2	0.8	1.9	0.8	1.7	0.8	1.7	0.8	**1.1**
0004	1.2	1.4	**1.1**	**1.4**	1.3	2e1	1.2	2e1	1.7	2.0	**1.1**	**1.4**
0005	0.9	1.2	1.0	1.2	**0.8**	**1.1**	1.0	2e2	**0.8**	1.2	**0.8**	1.2
0006	1.7	2.3	1.8	2.3	1.5	6e1	1.9	1e2	**1.5**	**1.9**	**1.5**	**1.9**
0007	1e2	3e2	2.0	8.2	7e2	1e3	7e2	9e2	3.1	2e2	**2.0**	**5.5**
0009	**0.9**	**1.0**	**0.9**	**1.0**	**0.9**	**1.0**	1.8	5e1	**0.9**	**1.0**	**0.9**	**1.0**
0010	5e1	9e1	2.5	4.7	2.1	3e2	4e2	6e2	5e1	9e1	**1.8**	**3.4**

Table 3. Camera position accuracy with different input image observations and different angle-based functions.

In conclusion, with robust initialization, non-bilinear formulations generally yield better performance than bilinear ones, and by incorporating camera-to-point constraints from feature tracks, hybrid methods achieve improved accuracy.

5. Conclusion

In this paper, we propose a multi-camera-driven global SfM pipeline that begins with decoupled multi-camera rotation averaging, followed by hybrid multi-camera translation averaging. Extensive experiments demonstrate that our approach surpasses many state-of-the-art SfM methods in terms of efficiency, accuracy, scalability, and robustness.

Acknowledgments This work was supported by the National Key R&D Program of China (No.2023YFB3906600), the National Natural Science Foundation of China (No.U22B2055, U23A20386 and 62273345), and the Beijing Natural Science Foundation (No.L223003).

References

[1] Sameer Agarwal, Keir Mierle, and The Ceres Solver Team. Ceres Solver. https://github.com/ceres-solver/ceres-solver, 2023. 6

[2] Federica Arrigoni, Tomas Pajdla, and Andrea Fusiello. Viewing graph solvability in practice. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8147–8155, 2023. 1

[3] Daniel Barath, Dmytro Mishkin, Ivan Eichhardt, Ilia Shipachev, and Jiri Matas. Efficient initial pose-graph generation for global sfm. In *IEEE Conference on Computer Vision and Pattern Recognitio (CVPR)*, pages 14546–14555, 2021. 1

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011. 6

[5] Qi Cai, Lilian Zhang, Yuanxin Wu, Wenxian Yu, and Dewen Hu. A pose-only solution to visual reconstruction and navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(1):73–86, 2021. 1, 3

[6] Carlos Campos, Richard Elvira, Juan J. Gómez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021. 3

[7] Avishek Chatterjee and Venu Madhav Govindu. Robust relative rotation averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):958–972, 2017. 1, 2, 4, 6

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 6, 7

[9] Xieyuanli Chen, Ignacio Vizzo, Thomas Läbe, Jens Behley, and Cyrill Stachniss. Range image-based lidar localization for autonomous vehicles. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5802–5808. IEEE, 2021. 1

[10] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized RANSAC. In *Joint Pattern Recognition Symposium*, pages 236–243. Springer, 2003. 1, 3

[11] David J Crandall, Andrew Owens, Noah Snavely, and Daniel P Huttenlocher. SfM with MRFs: Discrete-continuous optimization for large-scale structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(12):2841–2853, 2012. 2

[12] Hainan Cui and Shuhan Shen. MMA: Multi-camera based global motion averaging. In *AAAI Conference on Artificial Intelligence*, pages 490–498, 2022. 1, 2, 3, 4, 5, 6, 7

[13] Hainan Cui, Xiang Gao, and Shuhan Shen. MCSfM: Multi-camera-based incremental structure-from-motion. *IEEE Transactions on Image Processing*, 32:6441–6456, 2023. 1, 2, 3, 6, 7

[14] Zhaopeng Cui and Ping Tan. Global structure-from-motion by similarity averaging. In *IEEE International Conference on Computer Vision (ICCV)*, pages 864–872, 2015. 3

[15] Zhaopeng Cui, Nianjuan Jiang, Chengzhou Tang, and Ping Tan. Linear global translation estimation with feature tracks. In *British Machine Vision Conference (BMVC)*, pages 46.1–46.13, 2015. 3

[16] Frank Dellaert, David M Rosen, Jing Wu, Robert Mahony, and Luca Carlone. Shonan rotation averaging: Global optimality by surfing $SO(p)^n$. In *European Conference on Computer Vision (ECCV)*, pages 292–308. Springer, 2020. 2

[17] Yaqing Ding, Jian Yang, Viktor Larsson, Carl Olsson, and Kalle Åström. Revisiting the p3p problem. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4872–4880, 2023. 1

[18] Xiang Gao, Lingjie Zhu, Zexiao Xie, Hongmin Liu, and Shuhan Shen. Incremental rotation averaging. *International Journal of Computer Vision (IJCV)*, 129:1202–1216, 2021. 3

[19] Xiao-Shan Gao, Xiao-Rong Hou, Jianliang Tang, and Hang-Fei Cheng. Complete solution classification for the perspective-three-point problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 25(8):930–943, 2003. 1

[20] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research (IJRR)*, 32(11): 1231–1237, 2013. 6

[21] Marcel Geppert, Peidong Liu, Zhaopeng Cui, Marc Pollefeys, and Torsten Sattler. Efficient 2d-3d matching for multi-camera visual localization. In *International Conference on Robotics and Automation (ICRA)*, pages 5972–5978. IEEE, 2019. 1

[22] Venu Madhav Govindu. Lie-algebraic averaging for globally consistent motion estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I–I, 2004. 1, 2

[23] Xiaotian Han, Quanzeng You, Chunyu Wang, Zhizheng Zhang, Peng Chu, Houdong Hu, Jiang Wang, and Zicheng Liu. Mmptrack: Large-scale densely annotated multi-camera multiple people tracking benchmark. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4860–4869, 2023. 1

[24] Christian Häne, Lionel Heng, Gim Hee Lee, Friedrich Fraundorfer, Paul Furgale, Torsten Sattler, and Marc Pollefeys. 3d visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. *Image and Vision Computing*, 68:14–27, 2017. 3

[25] Richard Hartley, Khurrum Aftab, and Jochen Trumpf. L1 rotation averaging using the weiszfeld algorithm. In *CVPR 2011*, pages 3041–3048. IEEE, 2011. 4

[26] Richard Hartley, Jochen Trumpf, Yuchao Dai, and Hongdong Li. Rotation averaging. *International Journal of Computer Vision (IJCV)*, 103:267–305, 2013. 1, 4

[27] Lionel Heng, Mathias Bürki, Gim Hee Lee, Paul Furgale, Roland Siegwart, and Marc Pollefeys. Infrastructure-based calibration of a multi-camera rig. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4912–4919, 2014. 3

[28] Lionel Heng, Benjamin Choi, Zhaopeng Cui, Marcel Geppert, Sixing Hu, Benson Kuan, Peidong Liu, Rang Nguyen,

Ye Chuan Yeo, Andreas Geiger, et al. Project autovision: Localization and 3d scene perception for an autonomous vehicle with a multi-camera system. In *International Conference on Robotics and Automation (ICRA)*, pages 4695–4702. IEEE, 2019. 1

[29] Nianjuan Jiang, Zhaopeng Cui, and Ping Tan. A global linear method for camera pose registration. In *IEEE International Conference on Computer Vision (ICCV)*, pages 481–488, 2013. 1, 3, 5

[30] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944. 6

[31] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(3):3292–3310, 2022. 6, 8

[32] Liu Liu, Hongdong Li, Yuchao Dai, and Quan Pan. Robust and efficient relative pose with a multi-camera system for autonomous driving in highly dynamic environments. *IEEE Transactions on Intelligent Transportation Systems*, 19(8): 2432–2444, 2017. 1

[33] Liyang Liu, Teng Zhang, Brenton Leighton, Liang Zhao, Shoudong Huang, and Gamini Dissanayake. Robust global structure from motion pipeline with parallax on manifold bundle adjustment and initialization. *IEEE Robotics and Automation Letters*, 4(2):2164–2171, 2019. 3

[34] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60:91–110, 2004. 3

[35] Lalit Manam and Venu Madhav Govindu. Correspondence reweighted translation averaging. In *European Conference on Computer Vision (ECCV)*, pages 56–72. Springer, 2022. 3

[36] Lalit Manam and Venu Madhav Govindu. Sensitivity in translation averaging. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:62740–62763, 2023. 1

[37] Ra'ul Mur-Artal and Juan D. Tard'os. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 3

[38] Pha Nguyen, Kha Gia Quach, Chi Nhan Duong, Ngan Le, Xuan-Bac Nguyen, and Khoa Luu. Multi-camera multiple 3d object tracking on the move for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2569–2578, 2022. 1

[39] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(6):756–770, 2004. 3

[40] Onur Ozyesil and Amit Singer. Robust camera location estimation by convex programming. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2674–2683, 2015. 1, 3

[41] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, pages 58–77. Springer, 2024. 1, 2, 3, 5, 6, 7

[42] Jie Ren, Wenteng Liang, Ran Yan, Luo Mai, Shiwen Liu, and Xiao Liu. MegBA: A gpu-based distributed library for large-scale bundle adjustment. In *European Conference on Computer Vision (ECCV)*, pages 715–731. Springer, 2022. 1

[43] Francisco Rubio, Francisco Valero, and Carlos Llopis-Albert. A review of mobile robots: Concepts, methods, theoretical framework, and applications. *International Journal of Advanced Robotic Systems*, 16(2):1729881419839596, 2019. 1

[44] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 1, 2, 3, 6, 7

[45] Marco Sewtz, Xiaozhou Luo, Johannes Landgraf, Tim Bodenmüller, and Rudolph Triebel. Robust approaches for localization on multi-camera systems in dynamic environments. In *2021 7th International Conference on Automation, Robotics and Applications (ICARA)*, pages 211–215. IEEE, 2021. 1

[46] Peilin Tao, Hainan Cui, Mengqi Rong, and Shuhan Shen. Revisiting global translation estimation with feature tracks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20686–20696, 2024. 1, 2, 3, 5

[47] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms*, pages 298–372. Springer, 2000. 1

[48] Veeravalli S Varadarajan. *Lie groups, Lie algebras, and their representations*. Springer Science & Business Media, 2013. 2, 4

[49] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21729–21740, 2023. 1

[50] Kyle Wilson and Noah Snavely. Robust global translations with 1dsfm. In *European Conference on Computer Vision (ECCV)*, pages 61–75. Springer, 2014. 3, 5

[51] Kyle Wilson, David Bindel, and Noah Snavely. When is rotations averaging hard? In *European Conference on Computer Vision (ECCV)*, pages 255–270. Springer, 2016. 2

[52] Changhee Won, Hochang Seok, Zhaopeng Cui, Marc Pollefeys, and Jongwoo Lim. OmniSLAM: Omnidirectional localization and dense mapping for wide-baseline multi-camera systems. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 559–566. IEEE, 2020. 1

[53] Changchang Wu. Towards linear-time incremental structure from motion. In *International Conference on 3D Vision (3DV)*, pages 127–134. IEEE, 2013. 1

[54] Ganlin Zhang, Viktor Larsson, and Daniel Barath. Revisiting rotation averaging: Uncertainties and robust losses. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17215–17224, 2023. 3

[55] Bingbing Zhuang, Loong-Fah Cheong, and Gim Hee Lee. Baseline desensitizing in translation averaging. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4539–4547, 2018. 1, 2, 3, 5