

Article

Fast Intrinsic–Extrinsic Calibration for Pose-Only Structure-from-Motion

Xiaoyang Tian , Yangbing Ge, Zhen Tan , Xieyuanli Chen , Ming Li and Dewen Hu * 

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China; tianxy0709@nudt.edu.cn (X.T.); yangbing_ge@nudt.edu.cn (Y.G.); tanzhen1996@nudt.edu.cn (Z.T.); xieyuanli.chen@nudt.edu.cn (X.C.); liming78@nudt.edu.cn (M.L.)

* Correspondence: dwhu@nudt.edu.cn

Abstract

Structure-from-motion (SfM) is a foundational technology that facilitates 3D scene understanding and visual localization. However, bundle adjustment (BA)-based SfM is usually very time-consuming, especially when dealing with numerous unknown focal length cameras. To address these limitations, we proposed a novel SfM system based on pose-only adjustment (PA) for intrinsic and extrinsic joint optimization to accelerate computing. Firstly, we propose a base frame selection method based on depth uncertainty, which integrates the focal length and parallax angle under a multi-camera system to provide more stable depth estimation for subsequent optimization. We explicitly derive a global PA of joint intrinsic and extrinsic parameters to reduce the high dimensionality of the parameter space and deal with cameras with unknown focal lengths, improving the efficiency of optimization. Finally, a novel pose-only re-triangulation (PORT) mechanism is proposed for enhanced reconstruction completeness by recovering failed triangulations from incomplete point tracks. The proposed framework has been demonstrated to be both faster and comparable in accuracy to state-of-the-art SfM systems, as evidenced by public benchmarking and analysis of the visitor photo dataset.

Keywords: structure-from-motion; pose-only adjustment; bundle adjustment; camera calibration; mapping; pose estimation; 3D reconstruction



Academic Editor: Giuseppe Casula

Received: 20 May 2025

Revised: 20 June 2025

Accepted: 27 June 2025

Published: 30 June 2025

Citation: Tian, X.; Ge, Y.; Tan, Z.; Chen, X.; Li, M.; Hu, D. Fast Intrinsic–Extrinsic Calibration for Pose-Only Structure-from-Motion. *Remote Sens.* **2025**, *17*, 2247. <https://doi.org/10.3390/rs17132247>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Structure-from-motion (SfM) demonstrates significant potential for geometry reconstruction and visual localization [1–3], while the efficient estimation of cameras with unknown focal lengths remains critically important in contemporary data-dense scenarios [2,4–6]. Bundle adjustment (BA), which involves the joint non-linear refinement of camera and point parameters, is a critical component of most SfM systems. However, it is computationally expensive, especially when dealing with the focal length and extrinsic camera parameters multiple times along with 3D point optimization [7]. As the size of image collections grows exponentially, the efficiency of optimization algorithms becomes increasingly critical in order to handle the computational burden. It is well known that BA often requires multiple iterations in incremental SfM [7], particularly as the amount of data increases, leading to a significant computational cost. Although a number of approaches have attempted to address the optimization efficiency problem [2,8], these approaches have not reduced the high dimensionality of the BA parameter space, and the optimization efficiency issues remain limiting for reconstruction. In comparison, the global SfM

exhibits superior optimization efficiency by recovering all camera poses simultaneously. Nevertheless, this has resulted in challenges, including diminished robustness and reduced accuracy [9]. Consequently, the development of an optimization method that can enhance efficiency and reliability in unknown scenario is imperative.

A novel pose-only adjustment (PA) algorithm founded upon the pose-only imaging geometry has recently been proposed as a potential solution to the problem in this area [10,11]. In contradistinction to the BA, the PA attains a pose-only representation of 3D points by a pair of pose-only (PPO) constraints, which is an implicit representation of 3D points. It is noteworthy that the optimization process does not necessitate the incorporation of 3D points, as it is sufficient to optimize solely for the positional pose. This significantly reduces the dimensionality of the parameter space. As demonstrated in [12–16], SLAM experiments have further substantiated that PA optimization has the potential to replace BA optimization if efficiency improvements are needed. Interestingly, PA also mitigates the point-scattering phenomenon [10]. Unfortunately, these methods can only optimize the pose while ignoring the existence of a large number of unknown-focal-length cameras in real scenarios, and they are extremely dependent on the exact normalized coordinates, which are very sensitive to image noise and, thus, become inapplicable when the focal length is uncertain or varying. In real scenarios, the inevitable disturbances caused by motion inevitably affect the focal length of camera. The joint optimization of intrinsic and extrinsic parameters is, therefore, required to achieve the SfM task for real datasets.

To tackle the aforementioned challenges, this paper focuses on enhancing the efficiency of the SfM while improving its robustness to cameras of varying focal lengths from the perspective of the optimization process and without relying on any prior information. We explicitly derive a global PA for intrinsic and extrinsic joint optimization and propose a novel SfM system based on pose-only imaging geometry theory [10]. Specifically, we propose a base frame selection algorithm based on depth uncertainty, which takes into account the effect of camera focal length on depth estimation, improves the robustness of depth estimation, and facilitates the performance of the subsequent optimization process. Next, we adopt global PA instead of global BA to reduce the computational burden associated with the high-dimensional parameter space and enhance overall efficiency. Furthermore, we extend PA from extrinsic optimization only to joint intrinsic–extrinsic optimization, thus addressing the challenge of unknown camera parameters and improving robustness. Finally, we design a novel pose-only re-triangulation (PORT) module to continue trajectories of triangulation failure to improve reconstruction completeness. We evaluate our method using public 1DSfM [17], Cambridge Landmarks [18], and Mip-Nerf 360 [19] (MIP360), as well as self-collected multi-focal-length datasets. The results show that the system effectively improves efficiency while performing the high-precision reconstruction task, as illustrated in Figure 1. To further evaluate and demonstrate the capability of our SfM framework on challenging scenes, an unordered-lens SfM real-scene dataset for multiple focal lengths was collected.

The main contributions are threefold:

- A depth uncertainty-based base frame selection method improves the robustness of depth estimation in scenes with different focal length combinations.
- A new global pose-only SfM framework with combined focal length and extrinsic optimization improves reconstruction speed in complex scenes.
- A novel PORT mechanism enhances reconstruction completeness by recovering failed triangulations from incomplete point tracks.

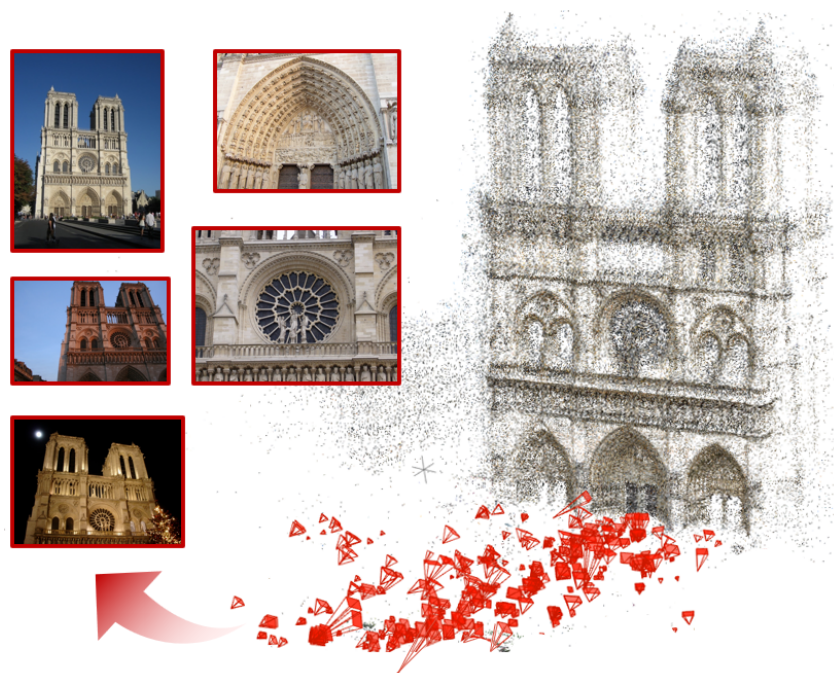


Figure 1. The proposed POMAP achieves satisfactory accuracy by sparse reconstruction for pose estimation on 1DSfM dataset of tourist photos with camera sensors with different focal lengths.

The paper is organized as follows. Section 2 introduces the related works on BA-based SfM and the PA optimization. In Section 3, the overall framework of the system is introduced, followed by an examination of the underlying frame selection principle based on depth uncertainty. The subsequent sections provide an explicit derivation process for PA-internal and external parametric optimization and a detailed delineation of the PORT process. Section 4 provides a detailed exposition of the performance of the proposed POMAP system with the optimization algorithm on both simulated and real datasets. In Section 5, the potential areas for enhancement of the current system are examined. The conclusions of the paper are summarized in Section 6.

2. Related Works

Optimization strategies have been shown to play a crucial role in SfM tasks [20,21]. In contrast, this paper focuses on the study of SfM optimization methods, combining traditional BA [7] and improved methods, while also exploring the more novel PA methods currently available.

2.1. Bundle Adjustment of SfM

Improving the speed of BA optimization must address three key issues [22]: high computational cost, limited efficiency, and susceptibility to input noise. The preconditioned conjugate gradient method was employed to accelerate BA solutions [23], while the sparse sparse BA (sSBA) algorithm proposed in [24] effectively addresses sparse BA problems through sparse linear subproblem formulations, substantially reducing computational complexity. Subsequent studies [2,4] developed adaptive Levenberg–Marquardt (LM) algorithms that dynamically select exact or truncated step sizes according to the problem scale, simultaneously reducing the memory footprint and computational costs during optimization. Additional enhancements include geometric estimation [25], batch or distributed processing [8,26], and re-triangulation (RT) [1,5], which collectively mitigate input noise effects, optimize memory efficiency, and improve system robustness. These approaches rely on iterative incremental SfM optimizations yet remain computationally intensive. In con-

trast, global SfM significantly reduces computational costs by recovering all camera poses simultaneously. However, it is more susceptible to input noise. To address the limitation of input, Moulon et al. [9] improved the robustness and precision of rotational averaging based on the trifocal tensor, while Zhuang et al. [27] focused on improving baseline insensitivity in translational averaging. GLOMAP [3] built on this by combining translational averaging and triangulation into a single global position step, improving efficiency while providing better-optimized initial values. Note that these works do not reduce the high dimensionality of the BA parameter space, and the issue of elevated memory utilization remains unresolved. Furthermore, hybrid SfM methodologies have been developed [28,29], including HSfM [28], which achieves an equilibrium between efficiency and robustness through global rotation estimation and incremental center estimation, although it relies on precise camera parameters.

2.2. Pose-Only Adjustment of SfM

The PPO constraint relationship decouples the camera pose from the three feature coordinates. This is actually an equivalent expression of the geometric constraints for dual-view imaging [11]. Building on this, Cai et al. [10] further extend the PPO problem to multiple views to obtain a depth-pose-only (DPO) constraint set for the 3D feature point. That is, the 3D feature point P^W can be described in terms of the pose if the normalized coordinates are known. The DPO-based optimization problem is referred to as PA. Moreover, Cai et al. utilized PA in conjunction with the proposed linear global translation constraint (LiGT) [10] to verify the effectiveness of PA optimization in 3D reconstruction tasks. Building on the aforementioned theoretical framework, Ge et al. [12] pioneered the application of PA optimization in conjunction with preintegration merging theory within a VI-SLAM system. This approach enabled the attainment of an expeditious and lightweight SLAM, circumventing the challenges posed by the dimensional explosion and substantial memory consumption inherent to traditional BA. Furthermore, an in-depth analysis of the characteristics of PA optimization was conducted from the perspective of Hessian matrix sparsity, resulting in the generation of a more compact optimization matrix under a uniform optimization objective. Furthermore, a substantial body of research [13–15] has empirically demonstrated that PA can significantly improve optimization efficiency and is a potential alternative to BA in computationally intensive scenarios. However, it should be noted that the aforementioned PA optimization using solely extrinsic parameters is overly reliant on the precise normalized coordinates, and is vulnerable to focus noise and image noise. Moreover, it is incapable of adapting to scenarios where the focal length is unknown or in constant fluctuation.

3. Methods

3.1. Framework Overview

Our framework focuses on improving the efficiency and robustness of cameras with different focal lengths in the optimization process during SfM. As shown in Figure 2, the initialization phase involves feature extraction and matching, relative pose estimation, camera calibration, and global pose estimation. Feature extraction, matching, and pose estimation are implemented using well-established algorithms from existing frameworks [1], aiming at obtaining accurate matching relations, feature trajectories, initial camera parameters, and relative poses. For the problem of reconstruction failure due to mirroring, the doppelganger [30] can be used to filter out false matches. Subsequently, the global information with more accurate global rotation, global translation, and observing points is obtained through the rotation average [31] and global positioning [3]. In the optimization stage, base frames need to be selected before global PA to achieve an accurate implicit representation of 3D points. We propose the depth uncertainty for base frame selection.

After global PA, PORT can be used further to improve the accuracy and completeness of the reconstruction. In addition, camera clustering [3] with larger error-point filtering can be combined to improve the robustness and accuracy of the system.

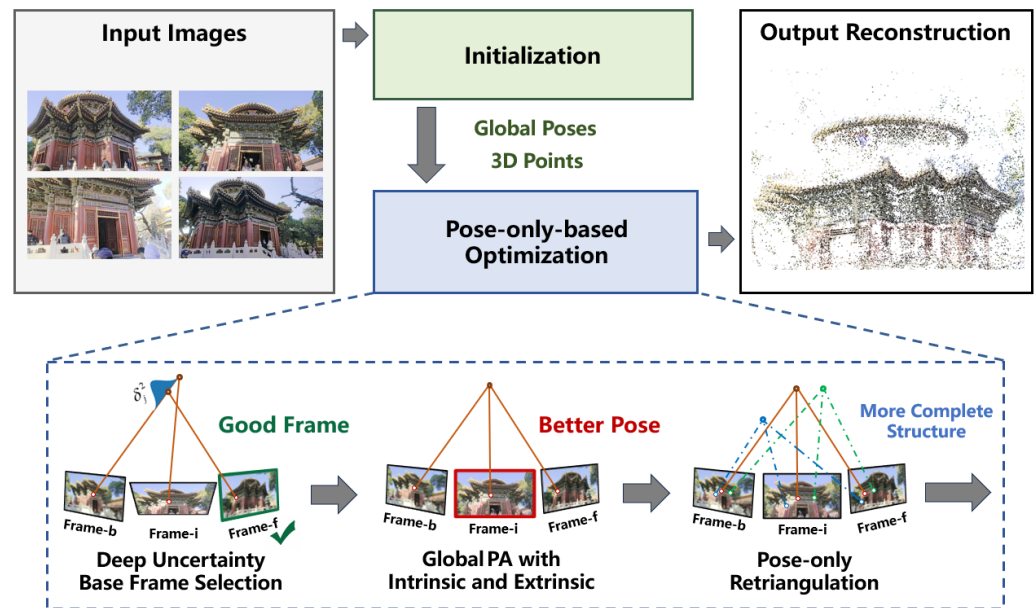


Figure 2. Pipeline of proposed POMAP system. Our work focuses on the optimization part, proposing a base frame selection method based on depth uncertainty, a global pose-only adjustment based on intrinsic and extrinsic, and a subsequent PORT module to improve the reconstruction integrity of the system.

3.2. Depth Uncertainty-Based Base Frame Selection

The selection of base frames critically influences subsequent optimization. In real-world scenarios with multiple focal lengths, parallax-based selection becomes biased by ignoring focal parameters [32]. To mitigate this, depth uncertainty evaluation incorporating camera focal length enables robust base frame selection, enhancing optimization robustness in complex situations.

Let P_j^W be the current estimation of the 3D feature point corresponding to the pixel coordinate p_{bj} or p_{fj} in frame b or f , and the vector representing the center of light of camera b to the 3D point is l_{bj} . R_{bf} and t_{bf} are the relative rotation and translation between frames b and f . $l_b = l_{bj}/\|l_{bj}\|$, and α and β denote the angles formed by the two rays and the baseline t_{bf} , shown in Figure 3; then,

$$l_{fj} = l_{bj} - t_{bf}, \quad (1)$$

$$\alpha = \arccos\left(\frac{l_b \cdot t_{bf}}{\|t_{bf}\|}\right), \quad (2)$$

$$\beta = \arccos\left(-\frac{l_{fj} \cdot t_{bf}}{\|l_{fj}\| \cdot \|t_{bf}\|}\right), \quad (3)$$

where l_{fj} denotes the vector from the optical center of frame f to P_j^W . Let $\hat{f}$ be the camera focal length of frame f . $\Delta\beta$, the angle spanning one pixel, can be added to β in order to compute parallax angle γ ; thus, by applying the law of sines, recover the norm of P_j^W :

$$\Delta\beta = 2 \tan^{-1}\left(\frac{1}{2\hat{f}}\right), \quad (4)$$

$$\gamma = \pi - \alpha - \beta - \Delta\beta, \quad (5)$$

$$\|\hat{l}_{bj}\| = \|t_{bf}\| \frac{\sin(\beta + \Delta\beta)}{\sin \gamma}, \quad (6)$$

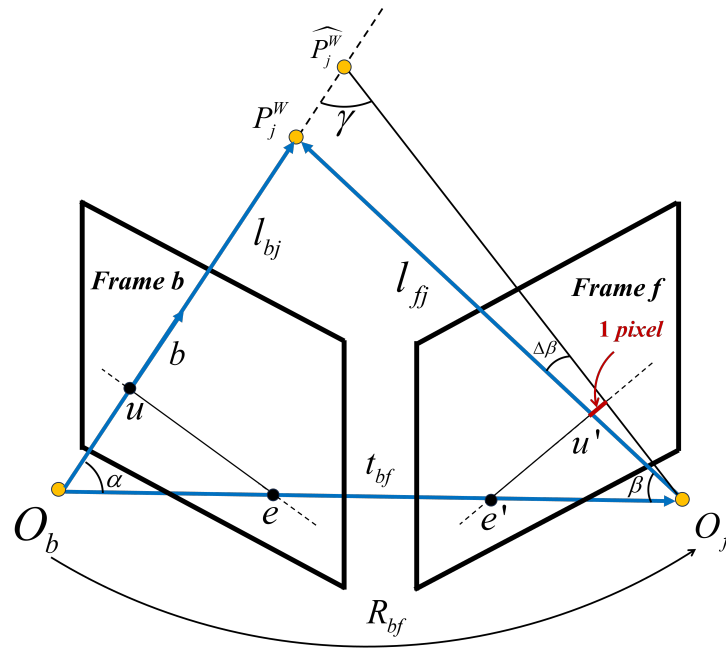


Figure 3. Calculation depth uncertainty. O_b and O_f denote the camera optical centers corresponding to frames b and f , respectively. u, u' is the projection of P_j^W in frames b and f , and e, e' is the epipoles. The mode of l_{bj} describes the depth information between the optical center of the camera at the currently predicted 3D point P_j^W . P_j^W describes the 3D points that may be estimated after possessing a pixel of image noise, and the Euclidean distance between him and P_j^W can describe the possibility of depth δ_j^2 .

If the focal length in the x -axis f_x differs from that in the y -axis f_y , a virtual focal length f_{new} can be introduced to simulate the angular change caused by moving 1 pixel along the polar line. The f_{new} is calculated from the fundamental matrix F as

$$l_{e'u'} = F p_{fj} = [a_r, b_r, c_r]^T, \quad (7)$$

$$\theta = \tan^{-1}\left(-\frac{a_r}{b_r}\right), \quad (8)$$

$$f_{new} = \sqrt{\cos^2 \theta \cdot f_x^2 + \sin^2 \theta \cdot f_y^2}, \quad (9)$$

where $l_{e'u'}$ is the polar line in frame f , θ is the angle between $l_{e'u'}$ and the x -axis, and p_{fj} is the j -th pixel point in frame f . a_r, b_r , and c_r are solvable variables. Therefore, the depth uncertainty δ_j^2 is computed as

$$\delta_j^2 = (\|\hat{l}_{bj}\| - \|l_{bj}\|)^2, \quad (10)$$

The aforementioned equation can be resolved in order to ascertain the depth uncertainty. The set exhibiting the minimum depth uncertainty is then selected as the base frame for subsequent optimization.

3.3. Global Pose-Only Adjustment

Consider the j -th pixel coordinate in frame i as p_{ij} , and P_{ij} is the j -th normalized coordinate of the 3D feature point P_j^W in frame i . We use the DPO constraints to obtain the depth S_{bf} of frame b , by which we can also get the pose-only description of the 3D feature point P_j^W :

$$\begin{aligned} P_j^W &= H_{po}(\{R_l, t_l, P_l\}_{l=b,f}) \\ &= \frac{\|[P_{fj}]_{\times}(t_f - R_f R_b^T t_b)\|}{\|[P_{fj}]_{\times} R_f R_b^T P_{bj}\|} P_{bj} - t_b \\ &= \frac{\lambda_{bf}}{\phi_{bf}} P_{bj} - t_b \end{aligned} \quad (11)$$

where P_l denotes the normalized coordinates from frames b or f , and $[P_{fj}]_{\times}$ is the antisymmetric matrix of P_{fj} . R_l, t_l denotes the global rotation and translation from frames b or f . $H_{po}(\cdot)$ denotes the pose-only representation function of 3D points P_j^W . $\lambda_{bf} = \|[P_{fj}]_{\times}(t_f - R_f R_b^T t_b)\|$, and $\phi_{bf} = \|[P_{fj}]_{\times} R_f R_b^T P_{bj}\|$. Then, the pixel point p_{ij} can be estimated by transforming P_{bj} from the normalization plane to frame i followed by reprojection. The progress shown in Figure 4 can be formulated as

$$p_{ij} = H_m(H_d(H_n(H_r(H_{po}(\{R_l, t_l, P_l\}_{l=b,f}), R_i, t_i)), K_i), M_i) \quad (12)$$

where H_m , H_d , H_n , and H_r denotes the perspective projection functions, the distortion function, the normalization function, and the rigid transformation function. K_i and M_i , respectively, denote the distortion parameter vectors and the intrinsic parameter vectors including the focal length of the camera for the pending frame i .

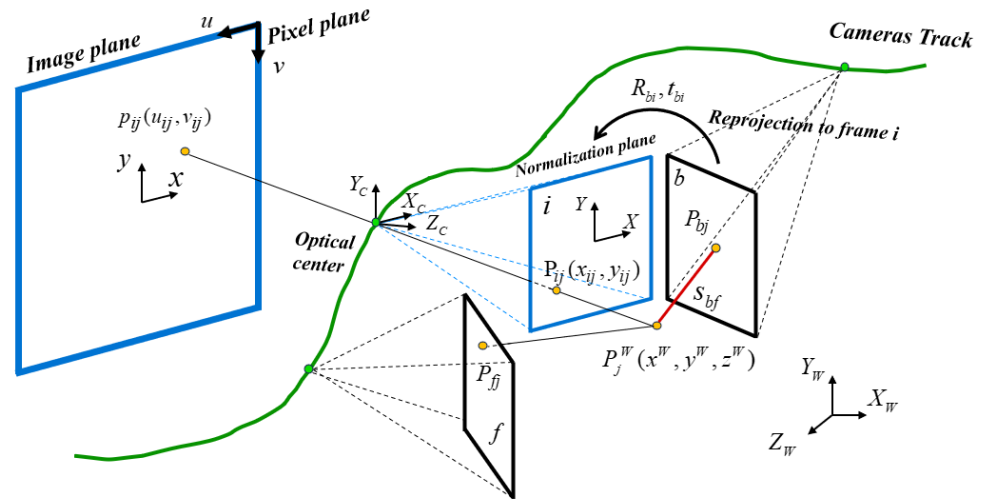


Figure 4. A simple demonstration of projection strategies. The depth S_{bf} can be obtained from the depth constraint relationship between frames b and f . Then, P_j^W can be directly represented by the pose between frames b and f . This is then projected into camera coordinates of frame i and onto its pixel plane via affine and perspective projection. For clarity, the image plane is shown after the optical centre, although the actual calculations use the one before.

The reprojection error can be also described as

$$e_{ij}(p_{ij}, \tilde{p}_{ij}) = p_{ij} - \tilde{p}_{ij} \quad (13)$$

where $\tilde{p}_{ij}$ denotes the measured values. Then, the optimization problem for m frames of individual images of n points can be described as

$$\arg \min_{\{R_i, t_i, K_i, M_i\}_{i=1 \dots m}} \sum_{i=1}^m \sum_{j=1}^n e_{ij}^T e_{ij} \quad (14)$$

In the above optimization problem, the optimization objectives are R_i , t_i , K_i , and M_i , which denote the rotation matrix, translation vectors, distortion parameters, and intrinsic parameters, respectively, excluding the 3D feature points. After obtaining the reprojection error, nonlinear optimization can be used to obtain the required parameters. In this paper, the Levenberg–Marquardt method is used to solve to obtain the update direction. Assuming $J_i = [J_{R_i}, J_{t_i}, J_{K_i}, J_{M_i}]$ is the Jacobian matrix of the frame i , where J_{R_i} , J_{t_i} , J_{K_i} , and J_{M_i} denote the Jacobian matrix of the rotation, the translation, the distortion parameters, and the intrinsic parameters, respectively; then, each part can be written in detail as follows:

$$\begin{aligned} J_{R_i} &= \frac{\partial e_{ij}}{\partial R_i} = \frac{\partial e_{ij}}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial P_{ij}^D} \frac{\partial P_{ij}^D}{\partial P_{ij}} \frac{\partial P_{ij}}{\partial P_{ij}^C} \frac{\partial P_{ij}^C}{\partial \varphi_i}, \\ J_{t_i} &= \frac{\partial e_{ij}}{\partial t_i} = \frac{\partial e_{ij}}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial P_{ij}^D} \frac{\partial P_{ij}^D}{\partial P_{ij}} \frac{\partial P_{ij}}{\partial P_{ij}^C} \frac{\partial P_{ij}^C}{\partial t_i}, \\ J_{K_i} &= \frac{\partial e_{ij}}{\partial K_i} = \frac{\partial e_{ij}}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial P_{ij}^D} \frac{\partial P_{ij}^D}{\partial K_i}, \\ J_{M_i} &= \frac{\partial e_{ij}}{\partial M_i} = \frac{\partial e_{ij}}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial M_i}. \end{aligned} \quad (15)$$

where φ denotes the right rotation perturbation. p_{ij} is the j -th pixel coordinate in the frame i . P_{ij}^D is the j -th distortion coordinate in the frame i . P_{ij} is the j -th normalized image coordinate in the frame i , and P_{ij}^C is the j -th camera coordinate in the frame i . Next, a detailed derivation of the intrinsic Jacobian matrix for frame i is presented as an example. The derivation of the Jacobian matrix with respect to the extrinsics can be found in [12].

(1) *Partial derivative of the normalized coordinates from the distortion coordinates:* In reality, given the inherent imperfections of the camera component assembly and the possibility of lens positional fluctuations due to movement, it becomes imperative to incorporate radial and tangential distortions to accurately depict the aforementioned scenario. This paper aims to facilitate the calculation of the distortion transformation by providing a detailed description function $H_d(\cdot)$, and P_{ij}^D is the j -th distortion coordinate in the frame i which can be expressed in the following manner:

$$\begin{aligned} P_{ij} &= H_n(H_r(H_{po}(\{R_l, t_l, P_l\}_{l=b,f}))) \\ &= \frac{\lambda_{bf} R_{bi} P_{bj} + \phi_{bf} t_{bi}}{(0, 0, 1)(\lambda_{bf} R_{bi} P_{bj} + \phi_{bf} t_{bi})} \end{aligned} \quad (16)$$

$$P_{ij}^D = H_d(P_{ij}, K_i) = \Delta P_{rad} + \Delta P_{tan} + P_{ij} \quad (17)$$

with

$$\begin{cases} \Delta P_{rad} = \begin{bmatrix} \Delta x_{rad} \\ \Delta y_{rad} \end{bmatrix} = (k_1 r^2 + k_2 r^4 + k_3 r^6) \begin{bmatrix} x_{ij} \\ y_{ij} \end{bmatrix} \\ \Delta P_{tan} = \begin{bmatrix} \Delta x_{tan} \\ \Delta y_{tan} \end{bmatrix} = \begin{bmatrix} 2p_1 x_{ij} y_{ij} + p_2 (r^2 + 2x_{ij}^2) \\ 2p_2 x_{ij} y_{ij} + p_1 (r^2 + 2y_{ij}^2) \end{bmatrix} \\ r^2 = x_{ij}^2 + y_{ij}^2 \end{cases}$$

where $R_{bi} = R_i R_b^T$ and $t_{bi} = t_i - R_b t_b$. ΔP_{rad} denotes the error generated by radial distortion, ΔP_{tan} denotes the error generated by tangential distortion, and r is the distance from the normalized plane point to the center. $K_i = [k_1, k_2, p_1, p_2, k_3]$ denotes the vector of distortion parameters; k_1, k_2 , and k_3 is the radial distortion coefficient; and p_1 and p_2 denotes the tangential distortion coefficient. Therefore, the partial derivative of the distortion point to the normalized coordinates is

$$\frac{\partial P_{ij}^D}{\partial P_{ij}} = \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix} \quad (18)$$

in which

$$\begin{cases} a_1 = R_d + 2k_1 x^2 + 4k_2 r^2 x^2 + 6k_3 r^4 x^2 + 2p_1 y + 6p_2 x \\ a_2 = 2k_1 xy + 4k_2 r^2 xy + 6k_3 r^4 xy + 2p_1 x + 2p_2 y \\ b_1 = 2k_1 xy + 4k_2 r^2 xy + 6k_3 r^4 xy + 2p_1 x + 2p_2 y \\ b_2 = R_d + 2k_1 y^2 + 4k_2 r^2 y^2 + 6k_3 r^4 y^2 + 2p_2 x + 6p_1 y \\ R_d = 1 + k_1 r^2 + k_2 r^4 + k_3 r^6 \end{cases}$$

where (x, y) is shorthand for the normalized coordinates (x_{ij}, y_{ij}) . In this paper, only three parameters are used to represent the radial distortion parameters. This is because higher-order parameters often do not bring significant performance improvement in practical applications and increase the computational complexity, so three parameters are sufficient to describe the majority of cases.

Similarly, according to (17), the partial derivative of the distortion coordinates to the distortion parameters vectors is

$$\frac{\partial P_{ij}^D}{\partial K_i} = \begin{bmatrix} xr^2 & xr^4 & 2xy & 3x^2 + y^2 & xr^6 \\ yr^2 & yr^4 & 3y^2 + x^2 & 2xy & yr^6 \end{bmatrix} \quad (19)$$

(2) *Partial derivative of the distortion coordinates from pixel coordinates:* Assume that the camera model is a pinhole camera model, and its perspective projection and affine transformation is expressed as a function $H_m(\cdot)$. For a square detector unit, the skew coefficient can be defaulted to 0. Then, the intrinsic parameters of the camera include the focal length of the x -axis f_x , the focal length of the y -axis f_y , and the principal point (u_0, v_0) . Let the intrinsic vector $M_i = [f_x, f_y, u_0, v_0]$; then, the pixel point p_{ij} can be described by the following equations:

$$p_{ij} = H_m(P_{ij}^D, M_i) = \begin{bmatrix} f_x x_{ij}^D + u_0 \\ f_y y_{ij}^D + v_0 \end{bmatrix} \quad (20)$$

where P_{ij}^D is (x_{ij}^D, y_{ij}^D) , and the partial derivative of the j -th pixel coordinates of frame i to the distortion coordinates can be described as

$$\frac{\partial p_{ij}}{\partial P_{ij}^D} = \begin{bmatrix} f_x & 0 \\ 0 & f_y \end{bmatrix} \quad (21)$$

According to (20), the partial derivative of the pixel coordinates to the intrinsic reference vector can be obtained:

$$\frac{\partial p_{ij}}{\partial M_i} = \begin{bmatrix} x_{ij}^D & 0 & 1 & 0 \\ 0 & y_{ij}^D & 0 & 1 \end{bmatrix} \quad (22)$$

Thus, we can solve the Jacobian matrix of frame i and use the above method to solve the Jacobian matrix of frames b and f . Note that the optimization of the intrinsic parameters of frames b and f is slightly different from that of frame i , as it is hidden in the process of normalizing the points P_{bj} and P_{fj} to pixel coordinates.

3.4. Pose-Only Re-Triangulation

Re-triangulation is used to further improve the completeness of the reconstruction by continuing to track points that were not successfully triangulated earlier due to camera pose inaccuracies. Unlike traditional RT, which uses pre-BA re-triangulation and post-BA re-triangulation to minimize error accumulation, this paper uses the normalization error of PA as the error metric and adjusts it only for the extrinsic parameters; then, the optimization problem of PORT can be described as

$$\arg \min_{\{R_i, t_i\}_{i=1 \dots m}} \sum_i^m \sum_j^n (e_{ij})^T e_{ij} \quad (23)$$

The process involves multiple iterations of triangulation, PA optimization, and filtering without changing the thresholds at the filtered points. PORT is a refinement module designed specifically for PA optimization. The traditional combination of bundle adjustment triangulation (BART) and PA may increase the optimization time due to the inconsistency of the optimization space, so PORT is used instead of BART in order to improve the consistency of the system, and secondly, the PORT module is richer in observation information due to the reintroduction of all the observation relationships, which is beneficial to the PA in choosing a more robust base frame. The advantage of using PORT is that it combines robustness while maintaining higher efficiency. The reason for optimizing only the extrinsic parameters is, firstly, to reduce the overfitting of the intrinsic parameters due to the error points. After the global initialization and global PA, the intrinsic parameters are already considered to be accurate. Refinement of the intrinsic parameters in the presence of a large number of uncertain 3D points may lead to the instability of the intrinsic parameters [33]. Secondly, we want the RT step to focus more on the accuracy and correctness of the 3D points, so only estimating the extrinsic parameters is better for visualizing the point characteristics to reduce point-scattering degrees. Finally, reducing the optimization parameters further improves the optimization efficiency. Since the RT process requires multiple iterations, fewer optimization parameters can significantly improve the optimization efficiency. After PORT, the reconstruction results can be obtained with richer information and a more complete structure.

4. Experiments

To demonstrate the optimization performance of our proposed POMAP, the experiments involved multiple datasets, including the publicly available Mip-NeRF360

dataset [19], Cambridge Landmarks dataset [18], and 1DSfM dataset [17], as well as the multi-focal-length (MFL) dataset captured and collected by our photography. These datasets cover a wide range of scenarios, ranging from scenes captured by a single camera in an ordered sequence to those captured by multiple cameras in an unordered manner, and from single object reconstruction to scene reconstruction. Specifically, we selected existing SOTA frameworks (COLMAP [1] and GLOMAP [3]) as benchmarks for comparison and performed extensive comparative analyses on the aforementioned datasets. Furthermore, we present ablations to study the behavior of the different components of our proposed system.

Metrics: We evaluate this in two downstream tasks, novel view synthesis and visual mapping and localization. For the Mip-NeRF360 dataset and our collection of multi-focal-length tourist data, both of which are small-scale object-centered reconstructions, this paper measures the quality of the reconstruction by rendering the image in the novel view synthesis that relies heavily on the accurate extrinsic and intrinsic parameters [34,35] where the quality of the rendered images measures the reconstruction. Specific image evaluation metrics include peak signal-to-noise ratio (PSNR), the structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS). The test method uses the popular 3DGS [34] algorithm. For evaluating the reconstruction quality of the Cambridge Landmarks dataset, the use of new viewpoint synthesis is very time consuming and computationally expensive, so we focus on the pose of the camera and directly compare each estimated pose with the real pose in the visual localization to judge the reconstruction effect. For the 1DSfM dataset, which is disorganized and features a large number of tourist photo datasets, we use the reprojection error and qualitative analysis, focusing on the comparison of solving the real and complex scenarios with optimization speed.

Implementation Details: For a fair comparison, the feature extraction and matching of COLMAP were consistently used as input to all methods in the novel view synthesis, while the HLOC tool [36] was used to ensure consistency in the visual localization. The running time of the experiments was recorded from the completion of feature matching, focusing on comparing the time taken for the optimization part. The hardware devices used in the experiments are a 13th Gen Intel(R) Core(TM) i9-13900HX CPU, and a DDR4 2200 MHz 64GB memory stick. The feature extraction and matching session was implemented using NVIDIA GeForce RTX 4060 Laptop GPU, and CUDA version 12.1. For both GLOMAP and COLMAP methods, the recommended default settings are used, and our method selects the g2o [37] and Ceres [38] optimization libraries in the optimization section and chooses a fixed parameters configuration based on the experimental conditions. Some of the experimental details are shown in the Table 1.

Table 1. Optimization methods and filtering criteria. The italics indicate whether an optimisation or filtering method was used. The filtering method indicates that the method is used to remove some bad matching pairs.

Method	Parameters	Maximum Iteration	Stopping Criteria
<i>Optimization:</i>			
Global PA	$\tau = 10^{-5}$	100	$\rho = 0$ or $\lambda = \text{NaN}$
PORT	—	5	$\Delta\text{Obs. points} < 0.05\%$
PA in PORT	$\tau = 10^{-5}$	50	$\rho = 0$ or $\lambda = \text{NaN}$
<i>Filtering Criteria:</i>			
Depth	—	—	$d \geq 0.0$
Angular	—	—	$\theta \geq 1.5^\circ$
Reprojection	—	—	$\epsilon \leq 4.0 \text{ px}$

4.1. Simulation Test Performance

In order to evaluate the convergence and robustness of the algorithm, we conducted simulated data experiments and synthetic noise experiments.

Convergence test: The simulation data experiments use 200 pan-generated cameras with arbitrary viewing angles, each with its own independent internal datum, and we randomly add 10% or 20% noise to the focal length, principal points, and distortion parameters without adding noise to the extrinsic parameters and feature points. Then, we test the PA with intrinsic optimization only, with each noise experiment being performed five times and averaged. The results of the experiment are shown in Figure 5. The error in the PA optimization we use decreases quickly after the first iteration, but the non-linearity due to distortion does not converge until six or seven iterations later. It still retains a small error, but the number of iterations required is much smaller than for the BA.

Robustness test: Synthetic noise experiments are used on the MLF dataset to test the robustness of the proposed optimization method in the presence of image noise. We synthesize perfect-image observations by projecting COLMAP triangulations onto the real camera, and then, for each observation, we add random Gaussian noise to the reprojection. The results are shown in Table 2. For perfect image observations with 0px noise, we reliably converge to the true value, underlining the effectiveness of the optimization. As the noise level increases to extreme values, the AUC (less than 5 per cent error in camera parameters) score gradually decreases, indicating the high robustness of our proposed optimization method.

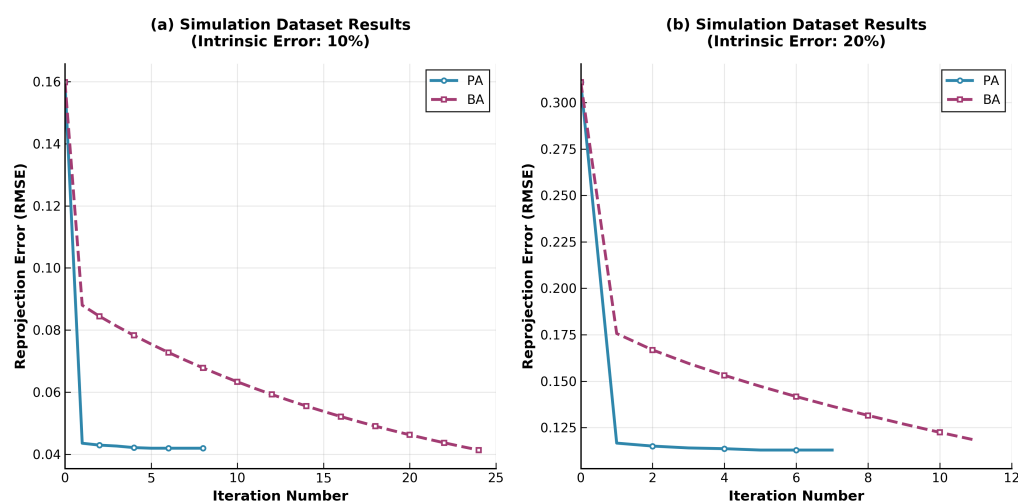


Figure 5. Experiments on the simulation of different intrinsic noise.

Table 2. Ablation on robustness of POMAP to different noise levels of points (%).

Scene	Noise Level (Pixels)							
	0	1	2	4	8	16	32	64
Table	100.00	99.49	99.49	98.75	97.88	96.60	94.44	87.12
Book	100.00	99.49	99.49	98.54	96.29	94.27	74.36	44.44
Cap	100.00	89.37	85.14	92.31	68.86	58.08	50.51	14.13
Luffy	100.00	99.49	99.38	98.90	97.59	95.04	88.80	68.69
Teapot	100.00	99.49	99.49	99.39	98.43	96.69	94.78	74.44
Flower	100.00	99.49	99.43	99.18	98.04	96.53	94.08	84.95
Pavilion	100.00	87.31	79.50	88.54	84.01	77.34	67.10	45.68
Dashufa	100.00	83.33	71.46	64.47	58.87	61.51	34.21	9.27
Average	100.00	94.68	91.67	92.51	87.50	75.76	74.79	53.59

4.2. Experiments on Base Frame Selection

To validate the effectiveness of the depth uncertainty-based base frame selection method, we performed a comparison between the simulated dataset and the real dataset, examining the baseline method from [12] that includes maximum parallax angle (PA-A) and depth averaging options (PA-D). The results are shown in Table 3. It can be seen that our proposed base frame selection method has stronger anti-interference ability and robustness than other methods. When the image noise reaches 10 pixels, it still maintains high accuracy. The other two methods have a rapidly increasing error due to noise. We analyze that this is because cameras with different focal lengths have various sensitivities to noise, and that cameras with larger focal lengths occupy a smaller world scale per pixel when photographing the same object, and are, therefore, less affected by random image noise. Our method can select a group of views with the least interference from noise, thereby reducing the uncertainty of depth estimation to some extent. Notably, our method still maintains high accuracy in the real dataset. This indicates that our proposed base frame selection algorithm has strong robustness and can provide better information for subsequent optimization.

Table 3. Performance comparison on simulated dataset (degrees/meter) and multi-focal-lengths dataset. As shown in the following tables, bolding indicates the best result for the indicator.

Method	Noise Level (Pixel)					MFL Dataset		
	0.0	1.5	3.0	5.0	10.0	SSIM	PSNR	LPIPS
PA-A	0.33/ 0.01	0.97/0.05	4.97/0.10	9.11/0.17	18.78/0.26	0.956	34.78	0.111
PA-D	0.36/0.02	1.69/0.10	0.91/0.11	1.75/0.14	5.26/0.17	0.949	34.33	0.118
Ours	0.25/0.02	0.36/0.03	0.35/0.03	0.68/0.05	0.69/0.03	0.958	35.84	0.110

4.3. Novel View Synthesis

Mip-NeRF360 comprises seven object-centric scenes with high-resolution images captured sequentially by a single camera. Since these data do not provide accurate real parameters of the camera, we also evaluate the reconstruction results in terms of the quality of the rendered images in the novel view synthesis. As shown in Table 4, our method is superior in PNSR, SSIM, and LPIPS metrics compared to the two existing SOTA methods. In reconstruction speed, our proposed method is almost 1.5 times faster than GLOMAP and 3 times faster than COLMAP. Overall, in the object-centered reconstruction task, our proposed method is basically in line with the SOTA in terms of accuracy with a slight improvement, while there is a very significant speed improvement.

Table 4. Rendering performance on Mip-NeRF360. * indicates data from the authors of the original paper [3].

	PSNR (dB) ↑			SSIM ↑			LPIPS ↓			Time (s) ↓		
	COLMAP *	GLOMAP *	Ours	COLMAP *	GLOMAP *	Ours	COLMAP	GLOMAP	Ours	COLMAP	GLOMAP	Ours
bicycle	23.15	23.13	23.12	0.5320	0.5310	0.7168	0.3677	0.3839	0.3068	127.1	83.2	56.8
bonsai	29.66	30.36	30.48	0.8960	0.9040	0.9338	0.2193	0.2107	0.2004	735.3	341.2	293.5
counter	26.81	26.72	27.99	0.8370	0.8350	0.8972	0.2292	0.2288	0.2112	736.9	164.2	85.1
garden	24.98	24.97	27.34	0.6530	0.6550	0.8444	0.1860	0.1890	0.1589	289.2	241.5	162.3
kitchen	29.23	29.35	29.84	0.8510	0.8550	0.9210	0.1503	0.1503	0.1388	595.6	467.3	336.9
room	29.14	29.41	31.07	0.8710	0.8760	0.9170	0.2356	0.2385	0.2170	258.3	122.0	76.4
stump	23.98	23.81	25.31	0.6020	0.5950	0.7540	0.3286	0.3294	0.3901	57.7	40.0	34.6
Average	26.71	26.82	27.88	0.7489	0.7501	0.8549	0.2452	0.2472	0.2319	400.0	208.5	149.4

The **multi-focal-lengths dataset** comprises high-resolution data we collected from multiple unordered shooting scenes, each shot with at least four cameras of different focal lengths. This dataset is used mainly to test the robustness of the optimization algorithm, including indoor and outdoor situations. The results tested on the new perspective synthesis task are shown in Table 5 and Figure 6. Our proposed method is faster in achieving the same accuracy. Compared to the best available result, we observe an improvement of 0.36 dB in PSNR, which indicates that our proposed method can cope with the complexity of multi-focal recordings and achieve a good joint optimization of intrinsic and extrinsic parameters. Moreover, the reconstruction speed of the whole system is nearly 1.5 times that of GLOMAP and more than 3 times that of COLMAP, since no optimization of 3D points is required.

Table 5. Rendering performance on multi-focal-length dataset.

	PSNR (dB) $\uparrow$			SSIM $\uparrow$			LPIPS $\downarrow$			Time (s) $\downarrow$		
	COLMAP	GLOMAP	Ours	COLMAP	GLOMAP	Ours	COLMAP	GLOMAP	Ours	COLMAP	GLOMAP	Ours
Table	32.85	32.91	33.00	0.9699	0.9708	0.9706	0.0531	0.0520	0.0522	4.62	1.14	1.08
Book	37.51	37.08	37.96	0.9711	0.9701	0.9718	0.1133	0.1148	0.1112	12.42	3.72	3.39
Cap	42.62	42.34	43.97	0.9908	0.9902	0.9916	0.0206	0.0227	0.0176	4.44	6.18	3.86
Luffy	37.48	38.42	38.80	0.9713	0.9742	0.9751	0.1022	0.0972	0.0933	9.24	2.93	1.93
Teapot	36.34	36.51	36.53	0.9723	0.9730	0.9727	0.0757	0.0748	0.0753	9.90	3.72	2.29
Flower	39.43	39.18	39.43	0.9619	0.9616	0.9631	0.1920	0.1947	0.1883	7.56	2.93	1.88
Pavilion	26.81	25.91	25.62	0.9319	0.9266	0.9190	0.0843	0.0845	0.0906	76.08	47.11	23.75
Dashufa	32.84	33.25	33.51	0.9591	0.9609	0.9619	0.0789	0.0782	0.0773	116.70	32.34	31.01
Average	35.74	35.70	36.10	0.9660	0.9659	0.9657	0.0900	0.0899	0.0882	30.12	12.51	8.65

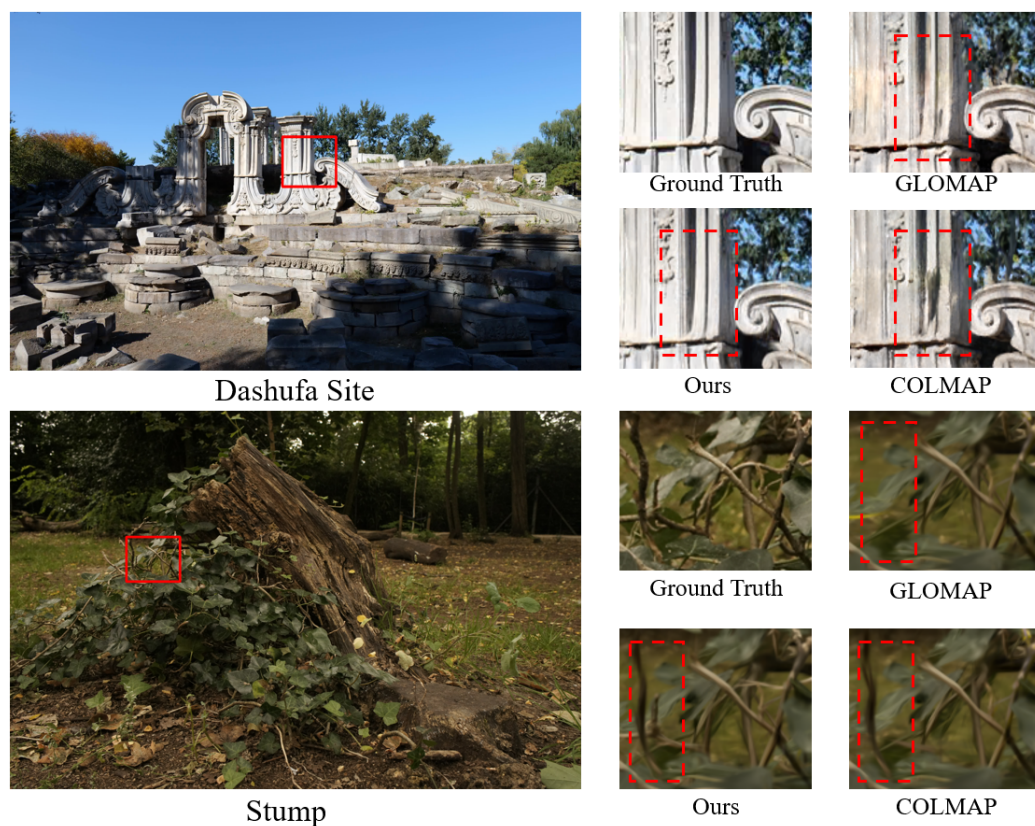


Figure 6. Qualitative results for novel view synthesis with 3DGS [34]. Visual comparisons across multiple scenes demonstrate that our method achieves superior camera parameters estimation. For example, the surface reconstruction of the Dashuifa site is closer to the real scene, and distant objects in Stump are clearly reconstructed, rather than being lost due to inaccurate parameters estimation as in GLOMAP and COLMAP.

4.4. Visual Mapping and Localization

The Cambridge Landmarks dataset is a visual location dataset taken outdoors around Cambridge University. A total of 26 gigabytes of data were selected from five complex scenarios for testing. Since these data have a real pose, we used the percentage of the number of cameras at three different error thresholds as well as the average distance error and angular error to assess the reconstruction quality. The results are summarized in Table 6 and show that our method achieves better results in estimating poses. With the position error less than 50 cm and angle error less than 5° , the number of correct camera poses estimated by our method is significantly higher than that of other methods, reaching 73.73%. In terms of average positional accuracy, our proposed method improves the accuracy by 42.8% compared to COLMAP, while it improves the tie angle accuracy by 28.6%, which is the highest camera position estimation accuracy among the tested methods. From the above experiments, it can be seen that our proposed method can effectively improve the accuracy of visual localization tasks in outdoor areas and remain robust in complex situations. Figure 7 shows a comparison of trajectory predictions for St Mary's Church.

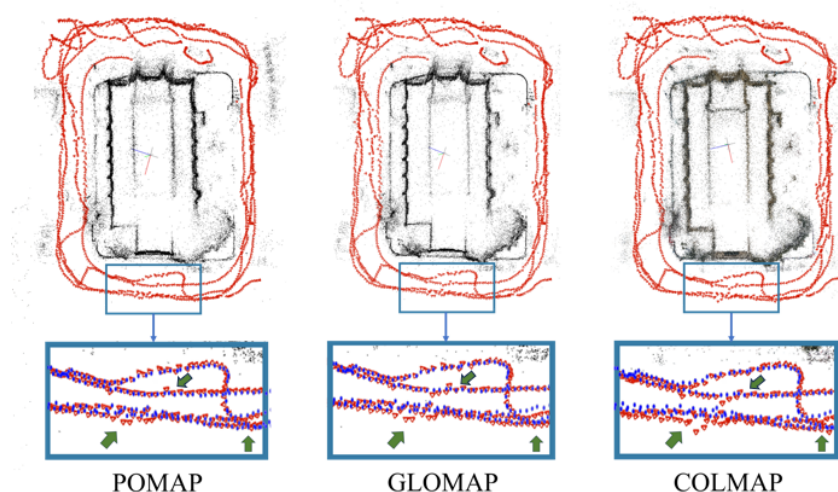


Figure 7. Schematic representation of some of the results in the visual relocalization task. Our proposed method achieves better results in position estimation, with red representing the estimated trajectory and blue representing the true value.

The 1DSfM dataset is a widely adopted benchmark for evaluating SfM algorithms, which are characterized by collections of unordered tourist photographic images with varying focal lengths and complex real-world scene variations. This dataset has no truth values, and we analyze it qualitatively in terms of the reconstruction effect, like with Figure 8, focusing on the speed of each process in the reconstruction process, and the results are shown in Table 7. We compare in detail the computation time of each step of GLOMAP vs. POMAP. COLMAP is an incremental SfM, and, thus, it is not appropriate to compare it with the global SfM on a case-by-case basis. Instead, we examine the overall time it takes. In Table 7, we start calculating the time after the matching phase, which includes the initialization, optimization, and RT phases. Since we use the same initialization method in [3], we focus on comparing the initialization and RT phases. From the results, it can be seen that POMAP is 24.16 times faster than COLMAP and 4.17 times faster than GLOMAP. In this particular instance, the global BA optimization accounts for a minimum of 50% of the overall process time, whereas POMAP with global PA is faster than GLOMAP with global BA, which is 55 times faster than BA optimization. This demonstrates that the proposed method can significantly enhance the optimization speed.

Table 6. Comprehensive evaluation on Cambridge Landmarks.

Scene	Method	Localization (%)			Pose Error		
		5 cm 5°	25 cm 2°	50 cm 5°	Position (m)	Angle (°)	Time (s)
Kings College	GLOMAP	3.77	28.28	48.69	0.531	0.689	394.9
	COLMAP	4.51	31.72	50.98	0.488	0.400	12,178.3
	Ours	4.84	38.52	72.38	0.320	0.721	263.5
Old Hospital	GLOMAP	0.34	4.80	43.35	0.561	1.335	358.6
	COLMAP	2.57	38.10	68.27	0.346	0.732	6421.4
	Ours	0.22	6.93	46.37	0.523	1.288	246.3
Shop Facade	GLOMAP	27.27	92.64	98.70	0.080	0.228	53.50
	COLMAP	27.71	95.24	98.70	0.082	0.254	229.1
	Ours	28.12	96.87	98.88	0.076	0.223	42.15
St. Marys Church	GLOMAP	5.99	49.90	80.97	0.251	0.268	371.2
	COLMAP	3.63	16.48	21.25	1.226	1.846	15,049.9
	Ours	7.33	51.45	83.86	0.241	0.259	305.2
Great Court	GLOMAP	0.78	23.43	50.98	0.493	0.212	506.6
	COLMAP	1.83	20.56	49.02	0.510	0.541	13,989.6
	Ours	1.83	35.90	67.17	0.357	0.203	486.5
Average	GLOMAP	7.63	39.81	64.54	0.383	0.546	337.0
	COLMAP	8.05	40.42	57.64	0.530	0.755	9573.7
	Ours	8.47	45.93	73.73	0.303	0.539	268.7

**Figure 8.** Example reconstructions from the proposed POMAP on 1DSfM datasets.**Table 7.** 3D Reconstruction performance comparison on 1DSfM dataset (time in seconds).

Scene	Images	COLMAP		GLOMAP		Ours		
		Total	Total	BA	BART	Total	PA	PART
Alamo	2915	12,935.7	1555.7	1070.8	174.5	224.0	34.1	134.6
Ellis Island	2587	8820.9	722.9	479.7	119.2	138.8	4.4	117.6
Gendarmenmarkt	1463	13,771.6	1577.2	663.4	719.2	354.7	9.7	288.3
Montreal Notre Dame	2298	12,679.1	584.1	411.7	65.7	96.5	15.8	42.2
Notre Dame	1430	35,619.4	1856.5	926.9	769.0	661.0	75.6	376.3
NYC Library	2550	9215.1	576.3	419.7	91.1	209.3	6.8	78.3
Piazza del Popolo	2377	11,448.0	2780.8	1951.8	179.1	222.1	13.7	115.1
Piccadilly	7347	37,352.1	6961.2	5139.0	1559.1	2016.4	194.6	1484.8
Roman Forum	2365	12,242.4	1404.5	666.2	598.8	453.3	27.8	271.6
Tower of London	1576	4709.0	645.2	453.4	46.1	388.5	13.5	293.3
Trafalgar Square	15,978	87,083.2	25,377.4	21,175.6	2369.8	6180.0	182.7	5434.8
Union Square	5961	12,205.6	1231.7	850.6	220.1	108.8	6.1	72.3
Vienna Cathedral	6288	24,791.9	2024.5	1232.9	261.0	414.1	32.4	214.3
Yorkminster	3452	15,032.9	4162.8	2601.0	833.1	861.2	70.0	361.3
Average	4185	21279.1	3675.8	2717.3	571.8	880.6	49.1	663.2

4.5. Ablation Study

In order to validate the effectiveness of the proposed system, an experimental study was conducted on the MFL dataset. This study involved the comparison of the quality of novel view synthesis for 11 different configurations under the same initialization conditions. These conditions included no optimization and an RT module (No optimization), the inclusion of PA with intrinsic optimization only (PA (intrinsic)), the inclusion of PA

with extrinsic optimization only (PA (pose)), the optimization of complete PA (PA) and BA (BA), PA with intrinsic optimization only, or PA with extrinsic optimization only in combination with PORT PA optimization combined with PORT/BART, and BA optimization combined with PORT/BART. The results are shown in Table 8. Overall, the combination of PA intrinsic and extrinsic optimization with PORT is most effective when the camera parameters are unknown. From an RT perspective, PORT outperforms BART, achieving a 0.76 dB improvement in PSNR, as well as 3.2% and 3.5% improvements in SSIM and LPIPS, respectively. In terms of optimization, when the focal length is unknown, optimizing the intrinsic is essential. The PNSR obtained from PA with intrinsic optimization is 1.09 dB higher than that from PA with extrinsic optimization alone. At the same time, PA with intrinsic optimization alone is effective and improves the PNSR by 0.24 dB compared to no optimization. Notably, we find that using PA/BA in combination with BART/PORT increases computation time. We speculate this to be due to the variability in the optimization space between PA and BA.

Table 8. Ablation studies on multi-focal-lengths dataset.

Method	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	Time (s) ↓
No optimization	33.15	0.9399	0.1441	–
PA (intrinsic)	33.39	0.9381	0.1615	3.17
PA (pose)	33.69	0.9404	0.1403	4.69
PA	34.78	0.9502	0.1214	6.74
BA	34.38	0.9420	0.1220	10.29
BA + BART	34.75	0.9497	0.1232	12.51
PA + BART	34.32	0.9488	0.1245	11.46
BA + PORT	34.36	0.9487	0.1227	16.07
PA (intrinsic) + PORT	34.54	0.9469	0.1289	12.27
PA (pose) + PORT	34.38	0.9496	0.1250	8.75
POMAP (PA + PORT)	35.08	0.9519	0.1201	8.65

5. Discussion

Our proposed POMAP achieves relatively good results in both novel view synthesis and visual localization tasks, with an accuracy almost comparable to that of current methods [1,3] as well as speed improvement, especially in reconstruction tasks facing large-scale datasets. It was evident that the proposed PA with joint intrinsic and extrinsic parameters optimization was capable of significantly reducing the time required for the optimization step. However, as the size of the input image increased, the time required for the feature extraction and matching part of the process became non-negligible. This may have a detrimental effect on the reconstruction speed gain brought about by the improvement of the optimization efficiency. Therefore, it would be advantageous to consider using the faster feature extraction and matching method to further enhance the reconstruction efficiency, for example, using methods such as Xfeat [39] with LightGlue [40] or Efficient LoFTR [41] with semi-dense feature matching.

6. Conclusions

This paper proposes a novel global SfM based on PA with joint intrinsic and extrinsic parameters optimization, addressing diverse scenarios while enhancing reconstruction speed and accuracy. Our global PA strategy jointly refines camera parameters, effectively handling challenges such as unknown-focal-length and multi-focal configurations. By avoiding 3D point optimization limitations, it significantly improves computational efficiency and system robustness. Additionally, our PORT mechanism enhances reconstruction quality and processing speed. Extensive experiments have shown that our framework is

faster and achieves an accuracy similar to that of existing SfM baselines. For example, our PA optimization method is 55 times faster than the BA optimization method on the 1DSfM dataset, and the proposed system is 4 times faster than GLOMAP and 24 times faster than COLMAP. We believe this work significantly advances reconstruction under cameras with unknown focal lengths, benefiting downstream tasks such as novel view synthesis and visual localization.

Author Contributions: Conceptualization, X.T., Y.G. and Z.T.; Methodology, X.T. and Y.G.; Investigation, X.T. and Z.T.; Resources, M.L., X.C. and D.H.; Supervision, D.H.; Writing—original draft, X.T.; Writing—review and editing, Z.T. and X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The Science and Technology Innovation Program of Hunan Province (Grant No. 2024QK2006).

Data Availability Statement: The original contributions presented in the study are included in the article, and further inquiries can be directed to the corresponding author.

Acknowledgments: The authors of this paper are very grateful for the help provided by the following experts and personnel: Zhang Jinpu provided valuable comments on the structure of the paper for touching up and plotting, and Lian Xiangkai provided valuable information on the mathematical theory and plotting of the paper, which is hereby acknowledged.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Schönberger, J.L.; Frahm, J.M. Structure-from-Motion Revisited. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4104–4113.
- Agarwal, S.; Snavely, N.; Simon, I.; Seitz, S.M.; Szeliski, R. Building Rome in a day. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision (ICCV), Kyoto, Japan, 29 September–2 October 2009; pp. 72–79.
- Pan, L.; Bar’ath, D.B.; Pollefeys, M.; Schönberger, J.L. Global Structure-from-Motion Revisited. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024.
- Agarwal, S.; Snavely, N.; Seitz, S.M.; Szeliski, R. Bundle Adjustment in the Large. In Proceedings of the European Conference on Computer Vision (ECCV), Heraklion, Greece, 5–11 September 2010.
- Wu, C. Towards Linear-Time Incremental Structure from Motion. In Proceedings of the International Conference on 3D Vision (3DV), Seattle, WA, USA, 29 June–1 July 2013; pp. 127–134.
- Zhu, S.; Zhang, R.; Zhou, L.; Shen, T.; Fang, T.; Tan, P.; Quan, L. Very Large-Scale Global SfM by Distributed Motion Averaging. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4568–4577.
- Triggs, B.; McLauchlan, P.F.; Hartley, R.I.; Fitzgibbon, A.W. Bundle adjustment—A modern synthesis. In Proceedings of the International Workshop on Vision Algorithms, Corfu, Greece, 21–22 September 2000.
- Cui, H.; Shen, S.; Gao, X.; Hu, Z. Batched Incremental Structure-from-Motion. In Proceedings of the International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 205–214.
- Moulon, P.; Monasse, P.; Marlet, R. Global Fusion of Relative Motions for Robust, Accurate and Scalable Structure from Motion. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 3248–3255.
- Cai, Q.; Zhang, L.; Wu, Y.; Yu, W.; Hu, D. A Pose-Only Solution to Visual Reconstruction and Navigation. *IEEE Trans. Pattern Analysis Mach. Intell. (TPAMI)* **2023**, *45*, 73–86. [[CrossRef](#)] [[PubMed](#)]
- Cai, Q.; Wu, Y.; Zhang, L.; Zhang, P. Equivalent Constraints for Two-View Geometry: Pose Solution/Pure Rotation Identification and 3D Reconstruction. *Int. J. Comput. Vis. (IJCV)* **2018**, *127*, 163–180. [[CrossRef](#)]
- Ge, Y.; Zhang, L.; Wu, Y.; Hu, D. PIPO-SLAM: Lightweight Visual-Inertial SLAM With Preintegration Merging Theory and Pose-Only Descriptions of Multiple View Geometry. *IEEE Trans. Robot. (TRO)* **2024**, *40*, 2046–2059. [[CrossRef](#)]
- Du, X.; Ji, C.; Zhang, L.; Luo, X.; Zhang, H.; Wang, M.; Wu, W.; Mao, J. SP-VIO: Robust and Efficient Filter-Based Visual Inertial Odometry with State Transformation Model and Pose-Only Visual Description. *arXiv* **2024**, arXiv:2411.07551.

14. Du, X.; Zhang, L.; Liu, R.; Wang, M.; Wu, W.; Mao, J. PO-MSCKF: An Efficient Visual-Inertial Odometry by Reconstructing the Multi-State Constrained Kalman Filter with the Pose-only Theory. *arXiv* **2024**, arXiv:2407.01888.
15. Tang, H.; Zhang, T.; Wang, L.; Wang, G.; Niu, X. PO-VINS: An Efficient and Robust Pose-Only Visual-Inertial State Estimator With LiDAR Enhancement. *arXiv* **2024**, arXiv:2305.12644.
16. Wang, L.; Tang, H.; Zhang, T.; Wang, Y.; Zhang, Q.; Niu, X. PO-KF: A Pose-Only Representation-Based Kalman Filter for Visual Inertial Odometry. *IEEE Internet Things J.* **2025**, *12*, 14856–14875. [[CrossRef](#)]
17. Wilson, K.; Snavely, N. Robust Global Translations with 1DSfM. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014.
18. Kendall, A.; Grimes, M.; Cipolla, R. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2938–2946.
19. Barron, J.T.; Mildenhall, B.; Verbin, D.; Srinivasan, P.P.; Hedman, P. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5460–5469.
20. Özyesil, O.; Voroninski, V.; Basri, R.; Singer, A. A survey of structure from motion. *Acta Numer.* **2017**, *26*, 305–364. [[CrossRef](#)]
21. Jiang, S.; Jiang, C.; Jiang, W. Efficient structure from motion for large-scale UAV images: A review and a comparison of SfM tools. *ISPRS J. Photogramm. Remote Sens. (JPRS)* **2020**, *167*, 230–251. [[CrossRef](#)]
22. Zhang, J.; Boutin, M.; Aliaga, D.G. Robust Bundle Adjustment for Structure from Motion. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2006; pp. 2185–2188.
23. Byröd, M.; Åström, K. Bundle Adjustment using Conjugate Gradients with Multiscale Preconditioning. In Proceedings of the British Machine Vision Conference (BMVC), London, UK, 7–10 September 2009.
24. Konolige, K. Sparse Sparse Bundle Adjustment. In Proceedings of the British Machine Vision Conference (BMVC), Wales, UK, 30 August–2 September 2010.
25. Sweeney, C.; Sattler, T.; Höllerer, T.; Turk, M.; Pollefeys, M. Optimizing the Viewing Graph for Structure-from-Motion. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 801–809.
26. Zhang, R.; Zhu, S.; Fang, T.; Quan, L. Distributed Very Large Scale Bundle Adjustment by Global Camera Consensus. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 29–38.
27. Zhuang, B.; Cheong, L.F.; Lee, G.H. Baseline Desensitizing in Translation Averaging. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4539–4547.
28. Cui, H.; Gao, X.; Shen, S.; Hu, Z. HSfM: Hybrid Structure-from-Motion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2393–2402.
29. Chen, Y.; Yu, Z.; Song, S.; Yu, T.; Li, J.; Lee, G.H. AdaSfM: From Coarse Global to Fine Incremental Adaptive Structure from Motion. In Proceedings of the IEEE International Conference on Robotics & Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 2054–2061.
30. Cai, R.; Tung, J.; Wang, Q.; Averbuch-Elor, H.; Hariharan, B.; Snavely, N. Doppelgangers: Learning to Disambiguate Images of Similar Structures. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; pp. 34–44.
31. Chatterjee, A.; Govindu, V.M. Efficient and Robust Large-Scale Rotation Averaging. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 521–528.
32. Pizzoli, M.; Forster, C.; Scaramuzza, D. REMODE: Probabilistic, monocular dense reconstruction in real time. In Proceedings of the IEEE International Conference on Robotics & Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 2609–2616.
33. Olsson, C.; Eriksson, A.; Hartley, R. Outlier removal using duality. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 1450–1457.
34. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* **2023**, *42*. [[CrossRef](#)]
35. Zwicker, M.; Pfister, H.; van Baar, J.; Gross, M. EWA splatting. *IEEE Trans. Vis. Comput. Graph.* **2002**, *8*, 223–238. [[CrossRef](#)]
36. Sarlin, P.E.; Cadena, C.; Siegwart, R.; Dymczyk, M. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
37. Kümmerle, R.; Grisetti, G.; Strasdat, H.; Konolige, K.; Burgard, W. G2o: A general framework for graph optimization. In Proceedings of the IEEE International Conference on Robotics & Automation (ICRA), Shanghai, China, 9–13 May 2011; pp. 3607–3613.
38. Agarwal, S.; Mierle, K.; The Ceres Solver Team. *Ceres Solver*, version 2.2; Ceres: Boston, MA, USA, 2023; Apache License 2.0.
39. Potje, G.; Cadar, F.; Araujo, A.; Martins, R.; Nascimento, E.R. XFeat: Accelerated Features for Lightweight Image Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 2682–2691.

40. Lindenberger, P.; Sarlin, P.E.; Pollefeys, M. LightGlue: Local Feature Matching at Light Speed. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 17581–17592.
41. Wang, Y.; He, X.; Peng, S.; Tan, D.; Zhou, X. Efficient LoFTR: Semi-Dense Local Feature Matching with Sparse-Like Speed. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 21666–21675.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.