

Article

Efficient Structure from Motion for Large-Size Videos from an Open Outdoor UAV Dataset

Ruilin Xiang , Jiagang Chen and Shunping Ji * 

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; xiang_ruilin@whu.edu.cn (R.X.); chenjiagang2015@whu.edu.cn (J.C.)

* Correspondence: jishunping@whu.edu.cn

Abstract: Modern UAVs (unmanned aerial vehicles) equipped with video cameras can provide large-scale high-resolution video data. This poses significant challenges for structure from motion (SfM) and simultaneous localization and mapping (SLAM) algorithms, as most of them are developed for relatively small-scale and low-resolution scenes. In this paper, we present a video-based SfM method specifically designed for high-resolution large-size UAV videos. Despite the wide range of applications for SfM, performing mainstream SfM methods on such videos poses challenges due to their high computational cost. Our method consists of three main steps. Firstly, we employ a visual SLAM (VSLAM) system to efficiently extract keyframes, keypoints, initial camera poses, and sparse structures from downsampled videos. Next, we propose a novel two-step keypoint adjustment method. Instead of matching new points in the original videos, our method effectively and efficiently adjusts the existing keypoints at the original scale. Finally, we refine the poses and structures using a rotation-averaging constrained global bundle adjustment (BA) technique, incorporating the adjusted keypoints. To enrich the resources available for SLAM or SfM studies, we provide a large-size (3840 × 2160) outdoor video dataset with millimeter-level-accuracy ground control points, which supplements the current relatively low-resolution video datasets. Experiments demonstrate that, compared with other SLAM or SfM methods, our method achieves an average efficiency improvement of 100% on our collected dataset and 45% on the EuRoc dataset. Our method also demonstrates superior localization accuracy when compared with state-of-the-art SLAM or SfM methods.

Keywords: structure from motion; SLAM; unmanned aerial vehicle; large-size videos; keypoint adjustment



Citation: Xiang, R.; Chen, J.; Ji, S. Efficient Structure from Motion for Large-Size Videos from an Open Outdoor UAV Dataset. *Sensors* **2024**, *24*, 3039. <https://doi.org/10.3390/s24103039>

Academic Editor: Hui Kong

Received: 3 April 2024

Revised: 29 April 2024

Accepted: 8 May 2024

Published: 10 May 2024



Copyright: © 2024 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

Modern unmanned aerial vehicles (UAVs) equipped with cameras have become crucial in several fields, such as surveying and mapping, geographic information systems (GIS), and digital city modeling. To achieve accurate localization and create 3D representations of real-world scenes, techniques like image or video-based structure from motion (SfM) and visual simultaneous localization and mapping (VSLAM) are utilized [1–10]. However, it is important to note that there is a relatively limited amount of research on large-size video-based SfM specifically designed for outdoor UAVs. On the one hand, a mainstream UAV camera has already reached a resolution up to 20 megapixels, thus providing more detailed information for all kinds of applications. However, the widely-used video datasets [11–15] provide a resolution below 1 megapixel. On the other hand, there is limited research on how to combine SfM and VSLAM for large-size video-based localization. For large-size videos, current video-based SfM methods extract keyframes from videos usually based on simple empirical rules, for example, Kurniawan et al. [16] performed SfM on the keyframes extracted from videos simply according to the overlap rate of images, instead of a more sophisticated VSLAM method, to achieve 3D terrain reconstruction. In fact, VSLAM designed for continuous image processing inherently suits video data better. To process large-size videos in real-time, SLAM systems estimate camera poses and build maps on

downsampled images, which is more efficient but results in lower localization accuracy than SfM methods. Some researchers [17,18] have attempted to utilize VSLAM to assist the SfM method for reconstruction. However, these methods only utilize estimated camera poses from a SLAM system. In fact, reusing feature extraction, matching, and keyframe covisibility graph results from SLAM can significantly reduce the computational cost for large-size video processing.

In this paper, we propose an efficient SfM pipeline designed to process high-resolution aerial videos. Additionally, we introduce a new outdoor UAV video dataset comprising images with a resolution of 3840×2160 pixels. Our approach maximizes the usefulness of initial outcomes provided by a speedy VSLAM system and incorporates a constrained bundle adjustment (BA) as a singular backend refinement step. The pipeline unfolds in the following steps. Initially, we subject the downsampled video data to a VSLAM system, which serves multiple purposes, including selecting keyframes and keypoints, as well as establishing preliminary camera poses and 3D scene structures. Secondly, to optimize the efficiency of the pipeline, we employ a coarse-to-fine two-step keypoint adjustment (TS-KA) method with rotation invariants, which adjusts the positions of matched keypoints projected onto the original high-resolution images instead of re-matching new feature points. This adjustment process begins by roughly aligning keypoint positions using normalized cross-correlation (NCC) [19]. Following the rough alignment, we apply direct image alignment [20] within a learned dense feature space to further refine matched points up to sub-pixel accuracy. Finally, the global bundle adjustment takes the initial camera poses from the VSLAM system as inputs and integrates a rotation averaging strategy [21]. Optionally, ground control point (GCP) constraints can be included to retain high-accuracy poses and 3D scene points at a centimeter-level precision.

The contributions of this paper are summarized as follows. (1) Efficient SfM pipeline. We propose an efficient pipeline specifically designed to process large-size aerial videos. By leveraging the strengths of a rapid VSLAM system and incorporating refined adjustment steps, our pipeline achieves impressive accuracy and efficiency in pose localization of video sequences. (2) Two-step keypoint adjustment (TS-KA) strategy. The novel strategy refines the positions of keypoints matched in downsampled images up to sub-pixel accuracy on the original high-resolution images. (3) High-resolution UAV video dataset. We provide a high-resolution UAV video dataset and supply high-accuracy GCPs to facilitate evaluation. This dataset fills a gap in the current availability of outdoor high-resolution video datasets for SLAM or SfM research.

2. Related Work

2.1. Unstructured, Sparsely Sampled Collection

Early works laid the foundation for internet photo collections [22]. Inspired by these works, reconstruction systems for increasingly high-resolution photo collections have been developed [15,23]. These methods can be classified into incremental SfM, global SfM, and hybrid SfM, based on the manner in which camera poses are estimated. Currently available open-source incremental SfM algorithms, such as Bundler [1], VisualSfM [2], and COLMAP [3,24,25], provide a solid foundation for SfM research. Mainstream global SfM methods [4,5,26,27] estimate all camera poses and perform a global BA to refine the camera poses and reconstruction scene, resulting in better scalability and efficiency. Rotation averaging [28–31] estimates all camera rotations from pairwise relative rotations, while translation averaging [32–34] calculates the translation of each camera pose. However, the latter may fail to estimate correct camera centers when the camera moves collinearly [4,5].

Indeed, these methods focusing on unordered, sparsely sampled images face challenges when dealing with coherent, densely sampled data. This difficulty arises from frame-wise matching and triangulation with very short parallax, which can result in high computation loads and unreliable geometric structures.

2.2. Coherent, Densely Sampled Collection

This type of study addresses continuous feature tracking and mapping on coherent, densely sampled image sequences. Specifically, VSLAM methods have been developed to estimate camera trajectories and reconstruct scene structures from video streams in real time [7–10,35]. However, these methods often prioritize speed and, as a result, face limitations when processing large-size high-resolution images. This restriction hampers their ability to produce fine-grained high-quality reconstructions.

Over the years, SfM methods have also been developed specifically for densely sampled image sequences or videos. For example, Shum et al. [36] introduced the concept of “virtual keyframes” in a hierarchical SfM approach to enhance efficiency. Resch et al. [37] proposed multiple SfM techniques based on the KLT tracker and linear camera pose estimation [38] for large-scale videos. Leotta et al. [39] accelerated feature tracking for aerial videos by exploiting temporal continuity and planarity of the ground. More recently, a deep learning-based approach [40] was proposed to select appropriate keyframes for videos. To resolve ambiguity arising from repetitive structures, Wang et al. [41] proposed a track-community structure to segment the scene. Gong et al. [42] proposed to disambiguate scenes in SfM by prioritizing pose consistency over feature consistency. However, it should be noted that these methods may rely on fixed camera calibration and could encounter significant drift issues in scenes without a loop. Different from these methods, our work proposes a hybrid SfM solution that combines the advantages of global SfM and feature-based VSLAM methods.

2.3. Keypoint Adjustment

Recently, there has been an increased focus on developing local search-based methods to enhance the efficiency and accuracy of keypoint matching. These methods employ both handcrafted [43,44] and learned features [45–47] to establish more accurate correspondences between keypoints. For example, Taira et al. [48] presented a method that achieves dense correspondence through a coarse-to-fine matching process using VGG-16 [49]. Li et al. [50] employed a dual-resolution approach to achieve reliable and accurate correspondences. Zhou et al. [51] proposed a detect-to-refine method, where initial matches are refined by regressing pixel-level matches in local regions. However, it should be noted that these methods [48,50,51] are primarily optimized for stereo pairs and may not be directly applied for multiple views.

In order to enhance the quality of multi-view keypoints for downstream tasks like SfM, Dusmanu et al. [52] incorporated a geometric cost with optical flow. However, this method has limitations in terms of accuracy and scalability for large scenes. Lindenberger et al. [20] addressed the alignment of keypoints by utilizing feature-metric representation to jointly adjust feature matches across thousands of images. However, this method suffers from a limited range of adjustment and may become less accurate when dealing with images exhibiting significant viewpoint changes. To address these challenges, we introduce an efficient two-step matching approach that takes into account errors in initial matching at a lower resolution and effectively handles large viewpoint changes.

3. Proposed Method

3.1. System Overview

We introduce a novel SfM pipeline that efficiently selects appropriate keyframes and calculates camera poses by utilizing rich information from high-resolution, high-frame-rate videos. As depicted in Figure 1, our proposed pipeline comprises three main steps.

In the first step, we begin by downsampling the original high-resolution video to improve efficiency. Then, we estimate the initial camera poses and select keyframes using visual odometry on the downsampled video. The output of this step includes three components: a set of N keyframes denoted as $\mathcal{I} = \{I_1, \dots, I_N\}$ along with their poses $\mathbf{T}_{I_i}^W = \{(\mathbf{R}_{I_i}^W, \mathbf{t}_{I_i}^W) \in \text{SE}(3)\}$, M sparse world points $\mathcal{P} = \{\mathbf{P}_l^W \in \mathbb{R}^3\}$ paired with their

corresponding 2D keypoints $\{\mathbf{p}_u\}$, and a view graph (VG) $\mathcal{G} = \{V, E\}$ with absolute rotation (\mathbf{R}_i^W) as vertices and relative rotation ($\mathbf{R}_{ij}^{I_j}$) as edges for image pairs (I_i, I_j) .

In the second step of the pipeline, we upsample the keypoints obtained from visual odometry to match the original resolution of the keyframes. To achieve sub-pixel accuracy, we employ a two-step keypoint adjustment method called TS-KA. TS-KA refines the position of the upsampled keypoints in a coarse-to-fine strategy. Initially, we utilize the NCC algorithm [19] to roughly adjust the keypoint positions considering view angle changes. Then, we introduce feature-metric optimization for further refinement.

Moving on to the third step, we perform rotation averaging on the VG obtained in the first step. This helps us estimate the global rotation of all keyframes. The obtained global rotation will be integrated into BA as a regularization measure, reducing cumulative errors. Then, we refine the camera intrinsic parameters, keyframe poses, and sparse point cloud coordinates through rotation-averaged BA. To handle outliers, we incorporate a reprojection error threshold to filter them out. Additionally, we enhance trajectory accuracy at the centimeter-level by including GCPs in the BA process.

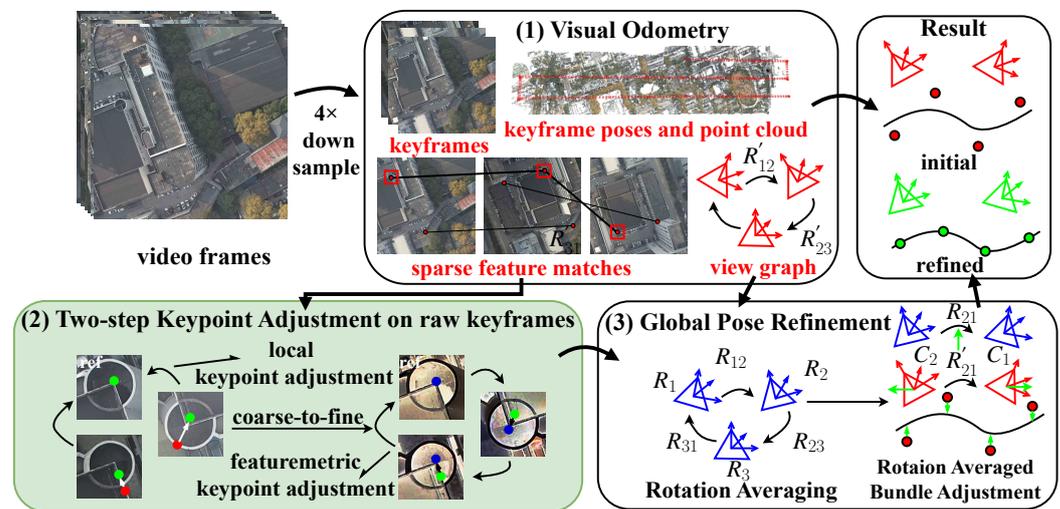


Figure 1. System overview. (1) The initial scene structure is obtained from $4\times$ downsampled video using visual odometry. (2) Keypoints are refined on full-resolution keyframes by a two-step keypoint adjustment method. **red**: original matching points; **green**: matching points after coarse keypoint adjustment; **blue**: matching points after sub-pixel refinement. (3) Global rotation is obtained by rotation averaging, and scene structure is finally refined using rotation-averaged bundle adjustment.

3.2. Initial Pose Estimation

We utilize visual odometry for both keyframe selection and initial scene reconstruction. Given the high-resolution aerial video used in this study, we initially downsample the raw video by a factor of 4. This downsampling step ensures real-time initial camera trajectory estimation. Visual odometry involves the detection and tracking of distinctive features in consecutive camera frames. By matching these features, it estimates the camera's relative motion and selects keyframes that represent significant viewpoints. In our proposed pipeline, we leverage the widely used OpenVSLAM for the initial camera pose estimation. OpenVSLAM includes three modules: tracking, local mapping and loop closing. The tracking module is primarily responsible for estimating the camera's pose in real time. This module estimates the camera's position and orientation by extracting and tracking feature points from consecutive video frames. It also determines whether to incorporate the current frame as a keyframe into the map based on specific rules. The local mapping module focuses on building and maintaining the map. It uses feature points from keyframes, creates new map points via triangulation, and performs local optimization of the map's structure to enhance its accuracy. The loop closing module detects and handles

loop closures. By recognizing revisited images and aligning them with previous map data, the module corrects cumulative navigational errors. More details can be found in [8].

3.3. Two-Step Keypoint Adjustment

As the matched points in VSLAM are obtained from $4\times$ downsampled images with limited precision in point coordinates, it is necessary to adjust the keypoint coordinates to sub-pixel accuracy on the original resolution. Inspired by the work of the keypoint adjustment method in Pixel-SfM [20], we propose a simple yet powerful two-step keypoint adjustment approach TS-KA.

3.3.1. Coarse Keypoint Adjustment

The coarse keypoint adjustment in Figure 2 aims to refine the keypoints within the given search area by utilizing a rotation-invariant similarity measure. This adjustment allows for accurate keypoint refinement over a large range. The first step involves determining the reference keypoint, \mathbf{p}_r , within a track, $\{\mathbf{p}_i\}$ ($i = 1, \dots, N$). A track refers to a collection of N keypoints corresponding to the same 3D world point. We calculate the accumulated matching scores for each keypoint in the track, and \mathbf{p}_r is selected as the reference keypoint with the highest accumulated matching score. The remaining points in the track are then adjusted and matched to \mathbf{p}_r . Second, we assign a consistent orientation to each keypoint based on local image characteristics; the keypoint can be represented relative to this orientation, thus ensuring invariance to image rotation. For a pixel, \mathbf{p} , within the search region of a keypoint, \mathbf{p}_i , we compute the gray centroid, \mathbf{p}_c , of its NCC window. The NCC window is a circular window with a radius of 15 pixels here. To achieve orientation invariance, we rotate the NCC window based on the angle between $\mathbf{p}\mathbf{p}_c$ and $\mathbf{p}_r\mathbf{p}_{rc}$, where \mathbf{p}_{rc} represents the gray centroid of \mathbf{p}_r . Finally, we obtain the best matching points through NCC matching.

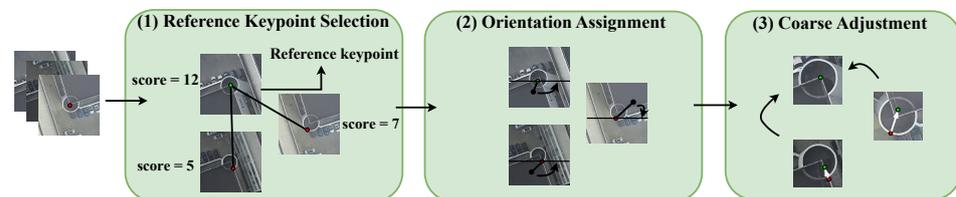


Figure 2. Coarse keypoint adjustment. (1) The reference keypoint is selected as the one having the highest score. (2) Each keypoint is assigned with a consistent orientation. (3) Best matching points (green ones) are obtained using NCC.

3.3.2. Sub-Pixel Refinement

The coarse keypoint adjustment primarily achieves feature matching accuracy at the pixel level. To meet the accuracy requirements of various downstream tasks, it is often necessary to refine keypoints to sub-pixel accuracy. For this purpose, we introduce the feature keypoint adjustment (FKA) method [20]. We first extract a dense feature map of 16×16 patches centered on the keypoints by S2DNet [53], and then we treat the refinement of the keypoints, $M(l)$, in a track belonging to the same landmark, l , as an energy minimization problem, as follows:

$$E_{FKA_l} = \sum_{(a,b) \in M(l)} \omega_{ab} \| \mathbf{F}_{i(a)}[\mathbf{p}_a] - \mathbf{F}_{j(b)}[\mathbf{p}_b] \|_{\gamma} \quad (1)$$

where ω_{ab} represents the confidence between matched points \mathbf{p}_a and \mathbf{p}_b , according to the similarity of the local features. $\mathbf{F}[\cdot]$ represents the feature map.

It should be noted that the original FKA [20] lacks a coarse adjustment step, which can result in numerous incorrect adjustments. To address this limitation, we have incorporated a coarse adjustment step in our approach. We also add a constraint $\| \mathbf{p}_{best}^{cf} - \mathbf{p}_{best}^c \| < K$,

where \mathbf{p}_{best}^{cf} denotes the position of the keypoint after fine adjustment, and K is set to be lower than the radius, r .

3.4. Global Pose Refinement

The trajectory derived from visual odometry often suffers from the drift accumulation problem, leading to significant deviations from the true trajectory. Inspired by [21], in order to enhance the precision of camera pose estimation, we incorporate the global camera pose obtained through rotation averaging as a regularizer into the BA process. Furthermore, when available, we include the GCPs in the BA equations.

3.4.1. Rotation Averaging

Rotation averaging (RA) is a method utilized for estimating global camera poses by simultaneously considering pairwise relative poses. The global rotation is computed by minimizing the cost function:

$$\min_{\mathbf{R}_{I_1}^W, \dots, \mathbf{R}_{I_N}^W} \sum_{(I_i, I_j) \in E} d^2 \left(\mathbf{R}_{I_i}^{I_j}, \mathbf{R}_{I_j}^{W^T} \mathbf{R}_{I_i}^W \right) \quad (2)$$

where d^2 represents the Euclidean norm. However, RA [21] is sensitive to outliers, which may result in inaccurate estimates.

Before performing RA, it is necessary to construct a view graph with edges being pairs of matched images. To avoid starting image matching from scratch, we leverage the co-visibility data derived from VO, as outlined in Section 3.2, transforming it into a view graph with candidate edges. Then, we assign higher weight values to edges with more visible points and a more uniform distribution of matches. Concurrently, we prune edges within a view graph under the following conditions: (1) the number of matches falls short of the predetermined threshold, N_m ; and (2) the angular error for a given edge, denoted as $\mathbf{R}_{I_i}^{I_j}$, is below a specified threshold, σ , as delineated by the following formula:

$$d^2 \left(\mathbf{R}_{I_i}^{I_j} \mathbf{R}_{I_k}^{I_j}, \mathbf{I}_3 \right) \leq \sigma \quad (3)$$

\mathbf{I}_3 represents the 3×3 identity matrix. σ is set as 0.01.

3.4.2. Rotation Averaged Bundle Adjustment

Rotation averaged BA is conducted to optimize camera poses, 3D points, and camera intrinsics. Since the observations are independent, the trajectory estimated by RA does not accumulate errors. Therefore, it can serve as a regularizer in BA to mitigate drift in the initial trajectory. The objective function for this optimization is as follows:

$$\sum_{I_i \in \mathcal{I}} \sum_{\mathbf{P}_l^W \in \mathcal{P}} \rho \left(\|\mathbf{r}_{i,l}\|^2 \right) + \sum_{(I_i, I_j) \in E} \omega_{i,j} \left(\|\mathbf{r}'_{i,j}\|^2 \right) \quad (4)$$

where $\rho(\cdot)$ is the loss function, and, in this paper, the huber loss function $\rho(x) = \frac{x^2}{x^2 + \delta^2}$ is used. $\omega_{i,j}$ represents the weight value for the known rotation term, $\mathbf{r}'_{i,j}$. The objective divides into two terms, explained as follows:

- Reprojection term: this term represents the reprojection error corresponding to all tie points in bundle adjustment, as follows:

$$\mathbf{r}_{i,l} = \pi \left(\mathbf{R}_{I_i}^W \mathbf{P}_l^W + \mathbf{t}_{I_i}^W, \mathbf{C} \right) - \mathbf{p}_{i,l} \quad (5)$$

where \mathbf{C} represents the intrinsic matrix, and \mathbf{P}_l^W is the 3D coordinate of a world point. Additionally, GCPs can be included in the reprojection term, as follows:

$$\mathbf{P}_l^W = s_G^W \mathbf{R}_G^W \mathbf{P}_l^G + \mathbf{t}_G^W \quad (6)$$

where \mathbf{P}_l^G denotes the position of GCP in the geodetic coordinate, and \mathbf{R}_G^W , \mathbf{t}_G^W , and s_G^W represent the rotation matrix, the translation, and the scale between the world and geodetic coordinates. The centimeter-level accuracy trajectory can be obtained by introducing the GCP term into BA.

- Known rotation term: this term is used as a regularizer to reduce the accumulated error, which is given by:

$$r'_{i,j} = \log \left(\widehat{\mathbf{R}}_{I_i}^{I_j} \mathbf{R}_{I_i}^{WT} \mathbf{R}_{I_j}^W \right) \quad (7)$$

where \log is logarithm mapping from the special orthogonal group $SO(3)$ to Lie algebra $\mathfrak{so}(3)$. $\widehat{\mathbf{R}}$ and \mathbf{R} denote estimated and global rotation, respectively.

4. Experiment

4.1. Dataset and Metrics

4.1.1. Dataset

The dataset consists of two aerial video sequences captured using the DJI M300 RTK drone with the DJI P1 camera, both manufactured by DJI in Shenzhen, China. Figure 3 illustrates the two sequences: one with a regular strip configuration and the other with an irregular configuration. These videos were recorded at the Informatics Department of Wuhan University at an altitude of 200 m. The recording frequency was set at 60 frames per second (fps), with a resolution of 3840×2160 pixels. The average ground resolution achieved was 0.03 m.

The regular sequence is a 1379-second video that contains evenly distributed air strips across the area. The between-strip overlapping is set at a degree of 40%. The coverage area of this sequence is $860 \times 460 \text{ m}^2$ and consists of buildings, trees, and playgrounds. On the other hand, the irregular sequence is a 345-s video that follows a heart-shaped loop trajectory. This sequence includes wooded areas, buildings, and a lake, with numerous texture-repeated regions, making it more challenging compared with the regular sequence.

Additionally, we collected 16 GCPs that are evenly distributed throughout the dataset area. Some of these GCPs were utilized to compute a high-accuracy trajectory, while the remainder served as checkpoints to assess the accuracy of the trajectory. The GCPs were measured using a high-accuracy GPS receiver and processed to achieve a localization accuracy of 9.0 mm.

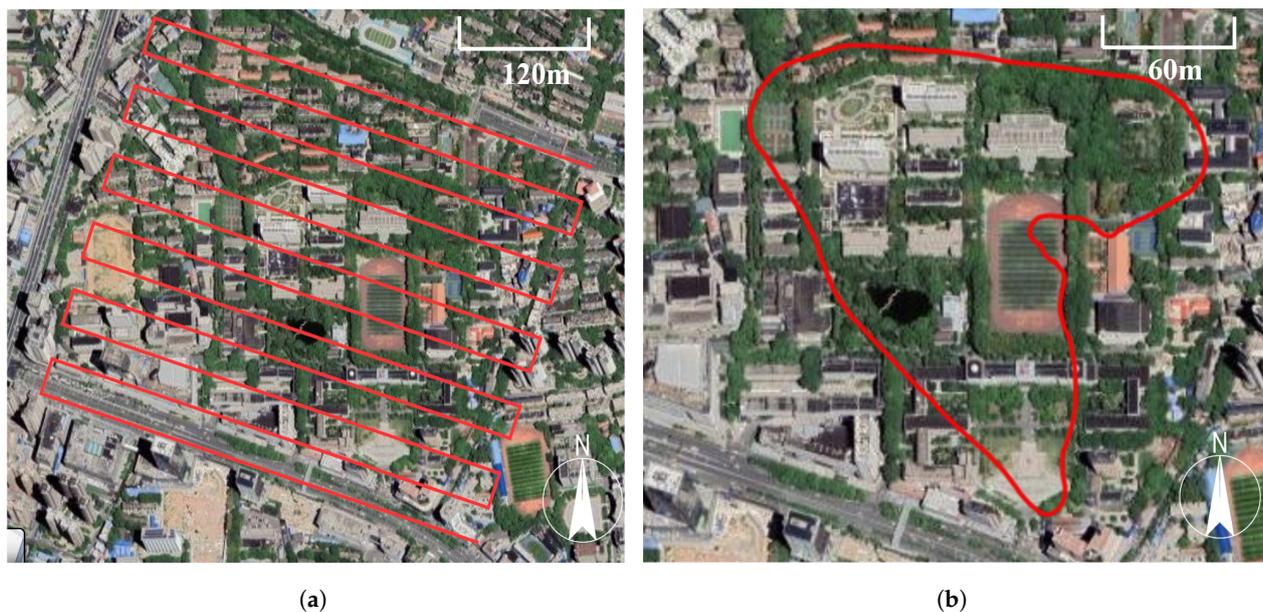


Figure 3. The visualization of dataset. (a) Regular scene. (b) Irregular scene. Red lines represent the trajectory of the drone.

4.1.2. Metrics

We use check points error (CPE) and absolute trajectory error (ATE) for evaluation.

- Check points error: the accuracy of triangulation is evaluated by utilizing surveyed points called check points (CPs) that were not used for georeferencing. Given a check point with coordinate $\tilde{\mathbf{P}}_l = \{\tilde{\mathbf{P}}_l(x), \tilde{\mathbf{P}}_l(y), \tilde{\mathbf{P}}_l(z)\}$, the root mean square errors (RMSEs) for plane (δ_{xy}), elevation (δ_z), and pixel (δ_p) in terms of m CPs are evaluated as follows:

$$\begin{aligned} \sigma_{xy} &= \frac{1}{m} \sum_l^m \sqrt{(\tilde{\mathbf{P}}_l(x) - \mathbf{P}_l(x))^2 + (\tilde{\mathbf{P}}_l(y) - \mathbf{P}_l(y))^2}, \\ \sigma_z &= \frac{1}{m} \sum_l^m \sqrt{(\tilde{\mathbf{P}}_l(z) - \mathbf{P}_l(z))^2}, \quad \sigma_p = \frac{1}{mN} \sum_i^N \sum_l^m \|\mathbf{r}_{i,l}\| \end{aligned} \quad (8)$$

- Absolute trajectory error: ATE is utilized to assess the drift in the position and rotation of the estimated trajectory. The estimated trajectory has been aligned with the ground truth trajectory using Umeyama's method [54], resulting in aligned poses represented as $\{\mathbf{T}_{I_i}^W\} = \{(\mathbf{R}_{I_i}^W, \mathbf{t}_{I_i}^W)\}$. RMSEs for the position (δ_{pos}) and rotation (δ_{rot}) are evaluated as follows:

$$\sigma_{pos} = \frac{1}{N} \|\mathbf{t}_{I_i}^W - \tilde{\mathbf{t}}_{I_i}^W\|, \quad \sigma_{rot} = \frac{1}{N} \sum_{I_i \in \mathcal{I}} d^2(\mathbf{R}_{I_i}^W, \tilde{\mathbf{R}}_{I_i}^W) \quad (9)$$

4.2. Results

In our method, the video input is set to 6 fps, and a downsampling rate of $4\times$ is applied. The search radius for the two-step keypoint adjustment is set to 20 pixels. In contrast, the other methods utilize the original-scale videos as input. However, for methods like COLMAP and Theia that require temporally sampled keyframes, we employ two strategies. One strategy involves sampling the video images every one second. The other strategy involves using our method, as described in Section 3.2, which entails applying OpenVSLAM on the $4\times$ downsampled videos to obtain keyframes.

We evaluate our method against the incremental SfM methods, namely COLMAP [3] and Theia [4], as well as the VSLAM method, OpenVSLAM [8], on the collected dataset, considering scenarios both with and without GCPs.

Performance on our collected dataset: As shown in Table 1, our pipeline has demonstrated significant improvements in accuracy when compared with COLMAP [3], Theia [4], and OpenVSLAM [8]. Specifically, in the regular sequence, our method outperforms the second-best method, OpenVSLAM, with improvements of 4.8 cm in δ_{xy} (a relative improvement of 137%), 1.7 cm in δ_z (a relative improvement of 40%), 0.12 pixels in δ_p (a relative improvement of 7%), as well as 0.7 m in δ_{pos} (a relative improvement of 175%), and 0.25° in δ_{rot} (a relative improvement of 178%).

Our method also exhibits significant improvements over other methods in all metrics for the irregular sequence. Compared with OpenVSLAM [8], the second-best performer, our method achieves improvements of 0.7 cm in δ_{xy} (a relative improvement of 14%), 6.2 cm in δ_z (a relative improvement of 293%), and 0.82 pixels in δ_p (a relative improvement of 103%). Additionally, our method demonstrates improvements of 0.24 m in δ_{pos} (a relative improvement of 48%) and 0.26° in δ_{rot} (a relative improvement of 96%). These results indicate that our pipeline effectively enhances reconstruction robustness and yields more accurate scene structure.

It is worth noting that when COLMAP [3] and Theia [4] are initialized with our selected keyframes, the accuracy improves across all metrics in both sequences against using per second sampling. This suggests that SLAM-based methods can effectively provide keyframes for subsequent SfM methods.

Table 1 presents a comparison of the efficiency of different methods. Our method requires the least amount of time compared with other methods, yielding a remarkable 200% enhancement over COLMAP [3], a 100% to 200% enhancement over OpenVSLAM [8], and a 50% to 100% improvement over Theia [4] for the regular sequence.

Table 1. Trajectory error metrics and efficiency on our dataset. δ_{xy} (cm), δ_z (cm), δ_p (pixel): RMSE error in (8). δ_{pos} (m), δ_{rot} (deg): RMSE error in (9). The 3840×2160 videos are $4 \times$ downsampled only on our method. ‘*’ indicates models are initialized using per second sampling.

Method	w/ GCPs			w/o GCPs		
	σ_{xy}	σ_z	σ_p	σ_{pos}	σ_{rot}	Time (s)
(a) Regular sequence						
COLMAP [3]	6.8	7.4	1.78	1.99	0.34	-
COLMAP * [3]	6.9	12	1.81	2.14	1.04	4650
Theia [4]	8.8	8.9	1.93	1.32	0.46	-
Theia * [4]	10	25	2.86	2.30	1.15	1413
OpenVSLAM [8]	8.3	6.3	1.67	1.10	0.39	1812
Ours	3.5	4.6	1.55	0.40	0.14	632
(b) Irregular sequence						
COLMAP [3]	10	13	3.12	0.64	0.52	-
COLMAP * [3]	11	11	3.95	1.87	1.83	550
Theia [4]	6.6	6.2	1.97	1.33	1.08	-
Theia * [4]	10	9	3.55	1.56	1.47	648
OpenVSLAM [8]	5.8	8.3	2.59	0.74	0.53	430
Ours	5.1	2.1	1.77	0.50	0.27	316

Bold represents the optimal metrics.

Performance on the EuRoc MAV: We test our proposed method on the small-scale and low-resolution EuRoc MAV Dataset [11], which consists of 11 sequences categorized into easy, medium, and difficult classes based on illumination and camera motion. In our method, we did not downsample the sequences from the EuRoc MAV Dataset since they already have a resolution of only 752×480 pixels. Additionally, the search radius for the two-step keypoint adjustment is set to 4 pixels.

In Table 2, we provide the σ_{pos} results. Given the small scale of the EuRoc sequences, our method shows slight improvements compared with OpenVSLAM [8]. For most sequences, our method delivers either better or comparable results to the state-of-the-art methods. Notably, COLMAP [3] demonstrates competitive accuracy with our method, but our approach is noticeably more efficient, as seen in Table 3. Our method also outperforms Theia [4] in terms of efficiency, except for sequences V102, V103, and V203. However, it is worth noting that, in sequences V103 and V203, Theia exhibits significantly lower accuracy compared with our method.

Table 2. σ_{pos} on EuRoc MAV [11]. The 752×480 EuRoc sequences are not downsampled in our method.

Method	M01	M02	M03	M04	M05	V101	V102	V103	V201	V202	V203	Mean
COLMAP [3]	0.037	0.033	0.052	0.073	0.053	0.089	0.063	0.088	0.064	0.056	0.058	0.061
Theia [4]	0.040	0.033	0.072	0.269	0.078	0.091	0.067	0.156	0.072	0.088	1.980	0.267
OpenVSLAM [8]	0.041	0.032	0.033	0.096	0.049	0.096	0.064	0.066	0.061	0.053	0.072	0.060
Ours	0.040	0.032	0.032	0.093	0.048	0.094	0.063	0.065	0.059	0.053	0.071	0.059

Bold represents the optimal metrics.

Table 3. A comparison of processing time (s) on EuRoc MAV [11]. ‘*’ indicates that in this scene Theia produces a very rough trajectory.

Method	M01	M02	M03	M04	M05	V101	V102	V103	V201	V202	V203
COLMAP [3]	534	463	279	187	226	371	63	254	202	144	564
Theia [4]	148	131	108	81	68	127	50	41 *	85	85	25 *
Ours	132	105	65	71	58	101	63	55	78	63	65

Bold represents the optimal metrics.

4.3. Ablation Experiment

We perform several ablation experiments on the collected dataset. Figure 4 illustrates the results for all metrics under different settings.

TS-KA: as shown in Figure 4, the incorporation of TS-KA enhances accuracy ranging from 2 to 5 times for all metrics when keypoints are extracted from downsampled images. Even when keypoints are obtained in the original scale, TS-KA still enhances matching performance, particularly for σ_{rot} . This emphasizes the significance of adjusting keypoints prior to global refinement.

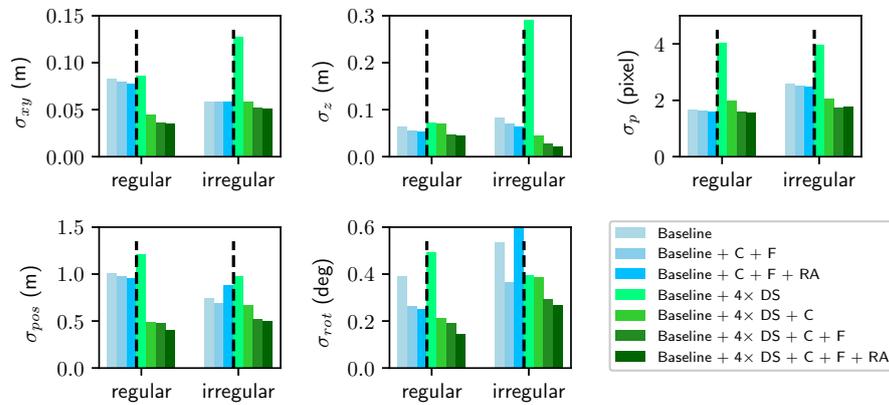


Figure 4. RMSE results. Baseline: bundle adjustment applied once at the original scale, and parameters and keyframes are initialized with OpenVSLAM on 4× downsampled video. DS: downsample; RA: rotation average; BA: global bundle adjustment; C: coarse keypoint adjustment; F: fine keypoint adjustment.

Rotation averaging: in the regular sequence, the introduction of global averaged rotations into BA results in a slight improvement for all metrics. However, in the case of the irregular sequence, there can be a slight decrease in accuracy for certain metrics like σ_{pos} and σ_{rot} on the original scale. This can be attributed to the fact that the view graph of the regular scene is denser, which necessitates the use of rotation averaging. Conversely, in the irregular sequence, the scene may have a sparser view graph, making the global averaged rotations less beneficial.

Accuracy vs precision: Figure 5 shows breakdown timings of each component to the total reconstruction in the regular scene. We see the coarse adjustment takes equal time but gain significantly more improvement in accuracy than fine keypoint adjustment, according to Figure 4. Therefore, it is a good choice to remove the fine adjustment components [20] instead of the coarse adjustment in an efficiency-first scenario.

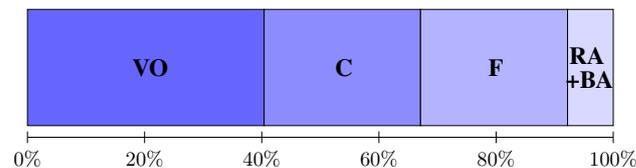


Figure 5. Breakdown timings of each component in regular sequence. RA: rotation averaging regularizer; C: coarse keypoint adjustment; F: fine keypoint adjustment; BA: bundle adjustment.

TS-KA vs FKA [20]: we further compare our TS-KA with the featuremetric keypoint adjustment (FKA) [20] for SfM tasks in six outdoor sequences from the ETH3D benchmark [15]. This benchmark provides ground-truth camera poses, intrinsic parameters, and highly accurate dense point clouds. To evaluate the matching effect, we follow the protocol introduced in [52]. We reconstruct a 3D sparse model using COLMAP [3], with fixed camera intrinsics and poses provided by the authors. We use four different features: SIFT [43], learning-based SuperPoint [45], D2-Net [47], and R2D2 [46] for extracting feature points in the original scale.

The results of applying two keypoint refinement methods on different feature points are presented in Table 4. It can be observed that our method consistently achieves better accuracy and completeness compared with [20] across all feature points in almost all

scenes. This consistent improvement confirms that our TS-KA method offers superior keypoint alignment.

Table 4. Results of 3D sparse reconstruction using our TS-KA or FKA [20] on different feature point extractors. We use metrics “accuracy” and “completeness” for threshold 1 cm, 2 cm, and 5 cm, as defined in [55].

Features Refinement	ETH3D Outdoor					
	Accuracy (%)			Completeness (%)		
	1 cm	2 cm	5 cm	1 cm	2 cm	5 cm
SIFT [43]	62.36	71.70	86.27	0.06	0.34	2.65
FKA [20]	65.63	76.25	91.19	0.07	0.40	2.86
Ours	66.48	78.75	92.12	0.07	0.40	2.90
SuperPoint [45]	49.19	64.34	82.74	0.09	0.49	3.46
FKA [20]	67.20	79.84	90.63	0.16	0.82	4.98
Ours	68.17	80.13	90.87	0.17	0.83	4.96
D2-Net [47]	34.66	51.38	72.12	0.02	0.13	1.77
FKA [20]	64.68	79.17	90.88	0.08	0.59	5.37
Ours	65.39	80.18	91.25	0.09	0.59	5.36
R2D2 [46]	42.71	59.81	80.71	0.05	0.36	3.02
FKA [20]	64.02	77.77	90.19	0.11	0.60	4.01
Ours	64.19	77.99	90.24	0.11	0.61	4.04

Bold represents the optimal metrics.

Figure 6 provides samples of feature point refinement, showcasing the ability of our method to adjust feature points in multi-view images to their correct positions. In comparison, FKA [20] is capable of correctly adjusting points under small view angle changes, as demonstrated in the first and second row. However, when faced with significant variations in view angles, as shown in the third row, FKA tends to produce a larger number of incorrect keypoints. This discrepancy largely contributes to the comparatively poorer performance of FKA, as evident in Table 4.

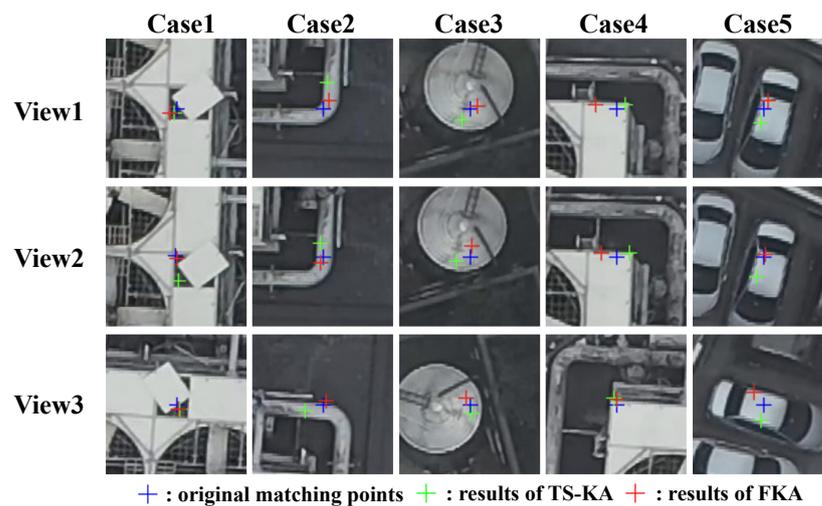


Figure 6. The comparison of keypoint adjustment methods. For each keypoint, we select three views. View 1 and View 2 have similar capture angles, whereas the viewing angle of View 3 varies significantly from them. For each view, we demonstrate the matching positions using different keypoint adjustment methods.

5. Conclusions

This paper introduces an efficient SfM pipeline for processing high-resolution, large-size videos. The pipeline utilizes visual odometry to select keyframes and obtain initial camera poses and reconstruction results efficiently by operating on downsampled video

data. A two-step keypoint adjustment method, TS-KA, is proposed to efficiently reuse and adjust the keypoints extracted during visual odometry, resulting in improved stability for subsequent global bundle adjustment. Experimental results demonstrate the superior performance and efficiency of our method compared with state-of-the-art SfM and VSLAM methods. Additionally, we have curated and introduced an outdoor high-resolution, large-size video dataset with high-accuracy GCPs, serving as a valuable supplement to existing public video datasets and offering considerable benefits to SfM and VSLAM research.

In this article, we focus exclusively on video-based SfM. With the increasing availability of onboard sensors, we plan to integrate our method with data from other sensors, such as IMU and GNSS, to further enhance the accuracy and robustness of our algorithm.

Author Contributions: Conceptualization, S.J. and J.C.; methodology, R.X. and J.C.; resources, S.J.; data curation, R.X.; writing—original draft preparation, R.X.; writing—review and editing, R.X., J.C. and S.J.; visualization, R.X.; supervision, S.J. funding acquisition, S.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (42171430).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study will be available at http://gpcv.whu.edu.cn/data/WHU_Aerial_Video_Dataset.html (accessed on 2 April 2024).

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

References

1. Snavely, N. Bundler: Structure from motion (SfM) for unordered image collections. *J. Inst. Image Inf. Telev. Eng.* **2008**, *65*, 479–482. [[CrossRef](#)]
2. Wu, C. Towards linear-time incremental structure from motion. In Proceedings of the 2013 International Conference on 3D Vision-3DV, Seattle, WA, USA, 29 June–1 July 2013; pp. 127–134.
3. Schonberger, J.L.; Frahm, J.-M. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4104–4113.
4. Sweeney, C.; Hollerer, T.; Turk, M. Theia: A fast and scalable structure-from-motion library. In Proceedings of the 23rd ACM International Conference on Multimedia, New York, NY, USA, 11–13 October 2015; pp. 693–696.
5. Ozyesil, O.; Singer, A. Robust camera location estimation by convex programming. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2674–2683.
6. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 1212–1221.
7. Mur-Artal, R.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM: A versatile and accurate monocular slam system. *IEEE Trans. Robot.* **2015**, *31*, 1147–1163. [[CrossRef](#)]
8. Sumikura, S.; Shibuya, M.; Sakurada, K. OpenVSLAM: A versatile visual slam framework. *ACM SIGMultimed. Rec.* **2019**, *11*, 10. [[CrossRef](#)]
9. Yu, L.; Yang, E.; Yang, B. AFE-ORB-SLAM: Robust monocular vslam based on adaptive fast threshold and image enhancement for complex lighting environments. *J. Intell. Robot. Syst.* **2022**, *105*, 26. [[CrossRef](#)]
10. Szántó, M.; Bogár, G.R.; Vajta, L. ATDN vSLAM: An all-through deep learning-based solution for visual simultaneous localization and mapping. *Period. Polytech. Electr. Eng. Comput. Sci.* **2022**, *66*, 236–247. [[CrossRef](#)]
11. Burri, M.; Nikolic, J.; Gohl, P.; Schneider, T.; Rehder, J.; Omari, S.; Achtelik, M.W.; Siegwart, R. The EuRoC micro aerial vehicle datasets. *Int. J. Robot. Res.* **2016**, *35*, 1157–1163. [[CrossRef](#)]
12. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [[CrossRef](#)]
13. Schubert, D.; Goll, T.; Demmel, N.; Usenko, V.; Stuckler, J.; Cremers, D. The TUM VI benchmark for evaluating visual-inertial odometry. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1680–1687.
14. Yin, J.; Li, A.; Li, T.; Yu, W.; Zou, D. M2DGR: A multi-sensor and multi-scenario slam dataset for ground robots. *IEEE Robot. Autom. Lett.* **2021**, *7*, 2266–2273. [[CrossRef](#)]
15. Schops, T.; Schonberger, J.L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; Geiger, A. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3260–3269.

16. Kurniawan, R.A.; Ramdan, F.; Furqon, M.T. Videogrammetry: A new approach of 3-dimensional reconstruction from video using sfm algorithm: Case studi: Coal mining area. In Proceedings of the 2017 International Symposium on Geoinformatics (ISyG), Malang, Indonesia, 24–25 November 2017; pp. 13–17.
17. Jeon, I.; Lee, I. 3D Reconstruction of unstable underwater environment with SFM using SLAM. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *43*, 957–962. [[CrossRef](#)]
18. Habib, Y.; Papadakis, P.; Fagette, A.; Le Barz, C.; Gonçalves, T.; Buche, C. From sparse SLAM to dense mapping for UAV autonomous navigation. *Geospat. Inform. XIII SPIE* **2023**, 12525, 110–120. [[CrossRef](#)]
19. Woodford, O.J.; Rosten, E. Large scale photometric bundle adjustment. *arXiv* **2020**, arXiv:2008.11762.
20. Lindenberger, P.; Sarlin, P.-E.; Larsson, V.; Pollefeys, M. Pixel-perfect structure-from-motion with featuremetric refinement. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 5987–5997.
21. Chen, Y.; Zhao, J.; Kneip, L. Hybrid rotation averaging: A fast and robust rotation averaging approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10358–10367.
22. Snavely, N.; Seitz, S.M.; Szeliski, R. Photo tourism: Exploring photo collections in 3D. *Semin. Graph. Pap. Push. Boundaries* **2023**, *2*, 515–526. [[CrossRef](#)]
23. Heinly, J.; Schonberger, J.L.; Dunn, E.; Frahm, J. Reconstructing the World in Six Days. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3287–3295.
24. Jiang, S.; Li, Q.; Jiang, W.; Chen, W. Parallel structure from motion for UAV images via weighted connected dominating set. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5413013. [[CrossRef](#)]
25. Chen, Y.; Shen, S.; Chen, Y.; Wang, G. Graph-based parallel large scale structure from motion. *Pattern Recognit.* **2020**, *107*, 107537. [[CrossRef](#)]
26. Barath, D.; Mishkin, D.; Eichhardt, I.; Shipachev, I.; Matas, J. Efficient initial pose-graph generation for global sfm. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14546–14555.
27. Wang, X.; Xiao, T.; Kasten, Y. A hybrid global structure from motion method for synchronously estimating global rotations and global translations. *ISPRS J. Photogramm. Remote Sens.* **2021**, *174*, 35–55. [[CrossRef](#)]
28. Zhu, S.; Shen, T.; Zhou, L.; Zhang, R.; Wang, J.; Fang, T.; Quan, L. Parallel structure from motion from local increment to global averaging. *arXiv* **2017**, arXiv:1702.08601.
29. Cui, Z.; Tan, P. Global structure-from-motion by similarity averaging. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 864–872.
30. 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 Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4568–4577.
31. Moreira, G.; Marques, M.; Costeira, J.P. Rotation averaging in a split second: A primal-dual method and a closed-form for cycle graphs. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 5452–5460.
32. Zhuang, B.; Cheong, L.-F.; Lee, G.H. Baseline desensitizing in translation averaging. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4539–4547.
33. Goldstein, T.; Hand, P.; Lee, C.; Voroninski, V.; Soatto, S. Shapefit and shapekick for robust, scalable structure from motion. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 10–16 October 2016; pp. 289–304.
34. Wilson, K.; Snavely, N. Robust global translations with 1dsfm. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 5–12 September 2014; pp. 61–75.
35. Yang, L.; Ye, J.; Zhang, Y.; Wang, L.; Qiu, C. A semantic SLAM-based method for navigation and landing of UAVs in indoor environments. *Knowl.-Based Syst.* **2024**, *293*, 111693. [[CrossRef](#)]
36. Shum, H.-Y.; Ke, Q.; Zhang, Z. Efficient bundle adjustment with virtual key frames: A hierarchical approach to multi-frame structure from motion. In Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), Fort Collins, CO, USA, 23–25 June 1999; pp. 538–543.
37. Resch, B.; Lensch, H.; Wang, O.; Pollefeys, M.; Sorkine-Hornung, A. Scalable structure from motion for densely sampled videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3936–3944.
38. Jiang, N.; Cui, Z.; Tan, P. A global linear method for camera pose registration. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 481–488.
39. Leotta, M.J.; Smith, E.; Dawkins, M.; Tunison, P. Open source structure-from-motion for aerial video. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–9.
40. Banterle, F.; Gong, R.; Corsini, M.; Ganovelli, F.; Gool, L.V.; Cignoni, P. A deep learning method for frame selection in videos for structure from motion pipelines. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 3667–3671.
41. Wang, L.; Ge, L.; Luo, S.; Yan, Z.; Cui, Z.; Feng, J. TC-SfM: Robust track-community-based structure-from-motion. *IEEE Trans. Image Process.* **2024**, *33*, 1534–1548. [[CrossRef](#)] [[PubMed](#)]

42. Gong, Y.; Zhou, P.; Liu, Y.; Dong, H.; Li, L.; Yao, J. A cluster-based disambiguation method using pose consistency verification for structure from motion. *ISPRS J. Photogramm. Remote Sens.* **2024**, *209*, 398–414. [[CrossRef](#)]
43. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
44. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to sift or surf. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
45. DeTone, D.; Malisiewicz, T.; Rabinovich, A. Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 224–236.
46. Revaud, J.; Weinzaepfel, P.; De Souza, C.; Humenberger, M. R2D2: Repeatable and reliable detector and descriptor. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 12414–12424.
47. Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; Sattler, T. D2-Net: A trainable cnn for joint detection and description of local features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8084–8093.
48. Taira, H.; Okutomi, M.; Sattler, T.; Cimpoi, M.; Pollefeys, M.; Sivic, J.; Pajdla, T.; Torii, A. Inloc: Indoor visual localization with dense matching and view synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7199–7209.
49. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
50. Li, X.; Han, K.; Li, S.; Prisacariu, V. Dual-resolution correspondence networks. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Online, 6–12 December 2020; pp. 17346–17357.
51. Zhou, Q.; Sattler, T.; Leal-Taixe, L. Patch2pix: Epipolar-guided pixel-level correspondences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4669–4678.
52. Dusmanu, M.; Schönberger, J.L.; Pollefeys, M. Multi-view optimization of local feature geometry. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 670–686.
53. Germain, H.; Bourmaud, G.; Lepetit, V. S2DNet: Learning Image Features for Accurate Sparse-to-Dense Matching. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 626–643.
54. Umeyama, S. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **1991**, *13*, 376–380. [[CrossRef](#)]
55. Knapitsch, A.; Park, J.; Zhou, Q.-Y.; Koltun, V. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph. (ToG)* **2017**, *36*, 78. [[CrossRef](#)]

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.