drones

MDPI

Article

A Visual Navigation Algorithm for UAV Based on Visual-Geography Optimization

Weibo Xu *, Dongfang Yang, Jieyu Liu, Yongfei Li and Maoan Zhou

Xi'an Research Institute of Hi-Tech, Xi'an 710025, China; yangdf@xjtu.edu.cn (D.Y.); liujieyu128@163.com (J.L.); liyf@xidian.edu.cn (Y.L.); zma@webmail.hzau.edu.cn (M.Z.)
* Correspondence: 18701260929@163.com

Abstract: The estimation of Unmanned Aerial Vehicle (UAV) poses using visual information is essential in Global Navigation Satellite System (GNSS)-denied environments. In this paper, we propose a UAV visual navigation algorithm based on visual-geography Bundle Adjustment (BA) to address the challenge of missing geolocation information in monocular visual navigation. This algorithm presents an effective approach to UAV navigation and positioning. Initially, Visual Odometry (VO) was employed for tracking the UAV's motion and extracting keyframes. Subsequently, a geolocation method based on heterogeneous image matching was utilized to calculate the geographic pose of the UAV. Additionally, we introduce a tightly coupled information fusion method based on visual-geography optimization, which provides a geographic initializer and enables real-time estimation of the UAV's geographical pose. Finally, the algorithm dynamically adjusts the weight of geographic information to improve optimization accuracy. The proposed method is extensively evaluated in both simulated and real-world environments, and the results demonstrate that our proposed approach can accurately and in real-time estimate the geographic pose of the UAV in a GNSS-denied environment. Specifically, our proposed approach achieves a root-mean-square error (RMSE) and mean positioning accuracy of less than 13 m.

Keywords: visual odometry; image matching; UAV geolocation; nonlinear optimization

Citation: Xu, W.; Yang, D.; Liu, J.; Li, Y.; Zhou, M. A Visual Navigation Algorithm for UAV Based on Visual-Geography Optimization. *Drones* **2024**, *8*, 313. https://doi.org/10.3390/drones8070313

Academic Editor: Chao Huang

Received: 8 April 2024
Revised: 4 July 2024
Accepted: 8 July 2024
Published: 10 July 2024

1. Introduction

In recent years, UAVs have been extensively utilized in various domains, including military operations, search and rescue missions, and terrain exploration [1–5]. While GNSS technology can provide precise geolocation information for UAVs, it is susceptible to signal blockage and interference from adversarial sources [6]. Therefore, it is imperative to develop methodologies for UAV localization in GNSS-denied environments [7]. In such challenging scenarios, UAVs must rely solely on their onboard camera to infer their pose without any external infrastructure. Hence, developing robust techniques for accurate and independent UAV positioning becomes paramount.

In GNSS-denied environments, UAVs can rely on visual sensors for precise pose estimation [8], enabling autonomous navigation and positioning. Visual sensors play a crucial role in environmental perception by capturing image information that provides UAVs with abundant contextual data. The technologies for UAV visual navigation and positioning can be primarily categorized into two approaches: VO, which operates without a map, and geolocation based on image matching, which utilizes an existing map.

The UAV can estimate its relative pose using VO, which solely relies on a cost-effective and lightweight vision sensor to accurately determine the local pose of the UAV in real-time. Depending on different approaches for associating features between frames, it can be categorized into feature-based, direct, and hybrid methods. The feature-based method is employed to extract features from the image for matching, while camera pose estimation is achieved by minimizing the reprojection error. PTAM [9] and ORB-SLAM [10] are both

classical SLAM systems that utilize the feature-based approach. Among them, ORB-SLAM leverages ORB [11] features in matching, tracking, relocation, and loop closing processes, thereby enhancing information interaction efficiency across different threads. Subsequently, ORB-SLAM2 [12] and ORB-SLAM3 [13] also employ ORB features. On the other hand, direct methods directly use the pixel intensities in the images and estimate motion by minimizing a photometric error. DSO [14] and VINS-Mono [15] are advanced SLAM systems that employ the direct method. SVO [16] is a hybrid VO approach that initially extracts FAST features but uses a direct method to track features and any pixel with a nonzero intensity gradient from frame to frame.

Currently, state-of-the-art VO algorithms employ keyframe-based techniques [17], which provide higher accuracy compared to filtering methods at the same computational cost [18]. Keyframe-based techniques utilize a multi-threaded approach for tracking the robot's motion, comprising a tracking thread responsible for the UAV's motion tracking and keyframe selection and a mapping thread that performs computationally expensive yet more accurate BA optimization at the keyframe rate. This enables the VO system to achieve real-time performance while enhancing its localization accuracy. Although VO has demonstrated reliable precision in short-term operations, it tends to accumulate drift over long-term operations. Moreover, VO fails to provide a more comprehensive geographic pose, which limits its applicability in aerial scenarios. Therefore, it is crucial to incorporate supplementary techniques to address these challenges.

With the continuous advancement of image-matching technology, sophisticated image-matching networks can effectively address the challenges posed by significant variations in seasons, lighting conditions, and other factors between satellite images and aerial images [19,20]. Consequently, these networks facilitate establishing correspondences between aerial and satellite images to accurately determine the geographic pose of the UAV. Various methods [21] have been developed for image matching and localization in GNSS-denied environments. Goforth et al. [22] employed a deep convolutional neural network (CNN) with an iterative closest landmark keypoints (ICLK) layer to align aerial images and satellite images. Chen et al. [23] proposed an image-based geolocation method that is adaptable to a downward-tilted camera, which necessitates prior offline dataset preparation.

However, a drawback of geolocation methods based on image matching is their significant computational cost, leading to substantial delays and rendering them unsuitable for meeting the real-time positioning requirements of UAVs [24]. To enhance the positioning performance of visual navigation algorithms, certain studies have focused on integrating VO with image-matching methods. Kinnari et al. [25] orthorectified UAV images based on VIO and planarity assumptions, which is also applicable to a downward-tilted camera configuration. The ortho-projected image is utilized for matching with the satellite images, and the geolocation results are combined with the pose estimation from VIO in a particle filter framework. Hao et al. [26] proposed a geolocation method that utilizes global pose graph optimization to integrate RVIO measurements and image registration. Zhang et al. [27] proposed an iterative trajectory fusion pipeline that integrates vSLAM measurements and image matching through solving a scaled pose graph problem.

The aforementioned studies employ a trajectory fusion method to integrate the measurements from VO and image matching, treating them as two relatively independent modules. While this design simplifies the system architecture, it also results in a disconnection between VO and image matching, impeding the immediate feedback of image-matching results in the VO estimation process. This limitation may potentially impact the overall accuracy and robustness of the system.

To fulfill the aforementioned requirements and address existing concerns, we have developed a tightly coupled visual navigation algorithm for UAVs based on visual-geography optimization. As depicted in Figure 1, the algorithm is structured into three concurrent threads inspired by the SLAM multi-threading concept: tracking thread, mapping thread, and image-matching thread. The tracking thread enhances real-time performance by monitoring the UAV's movement, while the mapping thread optimizes the VO map through

the fusion of visual information from VO and geographic positioning data obtained via image matching. Additionally, the image-matching thread provides globally unbiased geographic positioning information. The image-matching thread incorporates the existing coarse-to-fine image-matching method [23], while also introducing a prior-based approach. The prior-based image-matching method not only reduces computational costs but also enhances positioning accuracy, which holds significant implications for the fusion algorithm with VO. In the mapping thread, we have developed a robust geographic initializer and map fusion algorithm. The geographic initializer accurately estimates initialization parameters to align the VO world coordinate system with the geographic coordinate system. The map fusion algorithm utilizes geographical positioning information obtained from image matching for real-time optimization of the VO map. In summary, our algorithm enables real-time and precise estimation of a UAV's geographic pose through visual-geography optimization.

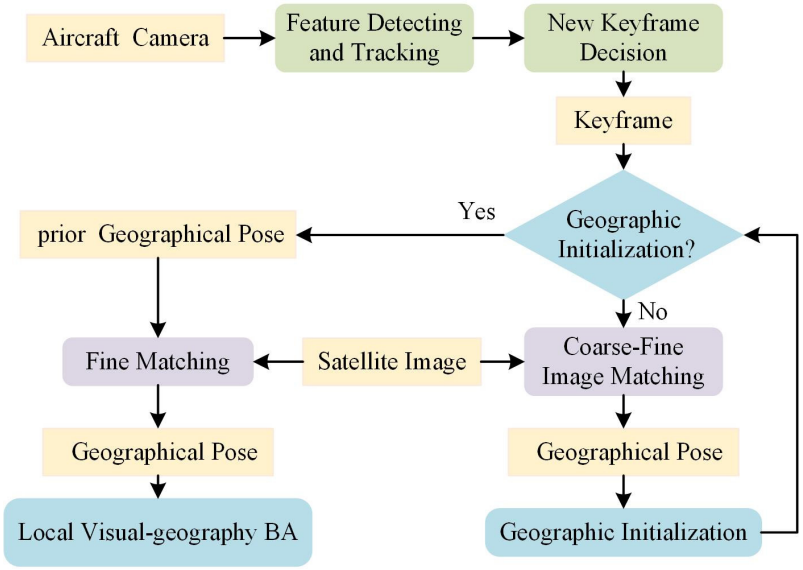


Figure 1. Pipeline of the proposed localization method. Aerial image refers to real-time images captured by a UAV, while satellite images need to be pre-stored in the UAV's onboard computer before the mission. The algorithm operates in real-time for tracking the movement of the UAV using aerial images and selects a subset of aerial images for matching with satellite images, facilitating geographic initialization and visual-geography BA.

This work makes the following contributions:

- We propose a UAV visual navigation algorithm that combines the merits of VO and image matching.
- We introduce a geolocation method based on heterogeneous image matching, which employs the coarse-to-fine and prior-based image-matching methods to enhance the accuracy of geolocation. This method effectively leverages the operational characteristics of VO to provide precise geolocation information.
- We present a tightly coupled information fusion method based on visual-geography optimization that jointly optimizes the visual and geolocation information of keyframes to facilitate tightly coupled geolocation in algorithms. Compared with the existing trajectory fusion method, our method achieves higher positioning accuracy.

The remaining sections of this paper are organized as follows. Section 2 presents a detailed explanation of the visual navigation algorithm for the UAVs proposed in this study. In Section 3, we discuss our real-world experiments conducted to validate our approach. Concluding remarks are provided in Section 4. Finally, an extensive discussion is presented in Section 5.

2. Method

The proposed visual navigation algorithm for the UAV is detailed in this section. This paper presents a visual-geography optimization method for integrating visual information (VO map consists of the UAV's poses and map points) and geographic information (the geographical poses of the UAV are determined through image matching), enabling geographic initializer and real-time pose estimation. More details are shown in Algorithm 1. The following section introduces the fusion of visual and geographic information with tight coupling as well as geolocation based on heterogeneous image matching.

Algorithm 1 UAV visual navigator algorithm

Input: UAV images I_U and satellite map tiles.
Output: Estimated UAV trajectory.

for all UAV images **do**
  Extract ORB features.
  VO initialization.
  Track UAV movement.
  if keyframe **then**
    if !S_{gw} **then**
      Visual BA optimization.
      $T_{gi} = CtoF\_Image\_Matching\left(T_U^i\right) \leftarrow$ Sec. 2.2.1
      Calculate the initial value of S_{gw} by Sec. 2.1.2(2)
      Reliability check of S_{gw}
    else if !GeoInitialized **then**
      Visual BA optimization.
      $T_{gi} = Prior\_Image\_Matching\left(T_U^i, T_{wi}\right) \leftarrow$ Sec. 2.2.2
      Geographic initialization. $\leftarrow$ Sec. 2.1.2(3)
    else
      $T_{gi} = Prior\_Image\_Matching\left(T_U^i, T_{wi}\right) \leftarrow$ Sec. 2.1.1
      Visual-geography BA optimization.
    end if
  end if
end for

2.1. Tightly Coupled Visual and Geographic Information Fusion

This paper proposes a tightly coupled visual-geographic fusion method based on visual-geography optimization, aiming to enhance the algorithm's navigation and positioning performance by effectively utilizing both visual and geolocation information from keyframes. The section of the code is available at https://github.com/XuWeibo-code/UAV_Visual_Navigation.git (accessed on 4 June 2024).

2.1.1. Visual-Geography Optimization

The visual-geography optimization proposed in this paper is essentially a nonlinear optimization, enabling the joint optimization of visual information and geographic information.

The 3D transformation involved in the UAV's motion primarily consists of the pose transformation matrix T and the similarity transformation matrix S, where $T \in \mathrm{SE}(3)$ and its tangent vector $\delta \in se(3)$. Similarly, $S \in \mathrm{Sim}(3)$ and its tangent vector $\omega \in sim(3)$. The specific form of this transformation is as follows:

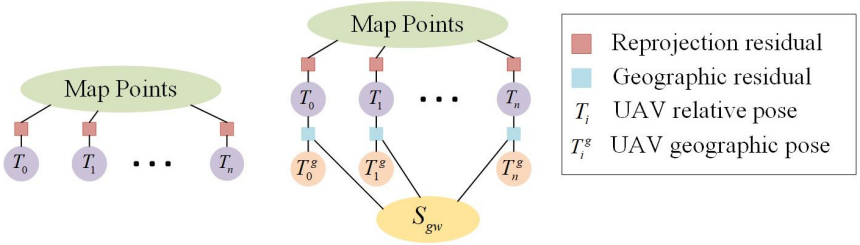
$$\begin{cases} T = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \in \mathrm{SE}(3); \quad \delta = \begin{bmatrix} \omega & v \end{bmatrix}^T \in se(3) \\ S = \begin{bmatrix} sR & t \\ 0 & 1 \end{bmatrix} \in \mathrm{Sim}(3); \quad \varepsilon = \begin{bmatrix} \omega & v & \lambda \end{bmatrix}^T \in sim(3) \end{cases}, \tag{1}$$

where R is the 3×3 rotation matrix, t is the 3×1 translation vector, and s represents the scale. The components in the tangent space are denoted as follows: ω is a 3×1 vector, v is a 3×1 vector, and λ is a scalar representing the rotation component, translation component, and scale component, respectively.

The optimization method employs four coordinate systems, namely the SLAM world coordinate system, geographic coordinate system, local camera coordinate system, and global camera coordinate system. In this paper, $S_{gw} \in Sim(3)$ is utilized to denote the similarity transformation matrix from the SLAM world coordinate system to the geographic coordinate system. $S_c \in Sim(3)$ represents the similarity transformation matrix from the local camera coordinate system to the global camera coordinate system. $T_{gi} \in SE(3)$ signifies the pose transformation matrix from the global camera coordinate system to the geographic coordinate system in frame i. $T_{wi} \in SE(3)$ denotes the pose transformation matrix from the local camera coordinate system to the SLAM world coordinates in frame i. The variables S_{gw} and S_c share a common scale parameter, which can be expressed as follows:

$$\begin{cases} S_{gw} = \begin{bmatrix} s_{gw}R_{gw} & t_{gw} \\ 0 & 1 \end{bmatrix} \\ S_c = \begin{bmatrix} s_{gw}I_{3\times3} & 0 \\ 0 & 1 \end{bmatrix} \end{cases}, \tag{2}$$

The conventional VO estimates the pose of a UAV, denoted as T_{wi}, by optimizing the visual information through visual BA (Figure 2a). In contrast, the proposed visual-geography BA (Figure 2b) jointly optimizes the parameters T_{wi} and S_{gw} using both visual and geolocation information from keyframes.



(a) Visual BA (b) Visual-Geography BA

Figure 2. Factor graph representation for different optimizations. The nodes in the factor graph correspond to optimization variables, including the UAV's relative pose in the SLAM world coordinate system, its geographic pose, map points, and initialization parameters. Meanwhile, the edges represent constraints through reprojection residuals and geographic residuals. By optimizing these nodes using BA, we aim to minimize these residuals.

First, the visual residual r_{ij}^v is defined as the reprojection error of the map point j in frame i.

$$r_{ij}^v = \hat{x}_j - \pi(T_{wi}X_j), \tag{3}$$

where $\hat{x}_j$ is the observed value of the map points, X_j represents the 3D location of the map point j, and $\pi(\cdot)$ corresponds to the camera projection function.

Then, the geographic residual r_i^g of frame i is defined as follows:

$$r_i^g = \mathrm{Log}\left(\hat{T}_{gi}^{-1}S_{gw}T_{wi}S_c^{-1}\right), \tag{4}$$

where $\mathrm{Log} : \mathrm{SE}(3) \to \mathbb{R}^6$ is a mapping from the Lie group $\mathrm{Sim}(3)$ to the vector space $sim(3)$, $\hat{T}_{gi}$ is an observability measure, and S_{gw} and T_{wi} are the optimization variables.

Given k keyframes and their states $\overline{T}_k \doteq \{T_0 \cdots T_{k-1}\}$, l 3D map points and their states $\chi_l \doteq \{X_0 \cdots X_{k-1}\}$, the cost function of this optimization is as follows:

$$\{T_i, X_j, S_{gw} | T_i \in \overline{T}_k, X_j \in \chi_l\} = \underset{T_i, X_j}{\operatorname{argmin}} \sum_{i=0}^{k-1} \sum_{j=0}^{l-1} \left(\rho\left(\left\| r_{kj}^c \right\|_{\sum_c}^2 \right) + \rho\left(\left\| r_k^g \right\|_{\sum_g}^2 \right) \right), \quad (5)$$

where $\sum_C$ and $\sum_g$ represent the weight factors of visual residual and geographical residual, respectively. The value of $\sum_C$ is obtained from [13], while the derivation of $\sum_g$ is elaborated in detail in Section 2.1.3. For visual and geographic residuals, we use a robust Huber kernel function $\rho(\cdot)$ to reduce the influence of spurious matchings.

To enhance the optimization efficiency of visual-geography BA, the algorithm must optimize the Jacobian matrix of the residual for the variable. Among them, the Jacobian matrix of the visual residual concerning the optimization variables T_{wi} and X_j can be obtained from [13]. Therefore, this paper only derives the Jacobian matrix of the geographic residual to the optimization variables S_{gw} and T_{wi}.

In Equation (4), the geographic residual r_i^g is a 7-dimensional vector, the tangent vector ε_{gw} of the S_{gw} is also a 7-dimensional vector, and the tangent vector δ_{wi} of T_{wi} is a 6-dimensional vector. The Jacobian matrices of the geographic residual concerning ε_{gw} and δ_{wi} are as follows:

$$\begin{cases} \frac{\partial r_i^g}{\partial \delta_{gw}} = Adj\left(\hat{T}_{gi}^{-1}\right) - Adj\left(r_i^g\right) \begin{bmatrix} 0_{6\times 6} & 0 \\ 0 & 1 \end{bmatrix} \\ \frac{\partial r_i^g}{\partial \delta_{wi}} = Adj\left(\hat{T}_{gi}^{-1} S_{gw}\right) \begin{bmatrix} I_{6\times 6} \\ 0 \end{bmatrix} \end{cases}, \quad (6)$$

where $Adj(S)$ is the adjoint matrix of S, which takes the following form:

$$Adj(S) = \begin{bmatrix} R & 0 & 0 \\ t_\times R & sR & -t \\ 0 & 0 & 1 \end{bmatrix}, \quad (7)$$

The optimization process of this BA requires the assignment of initial values. The initial values for the variables T_{wi} and X_j are provided by VO, while the initial value for the variable S_{gw} is calculated during the geographic initializer.

2.1.2. Geographic Initializer

The algorithm utilizes a [13]-based VO technique to track the motion of the UAV in real-time and perform a visual initializer. To enhance the real-time performance and accuracy of the SLAM, our algorithm employs a subset of representative frames as keyframes for BA optimization and image matching. During the selection process, temporal intervals, visual content disparities, camera movements, and co-visibility relationships between the current frame and other keyframes are all taken into consideration.

The algorithm then employs a rapid and robust geographic initializer, which aims to accurately estimate the initialization parameter S_{gw} that aligns the SLAM world coordinate system with the geographic coordinate system. The transformation relationship between the local pose in the SLAM world coordinate system and the geographical pose is described as follows:

$$T_{gi} = S_{gw} T_{wi} S_c^{-1}, \quad (8)$$

We state the geographic initializer as a map estimation problem, which can be divided into three sequential steps.

(1) Visual BA optimization

The algorithm utilizes a vision-only BA (Figure 2a) to optimize the UAV's pose and map points during the initial phase of its operation, when sufficient geolocation information is not yet available.

(2) Calculate the initial value of S_{gw}

The algorithm calculates the initial value of S_{gw} after receiving reliable geolocation information from the two keyframes. At this stage, the local poses of the two keyframes are denoted as T_{w1} and T_{w2}, while their geographical poses are represented by T_{g1} and T_{g2}. The scale component s_{gw} of the similarity transformation matrix S_{gw} is as follows:

$$s_{gw} = \frac{\left\| t_{g2} - t_{g1} \right\|}{\left\| t_{w2} - t_{w1} \right\|}, \tag{9}$$

The rotation component R_{gw} of the similarity transformation matrix S_{gw} is as follows:

$$\begin{cases} R_{gw} = R_{g1} R_{w1}^{T} \\ R'_{gw} = R_{g2} R_{w2}^{T} \end{cases}, \tag{10}$$

The translation component t_{gw} of the similarity transformation matrix S_{gw} is as follows:

$$\begin{cases} t_{gw} = -R_{g1} R_{w1}^{T} t_{w1} + s_{gw}^{-1} t_{g1} \\ t'_{gw} = -R_{g2} R_{w2}^{T} t_{w2} + s_{gw}^{-1} t_{g2} \end{cases}, \tag{11}$$

The initializer calculates the initial value of S_{gw} at this stage. It is important to note that the geolocation information obtained during this process relies on a coarse-to-fine image-matching method, which may result in mismatching. Therefore, the algorithm utilizes an error coefficient to check the reliability of S_{gw}. We construct S_{gw} and S'_{gw} using $\left(R_{gw}, t_{gw}, s_{gw}\right)$ and $\left(R'_{gw}, t'_{gw}, s_{gw}\right)$, respectively, and the error coefficient is calculated as follows:

$$r_S = \left\| \mathrm{Log}\left(S_{gw} S'^{-1}_{gw} \right) \right\|, \tag{12}$$

The reliability of S_{gw} is considered when $r_S < th$; otherwise, the algorithm will wait for the keyframe to continue receiving geolocation information of the UAV and repeat the aforementioned steps.

(3) Visual-geography BA

When the algorithm obtains a reliable initial value of S_{gw}, the prior geographic pose of the UAV can be calculated based on Equation (11). At this stage, the geolocation method based on heterogeneous image matching can directly extract the satellite image according to the prior geographic pose and perform more precise fine matching. Once geolocation information has been received for N keyframes, the algorithm optimizes the UAV pose, map points, and initialization parameter S using visual-geography BA (Figure 2a) of Equation (5).

2.1.3. Map Fusion Method and Geographic Weights Update

After the completion of the geographical initialization, it remains necessary for the system to output real-time and accurate geographical pose of the UAV. To achieve this, we employ visual-geography BA to optimize both UAV poses and map points. Unlike visual-geography BA with a geographical initializer, at this stage, our algorithm does not optimize the initialization parameter.

The geolocation information obtained by image matching is subject to errors, necessitating the assignment of weights to each geographic edge in the visual-geography BA. Therefore, we propose a calculation method for quantifying the magnitude of geographic residuals.

Refine the weight of the geographic edge i in BA according to the following criteria:

$$q = \beta_n \varphi_n + \beta_p \left\| \mathrm{Log}\left(T_{gc}^{-1} T'_{gc} \right) \right\|^2 + \beta_e \left\| e_i^g \right\|^2, \tag{13}$$

where the balance factors β_n, β_p, and β_e represent the weights assigned to different components; the variable φ_n represents the number of remaining inner points after removing outer

points from the matching points obtained by the RANSAC algorithm [28]; T_{gc} and T'_{gc} denote the prior geolocation information input in the image matching and the geolocation information output by image matching, respectively; and e_i^g represents the geographic residual of geographic edge i in BA.

The first two terms in Equation (13) are generated during the image matching, while the last term is dynamically updated through BA. In Equation (5), $qI_{7\times7}$ is introduced into $\sum_g$ to fine-tune the weighting of geographic residuals in the BA. By adjusting the weight of geographic edges in BA, this algorithm enables a more accurate estimation of the UAV's geographic pose.

2.2. Geolocation Based on Heterogeneous Image Matching

The geolocation method based on heterogeneous image matching is utilized to provide precise geolocation information for keyframes. Considering the characteristics of VO, appropriate adjustments are made to the geolocation method. In the initial stage of the algorithm, a coarse-to-fine image-matching method is employed for UAV geolocation. After the algorithm calculates the S_{gw}, the satellite image is intercepted using the prior pose of the UAV and then directly processed by the fine matching.

2.2.1. Coarse-to-Fine Image-Matching Method

The geolocation of the UAV is achieved in this paper by implementing the coarse-to-fine image-matching method utilized in [23], even in the absence of prior geolocation information. The pipeline of the coarse-to-fine image-matching method is shown in Figure 3.

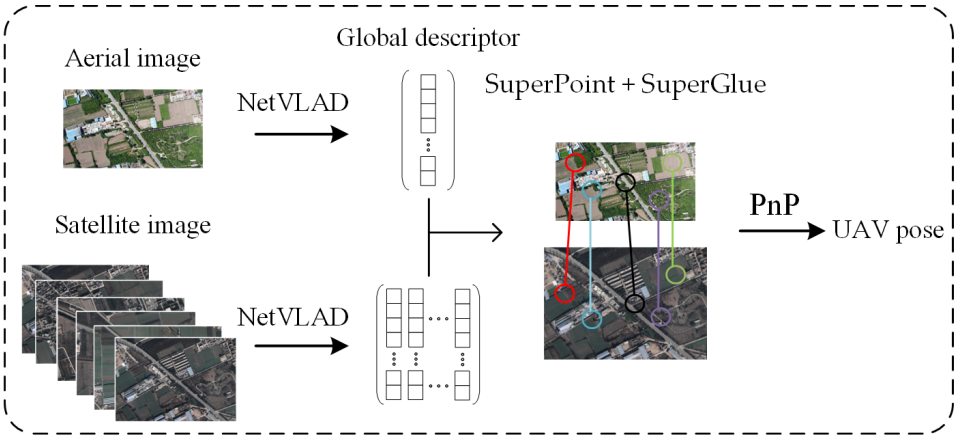


Figure 3. Pipeline of the coarse-to-fine image-matching method.

During the coarse matching stage, satellite images are preprocessed offline. These images, which have varying resolutions, are clipped into smaller map tiles and then mapped into 4096-dimensional global descriptors using NetVLAD [29]. During UAV execution, the extracted global descriptors from aerial images are also obtained using NetVLAD. The l2-norm of the differences between the aerial image's global descriptor and every single global descriptor in the prepared array is computed and sorted in ascending order. The top 3 candidates in the sorted list are selected for subsequent fine matching.

After the coarse matching stage, the algorithm proceeds to conduct fine matching between the aerial image and the satellite image. This process involves utilizing SuperPoint [30] for extracting local descriptors from both images, generating 2D matching points between the aerial image and the candidate matching images through SuperGlue [31]. The matching result of SuperPoint + SuperGlue is shown in Figure 4.

Assuming a planar ground with zero height, these 2D-2D correspondences are transformed into a camera projection relation from the 3D coordinates in the geographic coordinate system to the 2D-pixel coordinates in the aerial image. The EPnP+RANSAC algorithm [28] is then utilized to accurately determine the geographic pose of the UAV.

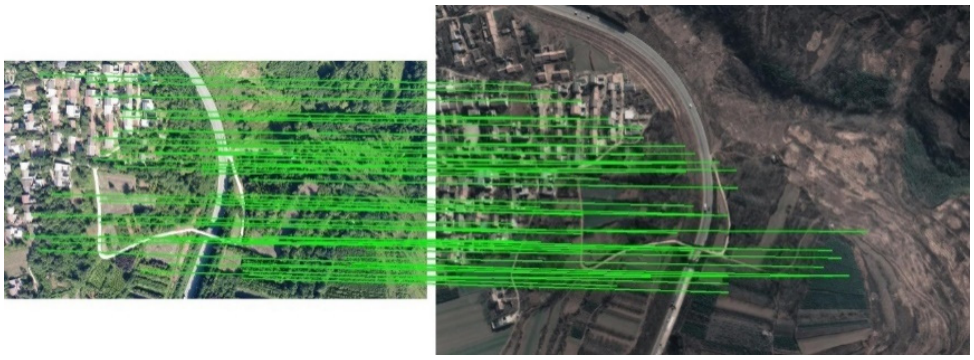


Figure 4. The matching result of SuperPoint + SuperGlue. The green lines depict the extracted feature correspondences between the aerial and satellite images using SuperPoint + SuperGlue.

2.2.2. Prior-Based Image-Matching Method

The coarse-to-fine image-matching method exhibits drawbacks, such as significant latency and a high mismatching rate. In this study, leveraging the operational characteristics of VO, we propose a prior-based image-matching approach for precise UAV geolocation.

After obtaining the geographic initialization parameter S_{gw}, the algorithm calculates the prior geographic pose $T_{ig}\left(T_{gi}^{-1}\right)$ of the keyframe i based on Equation (8). Assuming a planar ground with zero height, the transformation relationship between the geographic plane coordinates of the satellite image and pixel coordinates of the aerial image is described using a homography matrix H_{ig}.

$$H_{ig} = \frac{1}{\lambda} K\left[r_{ig1}, r_{ig2}, t_{ig}\right], \tag{14}$$

where R_{ig} and t_{ig} are the rotational and translational components of T_{ig}, r_{ig1} and r_{ig2} are the first and second columns of R_{ig}, respectively, K represents the camera internal reference, and λ denotes the depth of the feature point, which is equivalent to the absolute value of the z coordinate of $\left(-R_{ig}^{-1} t_{ig}\right)$.

The aerial image can be transformed into a geographic plane coordinate by applying the homography matrix H_{ig}, extracting a rectangular region in the geographic plane coordinate system for intercepting the satellite image. Subsequently, utilizing the homography matrix, the captured satellite image is projected onto the pixel coordinate of the aerial image. Figure 5 illustrates the flowchart depicting satellite image extraction.

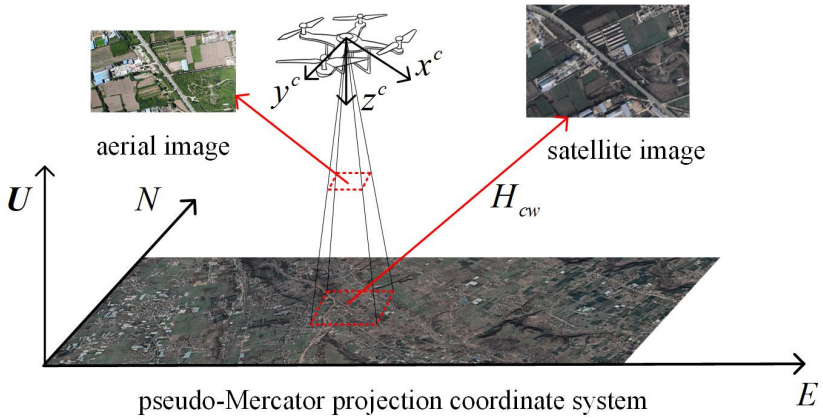


Figure 5. Flowchart of satellite image extraction.

After extracting the satellite image, the algorithm will perform fine matching, as described in Section 2.2.1, to accurately determine the geographic pose of the UAV.

3. Experimental Setups and Results

This section outlines the experimental setup and presents the test results. To simulate real-flight conditions, we have configured the Ubuntu environment on NVIDIA Jetson Xavier NX and conducted experimental validation.

3.1. Simulation Dataset

3.1.1. Setups

To evaluate the efficacy of the algorithm proposed in this paper, we generated a simulation dataset using a custom simulation program. This program employs the camera projection principle to transform satellite images captured in visible light mode into image sequences by configuring camera parameters and executing trajectories.

The trajectory of the UAV in the simulation dataset is depicted in Figure 6, with a satellite image serving as the background. Specific characteristics of the dataset are elaborated upon in Table 1.

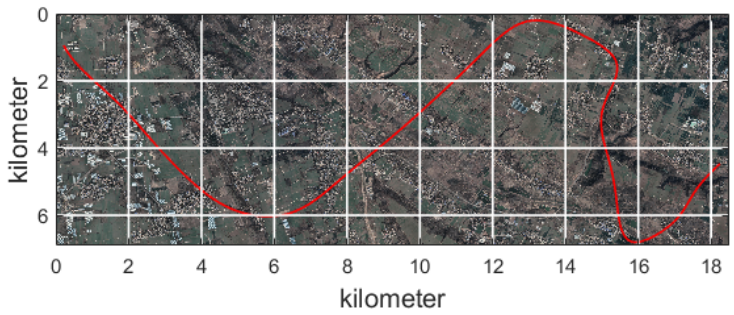


Figure 6. Trajectories of the UAV in the simulation dataset.

Table 1. Characteristics of the simulation datasets.

Dataset	Length (km)	Altitude (m)	Speed (m/s)	Duration (s)	Resolution	Frame Rate (fps)
simulation	29.1	500	30	960	960 × 540	20

3.1.2. Geolocation Performance

To evaluate the performance of our proposed algorithm, we compare it with various algorithms in the simulation dataset, including the VO algorithm (the monocular vision mode of ORB-SLAM3) [13], the VO algorithm after geographic initialization, the geolocation algorithm based on coarse-to-fine image matching [23], and our proposed algorithm. The 2D errors between the different algorithms and the ground-truth trajectory are depicted in Figure 7. The trajectory obtained from the VO algorithm is converted into the geographic coordinate through a similarity transformation matrix, which is derived using multiple ground truth poses by solving a least-squares problem [32].

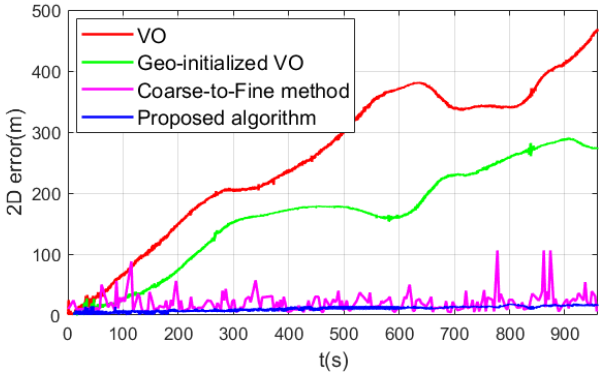


Figure 7. Localization errors of various algorithms in the simulated dataset.

3.2. Real-World Dataset

3.2.1. Setups

To evaluate the performance of the proposed algorithm in a real environment, we employed the DJI M300 RTK UAV equipped with an H20 pan-tilt camera to generate two distinct datasets for visual navigation and positioning of the UAV.

The trajectory of the UAV in real-world datasets is depicted in Figure 8, with the accurate trajectory obtained by RTK represented by the red curve. Furthermore, two videos were captured from a downward-facing perspective by the UAV, and their specific characteristics are detailed in Table 2.

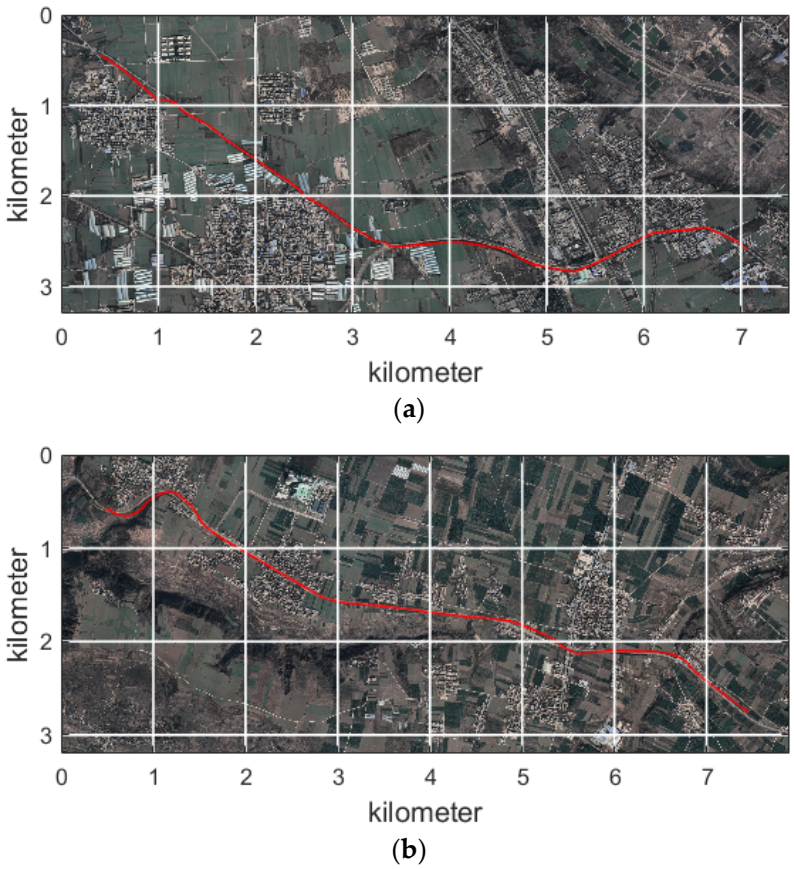


Figure 8. (**a**) The trajectories of UAVs in the real-world dataset 1; (**b**) The trajectories of UAVs in the real-world dataset 2. Red lines represent the accurate trajectory obtained by RTK.

Table 2. Characteristics of the real-world datasets.

Dataset	Length (km)	Altitude (m)	Speed (m/s)	Duration (s)	Resolution	Frame Rate (fps)
1	7.7	500	9.5	808	960 × 540	20
2	7.8	500	12.2	640	960 × 540	20

3.2.2. Geolocation Performance

To evaluate the performance of our proposed algorithm, we conducted a comparative analysis with various state-of-the-art algorithms using real-world datasets. These include the VO algorithm, the VO algorithm after geographic initialization, the geolocation algorithm based on coarse-to-fine image matching, and our proposed algorithm. The 2D errors between the different algorithms and the ground-truth trajectory are depicted in Figure 9.

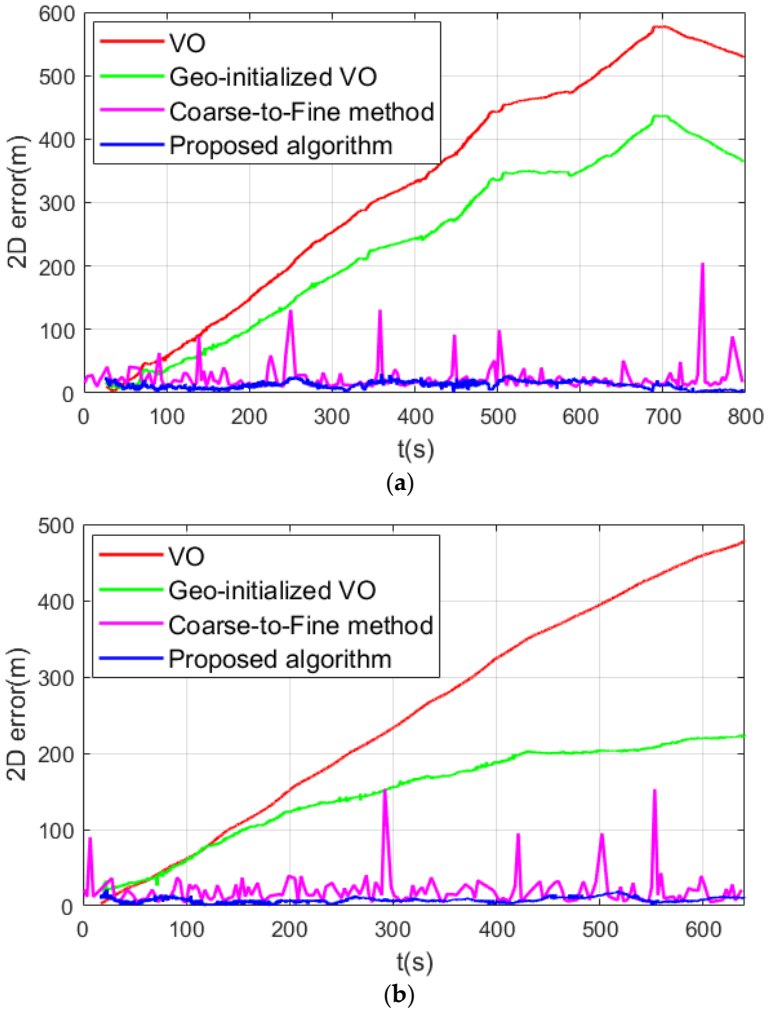


Figure 9. (**a**) Localization errors of various algorithms in the real-world dataset 1; (**b**) Localization errors of various algorithms in the real-world dataset.

3.3. Ablation Study

We conduct an ablation study to evaluate the contribution of each module in the proposed algorithm. The translation errors by removing different configurations are present in Table 3. We can observe that the result using the full proposed framework achieves the best performance.

Table 3. Ablation study on different configurations.

Dataset	Metric	Geo-Init	Image Matching	Trajectory Fusion	Full
Simulation	Mean (m)	161.12	14.36	15.36	10.32
	RMSE (m)	181.04	14.94	17.21	11.11
Real-World 1	Mean (m)	241.16	17.36	19.36	12.47
	RMSE (m)	275.79	19.21	21.33	12.72
Real-World 2	Mean (m)	150.37	16.33	14.13	7.53
	RMSE (m)	161.99	18.96	16.52	8.31

Geo-init: VO with geographic initialization; Image matching: pose estimation of UAV via SuperPoint + SuperGlue; Trajectory fusion: optimizing only trajectory in map fusion; Full: results using all proposed modules.

3.4. Analysis of Experimental Findings

We conducted geolocation experiments in a simulated dataset and two segments of real-world datasets. Throughout the experiment, we evaluated the proposed geographic

initialization method, which successfully accomplished geographic initialization within 5 s. Comparative analysis with the VO algorithm revealed that performing geographic initialization significantly enhances stability. The mean and RMSE of the different algorithms in the geolocation experiment are compared in Tables 4 and 5. It is evident that the algorithm proposed in this paper exhibits superior accuracy, enabling precise estimation of the geographic pose of the UAV.

Table 4. The mean and RMSE of various algorithms in the simulated dataset.

Metric	VO	Geo-Initialized VO	Coarse-to-Fine Method	Proposed Algorithm
Mean (m)	254.58	161.12	20.59	10.32
RMSE (m)	282.35	181.04	25.40	11.11

Table 5. The mean and RMSE of various algorithms in the real-world dataset.

Dataset	Metric	VO	Geo-Initialized VO	Coarse-to-Fine Method	Proposed Algorithm
1	Mean (m)	330.63	241.16	25.45	12.47
	RMSE (m)	376.23	275.79	37.11	12.72
2	Mean (m)	255.54	150.37	19.08	7.53
	RMSE (m)	293.14	161.99	26.47	8.31

The geolocation method based on coarse-to-fine image matching in the experiment exhibits high accuracy. However, the output positioning data exhibit a delay of 3 s and a mismatching probability not lower than 25%, making it unsuitable for real-time UAV positioning. The proposed algorithm achieves a data update frequency of 20 Hz/s, while the visual navigation algorithm presented in this paper enhances performance. Specifically, our proposed visual navigation algorithm enables accurate and real-time estimation of the geographic pose of the UAV in GNSS-denied environments.

4. Conclusions

This paper presents a visual navigation algorithm for UAVs based on visual-geography optimization, which effectively integrates visual and geolocation information from keyframes to achieve tightly coupled geolocation. Moreover, the algorithm incorporates a heterogeneous image-matching approach for geolocation, which combines coarse-to-fine image matching and prior-based methods. Based on the experimental results, the following conclusions can be drawn:

- In the geolocation method based on heterogeneous image matching, our proposed prior-based image-matching method utilizes the prior information to enhance the accuracy and efficiency of geolocation for the algorithm.
- The fusion method based on visual-geography optimization achieves stable and reliable estimation of geographic initialization parameters within 5 s, enabling real-time estimation of the UAV's geographic pose.
- We propose a tightly integrated fusion method that effectively combines the visual information from VO with the geolocation information obtained through the image-matching method. Experimental results demonstrate that our proposed algorithm accurately and in real-time estimates the UAV's geolocation information solely relying on the vision sensor, even in GNSS-denied environments.

5. Discussion

Although promising results have been achieved, there are several limitations. Firstly, the proposed algorithm is more suitable for scenarios within the range of tens to hundreds of meters. For higher or lower scenarios, significant errors may arise due to the image-matching method employed in the algorithm. Secondly, to calculate the geographic pose

of the UAV, we assume a planar ground with zero elevation. However, this assumption introduces a significant discrepancy in estimating the altitude of the UAV. Lastly, when confronted with low image texture or substantial variations in lighting conditions, inevitable disruptions occur in the visual navigation and positioning algorithm utilized by the UAV.

To enhance the positioning performance of UAVs in complex scenarios, our future research will focus on developing visual-inertial navigation algorithms for UAV positioning and incorporating barometers or laser rangefinders to accurately measure UAV altitude, thereby providing precise and reliable positioning information.

Author Contributions: Conceptualization, W.X. and Y.L.; methodology, W.X., D.Y. and Y.L.; software, W.X. and Y.L.; validation, Y.X, D.Y. and J.L.; formal analysis, W.X.; investigation, W.X. and D.Y.; resources, D.Y. and Y.L.; data curation, W.X., Y.L. and M.Z.; writing, original draft preparation, W.X.; writing, review and editing, W.X. and M.Z.; visualization, W.X.; supervision, D.Y. and J.L.; project administration, D.Y. and J.L.; funding acquisition, D.Y and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Natural Science Foundation of China (Grant Nos. 42301535, 61673017, 61403398) and the Natural Science Foundation of Shaanxi Province (Grant Nos. 2017JM6077).

Data Availability Statement: The data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

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

References

1. Scherer, J.; Yahyanejad, S.; Hayat, S.; Yanmaz, E.; Andre, T.; Khan, A.; Vukadinovic, V.; Bettstetter, C.; Hellwagner, H.; Rinner, B. An Autonomous Multi-UAV System for Search and Rescue. In Proceedings of the MobiSys'15: The 13th Annual International Conference on Mobile Systems, Applications, and Services, Florence, Italy, 18 May 2015.
2. Messinger, M.; Silman, M. Unmanned aerial vehicles for the assessment and monitoring of environmental contamination: An example from coal ash spills. *Environ. Pollut.* **2016**, *218*, 889–894. [CrossRef] [PubMed]
3. Liu, Y.; Meng, Z.; Zou, Y.; Cao, M. Visual Object Tracking and Servoing Control of a Nano-Scale Quadrotor: System, Algorithms, and Experiments. *IEEE/CAA J. Autom. Sin.* **2021**, *8*, 344–360. [CrossRef]
4. Ganesan, R.; Raajini, X.M.; Nayyar, A.; Sanjeevikumar, P.; Hossain, E.; Ertas, A.H. BOLD: Bio-Inspired Optimized Leader Election for Multiple Drones. *Sensors* **2020**, *20*, 3134. [CrossRef]
5. Yayli, U.; Kimet, C.; Duru, A.; Cetir, O.; Torun, U.; Aydogan, A.; Padmanaban, S.; Ertas, A. Design optimization of a fixed wing aircraft. *Int. J. Adv. Aircr. Spacecr. Sci.* **2017**, *4*, 65–80. [CrossRef]
6. Huang, K.W.; Wang, H.M. Combating the Control Signal Spoofing Attack in UAV Systems. *IEEE Trans. Veh. Technol.* **2018**, *67*, 7769–7773. [CrossRef]
7. Li, Y.; Yang, D.; Wang, S.; He, H.; Hu, J.; Liu, H. Road-Network-Based Fast Geolocalization. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6065–6076. [CrossRef]
8. Fragoso, A.T.; Lee, C.T.; McCoy, A.S.; Chung, S.-J. A seasonally invariant deep transform for visual terrain-relative navigation. *Sci. Robot.* **2021**, *6*, eabf3320. [CrossRef] [PubMed]
9. Klein, G.; Murray, D. Parallel Tracking and Mapping for Small AR Workspaces. In Proceedings of the 2007 6th IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Nara, Japan, 13–16 November 2007.
10. 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]
11. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011.
12. Mur-Artal, R.; Tardos, J.D. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. *IEEE Trans. Robot.* **2017**, *33*, 1255–1262. [CrossRef]
13. Campos, C.; Elvira, R.; Rodriguez, J.J.G.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM. *IEEE Trans. Robot.* **2021**, *37*, 1874–1890. [CrossRef]
14. Engel, J.; Koltun, V.; Cremers, D. Direct Sparse Odometry. *IEEE Trans. Pattern Anal.* **2018**, *40*, 611–625. [CrossRef] [PubMed]
15. Qin, T.; Li, P.; Shen, S. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. *IEEE Trans. Robot.* **2018**, *34*, 1004–1020. [CrossRef]
16. Forster, C.; Pizzoli, M.; Scaramuzza, D. SVO: Fast semi-direct monocular visual odometry. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014.
17. Leutenegger, S.; Lynen, S.; Bosse, M.; Siegwart, R.; Furgale, P. Keyframe-based visual–inertial odometry using nonlinear optimization. *Int. J. Robot. Res.* **2015**, *34*, 314–334. [CrossRef]

18. Strasdat, H.; Montiel, J.M.; Davison, A.J. Visual SLAM: Why filter? *Image Vis. Comput.* **2012**, *30*, 65–77. [CrossRef]
19. Sarlin, P.-E.; Unagar, A.; Larsson, M.; Germain, H.; Toft, C.; Larsson, V.; Pollefeys, M.; Lepetit, V.; Hammarstrand, L.; Kahl, F.; et al. Back to the Feature: Learning Robust Camera Localization from Pixels to Pose. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021.
20. Shetty, A.; Gao, G.X. UAV Pose Estimation using Cross-view Geolocalization with Satellite Imagery. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–22 May 2019.
21. Shi, Y.; Li, H. Beyond Cross-view Image Retrieval: Highly Accurate Vehicle Localization Using Satellite Image. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022.
22. Goforth, H.; Lucey, S. GPS-Denied UAV Localization using Pre-existing Satellite Imagery. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019.
23. Chen, S.; Wu, X.; Mueller, M.W.; Sreenath, K. Real-time Geo-localization Using Satellite Imagery and Topography for Unmanned Aerial Vehicles. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021.
24. Hu, S.; Lee, G.H. Image-Based Geo-Localization Using Satellite Imagery. *Int. J. Comput. Vis.* **2020**, *128*, 1205–1219.
25. Kinnari, J.; Verdoja, F.; Kyrki, V. GNSS-denied geolocalization of UAVs by visual matching of onboard camera images with orthophotos. In Proceedings of the 2021 20th International Conference on Advanced Robotics (ICAR), Ljubljana, Slovenia, 6–10 December 2021.
26. Hao, Y.; He, M.; Liu, Y.; Liu, J.; Meng, Z. Range–Visual–Inertial Odometry with Coarse-to-Fine Image Registration Fusion for UAV Localization. *Drones* **2023**, *7*, 540. [CrossRef]
27. Zhang, Y.; Shi, Y.; Wang, S.; Vora, A.; Perincherry, A.; Chen, Y.; Li, H. Increasing SLAM Pose Accuracy by Ground-to-Satellite Image Registration. In Proceedings of the 2024 International Conference on Robotics and Automation (ICRA), Tokyo, Japan, 13–17 May 2024.
28. Lepetit, V.; Moreno-Noguer, F.; Fua, P. EPnP: An Accurate O(n) Solution to the PnP Problem. *Int. J. Comput. Vis.* **2009**, *81*, 155–166. [CrossRef]
29. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1437–1451. [PubMed]
30. DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperPoint: Self-Supervised Interest Point Detection and Description. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018.
31. Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperGlue: Learning Feature Matching with Graph Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
32. Horn, B.K.P. Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Am.* **1987**, *4*, 629. [CrossRef]