Global Map Optimization based Real-time UAV Geolocation without GNSS

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

Posted Date: 9 January 2025

doi: 10.20944/preprints202501.0678.v1

Keywords: UAV autonomous localization; heterogeneous image matching; visual SLAM; graph optimization

Article

Global Map Optimization Based Real-Time UAV Geolocation Without GNSS

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

Xi'an Research Institute of Hi-Tech, Xi'an 710025, China

* Correspondence: liujieyu128@163.com

Abstract: Maintaining real-time and stable geolocation for unmanned aerial vehicles (UAVs) using pre-existing geolocation information remains a pivotal and pressing challenge in the realm of autonomous navigation and localization. To tackle this, we introduce a framework that integrates visual Simultaneous Localization and Mapping (vSLAM) with a deep-learning-based UAV-to-satellite (U2S) image matching technique, specifically tailored for downward-tilted camera setups. This framework incorporates a similarity-based heterogeneous image matching approach, enabling 6 degrees of freedom (DOF) U2S prediction for UAVs. Utilizing this prediction, we devise a single-shot global initialization method for rapid and robust global initialization. Furthermore, to bolster localization accuracy, we propose an iterative map fusion method that integrates U2S predictions with visual data from vSLAM in real-time. Evaluations conducted using aerial datasets demonstrate the proposed method's capability to facilitate rapid and effective global initialization, as well as real-time and accurate estimation of the UAV's geographical pose.

Keywords: UAV autonomous localization; heterogeneous image matching; visual SLAM; graph optimization

1. Introduction

With the advancement of computer vision technology, vision-based UAV navigation and positioning technology has gained significant attention in the field of UAVs [1-4]. UAV visual navigation technology utilizes a visual sensor to acquire visual information for precise positioning, offering advantages such as cost-effectiveness, low power consumption, and robust anti-interference capabilities. Additionally, UAV images captured by vision sensors provide extensive coverage, abundant information content, and high real-time performance, providing a rich and reliable UAV visual navigation technology data source. Furthermore, UAV visual navigation technology can be categorized into vSLAM and localization techniques based on heterogeneous image matching.

VSLAM leverages the inter-frame co-view relationship to estimate the UAV's motion. Davison [5,6] initially proposed a real-time monocular vSLAM called MonoSLAM, which utilizes Shi-Tomasi corner points and an Extended Kalman Filter (EKF). Klein introduced a keyframe-based PTAM [7], selecting only frames with significant co-view relationships for map construction. This SLAM system is divided into two threads: tracking and mapping, which employ nonlinear optimization for the first time. Strasdat [8] demonstrated that the keyframe-based technique outperforms the filter-based technique in terms of accuracy while maintaining the same computational cost. Furthermore, he proposed employing sliding window BA for monocular SLAM in large-scale scenes [9], as well as double-window optimization and covisibility graph SLAM [10]. Building on these ideas, ORB-SLAM [11-13] leverages ORB [14] features for matching, tracking, relocation, and loop closure, enhancing the information exchange efficiency among different threads within the system. Additionally, it utilizes the bag-of-words library DBoW2 [15] to facilitate image retrieval in relocation and loop closure.

Unlike the feature-based method, the direct method does not extract features; instead, it directly utilizes pixel intensity to estimate UAV motion by minimizing photometric error. Engel [16]

 doi:10.20944/preprints202501.0678.v1

proposed LSD-SLAM, which employs high gradient pixels for constructing semi-active maps. Subsequently, Forster introduced a hybrid system SVO [17], which extracts FAST features and utilizes the direct method to track features and pixels with non-zero intensity gradient from frame to frame. Subsequently, Engel [18] introduced the direct sparse odometer (DSO), which accurately calculates camera pose even in scenarios with limited features and enhances robustness in low-texture areas. This approach has been extended to stereo cameras [19], incorporating feature and DBoW2 for loop closure detection [20,21], as well as visual inertia odometry [22]. Zubizarreta [23] proposed a direct method of image reuse called DSM. However, vSLAM fails to provide geographical pose for UAVs. Moreover, prolonged UAV operation leads to cumulative drift issues.

Localization techniques based on heterogeneous image matching utilize image matching methods to establish the feature correspondences between UAV and satellite images, thereby determining the geographical pose of the UAV. Several studies have focused on UAV localization based on heterogeneous image matching [24]. Semantic features can effectively achieve heterogeneous image matching. Nassar [25] employed a semantic shape matching algorithm to extract and align significant shape information between UAV and satellite images, enhancing localization accuracy. Choi [26] extracted the building ratio from the UAV image and correlated it with building information on a numerical map to achieve precise UAV positioning within a specific range. Li [27] introduced a geolocalization technique that relies on road networks and employs global projective invariant features to align UAV images with reference road vector maps, facilitating successful localization across large-scale regions like entire cities. However, this methodology's practical applicability is constrained due to its reliance on semantic features, which refer to roads or buildings that are exclusively present in specific scenes.

With the development of deep learning technology, state-of-the-art image matching networks [28] have emerged as effective solutions to tackle challenges arising from significant variations in appearance caused by seasonal changes, lighting conditions, and other factors between UAV and satellite images. These networks establish feature correspondences between UAV and satellite images, facilitating precise geographical pose estimation for UAVs. Numerous methodologies have been developed for image matching based UAV localization in GPS-denied environments. Goforth [29] employed a deep convolutional neural network (CNN) with an iterative closest keypoint (ICLK) layer to achieve the alignment of UAV and satellite images while refining the geographical pose of all frames through a joint optimization approach that combines odometry and map alignment. Chen [30] proposed an image-based geolocation method for downward-tilted cameras, utilizing MobileNet [31] with a NetVLAD layer for global descriptor extraction. Additionally, SuperPoint [32] is used to extract local descriptors, followed by SuperGlue [33], which facilitates correspondence matching between local descriptors from both UAV and satellite images. Kinnari [34] conducted orthorectification of the UAV images using Visual-Inertial Odometry (VIO) and a planarity assumption, which is also applicable to downward-tilted cameras. The ortho-projected image was then employed for geolocation by matching it with satellite images, and the geolocation results were fused with tracking poses from VIO within a Monte-Carlo localization framework. Hao's geolocation method [35] integrates measurements from RVIO and image registration with global pose graph optimization, obviating the need for any prior information. Patel [36] presents a method for accurately estimating the global pose of a UAV in GPS-denied environments using pre-rendered images from a 3D reconstruction of the Earth, achieving position accuracy comparable to GPS, with particular strength in low-altitude flights and demonstrated robustness throughout the day despite significant lighting changes. Yao [37] present a UAV visual localization system integrating image matching, visual odometry, and terrain-constrained optimization for precise 3D positioning in GNSS-denied environments, day and night, validated through extensive real-world data across diverse terrains. However, a tightly coupled fusion algorithm for VO and geolocation based on heterogeneous image matching for UAV is currently lacking.

 doi:10.20944/preprints202501.0678.v1

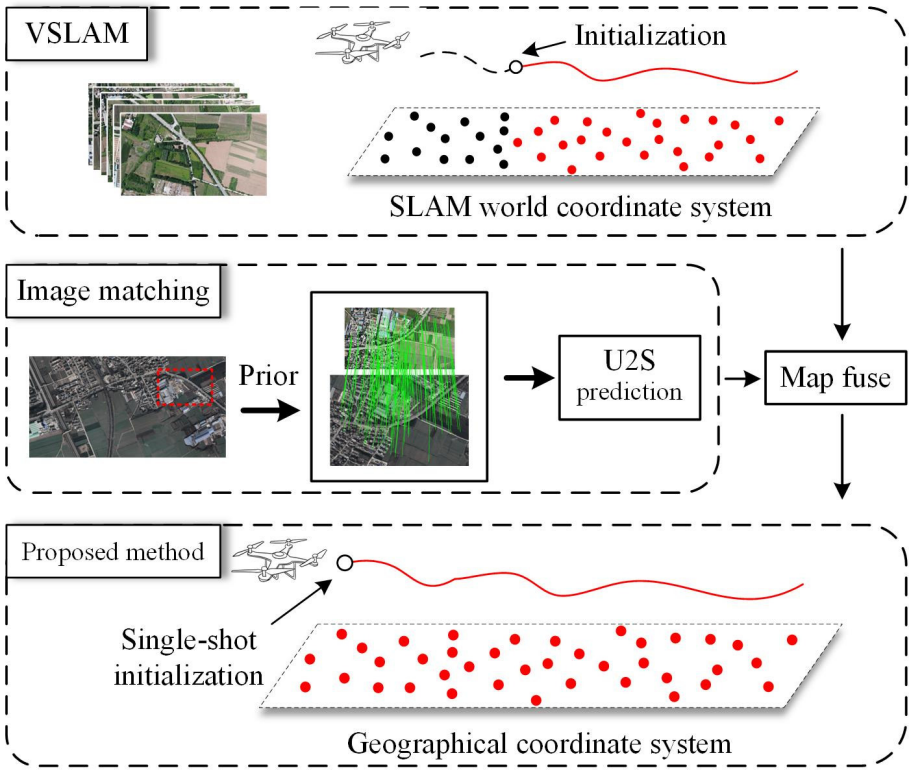


Figure 1. Framework of the proposed method.

As depicted in Figure 1, this paper proposes a novel framework that integrates the advantages of vSLAM with heterologous image matching. In UAV aerial scenarios, vSLAM requires a lengthy initialization process and only provides local localization information. Over prolonged UAV operations, significant cumulative drift is expected to occur. To address these challenges, this study introduces a novel approach called global map optimization based real-time UAV geolocation. This approach facilitates rapid global map initialization and provides more universally applicable geolocation information while effectively mitigating cumulative drift during long-term UAV missions.

In the framework, we propose a framework that utilizes heterologous image matching to provide a 6-DOF U2S prediction for UAVs. To achieve rapid and reliable global map initialization, we introduce a single-shot global initialization method for constructing the glabal map. Once the initialization is completed, to mitigate cumulative drift and enhance localization accuracy, the UAV geolocation method iteratively updates the global map whenever a valid U2S prediction is received.

This work makes the following contributions:

- Real-time UAV Geolocation framework: We propose a framework for real-time UAV geolocation that enables autonomous navigation and positioning at reduced cost, which has the ability to adapt to downward-tilted camera configurations.
- Similarity-Based Heterogeneous Image Matching: This method aligns satellite images with UAV images within a unified coordinate system by employing a homography matrix. Subsequently, it performs image matching between the aligned satellite and UAV images. Ultimately, similarity transformations are utilized to constrain the degrees of freedom in pose estimation. This comprehensive process enables precise U2S prediction for UAVs.
- Global Map Fusion Method: This method leverages U2S predictions derived from heterologous image matching and visual information sourced from vSLAM to rapidly construct and refine a global map that is precisely aligned with the geographic coordinate system. This method enables real-time estimation of UAVs' geographic poses.

The rest of this paper is structured as follows. In Section 2, we describe our proposed method; In Section 3, we describe experiments on datasets. The conclusions are summarized in Section 4.

 doi:10.20944/preprints202501.0678.v1

2. Methodology

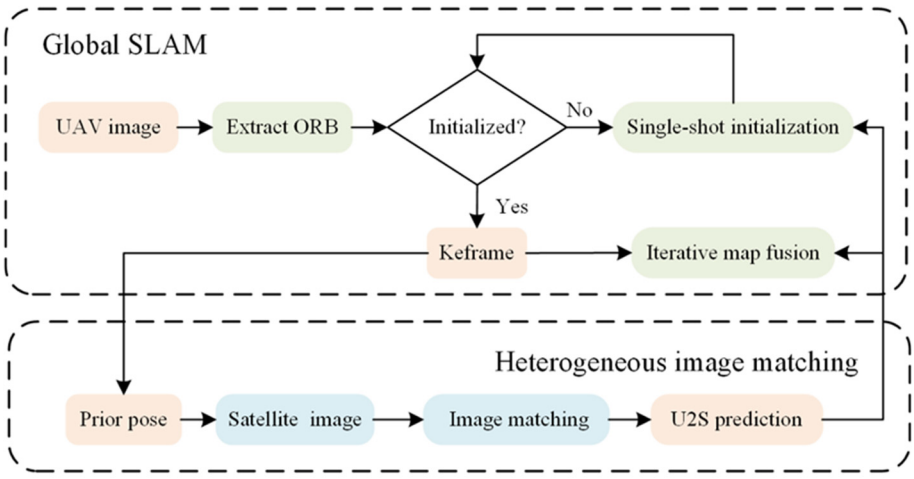


Figure 2. Flowchart of the proposed geolocation method.

The flowchart of the geolocation is illustrated in Figure 2. In this section, we present our methodology, which is primarily divided into two parts: similarity-based heterogeneous image matching and the global map fusion method.

2.1. Similarity-Based Heterogeneous Image Matching

To acquire geographical information, we employ a localization method based on heterogeneous image matching to provide geolocation data for UAV. The input for this localization method is derived from UAV images and their corresponding prior poses T_c^g obtained through SLAM algorithms. Leveraging the high accuracy of SLAM pose estimation during short-term operations, we extract satellite images in close proximity to the UAV images using prior pose. Subsequently, feature correspondences between UAV and satellite images are identified using Superpoint + SuperGlue. A similarity transformation matrix S_c is then computed and converted into a pose transformation matrix $\tilde{T}_c^g$ for the UAV. This pose transformation matrix, referred to as U2S prediction, provides crucial geolocation prediction information for the proposed geolocation method. Further details can be found in Algorithm 1.

Algorithm 1 Localization Based on Heterogeneous Image Matching

Input: prior pose $\hat{T}_c^g$, UAV image I_U and satellite map tiles.
Output: U2S prediction $\tilde{T}_c^g$.

$\hat{H}_c^g \leftarrow$ Calculate homography transformation matrix.
$Geo\_bound \leftarrow$ Calculate geographic boundary of satellite image.
$I_S^g \leftarrow$ Capture and splice satellite image.
$I_S = \hat{H}_c^g I_S^g$.
$Pairs = Match\_image\,(I_U, I_S)$.
if $length\,(Pairs) \geq Match\_threshold$ **then**
    $\left(H_g^c, inlier\right) = EstimateAffine\,(Pairs, method = RANSAC)$.
    if $length\,(inlier) \geq Match\_threshold$ **then**
        $\tilde{T}_c^g = EstimatePose\left(H_g^c\right)$.
        Consistency check of $\tilde{T}_c^g$.
    end if
end if

 doi:10.20944/preprints202501.0678.v1

2.1.1. Offline Preprocessing

The satellite maps and Digital Elevation Model (DEM) with a 30m resolution pertinent to the UAV's mission area are acquired from Google Earth. These satellite images are subsequently transformed into the pseudo-Mercator projection coordinate system utilizing QGIS, and are then cropped into map tiles with dimensions of $W \times H$ pixels for efficient storage. Simultaneously, the geographical boundaries $[g_{west}, g_{east}, g_{south}, g_{north}]$ and the ground resolution px of these map tiles are stored. Consequently, the formula for calculating the geographical coordinate (g_x, g_y) of a pixel (x, y) within these map tiles is as follows:

$$\begin{cases} g_x = g_{west} + x \times px \\ g_y = g_{north} - y \times px \end{cases}, \tag{1}$$

The proposed algorithm can extract the satellite image that corresponds to the demand by utilizing formula (1).

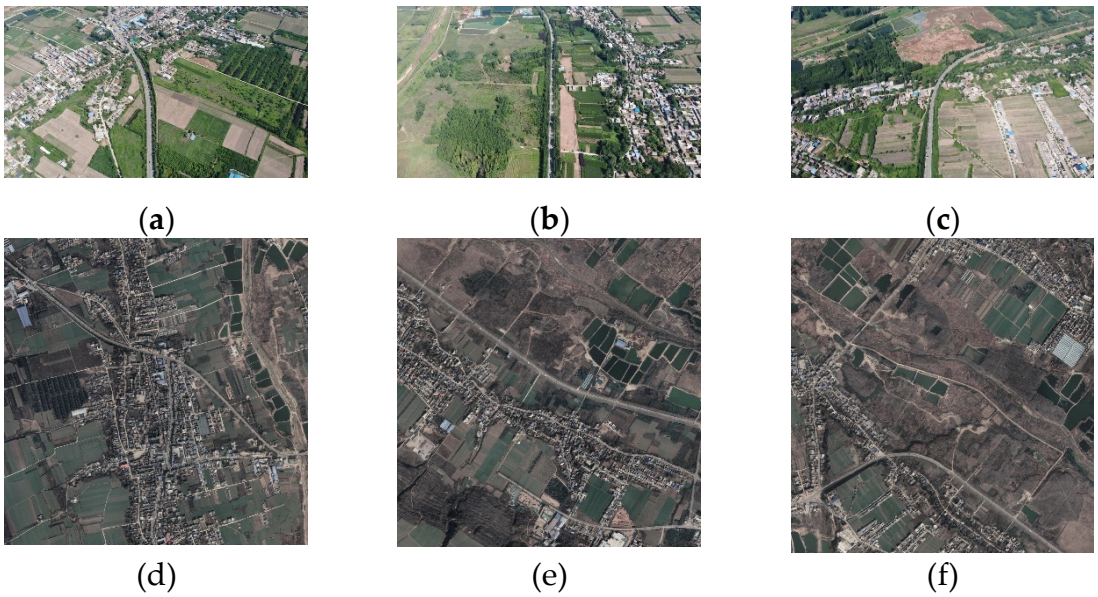
2.1.2. Extraction and Transformation of Satellite Images

Given the prior pose $\hat{T}_c^g$ of the UAV in the geographic coordinate system, its rotation and translation components are denoted as $\hat{R}_c^g$ and $\hat{t}_c^g$. The transformation relationship between the geographic plane coordinate system and the pixel coordinate system in the UAV image can be described by the homography matrix $\hat{H}_c^g$.

$$\hat{H}_c^g = \frac{1}{\lambda} K \left[\hat{R}_c^g e_1, \hat{R}_c^g e_2, \hat{t}_c^g \right], \tag{2}$$

Where $e_1 = [1,0,0]^T$ and $e_2 = [0,1,0]^T$, K represents the camera intrinsic parameter while λ denotes the depth of the feature point.

The pixel boundary of the UAV image is transformed into the geographical plane coordinate system utilizing the homography matrix $\hat{H}_c^g$. Following this transformation, a satellite image is acquired and subsequently converted to the pixel coordinate system of the UAV image, enabling precise image matching. Extraction and transformation of satellite images is presents in Figure 3.



(a) (b) (c)

(d) (e) (f)

 doi:10.20944/preprints202501.0678.v1

(g) (h) (i)

Figure 3. Extraction and transformation of satellite images. (a) ~ (c) represent UAV images，(d) ~ (f) depict the extracted satellite images，(g)~ (i) illustrate the transformed satellite images.

2.1.3. Computing U2S Prediction

The Superpoint + SuperGlue method extracts feature correspondences $\left\{\hat{p}_i^g, p_i^c\right\}_N$ between UAV and the transformed satellite images, where $\hat{p}_i^g = \left(\hat{x}_i^g, \hat{y}_i^g\right)^T$ represents the pixel coordinate of the feature i in the transformed satellite image, and $p_i^c = \left(u_i^c, v_i^c\right)^T$ denotes the pixel coordinate of the corresponding feature in the UAV image.

To enhance the accuracy of UAV geolocation and mitigate the impact of outliers, the feature correspondence between the UAV and the transformed satellite images can be approximated through the application of a 2D similarity transformation. Hence, the proposed algorithm employs $\left\{\hat{p}_i^g, p_i^c\right\}_N$ to compute the similarity transformation matrix, facilitating the mapping of pixel coordinates from UAV images onto the corresponding pixel coordinates in the transformed satellite images.

$$S_c = \begin{bmatrix} s\cos\theta & -s\sin\theta & t_x \\ s\sin\theta & s\cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

Where s, θ, t_x and t_y are unknown parameters. Let:

$$x_A = \begin{bmatrix} s\cos\theta & s\sin\theta & t_x & t_y \end{bmatrix}^T, \tag{4}$$

The parameters of S_c can be determined using the least squares method and the RANSAC algorithm is employed to remove outliers.:

$$\begin{bmatrix} u_c^i & -v_c^i & 1 & 0 \\ u_c^i & v_c^i & 0 & 1 \end{bmatrix} x_A = \begin{bmatrix} x_g^i \\ y_g^i \end{bmatrix}, \tag{5}$$

Provided that the prior pose information is accurate, the transformed satellite image will achieve precise alignment with the UAV image, with the resulting S_c being an identity matrix. However, it is often the case that the prior pose information contains errors, necessitating the updating of the homography matrix H_c^g:

$$H_c^g = S_c^{-1}\hat{H}_c^g, \tag{6}$$

The homography matrix H_c^g describes the precise transformation relationship between the geographic plane coordinate system and the pixel coordinate system in the UAV image. In satellite imagery, a uniform selection of 2D points $\left\{x_i^g, y_i^g\right\}_n$ is made. Their elevations $\left\{z_i^g\right\}_n$ are then

 doi:10.20944/preprints202501.0678.v1

obtained from a DEM using spatial interpolation. By employing the homography matrix H_c^g, corresponding matching points $\{x_i^c, y_i^c\}_n$ in the UAV image are determined. Consequently, a set of 3D-to-2D feature correspondences is established. Subsequently, the UAV's geographical pose $\tilde{T}_c^g$, known as U2S prediction, is calculated using the EPnP+RANSAC algorithm [38]. Compared to directly applying EPnP+RANSAC for pose estimation, incorporating a similarity transformation constrains the freedom of U2S prediction, leading to improved accuracy.

2.1.4. Consistency Check of U2S Prediction

The Superpoint + SuperGlue effectively addresses the significant variations among heterogeneous images and establishes feature correspondences. However, it may encounter situations where mismatches occur. To overcome this challenge, this paper designs a consistency check method to eliminate false U2S predictions by utilizing SLAM information.

To ensure consistency of U2S prediction, we examine whether the differences in rotation and translation between the input prior pose $\hat{T}_c^g$ and output U2S prediction $\tilde{T}_c^g$ are below a specified threshold. Let C_r and C_t denote the selected U2S orientation and translation respectively. We have:

$$\begin{cases} C_r = \left\{ \tilde{R}_c^g \middle| \arccos\left(\dfrac{\Delta r_{11} + \Delta r_{22} + \Delta r_{33} - 1}{2} \right) < th_\theta \right\} \\ C_t = \left\{ \tilde{t}_c^g \middle| \left\| \left(\tilde{t}_g^c - \hat{t}_g^c \right) \right\| < th_t \right\} \end{cases}, \tag{7}$$

Where Δr_{11}, Δr_{22} and Δr_{33} are the diagonal elements of $\Delta R = \left(\hat{R}_c^g\right)^T \tilde{R}_c^g$, $t_g^c = -R_c^g t_c^g$ and $\tilde{t}_g^c = -\tilde{R}_c^g \tilde{t}_c^g$.

2.2. Global Map Fusion Method

Algorithm 2 Global map fusion method

Input: Initial pose $initT_c^g$, UAV images I_U and satellite map tiles.
Output: Estimated UAV trajectory.

for all UAV images **do**
  Extract ORB features
  if !Initialized **then**
    $\left(H_{c,1}^g, \tilde{T}_{c,1}^g\right) = Image\_Matching\left(I_U^1, initT_c^g\right)$.
    $P_j^g = \left(H_{c,1}^g\right)^{-1} p_j^c$ ←Calculate map points.
    $T_{c,1}^g = \tilde{T}_{c,1}^g$ ←Calculate UAV pose.
    Reconstruct global map.
  else
    Track UAV movement.
    if keyframe **then**
      $\tilde{T}_{c,i}^g = Image\_Matching\left(I_U^i, T_{c,i}^g\right)$.
      Solve the nonlinear least squares problem (9).
      Update global map.
    end if
  end if
end for

 doi:10.20944/preprints202501.0678.v1

ORB-SLAM3[14] is an advanced keyframe-based visual navigation algorithm with a front end for extracting features and tracking camera movement, and a back end for optimizing maps. To reduce computational costs, the front end extracts keyframes based on the co-view relationship between images and the running state of the back end, which are then utilized for map optimization in the back-end. Based on the concept of keyframes, we propose a global map fusion method consisting of single-shot global initialization and iterative global map fusion. More details are shown in Algorithm 2.

2.2.1. Single-Shot Global Initialization

The objective of global initialization is to construct a global map comprising UAV poses and 3D map points. Conventional vSLAM initialization resolves the relative pose of the first two frames by decomposing the fundamental matrix or homography matrix but exhibits certain limitations in UAV aerial scenes: (1) The distance between the UAV and the ground makes it challenging to ensure sufficient parallax between consecutive frames. (2) Under conditions such as cloud cover or blurry imaging, ensuring an adequate number of matching features between adjacent frames becomes difficult. Consequently, vSLAM initialization may be delayed significantly when operating in the UAV aerial scene. (3) The vSLAM-generated map is established within a local coordinate system rather than a universal geographic coordinate system. To address these issues, we propose a single-shot initialization method that utilizes geolocation information obtained from heterogeneous image matching to facilitate fast and robust global initialization while constructing the global map within the geographic coordinate system.

Upon capturing the first image, the satellite image is obtained using a pre-set UAV pose and heterogeneous image matching is employed to obtain feature correspondences between the UAV and satellite images. The homography matrix $H_{c,1}^{g}$ and U2S prediction $H_{c,1}^{g}$ are then calculated, followed by the conversion of ORB features $\left\{\left(u_j^c, v_j^c\right)^T\right\}_N$ from the UAV image to geographic coordinates using $H_{c,1}^{g}$:

$$\begin{pmatrix} x_j^g \\ y_j^g \\ 1 \end{pmatrix} = \left(H_{c,1}^{g}\right)^{-1} \begin{pmatrix} u_j^c \\ v_j^c \\ 1 \end{pmatrix}, \tag{8}$$

Elevations $\left\{z_i^g\right\}_n$ of $\left\{\left(x_j^g, y_j^g\right)^T\right\}_N$ are then obtained from a DEM using spatial interpolation. The global map, consisting of UAV pose $T_{c,1}^{g} = \tilde{T}_{c,1}^{g}$ and map points $\left\{\left(x_j^g, y_j^g, z_j^g\right)^T\right\}_N$, is obtained at this stage. During the global map construction, only the initial UAV image is utilized, enabling rapid global initialization.

2.2.2. Iterative Global Map Fusion

Using the single-shot global initialization method, we obtain a global map in the geographic coordinate system. Subsequently, the proposed geolocation method estimates the pose of other frames using the PnP technique. To enhance the robustness of the global map and suppress cumulative drift, we adopt an iterative global map fusion method to merge global maps with U2S predictions.

After initialization, the proposed geolocation method selects keyframes based on the co-view relationship between images and the operational state of the back-end. The U2S prediction of the most recent keyframe is calculated based on the operational status of the heterogeneous image matching thread, and these specific keyframes are referred to as geo-keyframes.

 doi:10.20944/preprints202501.0678.v1

The proposed geolocation method optimizes the global map through a graph optimization method whenever a new geo-keyframe is generated. Given the poses of N keyframes, denoted as $\overline{T}_N \doteq \left\{T_{c,0}^g \cdots T_{c,N-1}^g\right\}$, and the U2S prediction of n geo-keyframes, denoted as $\overline{T}_n^g \doteq \left\{\tilde{T}_{c,0}^g \cdots \tilde{T}_{c,n-1}^g\right\}$. Additionally, it incorporates l map points observed by these keyframes and their 3D locations $\overline{P}_l^g \doteq \left\{P_0^g \cdots P_{l-1}^g\right\}$. the cost function is as follows:

$$\left\{T_{c,i}^g, P_j^g \mid T_{c,i}^g \in \overline{T}_N, P_j^g \in \overline{P}_l^g\right\} = \arg\min_{T_{c,i}^g, P_j^g} \sum_{i=0}^{N-1}\sum_{j=0}^{l-1} \rho\left(\left\|r_{ij}^c\right\|_{\Sigma_c}^2\right) + \sum_{k=0}^{n-1} \rho\left(\left\|r_k^g\right\|_{\Sigma_g}^2\right), \tag{8}$$

Where $\rho(\bullet)$ is robust kernel function and r_{ij}^v is visual reprojection error [14]:

$$r_{ij}^v = \hat{x}_j - \pi\left(T_{c,i}^g P_j^g\right), \tag{9}$$

Where $\hat{x}_j$ is the observed value of the map points P_j^g, is camera projection function.

In equation (8), r_k^g is geographical residual:

$$r_k^g = Log\left(\left(\tilde{T}_{c,k}^g\right)^{-1} T_{c,k}^g\right), \tag{10}$$

Where represents the mapping from the lie group to the tangent space.

In graph optimization, calculating the jacobian matrix of the residual for the variable is essential, playing a crucial role in achieving efficient and accurate optimization of UAV pose and map point locations. The jacobian matrix of the visual residual concerning the optimization variables can be obtained from [13]. The jacobian matrix of geographical residuals r_k^g for optimization variables $T_{c,k}^g$ is presented as follows:

$$\frac{\partial r_k^g}{\partial \delta_k} = Adj\left(\left(\tilde{T}_{c,k}^g\right)^{-1}\right), \tag{11}$$

Where $\delta_k = Log\left(T_{c,k}^g\right)$ and $Adj(T)$ is the adjoint matrix of T:

$$Adj(T) = \begin{bmatrix} R & t_\times R \\ 0 & R \end{bmatrix}, \tag{12}$$

Where $t_\times$ is the skew-symmetric matrix of t.

The proposed geolocation method utilizes iterative global map fusion to enhance the accuracy and robustness of the global map, enabling real-time and accurate estimation of UAV geographic pose. Consequently, it significantly enhances the localization performance of the proposed geolocation method.

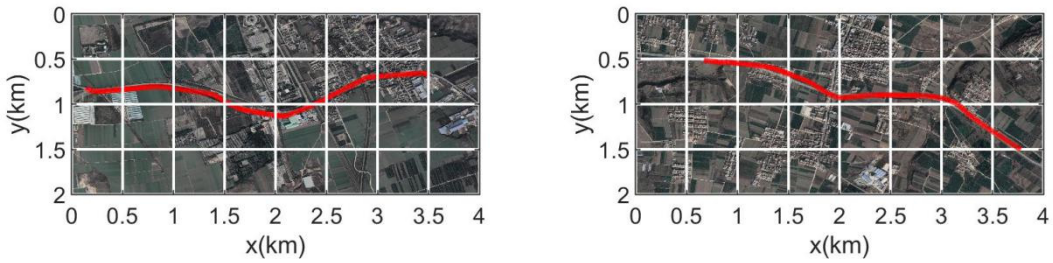
3. Experimental Setups and Results

In this section, the evaluations using real datasets are presented. We first introduce the experimental setup and evaluation metrics. The experiment results are then presented followed by an ablation study showing the effectiveness of each module in the framework.

3.1. Experimental Setup

Given the absence of publicly available UAV aerial datasets, we curated two datasets to assess the effectiveness of the proposed algorithm. A DJI M300 RTK UAV equipped with an H20 pan-tilt camera was employed to capture UAV images and obtain precise RTK trajectories.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2025 doi:10.20944/preprints202501.0678.v1



(**a**) dataset 1 (**b**) dataset 2

Figure 4. The trajectories of UAV in datasets.

Additionally, we downloaded the satellite map and DEM with a 30m resolution from Google Earth based on the UAV's trajectory and subsequently cropped the satellite map into $1k \times 1k$ map tiles with a ground resolution of $0.6 m/pixel$. Figure 4 illustrates the reference trajectories of the UAV on the satellite map, while Table 1 presents the characteristic of both datasets, Specifically, Dataset 1 consists of oblique imagery, and Dataset 2 comprises nadir imagery.

Table 1. Characteristics of datasets.

Dataset	Length (km)	Altitude (m)	Duration (s)	Pitch (◦)	Frame rate (fps)
1	3.56	600	300	60~80	20
2	3.40	520	300	90	20

3.2. Evaluation Metrics

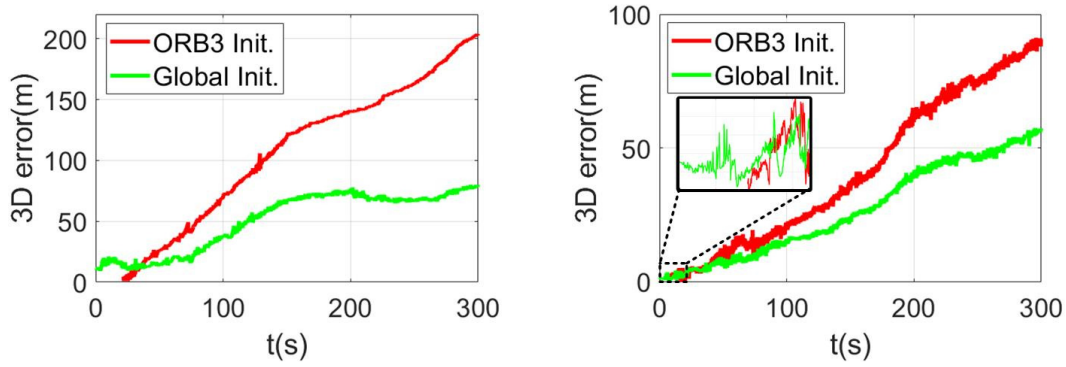
We evaluate the localization performance of different methods based on their 3D localization accuracy. The absolute max, mean, and root-mean-square error (RMSE) of translation are used to measure the error between the estimated trajectory and the ground truth [39]:

$$\delta T = T_{GT}^{-1} S T \quad , \tag{13}$$

Where T and T_{GT} are the estimated poses and the ground truth. S is the rigid body transformation mapping the estimated trajectory to the ground truth [40]. To obtain S, a least-squares problem is solved using multiple ground truth poses [41].

3.3. Global Initialization Experiment

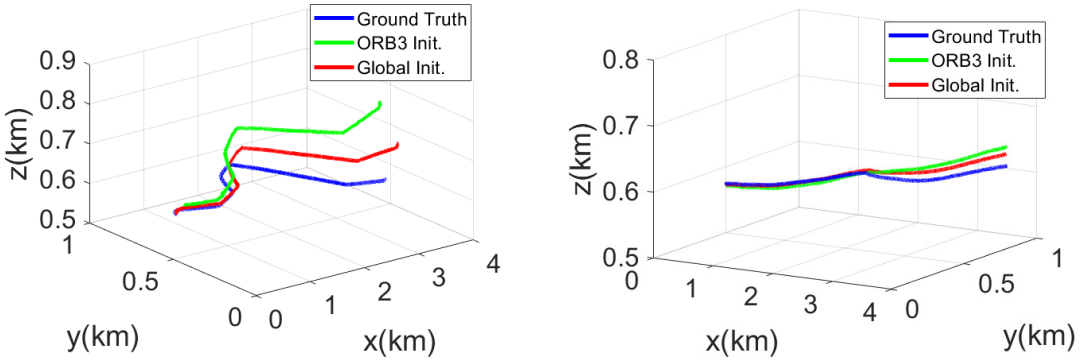
To verify the effectiveness of the proposed single-shot initialization method, we compare it with the visual initialization of ORB-SLAM3. Subsequently, UAV movement is tracked using the SLAM technique.



(**a**) dataset 1 (**b**) dataset 2

 doi:10.20944/preprints202501.0678.v1

Figure 5. Error of different methods in the initialization experiment.



(**a**) dataset 1 (**b**) dataset 2

Figure 6. Trajectories of different methods in the initialization experiment.

Figure 5 compares the 3D errors associated with various initialization methods, while Figure 6 demonstrates the estimated trajectories of these methods alongside the ground truth. Table 2 presents a comparison of the errors among different methods in the initialization experiment using aerial datasets. In contrast to the visual initialization of ORB-SLAM3, the proposed single-shot global initialization method eliminates the need for frame selection, thereby enabling the rapid construction of a global map. The time cost of this method in our experiments was 0.65s, significantly lower than that of the visual initialization of ORB-SLAM3. Furthermore, the global map constructed using our method exhibits robustness comparable to that of visual maps built by vSLAM, maintaining high accuracy during subsequent operations. Therefore, the proposed single-shot global initialization method efficiently accomplishes the initialization process and constructs a more robust global map. It is worth noting that, since the localization accuracy of heterogeneous image matching directly influences the robustness of the global map created through single-shot global initialization, the localization errors of the UAV at its initial position were measured as 10.78m and 1.57m in two aerial datasets during the experiment.

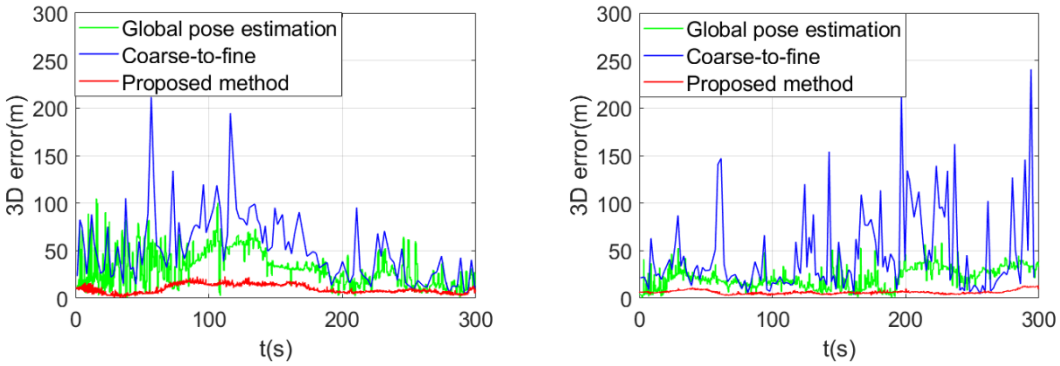
Table 2. Error comparison in initialization experiment.

Dataset	Method	Max (m)	Mean (m)	RMSE (m)	Time cost (s)
1	ORB3 Init.	103.94	50.25	64.01	11.70
	Global Init.	40.07	30.89	28.25	0.65
2	ORB3 Init.	91.22	41.99	50.65	4.85
	Global Init.	57.53	27.13	32.30	0.65

3.4. UAV Localization Experiment

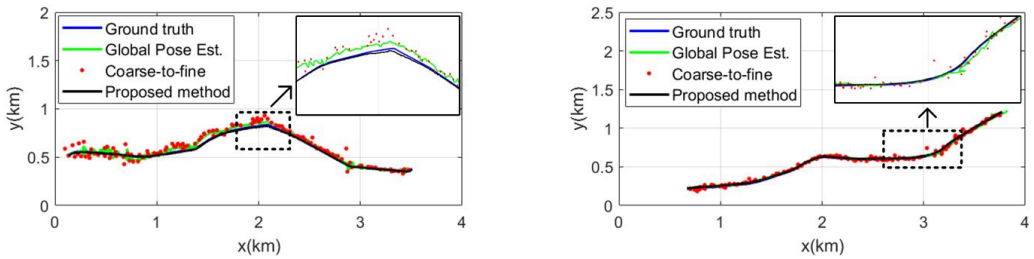
To assess the efficacy of the proposed geolocation method in UAV aerial scenes, we conducted UAV localization experiments using two aerial datasets, and compared the localization performance of our method against other vision-based localization method, including global pose estimation method [37] and coarse-to-fine geolocation method [31]. Both of these methods are vision-based advanced approaches for UAV geolocation. The global pose estimation method employs pre-rendered satellite imagery from known poses and UAV images for image matching. It integrates image matching with VO within a filtering framework to achieve precise geolocation of UAVs. On the other hand, the coarse-to-fine image matching method utilizes a combination of image retrieval and image matching to enable UAV geolocation without prior information. Notably, both localization methods are suitable for oblique-view UAV imagery.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2025 doi:10.20944/preprints202501.0678.v1



(**a**) dataset 1 (**b**) dataset 2

Figure 7. Error of different methods in UAV localization experiment.



(**a**) dataset 1 (**b**) dataset 2

Figure 8. Trajectories of different methods in UAV localization experiment.

Figure 7 compares the 3D localization errors of various methods in the UAV localization experiment, whereas Figure 8 demonstrates the estimated trajectories obtained by different methods alongside the ground truth. Table 3 presents a comparison of the mean and RMSE values achieved by various methods in UAV localization experiments utilizing aerial datasets. The experimental results indicate that the proposed method in this paper outperforms other approaches in terms of accuracy. During the image matching process, we ingeniously designed a similarity-based heterogeneous image matching algorithm, which significantly enhances the localization accuracy of image matching. Additionally, by constructing and continuously optimizing a global map, we achieve more precise geolocation of the UAV, capable of delivering an output frequency of 20Hz, thereby fulfilling the real-time localization requirements of UAVs.

Table 3. Error comparison in UAV localization experiment.

Dataset	Method	Max (m)	Mean (m)
1	Global pose estimation	31.03	36.52
	Coarse-to-fine	51.79	64.51
	Proposed Method	9.81	10.85
2	Global pose estimation	21.46	24.10
	Coarse-to-fine	31.29	46.83
	Proposed Method	6.32	6.63

3.5. Ablation Study

We conduct an ablation study to evaluate the contribution of each module in the proposed framework. The translation errors by removing different configurations are present in Tab. 4. The results show that using the full proposed framework achieves the best performance.

 doi:10.20944/preprints202501.0678.v1

Table 4. Ablation study on different configurations.

Dataset	Metric	Single-init	EPnP-U2S	Aff-U2S	Conc-U2S	Traj-fusion	Full
1	Mean	50.89	20.34	13.03	12.83	13.37	9.81
	RMSE	56.25	21.88	13.36	13.98	14.63	10.85
2	Mean	27.13	15.76	7.93	7.27	7.36	6.32
	RMSE	32.30	16.13	8.31	7.95	8.02	6.63

Single-init: SLAM with Single-shot global initialization；EPnP-U2S：using EpnP to calculate U2S poses; Aff-U2S: using Affine transformation to calculate U2S poses; Conc-U2S: using consistency to check U2S poses; Traj-fusion: optimizing only trajectory in U2S-SLAM fusion; Full: results using all proposed modules.

4. Conclusions

This paper introduces a comprehensive framework for enhancing UAV geolocation accuracy in aerial scenarios by integrating vSLAM with U2S prediction. Utilizing a heterogeneous image matching technique, U2S poses are forecasted and then validated for consistency. By harnessing U2S predictions, our method facilitates single-shot global initialization and iterative map fusion. The former allows for the rapid construction of a robust global map, while the latter integrates visual information from SLAM with globally drift-free geographic data derived from image matching. This integration enables real-time optimization of the global map, effectively mitigating cumulative drift in vSLAM. Experimental results on aerial datasets underscore the proposed method's proficiency in accurately estimating UAV geographical poses in real-time.

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. and Y.L.; writing, original draft preparation, W.X.; writing, review and editing, W.X.; visualization, W.X.; supervision, D.Y. and J.L.; project administration, D.Y. and J.L.; funding acquisition, D.Y. All authors 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).

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 conflict 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, doi:10.1016/j.envpol.2016.08.019.
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, doi:10.1109/JAS.2020.1003530.
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.
5. Davison, A.J.; Reid, I.D.; Molton, N.D.; Stasse, O. MonoSLAM: Real-Time Single Camera SLAM. *IEEE T. Pattern. Ansl.* **2007**, *29*, 1052-1067, doi:10.1109/TPAMI.2007.1049.
6. Civera, J.; Davison, A.J.; Montiel, J.M.M. Inverse Depth Parametrization for Monocular SLAM. *IEEE T. Robot.* **2008**, *24*, 932-945, doi:10.1109/TRO.2008.2003276.

7. 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.
8. A, H.S.; B, J.M.M.M.; A, A.J.D. Visual SLAM: Why filter? *Image. Vision. Comput.* **2012**, *30*, 65-77.
9. Strasdat, H.M.; Montiel, J.M.M.; Davison, A.J. Scale Drift-Aware Large Scale Monocular SLAM. In Proceedings of the Robotics: Science and Systems, 27 June 2010.
10. Strasdat, H.; Davison, A.J.; Montiel, J.M.M.; Konolige, K. Double window optimisation for constant time visual SLAM. In Proceedings of the 2011 International Conference on Computer Vision, 6-13 November 2011.
11. Mur-Artal, R.; Montiel, J.M.M.; Tardós, J.D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE T. Robot.* **2015**, *31*, 1147-1163, doi:10.1109/TRO.2015.2463671.
12. Mur-Artal, R.; Tardós, J.D. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE T. Robot.* **2017**, *33*, 1255-1262, doi:10.1109/TRO.2017.2705103.
13. Campos, C.; Elvira, R.; Rodríguez, J.J.G.; Montiel, J.M.M.; Tardós, J.D. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. *IEEE T. Robot.* **2021**, *37*, 1874-1890, doi:10.1109/TRO.2021.3075644.
14. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G.R. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, 6 November 2011.
15. Galvez-López, D.; Tardos, J.D. Bags of Binary Words for Fast Place Recognition in Image Sequences. *IEEE T. Robot.* **2012**, *28*, 1188-1197, doi:10.1109/TRO.2012.2197158.
16. Engel, J.J.; Schöps, T.; Cremers, D. LSD-SLAM: Large-Scale Direct Monocular SLAM. In Proceedings of the European Conference on Computer Vision, 6 September 2014.
17. Forster, C.; Zhang, Z.; Gassner, M.; Werlberger, M.; Scaramuzza, D. SVO: Semidirect Visual Odometry for Monocular and Multicamera Systems. *IEEE T. Robot.* **2017**, *33*, 249-265, doi:10.1109/TRO.2016.2623335.
18. Engel, J.; Koltun, V.; Cremers, D. Direct Sparse Odometry. *IEEE. T. Pattern. Anal.* **2018**, *40*, 611-625.
19. Wang, R.; Schwörer, M.; Cremers, D. Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), 22-29 October 2017.
20. Gao, X.; Wang, R.; Demmel, N.; Cremers, D. LDSO: Direct Sparse Odometry with Loop Closure. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 01-05 October 2018.
21. Lee, S.H.; Civera, J. Loosely-Coupled Semi-Direct Monocular SLAM. *IEEE Robot. Autom. Let.* **2019**, *4*, 399-406.
22. Stumberg, L.V.; Usenko, V.; Cremers, D. Direct Sparse Visual-Inertial Odometry Using Dynamic Marginalization. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), 21-25 May 2018.
23. Zubizarreta, J.; Aguinaga, I.; Montiel, J.M.M. Direct Sparse Mapping. *IEEE T. Robot.* **2020**, *36*, 1363-1370, doi:10.1109/TRO.2020.2991614.
24. Couturier, A.; Akhloufi, M.A. A review on absolute visual localization for UAV. *Robot. Auton. Syst.* **2021**, *135*, 103666, doi:https://doi.org/10.1016/j.robot.2020.103666.
25. Nassar, A.; Amer, K.; ElHakim, R.; ElHelw, M. A Deep CNN-Based Framework For Enhanced Aerial Imagery Registration with Applications to UAV Geolocalization. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 18-22 June 2018.
26. Choi, J.; Myung, H. BRM Localization: UAV Localization in GNSS-Denied Environments Based on Matching of Numerical Map and UAV Images. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 24 Oct.-24 Jan. 2021, 2020.
27. Li, Y.; Yang, D.; Wang, S.; He, H.; Hu, J.; Liu, H. Road-Network-Based Fast Geolocalization. *IEEE T. Geosci. Remote.* **2021**, *59*, 6065-6076, doi:10.1109/TGRS.2020.3011034.
28. Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; Zhou, X. LoFTR: Detector-Free Local Feature Matching with Transformers. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20-25 June 2021.

29. 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), 20-24 May 2019.
30. 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), 27 September - 01 October 2021.
31. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18-23 June 2018.
32. 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), 18-22 June 2018.
33. 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), 13-19 June 2020.
34. 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), 06-10 December 2021.
35. 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.
36. Patel, B.; Barfoot, T.D.; Schoellig, A.P. Visual Localization with Google Earth Images for Robust Global Pose Estimation of UAVs. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), 31 May 2020 - 31 August 2020.
37. Yao, F.; Lan, C.; Wang, L.; Wan, H.; Gao, T.; Wei, Z. GNSS-denied geolocalization of UAVs using terrain-weighted constraint optimization. *Int. J. Appl. Earth. Obs.* **2024**, *135*, 104277, doi:https://doi.org/10.1016/j.jag.2024.104277.
38. Lepetit, V.; Moreno-Noguer, F.; Fua, P. EPnP: An Accurate O(n) Solution to the PnP Problem. *Int. J. Comput. Vision.* **2009**, *81*, 155-166, doi:10.1007/s11263-008-0152-6.
39. Sarlin, P.-E.; DeTone, D.; Yang, T.-Y.; Avetisyan, A.; Straub, J.; Malisiewicz, T.; Bulò, S.R.; Newcombe, R.A.; Kontschieder, P.; Balntas, V. OrienterNet: Visual Localization in 2D Public Maps with Neural Matching. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 17-24 June **2023**.
40. Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; Cremers, D. A benchmark for the evaluation of RGB-D SLAM systems. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 7-12 October 2012.
41. Horn, B.K.P. Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Am. A.* **1987**, *4*, 629-642.