IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 10, 2020, accepted December 28, 2020, date of publication December 31, 2020,
date of current version January 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3048342

UAV Pose Estimation in GNSS-Denied Environment Assisted by Satellite Imagery Deep Learning Features

HUITAI HOU[1], QING XU[1], CHAOZHEN LAN[1], WANJIE LU[1], YONGXIAN ZHANG[2], ZHIXIANG CUI[3], AND JIANQI QIN[1]

[1]Information Engineering University, Zhengzhou 450052, China
[2]Wuhan University, Wuhan 430000, China
[3]31682 Troops, Lanzhou 730020, China

Corresponding author: Chaozhen Lan (lan_cz@163.com)

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 42001338.

ABSTRACT With the growing maturity of unmanned aerial vehicle (UAV) technology, its applications have widened to many spheres of life. The prerequisite for a UAV to perform air tasks smoothly is an accurate localization of its own position. Traditional UAV navigation relies on the Global Navigation Satellite System (GNSS) for localization; however, this system has disadvantages of instability and susceptibility to interference. Therefore, to obtain accuracy in UAV pose estimation in GNSS-denied environments, a UAV localization method that is assisted by deep learning features of satellite imagery is proposed. With the inclusion of a top-view optical camera to the UAV, localization is achieved based on satellite imageries with geographic coordinates and a digital elevation model (DEM). By utilizing the difference between the UAV frame and satellite imagery, the convolutional neural network is used to extract deep learning features between the two images to achieve stable registration. To improve the accuracy and robustness of the localization method, a local optimization method based on bundle adjustment (BA) is proposed. Experiments demonstrate that when the UAV's relative altitude is 0.5 km, the average localization error of this method under different trajectories is within 15 m.

INDEX TERMS Bundle adjustment, deep learning, localization, satellite imagery, unmanned aerial vehicle.

I. INTRODUCTION

With the growing maturity of unmanned aerial vehicle (UAV) technology, UAVs are now used in various military and civilian fields, including intelligence reconnaissance, military strikes, search and rescue, land surveying and mapping, precision agriculture, and environmental monitoring [1]–[4]. Similar to other types of robots, accurate localization of their position is the prerequisite for UAVs to perform tasks smoothly [5]. Traditional navigation technology relies on GNSS; however, it has disadvantages, such as instability and susceptibility to interference [6]. Carroll [7] and Caballero *et al.* [8] pointed out that the number of satellites and signal quality received by GNSS are important for calculating the position of a UAV. Radio effects such as very few satellites or multipath propagation will cause the degradation of position estimation. Conte and Doherty [9] and Viswanathan *et al.* [10] believe that GNSS jammers will cause signal interference, making position estimation unreliable. UAVs rely only on GNSS signals that may be maliciously interfered with, which will seriously affect mission execution and even cause catastrophic consequences.

When GNSS cannot be used, additional airborne sensors are required to assist navigation. Compared with laser rangefinders, optical cameras are reduced in size, weight, and cost, thereby making them more portable. The image captured by the camera contains a large amount of environmental information, and comprehensive use of this information can achieve accurate localization of the UAV.

Visual Odometry (VO) is used to study the estimation of the pose of a UAV through a series of images [11]–[13]. However, in the absence of global corrections, the offset accumulated by the odometer has a considerable influence on the pose estimation of long trajectories. To reduce the impact of drift, Liu and Zhang [14] and Glover *et al.* [15] proposed the loop closure method. However, in general, the flight

The associate editor coordinating the review of this manuscript and approving it for publication was Xian Sun.

trajectory of the UAV is not closed. For the drift problem, one possible solution is to register the UAV frame with satellite imagery of a known geographic location and use the ground reference to eliminate the accumulated error. This kind of satellite imagery has a wide range of sources and almost covers the entire surface of the earth.

Yol *et al.* [16] proposed a method for UAV localization using a map stitched by a series of georeferenced images. This method is based on a similarity function similar to mutual information [17] and it is only suitable for localization in a textured urban environment. Shan *et al.* [18] used the histogram of oriented gradient (HOG) [19] features to register UAV frames and satellite imageries, and subsequently used particle filter algorithms to locate UAVs. This method is not robust and relies heavily on clear buildings, roads, and other textural characteristics. Jurevicius *et al.* [20] studied the effect of the similarity of the image likelihood conversion function on the results of the particle filter localization algorithm [21] and proposed two parametric image similarity of likelihood conversion functions to improve the accuracy and robustness of the localization algorithm. This method is more dependent on the real-time of the map, and the use of an outdated map will increase the localization error. Mantelli *et al.* [6] proposed a new binary robust independent elementary feature (BRIEF) descriptor. Based on this descriptor, the measurement method in a Monte Carlo localization system is used for localization. The method assumes that the flying attitude of the UAV is always horizontal, thereby reducing the degree of freedom to four dimensions and finally conducts experiments on satellite maps of different years.

In the above methods, the flying environment of the UAV was an urban area with richer textural features; however, for areas with sparse texture, such as suburbs and rural areas, traditional feature descriptors (such as SIFT [22] and SURF [23]) often fail to extract effective features. In addition, the acquisition time span of satellite imageries is large, and it is difficult to overcome traditional feature descriptors for changes in lighting conditions, atmospheric conditions, and seasons.

In recent years, deep learning methods, especially the convolutional neural network (CNN) [24], have made tremendous progress and performance improvements in computer vision tasks such as image classification, target detection, and segmentation. Using the continuous layers of CNN, complex image features can be acquired easily, and specific deep learning features can be learned [25]. Based on these stable deep learning features, the registration of UAV frames, satellite maps, and the estimation of UAV pose can improve the robustness and adaptability of UAV visual localization.

Presently, some researchers have used the deep learning features of satellite imageries to realize the pose estimation of UAVs in various environments. Nassar *et al.* [26] first used scale invariant feature transform (SIFT) to register UAV frames and satellite maps and correlated the UAV movement with the actual map position for preliminary localization. A semantic shape matching algorithm was then used to extract and match meaningful shape information from the two images, and this information was used to improve the localization accuracy. Shetty *et al.* [27] used a satellite image-based cross-vision geolocation method to estimate the UAV's pose. This method is composed of two parts: scene localization network and camera localization network. After being combined with a VO, the defect of serious error accumulation of VO is improved to a certain extent. Goforth *et al.* [28] regard the ground as a plane and use a CNN to solve the homography between the ground and UAV frame to estimate the UAV's pose. This method overcomes the difference between UAV frames and satellite maps to a certain extent, but simply regards the ground as a plane. In the case of complex terrain, UAV localization accuracy will be seriously reduced.

Although the above methods have made certain explorations on the visual localization of the UAV, they all simplify the flight attitude or specify a certain flight environment, and the UAV may perform tasks in various complex environments. This study proposes a UAV visual localization method based on deep learning features of satellite imageries, which overcomes the difference between UAV frames and satellite imageries by extracting stable deep learning features. To adapt to the complex terrain, a global digital elevation model (DEM) is used. Finally, the local optimization method is used to achieve higher precision positioning.

The remainder of the paper is organized as follows: Section II presents the proposed method in detail; Section III presents and analyzes the data and experimental results with discussion, and conclusions are provided in Section IV.

II. MATERIALS AND METHODS

We assume that the position of the UAV is approximately known at the time of capturing the initial frame. This is a reasonable assumption in applications where the approximate take-off position is known, or where a single GNSS data point is given, before beginning the GNSS-denied flight. UAV frames are matched to achieve inter-frame pose transfer. To eliminate accumulated errors in the process of inter-frame pose transfer, UAV frames need to be registered with satellite imageries possessing known coordinates. However, the process of feature extraction and matching based on deep learning is computationally expensive and time-consuming. To reduce the amount of calculation and improve real-time performance, this study uses the concept of ORB-SLAM [29] to define keyframes in the UAV sequence. The keyframe is a frame which used to match the satellite imagery to obtain the absolute pose of the UAV, to eliminate the accumulated error in the process of transferring the pose between frames. The precise choice of keyframes depends on many characteristics of the particular flight hardware and the speed of the vehicle, and must be tuned for a given application. Frames other than the keyframe are matched with the nearest keyframe to convey the pose. The UAV pose estimation process is shown in Fig. 1 and *Ti* is the keyframe.

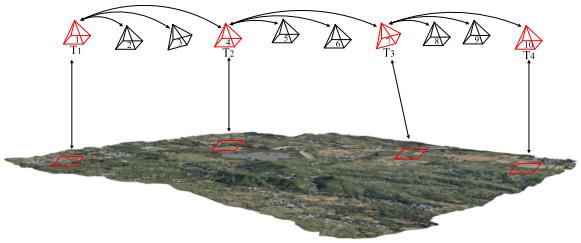


FIGURE 1. Schematic diagram of the UAV pose estimation process.

A. KEYFRAME POSE ESTIMATION

First, the local satellite imagery of the corresponding area is determined according to the initial position of the UAV. The deep learning features of the UAV frame and local satellite imagery are extracted for matching. To overcome the huge difference between UAV frames and satellite imageries, this study chooses the D2-Net [30] network that simultaneously performs feature point detection and descriptor extraction. This method (proposed by Mihai Dusmanu *et al.* of ETH Zurich in 2019) has shown great potential in solving road sign recognition for the visual navigation of ground vehicles in changing scenarios. Most of the images processed by D2-Net are ground close-range images. This study used satellite imageries of different years obtained from Google Maps to fine-tune the network. Based on the introduction of the basic idea of D2-Net feature extraction, the robust matching of UAV frames and satellite imageries is achieved. The algorithm flow is shown in Fig. 2.

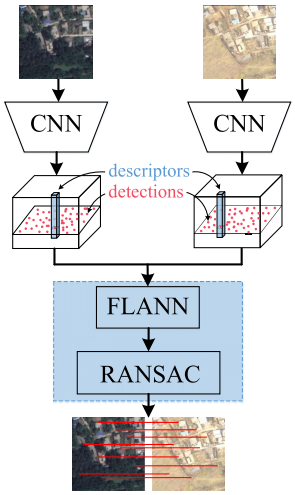


FIGURE 2. Matching algorithm flow.

Fine-tuning the D2-Net network only requires a few representative UAV frames and satellite imageries of the corresponding area. This article selects imageries from Dengfeng City, Henan Province, China, covering an area of approximately 15 km^2. The reason for choosing this area is that its textural characteristics vary greatly. It is close to the Songshan Mountains and has conspicuous terrain undulations, including typical flying environments such as cities, suburbs, rural areas, and jungles. The imagery is gathered during the years 2009, 2012, 2013, 2014, 2015, 2016, 2017, 2018, and 2019, across spring, summer, fall, and winter. The size of each imagery is 7247 × 9033 pixels, and the ground resolution is 0.5 m. The UAV frame acquisition time is the winter of 2019. The open source motion recovery structure (SfM) software COLMAP [31] was used to establish an accurate three-dimensional model, and the UAV frame and the satellite imagery were mapped at the pixel level in geographic coordinates. During training, a random selection was made from all satellite imageries and randomly specified latitude and longitude, and subsequently, 256 × 256-pixel image patches were centered on that location from the satellite imageries and UAV frames, as shown in Fig. 3. Finally, the image patches were input to the D2-Net network, and the weight was updated according to the loss function.

FIGURE 3. Examples from the alignment training dataset. Satellite imagery patches (top) and their corresponding UAV frame patches (bottom).

The fine-tuned D2-Net network was used to extract feature points and descriptors from UAV frames and satellite imageries, respectively. The feature matching method uses the fast-approximate nearest neighbor (FLANN) method. Due to the large image difference, there are several mismatches. In this study, the random sample consensus (RANSAC) constraint is used to eliminate mismatched points, and the geometric model chooses the affine transformation model.

After determining the two-dimensional feature matching relationship between the keyframe and satellite imagery, the keyframe feature points correspond well to the satellite imagery feature points. The plane coordinates of each pixel of the satellite imagery are known, and the elevation value is obtained by DEM interpolation. The three-dimensional coordinates of the satellite imagery feature points are the ground point coordinates corresponding to the keyframe feature points. Knowing the keyframe feature point image point coordinates and the corresponding ground point coordinates, the efficient perspective-n-points (EPnP) [32] algorithm was used to solve the UAV's pose.

B. CURRENT FRAME POSE ESTIMATION

For the current frame (other than the keyframe), the ORB [33] feature with a small calculation amount is selected to match the keyframe to achieve the pose transfer between frames, as shown in Fig. 4. Because the feature extraction algorithm for inter-frame matching is different from the feature extraction algorithm for keyframe and satellite imagery matching, the location of feature points is not repeatable, and therefore, the feature point tracking method in ORB-SLAM [29] cannot

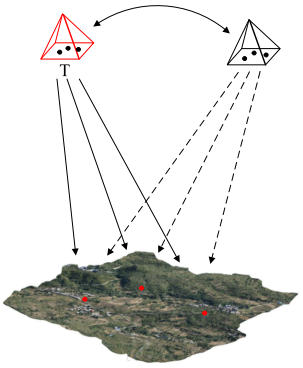


FIGURE 4. Pose transfer between frames.

be used. To obtain the ground point coordinates corresponding to the feature points of the current frame, this study adopted the iterative photogrammetric method [34], [35].

After determining the two-dimensional feature matching relationship between the keyframe and general frame, the pose of the keyframe and the corresponding ground DEM are known, and the ground point coordinates corresponding to the keyframe feature points are obtained through the iterative photogrammetric method.

Suppose the pixel coordinate of the feature point is (u_s, v_s), according to the elements of interior orientation, the normalized plane coordinate $(x', y', 1)$ of the feature points can be obtained in the image space coordinate system:

$$\begin{cases} x' = \dfrac{u_s - c_x}{f_x} \\ y' = -\dfrac{v_s - c_y}{f_y} \end{cases} \tag{1}$$

where f_x and f_y are the normalized focal length on u axis and v axis respectively, and $(x, y, 1)$ is the pixel coordinate of the optical center of the camera. For normalized coordinate de-distortion, this study only considers radial distortion, and the de-distorted coordinate $(x, y, 1)$ is obtained:

$$\begin{cases} x = x' \big/ \left(1 + k_1 r^2 + k_2 r^4\right) \\ y = y' \big/ \left(1 + k_1 r^2 + k_2 r^4\right) \end{cases} \tag{2}$$

Among them, $r^2 = x'^2 + y'^2$, the distortion parameter k_1, and k_2 are obtained by camera calibration. If the elevation value of the camera's outer azimuth elements and the corresponding ground point coordinates are known, the corresponding ground point plane coordinates can be easily solved according to the collinear condition equation as follows:

$$\begin{cases} X = (Z - Z_S) \dfrac{a_1 x + a_2 y - a_3}{c_1 x + c_2 y - c_3} + X_S \\ Y = (Z - Z_S) \dfrac{b_1 x + b_2 y - b_3}{c_1 x + c_2 y - c_3} + Y_S \end{cases} \tag{3}$$

Among them, (X, Y, Z) is the coordinate of the feature point corresponding to the ground point p in the geodetic coordinate system, $t = (X_S, Y_S, Z_S)$ is the line element in the elements of exterior orientation, and $a_i, b_i, c_i (i = 1, 2, 3)$ is the element in the rotation matrix R composed of the corner elements in the elements of exterior orientation.

However, the elevation value of the ground point coordinates is unknown. Only the DEM of the corresponding area is known, and therefore, an iterative solution process is required. As shown in Fig. 5, first provide the initial elevation value Z_0 of the ground point coordinates. According to this initial value, the plane coordinate (X_1, Y_1) is solved through the collinear condition equation to obtain the initial value position A_1 of the ground point. Project A_0 vertically onto the DEM to get the projection point B_1, and update the initial elevation value Z_0 with the elevation value Z_1 of B_1. Repeat this process until the difference between the two iterations before and after the iteration is less than the given threshold. At this time, the coordinates of A_n are considered to be ground point coordinates.

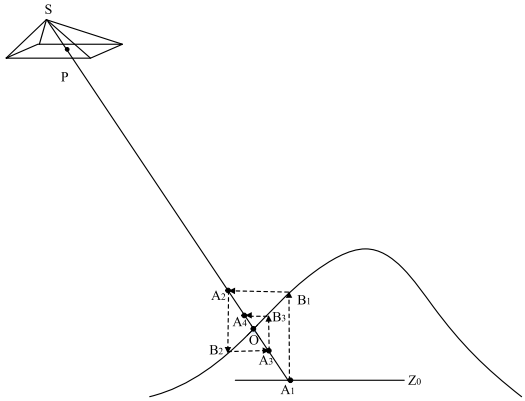


FIGURE 5. Principle of iterative photogrammetric method.

After calculating the ground point coordinates corresponding to the keyframe feature points according to the iterative photogrammetric method, the ground point coordinates corresponding to the current frame feature points matching the keyframe are obtained. In the same state, the pose of the current frame is solved according to the method of keyframe pose estimation. In each current frame pose estimation, two sets of 3D–2D points are obtained: keyframe feature points and their corresponding ground point coordinates; and current frame feature points and their corresponding ground point coordinates. The ground point coordinates corresponding to the two sets of feature points are the same, and these points will be adjusted uniformly in the subsequent optimization process.

C. LOCAL OPTIMIZATION

In VO, BA [36] is usually used to simultaneously optimize multiple consecutive motion poses. To improve the accuracy and robustness of the algorithm, this study designed a local optimization method to simultaneously optimize the poses of all UAV frames in the local interval.

From Section 2.2, the relationship between the pixel coordinate (u_s, v_s) of the image feature point and the corresponding ground point coordinate p is shown in Fig. 6.

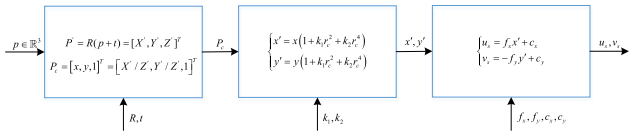


FIGURE 6. Schematic diagram of calculation process.

Abstractly record the entire process as an observation equation:

$$z = h(\xi, p) \tag{4}$$

Among them, $\boldsymbol{\xi}$ is the external orientation element of the camera represented by Lie algebra, p is the ground point coordinates, and the observation data is the feature point pixel coordinates $z \triangleq [u_s, v_s]^T$. The observation error is as follows:

$$e = z - h(\xi, p) \tag{5}$$

For the observations in the entire optimization interval, let $\mathbf{z}_{ij}$ be the data generated by the camera observing the ground point $\mathbf{p}_j$ at the pose $\boldsymbol{\xi}_i$, then the overall cost function is as follows:

$$\min_{\boldsymbol{\xi},\mathbf{p}} \sum_{k=1}^{t} \sum_{i \in V(k)} \sum_{j=1}^{n_1} \left\| e_{ij} \right\|_2^2 + \sum_{k=1}^{t} \sum_{j=1}^{n_2} \left\| e_{kj} \right\|_2^2 \tag{6}$$

The cost function is divided into two parts. The first half is the observation error of traditional feature points, and the second half is the observation error of keyframe deep learning feature points. Among them, n_1 is the number of traditional feature points, and n_2 is the number of deep learning feature points. $V(k)$ is a set of matching pairs formed by all frames directly matching the k-th keyframe in the optimization interval. For example, in Fig. 1, $V(1) = \{1, 2, 1, 3, 1, 4\}$ and t are the number of keyframes in the optimization interval.

Substituting (6) into (7), we get:

$$\sum_{k=1}^{t} \sum_{i \in V(k)} \sum_{j=1}^{n_1} \left\| \mathbf{z}_{ij} - h\left(\boldsymbol{\xi}_i, \mathbf{p}_j\right) \right\|_2^2 + \sum_{k=1}^{t} \sum_{j=1}^{n_2} \left\| \mathbf{z}_{kj} - h\left(\boldsymbol{\xi}_k, \mathbf{p}_j\right) \right\|_2^2 \tag{7}$$

With the addition of a small perturbation to the parameters $\boldsymbol{\xi}$ and $\mathbf{p}$, linearize the cost function as follows:

$$\sum_{k=1}^{t} \sum_{i \in V(k)} \sum_{j=1}^{n_1} \left\| e_{ij} + \mathbf{F}_{ij} \Delta\boldsymbol{\xi}_i + \mathbf{E}_{ij} \Delta\mathbf{p}_j \right\|_2^2 + \sum_{k=1}^{t} \sum_{j=1}^{n_2} \left\| e_{kj} + \mathbf{F}_{kj} \Delta\boldsymbol{\xi}_k + \mathbf{E}_{kj} \Delta\mathbf{p}_j \right\|_2^2 \tag{8}$$

where $\mathbf{F}_{ij}$ represents the partial derivative of $\sum_{k=1}^{t} \sum_{i \in V(k)} \sum_{j=1}^{n_1} \left\| e_{ij} \right\|_2^2$ with respect to the camera attitude of the i-th frame, and $\mathbf{E}_{ij}$ represents the partial derivative of $\sum_{k=1}^{t} \sum_{i \in V(k)} \sum_{j=1}^{n_1} \left\| e_{ij} \right\|_2^2$ with respect to the position of the ground point. $\mathbf{F}_{kj}$ represents the partial derivative of $\sum_{k=1}^{t} \sum_{j=1}^{n_2} \left\| e_{kj} \right\|_2^2$ with respect to the camera attitude of the k-th frame, and $\mathbf{E}_{kj}$ represents the partial derivative of $\sum_{k=1}^{t} \sum_{j=1}^{n_2} \left\| e_{kj} \right\|_2^2$ with respect to the position of the ground point.

Put all the pose parameters together:

$$\mathbf{x}_c = \left[\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_m\right]^T \in \mathbb{R}^{6m} \tag{9}$$

Put all ground point parameters together:

$$\mathbf{x}_p = \left[\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n\right]^T \in \mathbb{R}^{3n} \tag{10}$$

Formula (8) can be simplified as follows:

$$\frac{1}{2} \left\| e + \mathbf{F}\Delta\mathbf{x}_c + \mathbf{E}\Delta\mathbf{x}_p \right\|^2 \tag{11}$$

The Jacobian matrix $\mathbf{F}$ and $\mathbf{E}$ are the derivatives of the overall objective function to the overall parameters.

Afterwards, use Gauss–Newton iteration to solve the linearized objective function. Let this formula be zero, and solve the incremental linear equation as follows:

$$H\Delta\mathbf{x} = \begin{bmatrix} \mathbf{F}^T\mathbf{F} & \mathbf{F}^T\mathbf{E} \\ \mathbf{E}^T\mathbf{F} & \mathbf{E}^T\mathbf{E} \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x}_c \\ \Delta\mathbf{x}_p \end{bmatrix} = - \begin{bmatrix} \mathbf{F}^T \\ \mathbf{E}^T \end{bmatrix} e \tag{12}$$

Thousands of feature points can be proposed on each image, which increases the scale of this linear equation. If you directly invert the matrix to calculate the incremental equation, it will consume huge computing resources and is difficult to solve quickly. However, the sparse structure of the matrix can be used to marginalize [37] the equation and speed up the solution process.

III. RESULTS AND DISCUSSION

The frame of the UAV used in the experiment was acquired by the onboard top–down camera. The flying speed of the UAV was 14 m/s, and the frame rate was 1 fps. Each frame has accurate location information obtained by differential GPS, which is regarded as a true value in the experiment. The data contains two trajectories, covering sparsely textured jungle and richly textured buildings. The first is a straight line, as shown in Fig. 6, and the trajectory is 0.75 km long. The second is a curve, as shown in Fig. 7, and the trajectory is 0.54 km long. The relative altitude of the two trajectories is 0.5 km. The UAV frame acquisition season was winter. To verify the adaptability of the employed method to seasonal changes, the 2018 summer satellite imagery with a large difference was selected as the localization reference. The satellite imagery was downloaded from Google Maps; the ground resolution was 0.5 m, and the plane coordinate accuracy was approximately 5 m. DEM is the ASTER GDEM V2 global digital elevation data jointly developed by Japan METI and NASA and distributed to the public for free, with a spatial resolution of 30 m.

The test was divided into four parts. First, the performance of the D2-Net matching algorithm was evaluated. Second, the effectiveness of the optimization algorithm was verified. The influence of the keyframe selection interval on UAV localization was then analyzed and the keyframe interval

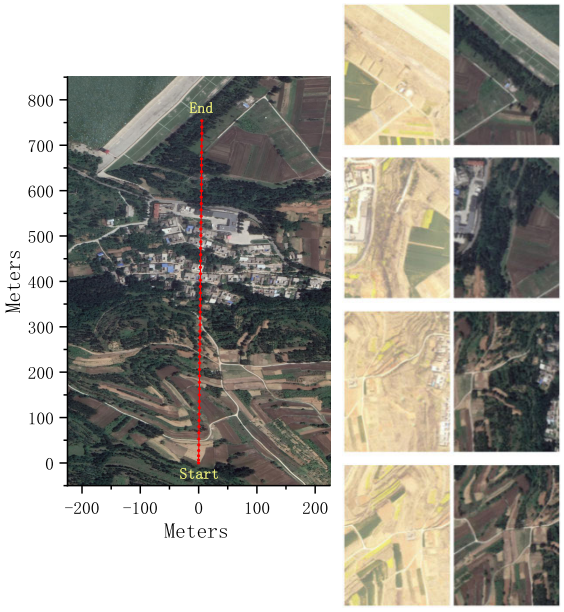


FIGURE 7. (a) Overview of the straight flight trajectory (b) Some examples of UAV frames (left) and their corresponding satellite imagery patches(right) for the straight flight trajectory.

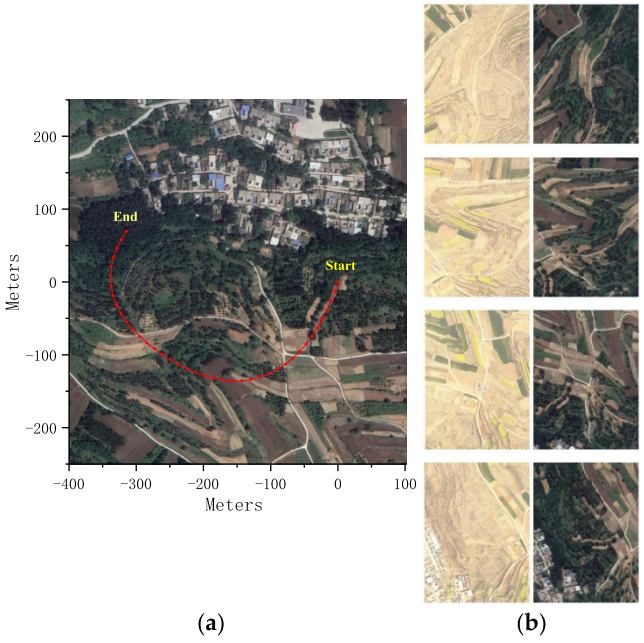


FIGURE 8. (a) Overview of the curved flight trajectory (b) Some examples of UAV frames (left) and their corresponding satellite imagery patches(right) for the curved flight trajectory.

applicable in different environments was determined. Finally, a comparative experiment with ORB-SLAM and the method specified in [28] was designed. The localization error in the experiment is the Euclidean distance between the real position of the UAV in the three-dimensional space and the estimated position.

The methodology in this study was implemented based on the Python language, and the deep learning model was constructed under the PyTorch 1.3.0 framework. The computer used for the test was the MSI P65 notebook with the i9-9880H CPU, GeForce RTX 2070Max-Q (8G video memory) graphics card, and a memory of 32 GB. The implementation language was Python, and the operating system used was Ubuntu 16.04.

A. MATCHING RESULTS

To evaluate the performance of the D2-Net matching algorithm, this study established a test data set containing UAV frames and their corresponding regional satellite imageries. Image pairs with large seasonal changes and conspicuous changes in ground features were selected from the original data. SIFT [22], SURF [23] traditional features, and DELF [38] deep learning features were used to conduct comparative experiments with this method to test its accuracy and adaptability.

As shown by Table 1 and Fig. 9, for images with different time phases and seasons, although traditional methods can extract several feature points, they cannot describe the feature stably. In contrast, deep learning features have achieved good results. Although DELF has successfully extracted some stable features, it takes a long time and the number is limited. Moreover, D2-Net has a greater ability to extract stable deep learning features, higher efficiency, and more matching pairs with more uniform distribution after matching.

TABLE 1. Matching Accuracy and Time-Consuming Statistics.

Test number	Algorithm	Initial matches	Number of matches after filtering	Correct rate after screening (< 4 pix, %)	Average matching time (s)
1	SIFT	12764	32	21.88	16.58
2	SURF	5462	26	15.38	3.89
3	DELF	18902	67	95.52	12.97
4	D2-Net	10536	1332	94.29	5.15

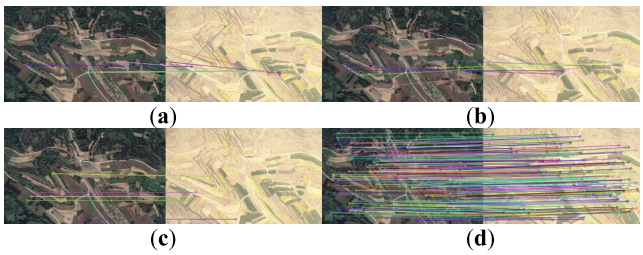


FIGURE 9. Examples of matching results: (a) SIFT; (b) SURF; (c) DELF; (d) D2-Net.

B. OPTIMIZATION METHOD EFFECT TEST

To verify the effectiveness of the local optimization method, a comparative test was conducted on two routes. The localization accuracy under different keyframe selection conditions before and after optimization were calculated separately, and the test results are shown in Fig. 10.

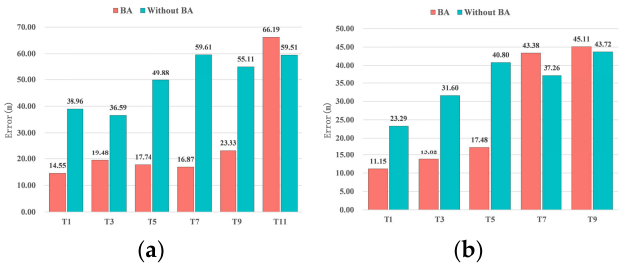


FIGURE 10. Average localization error of the entire track before and after optimization, where Ti, $(i = 1, 2, \cdots, 11)$ indicates that a keyframe is selected for every i UAV frames: (a) Flight path 1; (b) Flight path 2.

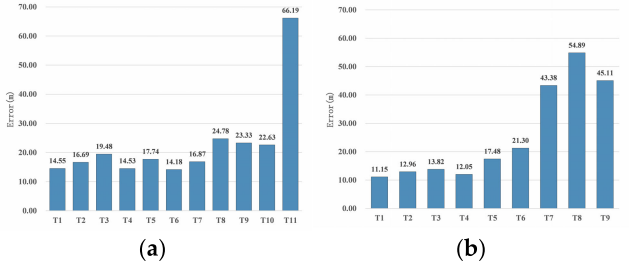


FIGURE 11. Average localization error of the entire track when the keyframe interval is different: (a) Flight path 1; (b) Flight path 2.

From the experimental results, it can be concluded that when the keyframe spacing is too large, the optimization algorithm does not converge, and it even increases the localization error. However, when the keyframe spacing is selected appropriately, the optimization method can effectively reduce the error and increase the localization accuracy by more than 50 %.

C. TEST OF THE INFLUENCE OF KEYFRAME INTERVAL ON LOCALIZATION ACCURACY

In the entire pose estimation process, deep learning feature extraction, and matching have the most number of calculations to be performed and the longest time-consumption. To reduce the amount of calculation, the keyframe spacing should be as large as possible. However, it can be concluded from the experiment in Section 3.2 that when keyframe spacing is too large the optimization algorithm fails to converge, the localization error increases rapidly, and the localization even fails. Therefore, under the premise of ensuring localization accuracy, it is very important to determine the universal keyframe interval for different trajectories. For this purpose, detailed experiments were conducted on different keyframe selection methods under different trajectories. The experimental results are shown in Table 2 and Fig. 11.

TABLE 2. Time-Consuming Statistics.

Keyframe selection conditions	Average keyframe localization time (s)	Average current frame localization time (s)	Average optimization time (s)
T1	5.28	0.14	0.11
T2	5.29	0.15	0.35
T3	5.27	0.14	0.93
T4	5.27	0.14	1.32
T5	5.28	0.14	2.13
T6	5.27	0.14	3.81

The experimental results reveal that the localization error can be maintained within a relatively stable range with an initial gradual increase of keyframe spacing. When the distance increases to a certain extent, the local optimization algorithm fails, and the localization error increases rapidly. As shown by

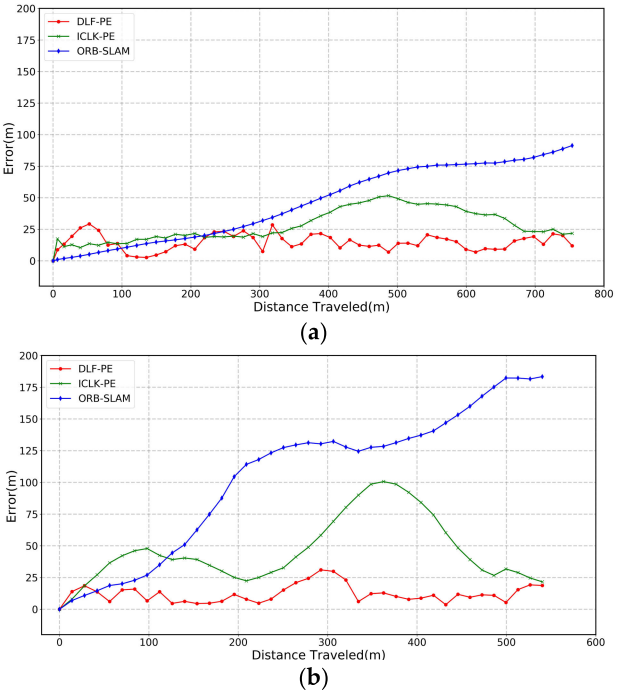


FIGURE 12. Localization errors of three methods under different trajectories: (a) Flight path 1; (b) Flight path 2.

Table 2, the keyframe localization time is much longer than the current frame localization time, which is caused by the time-consuming process of deep learning feature extraction and matching. When the optimization algorithm is effective, the Average optimization time will increase rapidly as the keyframe spacing increases. Compared with the curved trajectory, the UAV can adapt to a larger keyframe interval when flying in a straight line, but for universal applicability, it is determined that one keyframe every four frames is the optimal solution.

D. COMPARATIVE TEST OF DIFFERENT LOCALIZATION METHODS

The method in this study is named DLF-PE. To further test the performance of DLF-PE, a comparative test between DLF-PE and ORB-SLAM and the method in literature [28] was designed. ORB-SLAM is a more advanced method in the SLAM algorithm. The method in [28] has achieved relatively accurate localization results in a flat

ground environment and has strong adaptability to a sparse texture environment. For the convenience of presentation, the method is named ICLK-PE. Fig. 12 shows the errors of the three localization methods. The real trajectory and the predicted trajectories of the three methods are shown in Fig. 13.

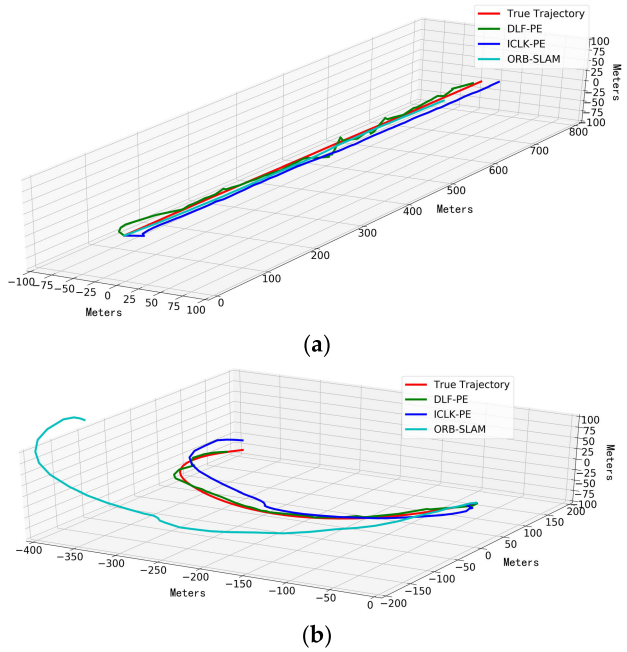


FIGURE 13. UAV true trajectory and estimated trajectory: (a) Flight path 1; (b) Flight path 2.

The average localization error is shown in Table 3. From the experimental results, it can be concluded that the traditional ORB-SLAM method has serious error accumulation and the lowest accuracy. Due to the large terrain fluctuations, the accuracy of the ICLK-PE method has decreased, and it cannot adapt to the curve trajectory well. The average localization error of DLF-PE under the two trajectories is controlled within 15 m, and it has the highest accuracy and displays strong adaptability to different flight trajectories.

TABLE 3. Average Localization Error of the Three Methods Under Different Trajectories.

Flight Path	DLF-PE	ICLK-PE	ORB-SLAM
1	14.53 m	27.59 m	45.03 m
2	12.05 m	45.88 m	104.30 m

IV. CONCLUSION

The localization of UAVs in a GNSS-denied environment is of great significance for ensuring efficiency of UAVs performance during various tasks. This study proposes a UAV visual localization method assisted by deep learning features of satellite imageries with good adaptability to different flight environments. To further improve localization accuracy, a local optimization method was designed to simultaneously optimize the poses of all UAV frames in the local interval. Tests have verified that the optimized average localization error can meet the UAV localization requirements. Although the method in this study can almost achieve real-time performance on a laptop computer, the amount of calculation is substantial for the processor on the UAV. Therefore, it is necessary to further reduce the amount of calculation and improve the real-time performance of the algorithm. A more robust and efficient deep learning feature extraction algorithm is required. In addition, a low-precision internal measurement unit is also a common UAV payload. Comprehensive utilization of the posture data provided by the IMU to improve accuracy and real-time performance of the algorithm is a problem that needs to be solved in future research.

REFERENCES

[1] J. Scherer, S. Yahyanejad, S. Hayat, E. Yanmaz, T. Andre, A. Khan, V. Vukadinovic, C. Bettstetter, H. Hellwagner, and B. Rinner, "An autonomous multi-UAV system for search and rescue," in *Proc. 1st Workshop Micro Aerial Vehicle Netw., Syst., Appl. Civilian Use*, May 2015, pp. 33–38, doi: 10.1145/2750675.2750683.

[2] S. Siebert and J. Teizer, "Mobile 3D mapping for surveying earthwork projects using an unmanned aerial vehicle (UAV) system," *Autom. Construct.*, vol. 41, pp. 1–14, May 2014, doi: 10.1016/j.autcon.2014.01.004.

[3] P. Tokekar, J. V. Hook, D. Mulla, and V. Isler, "Sensor planning for a symbiotic UAV and UGV system for precision agriculture," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1498–1511, Dec. 2016, doi: 10.1109/TRO.2016.2603528.

[4] Y. Lu, D. Macias, Z. S. Dean, N. R. Kreger, and P. K. Wong, "A UAV-mounted whole cell biosensor system for environmental monitoring applications," *IEEE Trans. Nanobiosci.*, vol. 14, no. 8, pp. 811–817, Dec. 2015, doi: 10.1109/TNB.2015.2478481.

[5] T. Senlet and A. Elgammal, "Satellite image based precise robot localization on sidewalks," in *Proc. IEEE Int. Conf. Robot. Autom.*, Saint Paul, MN, USA, May 2012, pp. 2647–2653.

[6] M. Mantelli, D. Pittol, R. Neuland, A. Ribacki, R. Maffei, V. Jorge, E. Prestes, and M. Kolberg, "A novel measurement model based on abBRIEF for global localization of a UAV over satellite images," *Robot. Auto. Syst.*, vol. 112, pp. 304–319, Feb. 2019, doi: 10.1016/j.robot.2018.12.006.

[7] J. V. Carroll, "Vulnerability assessment of the U.S. transportation infrastructure that relies on the global positioning system," *J. Navigat.*, vol. 56, no. 2, pp. 185–193, May 2003, doi: 10.1017/S0373463303002273.

[8] F. Caballero, L. Merino, J. Ferruz, and A. Ollero, "Improving vision-based planar motion estimation for unmanned aerial vehicles through online mosaicing," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Orlando, FL, USA, May 2006, pp. 2860–2865.

[9] G. Conte and P. Doherty, "An integrated UAV navigation system based on aerial image matching," in *Proc. IEEE Aerosp. Conf.*, Big Sky, MT, USA, Mar. 2008, pp. 1–10.

[10] A. Viswanathan, B. R. Pires, and D. Huber, "Vision-based robot localization across seasons and in remote locations," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Stockholm, Sweden, May 2016, pp. 4815–4821.

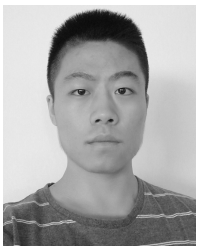
[11] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *IEEE Robot. Autom. Mag.*, vol. 18, no. 4, pp. 80–92, Dec. 2011, doi: 10.1109/MRA.2011.943233.

[12] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Hong Kong, May 2014, pp. 15–22, doi: 10.1109/ICRA.2014.6906584.

[13] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Hong Kong, May 2014, pp. 1524–1531, doi: 10.1109/ICRA.2014.6907054.

[14] Y. Liu and H. Zhang, "Visual loop closure detection with a compact image descriptor," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Hong Kong, Oct. 2012, pp. 1051–1056, doi: 10.1109/IROS.2012.6386145.

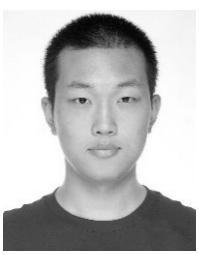
[15] A. Glover, W. Maddern, M. Warren, S. Reid, M. Milford, and G. Wyeth, "OpenFABMAP: An open source toolbox for appearance-based loop closure detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, Saint Paul, MN, USA, May 2012, pp. 4730–4735, doi: 10.1109/ICRA.2012.6224843.
[16] A. Yol, B. Delabarre, A. Dame, J.-E. Dartois, and E. Marchand, "Vision-based absolute localization for unmanned aerial vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Chicago, IL, USA, Sep. 2014, pp. 3429–3434, doi: 10.1109/IROS.2014.6943040.
[17] A. Dame and E. Marchand, "Second-order optimization of mutual information for real-time image registration," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4190–4203, Sep. 2012, doi: 10.1109/TIP.2012.2199124.
[18] M. Shan, F. Wang, F. Lin, Z. Gao, Y. Z. Tang, and B. M. Chen, "Google map aided visual navigation for UAVs in GPS-denied environment," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Zhuhai, China, Dec. 2015, pp. 114–119, doi: 10.1109/ROBIO.2015.7418753.
[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, USA, Jun. 2005, pp. 886–893, doi: 10.1109/CVPR.2005.177.
[20] R. Jurevičius, V. Marcinkevičius, and J. Šeibokas, "Robust GNSS-denied localization for UAV using particle filter and visual odometry," *Mach. Vis. Appl.*, vol. 30, nos. 7–8, pp. 1181–1190, Oct. 2019, doi: 10.1007/s00138-019-01046-4.
[21] S. Thrun, "Probabilistic robotics," *Commun. ACM*, vol. 45, no. 3, pp. 52–57, Mar. 2002.
[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94.
[23] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008, doi: 10.1016/j.cviu.2007.09.014.
[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[25] A. Yang, X. Yang, W. Wu, H. Liu, and Y. Zhuansun, "Research on feature extraction of tumor image based on convolutional neural network," *IEEE Access*, vol. 7, pp. 24204–24213, 2019, doi: 10.1109/ACCESS.2019.2897131.
[26] A. Nassar, K. Amer, R. ElHakim, and M. ElHelw, "A deep CNN-based framework for enhanced aerial imagery registration with applications to UAV geolocalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 1513–1523, doi: 10.1109/CVPRW.2018.00201.
[27] A. Shetty and G. X. Gao, "UAV pose estimation using cross-view geolocalization with satellite imagery," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, Montreal, QC, Canada, May 2019, pp. 1827–1833, doi: 10.1109/ICRA.2019.8794228.
[28] H. Goforth and S. Lucey, "GPS-denied UAV localization using pre-existing satellite imagery," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, Montreal, QC, Canada, May 2019, pp. 2974–2980, doi: 10.1109/ICRA.2019.8793558.
[29] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015, doi: 10.1109/TRO.2015.2463671.
[30] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable CNN for joint description and detection of local features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8092–8101.
[31] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 4104–4113, doi: 10.1109/CVPR.2016.445.
[32] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *Int. J. Comput. Vis.*, vol. 81, no. 2, pp. 155–166, Feb. 2009, doi: 10.1007/s11263-008-0152-6.
[33] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 2564–2571, doi: 10.1109/ICCV.2011.6126544.
[34] W. Linder, *Digital Photogrammetry*. Berlin, Germany: Springer, 2009.
[35] Y. Sheng, "Theoretical analysis of the iterative photogrammetric method to determining ground coordinates from photo coordinates and a DEM," *Photogramm. Eng. Remote Sens.*, vol. 71, no. 7, pp. 863–871, Jul. 2005, doi: 10.14358/PERS.71.7.863.
[36] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—A modern synthesis," in *Proc. Int. Workshop Vis. Algorithms*, in Lecture Notes in Computer Science. Berlin, Germany: Springer, 2000, pp. 298–372, doi: 10.1007/3-540-44480-7_21.
[37] G. Sibley, L. Matthies, and G. Sukhatme, "Sliding window filter with application to planetary landing," *J. Field Robot.*, vol. 27, no. 5, pp. 587–608, Jul. 2010, doi: 10.1002/rob.20360.
[38] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3456–3465.



HUITAI HOU received the B.S. degree in photogrammetry and remote sensing from the Zhengzhou Institute of Surveying and Mapping, China, in 2018. He is currently pursuing the Ph.D. degree in surveying and mapping with the Institute of Geospatial Information, Information Engineering University. His research interests include photogrammetry and UAV remote sensing.

QING XU received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from the Zhengzhou Institute of Surveying and Mapping, China, in 1985, 1990, and 1995, respectively. He is currently a Professor and a Doctoral Supervisor with Information Engineering University. His research interests include remote sensing and digital photogrammetry.

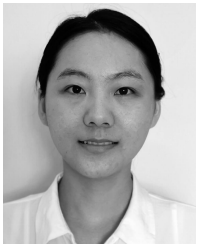
CHAOZHEN LAN received the B.S. and M.S. degrees in photogrammetry and remote sensing and the Ph.D. degree in surveying and mapping from the Zhengzhou Institute of Surveying and Mapping, China, in 2002, 2005, and 2009, respectively. He is currently an Associate Professor and a Master Supervisor with Information Engineering University. His research interests include photogrammetry and UAV remote sensing.



WANJIE LU received the B.S. degree in photogrammetry and remote sensing from the Zhengzhou Institute of Surveying and Mapping, China, in 2013, and the M.S. degree in surveying and mapping from Information Engineering University, China, in 2016, where he is currently pursuing the Ph.D. degree in surveying and mapping with the Institute of Geospatial Information. His research interests include photogrammetry, space situational awareness, and spatial information services.

YONGXIAN ZHANG received the B.S. degree in geographic information system from Xinyang Normal University, China, in 2016, and the M.S. degree in photogrammetry and remote sensing from the Information Engineering University, China, in 2020. He is currently pursuing the Ph.D. degree with Wuhan University. His research interests include photogrammetry and visible spectral remote sensing.

ZHIXIANG CUI received the B.S. degree in remote sensing science and technology from the University of Information Engineering, in 2014. He is currently an Assistant Engineer with 31682 Troops. His research interest includes UAV image processing and application.



JIANQI QIN received the B.S. degree in remote sensing science and technology from Information Engineering University, in 2015, where she is currently pursuing the master's degree. Her research interests include photogrammetry and the remote sensing of unmanned aerial vehicles.