

JointLoc: A Real-time Visual Localization Framework for Planetary UAVs Based on Joint Relative and Absolute Pose Estimation

Xubo Luo^{1,2,3}, Xue Wan^{2,3}, Yixing Gao⁴, Yaolin Tian^{1,2,3}, Wei Zhang^{2,3} and Leizheng Shu^{2,3}

Abstract—Unmanned aerial vehicles (UAVs) visual localization in planetary aims to estimate the absolute pose of the UAV in the world coordinate system through satellite maps and images captured by on-board cameras. However, since planetary scenes often lack significant landmarks and there are modal differences between satellite maps and UAV images, the accuracy and real-time performance of UAV positioning will be reduced. In order to accurately determine the position of the UAV in a planetary scene in the absence of the global navigation satellite system (GNSS), this paper proposes JointLoc, which estimates the real-time UAV position in the world coordinate system by adaptively fusing the absolute 2-degree-of-freedom (2-DoF) pose and the relative 6-degree-of-freedom (6-DoF) pose. Extensive comparative experiments were conducted on a proposed planetary UAV image cross-modal localization dataset, which contains three types of typical Martian topography generated via a simulation engine as well as real Martian UAV images from the Ingenuity helicopter. JointLoc achieved a root-mean-square error of 0.237m in the trajectories of up to 1,000m, compared to 0.594m and 0.557m for ORB-SLAM2 and ORB-SLAM3 respectively. The source code will be available at <https://github.com/LuoXubo/JointLoc>.

I. INTRODUCTION

Planetary exploration, vital for expanding human habitation beyond Earth, has seen a rise in interest in unmanned aerial vehicles (UAVs) due to their portability and exploration capabilities [21]. Notably, the Ingenuity helicopter of NASA deployed to Mars in September 2021 has completed 72 flights, capturing valuable imagery of the Martian surface [28]. Accurate positioning of UAVs is crucial for mission success, as evidenced by a January 2024 incident where the Ingenuity experienced communication loss after a hard landing due to positioning challenges in a featureless area [29], concluding its pioneering flight on Mars.

In contrast to the highly developed navigation system, such as the Global Navigation Satellite System (GNSS), on Earth, autonomous UAV navigation in a planetary environment presents several challenges. Firstly, the absence of GNSS means that the geographic position of the UAV cannot be

*This work was supported by the National Natural Science Foundation of China (NSFC, No. 42171445, and No. 62203184).

¹Xubo Luo and Yaolin Tian are with the School of Aeronautics and Astronautics, University of Chinese Academy of Sciences, Beijing 101408, China {luoxubo23, tianyaolin21}@mailsucas.ac.cn

²Xubo Luo, Yaolin Tian, Xue Wan, Wei Zhang, and Leizheng Shu are with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China {wanxue, zhangwei, shuleizheng}@csu.ac.cn

³Xubo Luo, Yaolin Tian, Xue Wan, Wei Zhang, and Leizheng Shu are with the Key Laboratory of Space Utilization, Chinese Academy of Sciences, Beijing 100094, China

⁴Yixing Gao is with the School of Artificial Intelligence, Jilin University, Changchun 130015, China gaoyixing@jlu.edu.cn

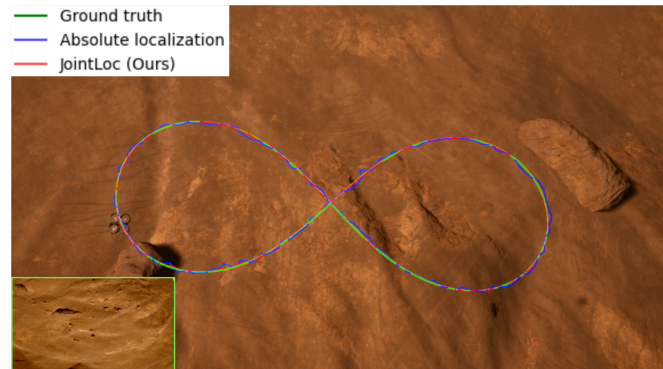


Fig. 1. JointLoc accurately estimates the UAV poses. In contrast, the results of absolute localization are rough and pose a great challenge to the control algorithm. The results of relative localization are in the local coordinate system and require manual alignment to the world coordinate system.

directly determined. Secondly, the planetary environment usually lacks salient targets and rich texture information which prevents some visual position estimation methods commonly used in Earth scenes from being directly applied in planetary exploration missions. In addition, the limited computing power of mobile computing platforms prevents the most popular methods from being used for positioning.

Mainstream localization methods in planetary exploration missions are typically categorized into absolute localization methods and relative localization methods [31]. Absolute localization involves estimating the 2-DoF position of the vehicles within the world coordinate system by comparing visual data collected by onboard sensors with existing satellite remote sensing maps or terrain maps [30]. For instance, Geromichalos et al. [26] employed particle filters and scan matching with existing maps to reduce positioning errors for planetary rovers. It is evident that most current absolute localization methods are primarily utilized for planetary rovers at relatively low speeds. However, this presents a challenge for absolute localization, as large-scale map positioning necessitates considerable time and computational resources, rendering it unsuitable for high-speed moving UAVs. Furthermore, modal differences between satellite maps and UAV images, such as scale, viewing angle, and lighting, mean that absolute localization cannot guarantee the accuracy of the results.

Relative localization estimates the 6-DoF pose of the vehicles relative to the starting point. The main difference between relative and absolute localization is that the geographic position, such as latitude and longitude, can not be determined in the relative localization method. The most widely

used method for relative localization in terrestrial scenes is the visual odometry (VO) [23]–[25], which provides rapid positioning and high success rates. However, VO primarily estimates the local 6-DoF poses of the vehicles through inter-frame matching, rather than in the world coordinate system. Particularly in planetary environments like textureless Mars deserts, traditional operator-based VO struggles to extract a sufficient number of map points, leading to tracking lost [27]. By fusing the 2-DoF absolute pose in the world coordinate system with the relative 6-DoF pose generated by VO, the 6-DoF pose in the world coordinate system can be estimated.

Although several studies have been carried out on either absolute or relative localization for planetary UAV, they do not take into account the consumption of computing resources by the absolute and relative poses fusion frequency [1] and the global localization methods based on large-scale maps [31]. Besides, in the case of textureless areas on a planetary surface, there is no effective method to deal with the deviation between absolute localization and relative localization. Therefore, this paper introduces JointLoc, a localization framework that estimates the 6-DoF pose of the planetary UAVs in world coordinate systems. JointLoc adaptively fuses absolute and relative poses, reducing the time consumption of absolute localization and also converting relative poses into the world coordinate system. A simple comparison of the UAV trajectories estimated by JointLoc and the absolute localization method is shown in Fig. 1.

The contributions of this paper can be summarized as follows:

- A loosely coupled localization framework is proposed for planetary UAVs, which adaptively fuses relative 6-DoF and absolute 2-DoF poses, to estimate the 6-DoF poses of the UAV within the world coordinate system.
- For UAV localization in large-scale satellite maps, a coarse-to-fine absolute localization module is proposed to reduce time consumption. The deep features are used to increase the robustness of the module in textureless environments with few landmarks.
- In order to enable JointLoc to accurately estimate the pose even when the absolute and relative pose results are inconsistent, an adaptive confidence mechanism is designed for the absolute pose so that the pose fusion module can adjust the fusion weights of different poses based on the confidence.
- A planetary UAV dataset based on simulation and real-shot data is proposed. The dataset includes 6 sets of trajectories of three types of terrain: gravels, mountains, and craters, with a total of 3,000 images from a UAV perspective. At the same time, manual calibration generated 105 real-shot images and 95 simulated images of the Ingenuity.

II. RELATED WORKS

A. Image matching based absolute localization

Depending on the method of image matching, image-based absolute localization can be categorized into template-

based and feature-based approaches. Template-based absolute localization has limitations, which uses the UAV image as a template and slides it over a geo-referenced map to determine the current position of the UAV [32]. It relies solely on pixel differences between the UAV and satellite images, neglecting style, scale, and resolution differences. Researchers have explored using Scale-Invariant Feature Transform [3] (SIFT) descriptors to improve robustness. SIFT features are scale-invariant and remain consistent under various transformations, enabling accurate correspondence between UAV and satellite images. This allows for estimating the position within the remote sensing map based on SIFT features and homography estimation.

Subsequently, more feature descriptors for image matching were developed, such as SURF [16], ORB [14], and AKAZE [15], among others. However, these methods rely on pixel variations in images and often struggle to extract features effectively in regions with limited texture information, such as plains and deserts. Hence, some researchers explored deep features for robust inter-domain image matching. In 2020, Sarlin et al. enhanced matching accuracy with SuperGlue [19], using local and global attention mechanisms on SuperPoint [18] features. However, direct image matching struggled due to UAV-satellite scale differences. In 2022, Luo [17] proposed a scale-adaptive image-based localization method, focusing on global coarse and local precise localization to boost accuracy. These methods prioritize image matching but neglect 3D spatial data, limiting 6-DoF pose estimation. In addition, without utilizing the temporal relationship of the UAV image sequence, the trajectory estimated by the absolute localization method is not smooth, which poses a great challenge to the UAV control system.

B. SLAM for UAV localization

Simultaneous Localization and Mapping (SLAM) has garnered considerable attention as a technology capable of conducting real-time positioning and map construction, independent of external sensors such as the Global Positioning System (GPS). In 2017, Mur-Artal et. al proposed ORB-SLAM2 [12], a monocular visual SLAM system, utilizing ORB features for robust mapping and localization with a single camera. Renowned for its real-time performance and versatility, it is a popular choice in computer vision and robotics for both indoor and outdoor applications. Researchers have proposed various SLAM algorithms for UAV localization, but they face limitations due to the absence of depth information, making it challenging to estimate the 6-DoF pose in the world coordinate system. This hinders the use of SLAM for UAV localization tasks. To address this, Wan et al. [1] introduced a terrain-assisted SLAM system in 2022 that matches UAV images with terrain maps to obtain the world coordinates of the UAV. This system then combines absolute localization results with visual odometry from SLAM using the LBA algorithm, ultimately determining the UAV 6-DoF pose in the world coordinate system.

However, there are still some unresolved issues. The absolute localization algorithm in [1] uses phase correlation

which is unreliable in planetary imagery with large scale differences. In addition, it uses fixed weights when fusing absolute and relative poses, without considering that wrong absolute poses will affect positioning accuracy.

III. METHOD

A. Localization framework for planetary UAVs

In order to accurately estimate the 6-DoF pose of the UAV in real time, this paper proposes JointLoc based on joint relative and absolute pose estimation, as shown in Fig. 2. Performing absolute localization for each frame of the UAV images will consume a lot of computing resources. Therefore, in order to ensure the real-time performance of JointLoc, this paper performs absolute localization at a fixed frequency, and the localization results are passed to the back-end module of relative localization for pose fusion. The details of pose fusion will be presented in Sec.III-C.

When the absolute localization result is wrong, adding the absolute pose to the back-end module for pose fusion will reduce the positioning accuracy of JointLoc. In order to assign appropriate weights to different absolute localization results, a novel confidence mechanism based on global and local distribution characteristics for absolute localization results is proposed. Based on the confidence, soft pose fusion can still be performed when the absolute pose and relative pose deviations are large. The detailed design of the confidence mechanism can be seen in Sec.III-B.4.

B. Coarse-to-fine absolute localization

To swiftly and accurately locate the given UAV image \mathcal{I}_{uav} within the world coordinate system based on the satellite map \mathcal{I}_{sat} , we propose a coarse-to-fine absolute localization module. This module initially partitions the satellite map into several local maps. Utilizing the matching results, it anticipates the local map where the UAV may be positioned in the subsequent moment, thereby minimizing the need to re-search all local maps and consequently reducing the positioning time.

1) *Local satellite maps partitions*: To address the challenge posed by the large scale of the satellite map \mathcal{I}_{sat} and its direct matching with the UAV image \mathcal{I}_{uav} - an approach that adversely affects localization time and accuracy - we adopt a preprocessing step. Prior to localization, \mathcal{I}_{sat} is divided into several local satellite maps $\mathcal{I}_{\text{cand}} = \{\mathcal{I}_{\text{sat}_1}, \mathcal{I}_{\text{sat}_2}, \dots, \mathcal{I}_{\text{sat}_N}\}$, where $\mathcal{I}_{\text{sat}_i}$ denotes the i -th cropped local map and N represents the total number of local maps. In the experiment, N is set to 40. Each local map $\mathcal{I}_{\text{sat}_i}$ shares the same dimensions as \mathcal{I}_{uav} . To preserve salient features and avoid information loss during cropping, this study adopts a sliding window cropping approach with a duplication rate r . In our experiments, r is set to 0.5.

2) *SuperPoint - Local feature extraction*: During the matching process, the absolute localization module employs SuperPoint [11] as the local feature extractor to extract feature points p and descriptors d from the candidate local satellite map $\mathcal{I}_{\text{sat}_c}$ and the UAV image \mathcal{I}_{uav} . Specifically, the SuperPoint network utilizes a VGG13 architecture to

initially encode the input image, producing the local feature representation. Subsequently, this local feature is fed into both the feature point decoder and the descriptor decoder. The feature point decoder, consisting of a convolutional layer followed by a softmax layer, generates the feature points p . Concurrently, the local feature is processed by the descriptor decoder, comprising a convolutional layer, an interpolation operation, and a normalization layer, to produce the descriptor d .

3) *LightGlue - Local feature matching*: After extracting the feature points and descriptors from \mathcal{I}_{uav} and $\mathcal{I}_{\text{sat}_c}$, the subsequent task involves feature matching to establish correspondences between feature points and estimate the homography matrix, which represents the transformation between \mathcal{I}_{uav} and $\mathcal{I}_{\text{sat}_c}$. In this regard, this paper utilizes the LightGlue network [10] for matching the UAV image with the local satellite map, as illustrated in Fig. 3. The objective is to derive local matching degree $\sigma_c = [\sigma_1, \sigma_2, \dots, \sigma_M]$ in the association and distribution of feature point pairs. σ_i denotes the matching degree of the i -th pair of feature points and M is the total number of the pairs of the feature points.

Upon the termination of the matching procedure, an assignment matrix and the matchability between the two sets of pruned feature points \hat{p}^{uav} and \hat{p}^{sat} are computed [33]–[35]. The estimated homography matrix \hat{H} between the UAV image and the local satellite map is then obtained via OpenCV [20], according to the feature points. Then the coordinate of the UAV (P_x^G, P_y^G) in the global map can be estimated as

$$\begin{pmatrix} P_x^G \\ P_y^G \\ 1 \end{pmatrix}^T = \begin{pmatrix} x_c \\ y_c \\ 1 \end{pmatrix}^T \times \hat{H} + \begin{pmatrix} x_0 \\ y_0 \\ 1 \end{pmatrix}^T \quad (1)$$

where (x_c, y_c) and the (x_0, y_0) are the coordinates of the center point and upper left corner of $\mathcal{I}_{\text{sat}_c}$, respectively. The altitude P_z^G of the UAV is obtained through the altimeter. To simulate real flight conditions, this paper introduces Gaussian error on the ground truth z-axis value to represent the altitude of the UAV. Therefore, the 3-DoF pose of the UAV estimated by absolute localization is

$$P^G = (P_x^G, P_y^G, P_z^G + \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt) \quad (2)$$

where the $\frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$ is the Gaussian noise.

4) *Adaptive subsequent area prediction*: To expedite localization, we introduce an adaptive subsequent area prediction strategy tailored for large-scale image localization. Specifically, we devise an adaptive confidence for absolute localization results, which considers both global and local distribution characteristics among images. The local satellite map with the highest confidence is considered to be the location where the UAV is currently located. By utilizing the currently determined local satellite map as the center and defining the search range, generated based on confidence level w , as the radius, we predict the area of the UAV in the subsequent step, as shown in Fig. 4.

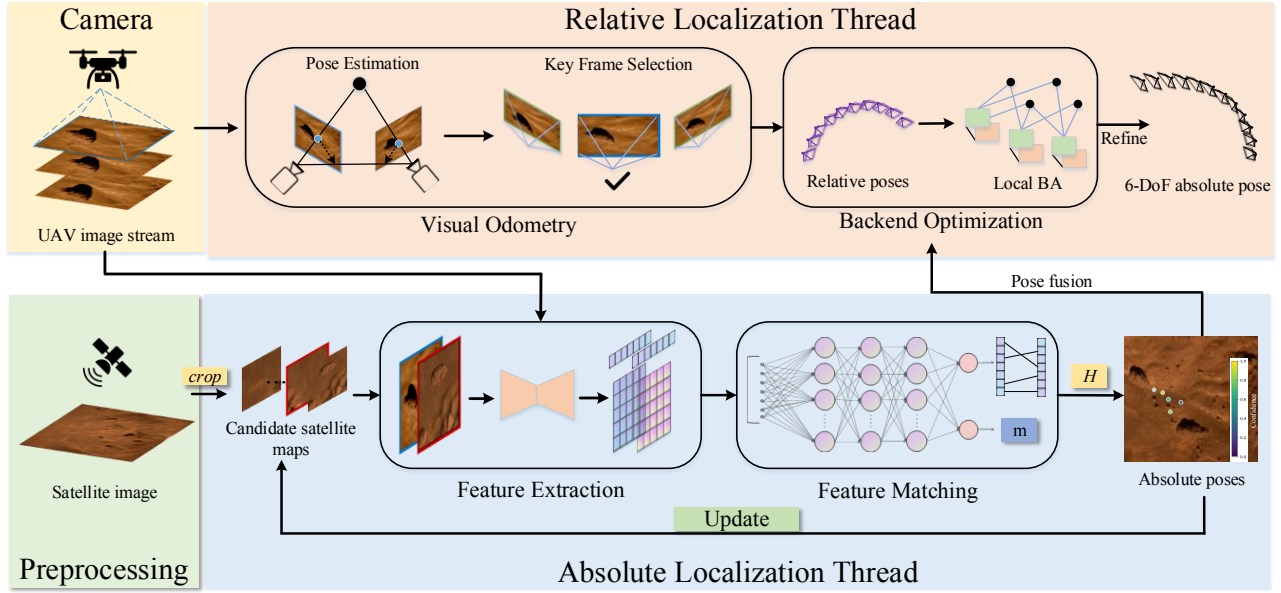


Fig. 2. The pipeline of the proposed JointLoc consists of two threads: the absolute localization thread and the relative localization thread. In the relative localization thread, the UAV image is fed into visual odometry for relative pose estimation and keyframe selection. Simultaneously, the absolute localization module matches the UAV image with the candidate satellite map to estimate the absolute 2-DoF pose of the UAV. These absolute localization and visual odometry results are jointly processed by a pose fusion-based local bundle adjustment (LBA) algorithm. Then the transformation matrix from the relative coordinate system to the world coordinate system is calculated, resulting in the motion trajectory in the world coordinate system.

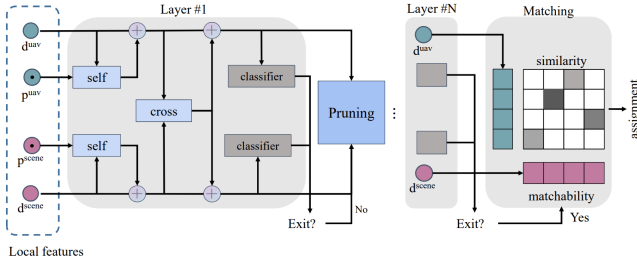


Fig. 3. LightGlue matches the feature points.

The structural similarity (SSIM) between the shared regions of $\mathcal{I}_{\text{sat}_c}$ and \mathcal{I}_{UAV} is utilized to gauge the similarity of the overall image distribution. Additionally, the average of the matching degrees of the feature points is employed to assess the similarity of the local feature points. Hence, the adaptive localization confidence is expressed as follows:

$$w = SSIM(\mathcal{I}_{\text{sat}_c}, \mathcal{T}(\hat{H}, \mathcal{I}_{\text{uav}})) \cdot \sum_{i=1}^M \sigma_i / M \quad (3)$$

, where $\mathcal{T}(\hat{H}, \mathcal{I}_{\text{uav}})$ represents the image obtained by perspective transforming \mathcal{I}_{uav} according to \hat{H} . Based on the confidence level w , the search radius $r \in [1, 10]$ of the potential area where the UAV may be located next is determined as follows:

$$r = \lceil 10^{1-w} \rceil \quad (4)$$

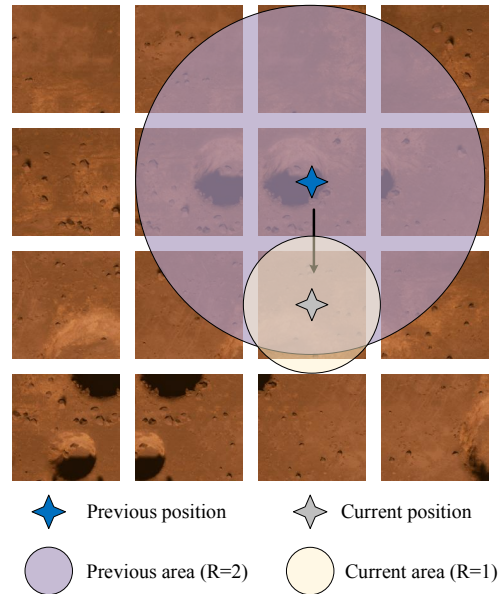


Fig. 4. Schematic diagram of the subsequent area prediction strategy for large-scale image localization. The confidence level from the last UAV localization is 0.7. Accordingly, a search radius of $\lceil 10^{1-0.7} \rceil = 2$ is employed, with the local map containing the current location serving as the center. This search aims to identify the local map that offers the closest match to the current UAV image. Subsequently, with the confidence of the current positioning standing at 0.9, the search radius is reduced to 1 when seeking the future UAV position.

C. Absolute and relative poses fusion

The relative localization module employs ORB-SLAM2 with UAV images as input to estimate the local 6-DoF poses of the UAV. Upon receiving a UAV image, it is simultaneously processed by both the absolute localization module and the visual odometry tracking module. The outcomes from absolute localization, in conjunction with the pose estimations from visual odometry, are integrated into the LBA for back-end optimization. Through this process, rotation and translation matrices from the SLAM coordinate system to the world coordinate system are estimated using the trajectories obtained from absolute localization and visual odometry. This culminates in the derivation of a comprehensive UAV motion trajectory within the world coordinate system.

1) *Visual odometry-based feature tracking*: Visual odometry (VO) entails estimating UAV motion from UAV camera images by comparing consecutive image pairs. Meanwhile, ORB-SLAM2 leverages ORB features for real-time tracking. In VO, UAV pose estimation is relative to the initial pose within the SLAM coordinate system. However, optimizing the pose for each frame individually in the backend can considerably slow down the entire localization system. Therefore, only keyframes are forwarded to the backend of the local mapping thread. To ensure the adaptive selection of keyframes that accommodates both inter-frame matching in SLAM and absolute localization, this paper introduces an enhanced keyframe selection strategy based on the existing keyframe selection approach.

- More than 10 frames have passed since the last keyframe.
- The information entropy of the image satisfies

$$H = - \sum_{i=0}^{255} \sum_{j=0}^{255} P_{i,j} / \log P_{i,j} > 0.5 \quad (5)$$

where $P_{i,j} = \frac{f(i,j)}{W \times H}$ denotes the probability distribution of pixels.

- The current keyframe has more than 90% more points than the reference keyframe.

2) *Pose fusion-based LBA*: LBA serves as an optimization technique employed to refine the accuracy of estimated camera poses and scene structures within the SLAM system. Its fundamental aim is to enhance the estimates of camera poses and map points by minimizing reprojection errors, thereby enhancing the precision and coherence of the SLAM system. The conventional BA algorithm optimizes the observed camera poses by constraining the disparity between the coordinates of the observed feature points and their corresponding reprojected coordinates:

$$e_{i,j} = z_{i,j} - h(T_i, M_j) \quad (6)$$

where $e_{i,j}$ denotes the reprojection error of the 3D map point M_j with the UAV camera extrinsic parameter T_i corresponding to the i -th image. $z \triangleq [u_s, v_s]^T$ denotes the observed pixel coordinates of the feature point. $h(T, p)$ is the reprojection function, which reprojects the 3D point

P corresponding to the feature points into the camera plane according to the external matrix T of the UAV camera.

To mitigate trajectory drift in the tracking module of monocular ORB-SLAM2 using the outcomes of absolute localization, this study adopts a pose fusion-based BA algorithm capable of integrating absolute coordinates. Based on the acquired absolute localization trajectory $q_{1,2,\dots,m}$ and visual odometry trajectory $p_{1,2,\dots,m}$, where m represents the total number of poses in each trajectory, the rotation and translation matrices from the SLAM coordinate system to the world coordinate system can be estimated. Specifically, the objective is to estimate the optimal rotation matrix R and offset t such that the objective function is optimized:

$$(R, t) = \operatorname{argmin} \frac{1}{n} \sum_{i=1}^T \|q_i - (Rp_i + t)\|_2^2 \quad (7)$$

where q_i and p_i denote the i -th keyframe poses in global trajectory and visual odometry trajectory, respectively. Our method maintains two sets of trajectories as continuously updated global variables. This ensures that the transformation matrix from the SLAM coordinate system to the world coordinate system can be dynamically adjusted based on the latest pose relationships. By utilizing the estimated rotation and translation matrices, all poses estimated by the visual odometry can be transformed into the world coordinate system.

Consequently, the cost function $f(x)$ to be optimized in the BA algorithm is formulated as follows:

$$f(x) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^M (e_{i,j}^T w_i e_{i,j} + e_i^{G^T} w_i^G e_i^G) \quad (8)$$

where w_i is the confidence of observation z_i . In contrast to the traditional LBA algorithm, the improved LBA introduces global coordinate constraints as an additional factor in the optimization process by incorporating the deviation $e_i^G = [e_x, e_y, e_z]$ with the absolute localization confidence w_i^G between the results of the absolute localization and the visual odometry that is defined as $e_i^G = |P^G - P^L|$, where P^G and P^L are the results of absolute localization and visual odometry, respectively.

It is worth noting that due to the influence of the reprojection error function, this cost function is not linear. Therefore, in the actual optimization process, $f(x)$ is first linearized through a Taylor expansion and then refined step by step using the Levenberg-Marquard method.

IV. EXPERIMENT RESULTS

A. Experiment platform and hyperparameter settings

The experiment platform operates within a Linux-based environment, utilizing Ubuntu 20.04 as the operating system. The experiments are conducted on a research server that features an Intel Core i9-9900K processor clocked at 3.6 GHz, 64GB of DDR4 RAM, and an NVIDIA GeForce RTX 2080Ti GPU.

B. Dataset

This paper produced a dataset including simulation data and real-shot data on Mars, some samples of which are shown in Fig. 5. We simulated three typical Martian terrains based on Unreal Engine 4 and Airsim, namely impact craters, gravel, and mountains. The UAV recorded overhead image sequences over these three terrains, including 925, 760, and 773 images respectively, with an average trajectory length of 149.63 meters. Our real-shot Mars data comes from images of the Martian surface taken by the lander of the Perseverance, including impact craters, gravels, and mountains. In each terrain, we selected 35 representative lander images and manually calibrated the homography matrix between them and the local map of Mars.

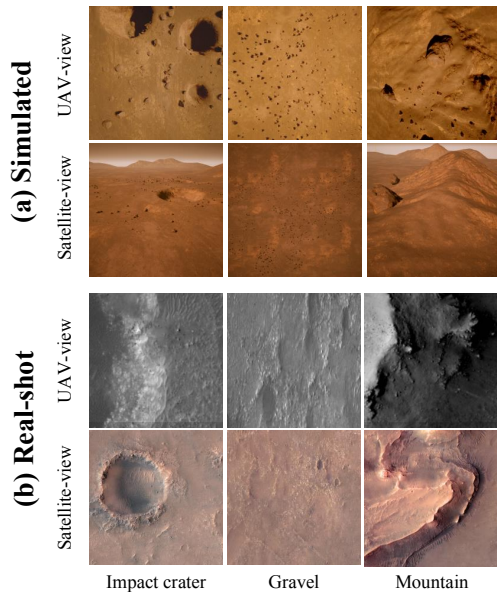


Fig. 5. Samples of the proposed planetary UAV localization datasets. The first two rows are simulated satellite maps and UAV images, and the latter two rows are Martian surface images taken by the Perseverance lander².

C. Evaluation metrics

1) *Metrics for image matching-based absolute localization*: For comparing image matching algorithms, we assess precision in absolute localization using homography estimation. We record error proportions for each method within 5 pixels (@5px), 10 pixels (@10px), 50 pixels (@50px) and 100 pixels (@100px). Furthermore, to assess the applicability of a localization algorithm in real-time systems, we also measured the time required for each algorithm to conduct a single matching operation.

2) *Metrics for visual localization*: To comprehensively compare visual localization algorithms, we use evo [2] for accuracy evaluation. By computing differences between estimated and ground truth trajectories, we derive statistics for five key metrics: standard deviation (std), root mean square error (rmse), minimum error (min), maximum error

(max), and median error (median). These metrics offer a thorough assessment of various visual localization algorithm performances.

D. Ablation study

1) *Ablation study on pose fusion frequency*: To investigate the optimal fusion frequency for absolute and relative localization to ensure the safety of planetary exploration while saving energy as much as possible, we set the intervals for absolute localization to 1, 10, 50, and 100. The experimental results are shown in Table I.

TABLE I

THE RESULTS OF ABLATION STUDY ON THE POSE FUSION FREQUENCY. - INDICATES THAT POSITIONING FAILED AT THIS FREQUENCY.

Frequency	Max(m)	Rmse(m)	Std(m)	Mean(m)	Speed(Hz)
1	1.003	0.219	0.133	0.184	17.887
10	0.947	0.237	0.113	0.209	20.995
50	44.374	29.303	10.158	27.486	37.280
100	-	-	-	-	62.813

It can be found that the positioning error obtained with a pose fusion frequency of 50 is too large, and SLAM failed to be successfully initialized in the experiment with a fusion frequency of 100. Therefore, this paper chooses 10 as the fusion frequency, which can accurately estimate the absolute 6-DoF pose at a speed of 20.995 Hz.

2) *Ablation study on adaptive area prediction*: To confirm the increased absolute localization speed from the area prediction strategy, we performed ablation experiments targeting the search range. Detailed results are in Table II. The strategy significantly improved speed without significant accuracy loss, suggesting its suitability for real-time systems, enhancing speed while maintaining accuracy.

TABLE II

THE RESULTS OF ABLATION STUDY ON THE AREA PREDICTION. *w/* AND *w/o* DENOTE THE ABSOLUTE LOCALIZATION METHODS WITH AND WITHOUT AREA SEARCHING, RESPECTIVELY.

Terrain	Method	@1px / @5px / @10px	Time(ms)
Crater	w/	0.2 / 0.5 / 0.9	81.91
	w/o	0.2 / 0.5 / 0.9	793.12
Gravel	w/	0.2 / 0.4 / 1.0	90.41
	w/o	0.3 / 0.6 / 1.0	868.11
Mountain	w/	0.0 / 0.5 / 0.8	81.91
	w/o	0.2 / 0.7 / 1.0	770.92

3) *Ablation study on absolute localization confidence*: In order to explore the impact of the proposed confidence mechanism on positioning accuracy, we compare the proposed confidence with 5 fixed confidences, SSIM-based confidence, and matching degree-based confidence. The results of this comparative analysis are presented in Table. III. The approach using absolute localization confidence as LBA reference frame weighting outperforms methods with fixed confidence.

²<https://mars.nasa.gov/mars2020/multimedia/raw-images/>

TABLE III

THE RESULTS OF THE ABLATION STUDY ON THE ABSOLUTE LOCALIZATION CONFIDENCE. *Adaptive* DENOTES OUR METHOD THAT EMPLOYS THE ADAPTIVE ABSOLUTE LOCALIZATION CONFIDENCE.

Confidence	Max(m)	Rmse(m)	Std(m)	Mean(m)
0.2	8.260	3.784	1.872	3.281
0.4	2.898	0.798	0.462	0.643
0.6	4.344	1.332	0.675	1.139
0.8	4.631	2.278	1.037	2.016
1.0	3.655	1.084	0.581	0.906
SSIM	0.948	0.256	0.124	0.225
Matching degree	0.896	0.266	0.122	0.235
Adaptive	0.947	0.237	0.113	0.209

E. Comparison experiments

1) *Image matching-based absolute localization experiment*: The absolute localization module relies heavily on matching UAV and satellite images, making the choice of a robust image-matching method crucial. We select 30 pairs of simulated UAV-local satellite images from different terrains and 20 pairs of real-shot ones from the lander of the Ingenuity as test samples.

We opt for representative image matching methods, encompassing algorithms founded on traditional operators like Scale Invariant Feature Transform (SIFT) [3], sparse feature-based methods like D2Net [4], R2D2 [5], SuperGlue [6], and LightGlue [10], and semi-dense matching methods like DFM [7], LoFTR [8], and COTR [9]. The experimental results of different image matching methods in various terrains are shown in Table IV. Evidently, SIFT, SP+SG, and SP+LG outperform other methods in terms of matching accuracy. Among these three methods, SP+LG has the fastest calculation speed, which is more obvious in the real-shot scene. Therefore, this paper adopts the SP+LG approach as the backbone of the absolute localization module.

TABLE IV

EXPERIMENTAL RESULTS OF DIFFERENT IMAGE MATCHING METHODS ON SIMULATED AND REAL-SHOT SCENES, WHEREAS @X MEANS THE PROPORTION OF MATCHING ERROR IN X PIXELS. THE SP, SG, AND LG DENOTE THE SUPERPOINT, SUPERGLUE, AND LIGHTGLUE, RESPECTIVELY.

Method	Simulation		Mars		
	@5 /@10	Time(ms)	@10 /@50 /@100	Time(ms)	
SIFT + nn	0.87 /0.97	90.47	0.30 /0.80 /0.80	737.12	
D2Net	0.70 /0.83	270.01	0.00 /0.20 /0.20	2983.52	
R2D2	0.60 /0.63	390.71	0.00 /0.00 /0.00	-	
DFM	0.83 /0.93	122.44	0.13 /0.38 /0.38	604.34	
LoFTR	0.60 /0.63	191.13	0.00 /0.00 /0.00	-	
SP+SG	0.93 /0.97	183.19	0.10 /0.60 /0.60	328.60	
SP+LG	0.90 /0.97	74.87	0.20 /0.60 /0.80	269.95	

2) *Visual localization experiment*: We compare JointLoc against SLAM-based ORB-SLAM2 [12] and ORB-SLAM3 [13], and image matching-based SIFT, SuperGlue, LoFTR, and LightGlue across 3 different terrains. We compared these algorithms from the perspective of positioning accuracy, such as mean (m), rmse (m) and std (m), and positioning speed (Hz). The experimental results can be seen in Table.V.

TABLE V

THE AVERAGE OF POSITION ACCURACY (M) AND SPEED (HZ) OF DIFFERENT LOCALIZATION METHODS IN 3 SIMULATED TERRAINS.

Category	Method	Mean / Rmse / Std	Speed
Monocular SLAM	ORB-SLAM2	0.485 /0.594 /0.342	22.191
	ORB-SLAM3	0.420 /0.557 /0.361	30.380
Absolute localization	SIFT	5.776 /6.532 /3.051	1.357
	SuperGlue	5.289 /6.024 /2.855	3.043
	LoFTR	-	-
	LightGlue	5.562 /6.253 /2.856	3.704
Fusion localization	JointLoc	0.209 /0.237 /0.113	20.995

The experimental results reveal that JointLoc surpasses both ORB-SLAM2 and ORB-SLAM3 on most metrics. Meanwhile, LoFTR relying on depth features, encounters localization failures in textureless regions due to its dependence on feature extraction. Moreover, the entirely matching-based approach of SIFT, SuperGlue, and LightGlue exhibits inadequate processing speed during testing, rendering real-time performance unattainable, and their localization errors are notably higher. In contrast, JointLoc maintains a processing speed of 20.995 Hz, accurately estimating the 6-DoF pose.

F. Timing results

We have also counted the average running time of each module of JointLoc across various scenarios, as depicted in Table VI.

TABLE VI

THE AVERAGE RUNNING TIME (MS) OF EACH MODULE OF JOINTLOC IN 3 SIMULATED TERRAINS. ABSLOC AND RELLOC DENOTES THE ABSOLUTE LOCALIZATION MODULE AND THE RELATIVE LOCALIZATION MODULE, RESPECTIVELY.

Module	Unit	Crater / Gravel / Mountains
AbsLoc	Feature extraction	18.922 / 14.768 / 12.505
	Feature matching	42.773 / 29.762 / 37.665
RelLoc	Tracking	28.599 / 29.364 / 26.910
	Mapping	212.550 / 211.579 / 210.452
	Looping	207.995 / 212.817 / 223.534
	Pose fusion-based LBA	403.045 / 398.162 / 412.116

V. CONCLUSIONS

This paper introduces JointLoc, a robust and dependable localization framework used for UAVs to cope with the absence of the GNSS in planetary UAV navigation. JointLoc stands out for its integration of both relative and absolute localization modules, showcased within this paper, while also offering compatibility with a wide array of existing localization techniques for effortless module replacement. Through the adaptive fusion of relative and absolute localization modules, JointLoc enables real-time estimation of

the 6-DoF pose of the UAV in the planetary coordinate system. JointLoc achieves an average positioning error of 0.237 meters at 21 Hz, better than ORB-SLAM2 (0.594 m) and ORB-SLAM3 (0.557 m). Moreover, a specialized UAV dataset within a planetary context is proposed, including both simulation data from AirSim and the images of the surface of Mars captured by the lander of the Perseverance, further enriching related research fields.

VI. ACKNOWLEDGEMENTS

Thanks to the Mars Exploration Program of NASA, the real-shot part of our dataset is based on the images captured by the Ingenuity helicopter.

REFERENCES

- [1] X. Wan, Y. Shao, S. Zhang, and S. Li, "Terrain aided Planetary UAV Localization based on Geo-referencing," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–18, Jan. 2022, doi: 10.1109/tgrs.2022.3198745.
- [2] M. Grupp, evo: Python package for the evaluation of odometry and SLAM. <https://github.com/MichaelGrupp/evo>, 2017. [Online]. Available: <https://github.com/MichaelGrupp/evo>
- [3] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, 1999, vol. 2, pp. 1150–1157.
- [4] M. Dusmanu et al., "D2-net: A trainable cnn for joint description and detection of local features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8092–8101.
- [5] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, "R2d2: Reliable and repeatable detector and descriptor," *Advances in neural information processing systems*, vol. 32, 2019.
- [6] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.
- [7] U. Efe, K. G. Ince, and A. Alatan, "Dfm: A performance baseline for deep feature matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4284–4293.
- [8] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.
- [9] W. Jiang, E. Trulls, J. Hosang, A. Tagliasacchi, and K. M. Yi, "Cotr: Correspondence transformer for matching across images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6207–6217.
- [10] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "LightGlue: Local Feature Matching at Light Speed," *arXiv preprint arXiv:2306.13643*, 2023.
- [11] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.
- [12] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.
- [13] C. Campos, R. Elvira, J. J. G. Rodriguez, J. M. Montiel, and J. D. Tardos, "Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.
- [14] Rublee, E., Rabaud, V., Konolige, K. & Bradski, G. ORB: An efficient alternative to SIFT or SURF. *2011 International Conference On Computer Vision*. pp. 2564-2571 (2011)
- [15] Alcantarilla, P. & Solutions, T. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell.* **34**, 1281-1298 (2011)
- [16] Bay, H., Ess, A., Tuytelaars, T. & Van Gool, L. Speeded-up robust features (SURF). *Computer Vision And Image Understanding*. **110**, 346-359 (2008)
- [17] Luo, X., Tian, Y., Wan, X., Xu, J. & Ke, T. Deep learning based cross-view image matching for UAV geo-localization. *2022 International Conference On Service Robotics (ICoSR)*. pp. 102-106 (2022)
- [18] DeTone, D., Malisiewicz, T. & Rabinovich, A. Superpoint: Self-supervised interest point detection and description. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition Workshops*. pp. 224-236 (2018)
- [19] Sarlin, P., DeTone, D., Malisiewicz, T. & Rabinovich, A. Superglue: Learning feature matching with graph neural networks. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 4938-4947 (2020)
- [20] DeTone, D., Malisiewicz, T. & Rabinovich, A. Deep image homography estimation. *ArXiv Preprint ArXiv:1606.03798*. (2016)
- [21] Sharma, M., Gupta, A., Gupta, S., Alsamhi, S. & Shvetsov, A. Survey on Unmanned Aerial Vehicle for Mars Exploration: Deployment Use Case. *Drones*. **6** (2022), <https://www.mdpi.com/2504-446X/6/1/4>
- [22] Schuster, M. Collaborative localization and mapping for autonomous planetary exploration: Distributed stereo vision-based 6D SLAM in GNSS-denied environments. (University of Bremen,2019)
- [23] Hong, S., Bangunharcana, A., Park, J., Choi, M. & Shin, H. Visual SLAM-based robotic mapping method for planetary construction. *Sensors*. **21**, 7715 (2021)
- [24] Giubilato, R., Chiodini, S., Pertile, M. & Debei, S. An experimental comparison of ros-compatible stereo visual slam methods for planetary rovers. *2018 5th IEEE International Workshop On Metrology For AeroSpace (MetroAeroSpace)*. pp. 386-391 (2018)
- [25] Giubilato, R., Gentil, C., Vayugundla, M., Schuster, M., Vidal-Calleja, T. & Triebel, R. GPGM-SLAM: A robust SLAM system for unstructured planetary environments with Gaussian process gradient maps. *ArXiv Preprint ArXiv:2109.06596*. (2021)
- [26] Geromichalos, D., Azkarate, M., Tsardoulias, E., Gerdes, L., Petrou, L. & Perez Del Pulgar, C. SLAM for autonomous planetary rovers with global localization. *Journal Of Field Robotics*. **37**, 830-847 (2020)
- [27] Colosi, M., Haug, S., Biber, P., Arras, K. & Grisetti, G. Better lost in transition than lost in space: Slam state machine. *2019 IEEE/RSJ International Conference On Intelligent Robots And Systems (IROS)*. pp. 362-369 (2019)
- [28] Balaram, J., Aung, M. & Golombek, M. The ingenuity helicopter on the perseverance rover. *Space Science Reviews*. **217**, 56 (2021)
- [29] Fisher, A. & Johnson, A. After Three Years on Mars, NASA's Ingenuity Helicopter Mission Ends. (2024,1,25), <https://mars.nasa.gov/news/9540/after-three-years-on-mars-nasas-ingenuity-helicopter-mission-ends/>
- [30] Furgale, P., Carle, P. & Barfoot, T. A comparison of global localization algorithms for planetary exploration. *2010 IEEE/RSJ International Conference On Intelligent Robots And Systems*. pp. 4964-4969 (2010)
- [31] Pavlenko, T., Schütz, M., Vossiek, M., Walter, T. & Montenegro, S. Wireless local positioning system for controlled UAV landing in GNSS-denied environment. *2019 IEEE 5th International Workshop On Metrology For AeroSpace (MetroAeroSpace)*. pp. 171-175 (2019)
- [32] Cao, Y., Ren, K. & Chen, Q. Template matching based on convolution neural network for UAV visual localization. *Optik*. **283** pp. 170920 (2023)
- [33] Wang, X., Xu, R., Cui, Z., Wan, Z. & Zhang, Y. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. *Advances In Neural Information Processing Systems*. **36** (2024)
- [34] Zhao, C. & Zhao, H. Accurate and robust feature-based homography estimation using HALF-SIFT and feature localization error weighting. *Journal Of Visual Communication And Image Representation*. **40** pp. 288-299 (2016)
- [35] Han, K., Aeschliman, C., Park, J., Kak, A., Kwon, H. & Pack, D. Uav vision: Feature based accurate ground target localization through propagated initializations and interframe homographies. *2012 IEEE International Conference On Robotics And Automation*. pp. 944-950 (2012)