Received 21 February, 2025; revised 21 July, 2025; accepted 12 September, 2025.

Digital Object Identifier 10.1109/TFR.2025.3612364

A Map-based Localization System for Ingenuity using Deep Image Matching

Georgios Georgakis[1], Dario Pisanti[2]*, Nathan Williams[1], Cecilia Mauceri[1], Gerik Kubiak[1], Adnan Ansar[1], Roland Brockers[1]

[1]Jet Propulsion Laboratory / California Institute of Technology, Pasadena, California, USA
[2]Space Robotics Research Group, SnT, University of Luxembourg, Luxembourg.

Corresponding author: Georgios Georgakis (email: georgios.georgakis@jpl.nasa.gov).

This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004).
Government sponsorship acknowledged.

ABSTRACT The success of NASA's Mars Helicopter Ingenuity has paved the way for aerial planetary exploration with future mission concepts that will require advanced autonomous capabilities to enable long-range navigation. In the absence of a Global Navigation Satellite System (GNSS) on Mars, a critical capability is localization within the global frame to eliminate pose estimation drift, which typically involves registering onboard images to orbital maps - e.g. derived from High-Resolution Imaging Science Experiment (HiRISE) data. However, the registration process poses several challenges including texture-less terrain, illumination variations, and most relevant to Ingenuity, large resolution difference between low altitude observations and HiRISE. With current registration methods using template-matching and hand-crafted features struggling under the aforementioned challenges, we turn our attention to deep learning-based image matchers that have shown impressive generalization potential, but failed to be widely adopted for space applications due to the lack of large-scale annotated datasets for training. In this paper, we present a Map-based Localization (MbL) system for Ingenuity that incorporates a state-of-the-art deep image matcher model. We justify the feasibility of this approach for future missions by demonstrating a training strategy that: 1) rapidly adapts the deep image matcher in a self-supervised manner using minimal amount of Ingenuity navigation images, 2) generalizes to previously unseen flights, and 3) is robust to the large resolution difference and outperforms prior template and hand-crafted registration methods in terms of localization accuracy.

INDEX TERMS Deep Learning, Map-based Localization, Mars exploration, Mars Helicopter, Vision-based Navigation

I. INTRODUCTION

NASA's Mars Helicopter Ingenuity was a resounding success as a technology demonstration, proving that Unmanned Aerial Vehicles (UAVs) can fly in the thin Martian atmosphere. This success is encouraging rotorcraft-enabled science on Mars and enables new mission concepts such as the Mars Science Helicopter (MSH) [1] and Dragonfly [2], by offering several advantages compared to rovers. First, aerial vehicles can significantly increase the area of operations, facilitating science investigations multiple kilometers away from the landing site in a short period of time. Second, rovers are restricted by the terrain that they can safely traverse. They have difficulty climbing slopes more than 30° and run the risk of getting stuck when traversing fine-grained sand. Aerial assets, on the other hand, are able to avoid troublesome surfaces and land on steep slopes and high elevation terrain.

In spite of its success flying seventy two flights, Ingenuity was limited in several ways. It lacked a safe landing site detection capability, which required engineers to target landing sites with a low risk of landing hazards [3]. The local position estimator relied on a Laser Range Finder (LRF), which limited the maximum altitude of flights to 24m [4]. The local position estimator also assumed a flat

* Work done as a Visiting Student Researcher with the Jet Propulsion Laboratory, California Institute of Technology, while being a Ph.D. Student with Scuola Superiore Meridionale, Naples, Italy.

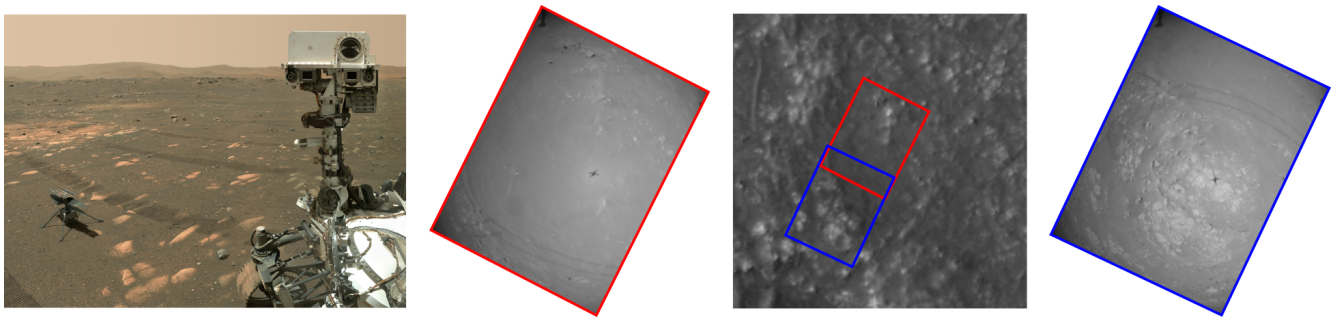

FIGURE 1. Perseverance's selfie with Ingenuity (left). Ingenuity was deployed in April 2021 as a flight tech demo and went on to complete 72 flights on Mars. Examples of the resolution difference, posing a significant challenge for registration, between two Ingenuity navigation images (outlined in red and blue colors) and a 100m $\times$ 100m crop of the HiRISE map (right). Detailed features such as rocks and small elevation changes which are visible in the navigation images are very difficult to associate to the map without taking into consideration the overall context of the scene.

Martian surface [5], which became limiting as Ingenuity began flying over more complex terrain [3]. Without a global position estimate, engineers needed to downlink landing images from the previous flight to re-localize Ingenuity before the next flight could be performed to avoid drift from the onboard state estimator from accumulating. The downlink and manual relocalization is time consuming and can hinder performance evaluation. Many of these limitations were partially addressed throughout the mission through software updates [3], but an on-board global localization capability was never introduced.

In the absence of a Global Navigation Satellite System (GNSS) on Mars, on-board navigation images need to be registered to a global reference frame provided by orbital maps created using either the High-Resolution Imaging Science Experiment (HiRISE) [6] or the Context Camera (CTX) [7] from the Mars Reconnaisance Orbiter (MRO). For Ingenuity, the registration process was performed manually. At the end of each flight, a human annotator would align the navigation image to the HiRISE map. This process provided accurate poses of the UAV only at the landing sites of each flight, with the on-board visual odometry system producing a position error drift of 2 - 12% depending on the difficulty of the terrain during flights of up to $\sim$700m distance [8]. The MSH mission concept is planning for long-range traverses of $\sim$10km where drift is going to be substantially higher. Therefore, periodic drift correction, performed online by registering navigation images to orbital maps, is crucial for enabling safe long-range navigation. The frequency of this operation depends on factors such as the amount of drift, the speed of the UAV, and the mission safety requirements.

However, there are several challenges to registering an on-board image to the orbital map. First, the UAV flights will not always occur at the same time of day as when the orbital map was captured, causing variations in lighting and shadow casting that can significantly alter terrain appearance. Second, scale variation between the map and the on-board navigation image may prove a challenge when matching terrain features, especially when the UAV is operating at low altitudes. Finally, the terrain may be relatively bland or it may contain repetitive patterns (e.g., dune fields) both hindering the identification of distinctive features necessary for registration.

Current methods rely primarily on template-matching techniques, such as Normalized Cross Correlation (NCC) [9], or on classical hand-crafted features, such as the Scale Invariant Feature Transform (SIFT) [10] to solve the registration problem. Specifically, template-matching was part of the Lander Vision System (LVS) [11] for the Mars2020 mission due to its low latency and accurate results when on-board images are ortho-corrected. SIFT showed robust matching in an early study towards a Map-based Localization (MbL) system for a future Mars Helicopter [12]. However, both of these applications made assumptions with regards to scale and illumination variations and underperformed in the absence of texture-rich terrain. While these classical approaches to registration are somewhat robust to scale variation, in-plane rotation, and linear brightness changes (such as intensity shift), they tend to struggle in low-texture situations and under non-linear lighting changes (such as shadows). Specifically for local hand-crafted features, these limitations stem mainly from two sources: 1) They are constructed from local information, and 2) they cannot be tuned to specific data distributions and as a result they may ignore useful patterns in the data.

On the other hand, deep learning-based methods have recently shown impressive performance gains across a wide-range of challenging vision tasks such as image recognition [13], object detection [14], semantic segmentation [15], and monocular depth estimation [16]. This recent success has been fueled by the emergence of Transformers [17] that has also led to the introduction of powerful vision foundation models [18]. For image registration, deep learning has also provided a principled framework for formulating a data-driven supervised-learning task for training robust image matchers [19]–[21]. In contrast to classical methods, deep learning-based approaches are capable of deriving discriminative representations for features by consuming large amounts of data that can bridge appearance variation including scale, observation angle, and, also, non-

linear lighting changes. In spite of their advantages, one of the shortcomings of data-driven methods is that they often exhibit poor performance in dissimilar or out-of-distribution data compared to the model's training data domain. Space-based domains, such as aerial and orbital imagery, are out-of-distribution for many publicly available models, which are typically trained on human-built environments [22]. There is a lack of appropriate large-scale annotated datasets with which these models could be fine-tuned to register on-board UAV images to a reference map.

In this work, we provide a solution for global localization of UAVs on Mars. We focus, specifically, on investigating a fine-tuned, deep learning-based image matcher as part of a Map-based Localization (MbL) system for Ingenuity. We recognize the large resolution difference between navigation images captured at low altitude (8m - 12m) and imagery from HiRISE captured from orbit (see Fig. 1) as the most challenging aspect of this problem that local hand-crafted features are unable to resolve. To address this challenge, we propose to use the Transformer-based LoFTR model [19] that can learn discriminative representations for feature matching by incorporating context from the entire image. In addition, we devise a training strategy that reduces the dependency on large-scale in-domain annotated datasets by first pre-training on rendered data and then fine-tuning on a small number of Ingenuity images in a self-supervised manner. The proposed LoFTR-based MbL system is intended to provide periodic drift corrections for the on-board visual odometry. We evaluate our approach by performing pose estimation on independent navigation frames and demonstrate its superior performance over standard template-matching registration approach for flight systems. In summary, our contributions are as follows.

- An MbL framework incorporates a state-of-the-art deep image matching model that can robustly register Ingenuity navigation images to HiRISE maps.
- A training strategy for rapidly fine-tuning such a deep model on the Ingenuity navigation images.
- Evaluation of our framework in terms of localization accuracy on the Ingenuity flights with 89.4% accuracy at 5m and an almost perfect 99.8% Acc@ 10m.
- Ablation studies show the generalization ability of the model to unseen flights and its capacity to adapt using minimal data for fine-tuning.

II. RELATED WORK

Map-based localization techniques have been extensively studied for UAVs in earth-based applications, leveraging the abundance of geo-referenced satellite imagery and GNSS-based ground truths. In contrast, research into onboard MbL strategies for Mars has been mainly focused on Terrain Relative Navigation (TRN) for Entry, Descend, and Landing (EDL), or on rover-based navigation, with limited attention given to aerial platforms. This is partially due to the scarcity of annotated image datasets that bridge the scale difference between orbital and surface imagery, leaving a critical gap in the altitude range relevant to aerial applications. The successful deployment of Ingenuity sparked interest in vision-based localization methods for UAV in GNSS-denied extraterrestrial environments.

In this section, we categorize existing work based on the image registration methods used in map-based localization pipelines. We first examine template-matching techniques, highlighting successful real-time operations during EDL and rover navigation on Mars, as well as recent research efforts for UAVs. We then review traditional hand-crafted feature matching methods adopted in space applications, followed by learning-based approaches.

A. TEMPLATE MATCHING

MbL in space applications has traditionally been driven by template-matching methods, aligning an ortho-rectified onboard image with a geo-referenced orbital map to compute pixel-wise similarity estimates. The process typically relies on similarity measures such as Normalized Cross-Correlation [9], Phase-Correlation [23], and Mutual Information [24]. The Mars2020 Lander Vision System (LVS) successfully performed onboard and autonomous global localization on Mars during the mission's EDL phase [11]. Their TRN pipeline integrated a coarse-to-fine template-matching approach to register the navigation camera images onto a CTX map (6 m/pixel) of the Jezero crater landing site. The geological diversity of the site, chosen to maximize scientific return, resulted in complex terrain morphology with potential hazards, making autonomous TRN the most critical component for a safe and precise landing [25].

Global localization has also been successfully executed onboard the Perseverance rover with the Censible framework proposed by Nash et al. [26]. The method consists of registering an ortho-mosaic of panoramic stereo images collected by the rover's navigation camera onto a HiRISE map (0.25 m/pixel) using a modified census transform [27]. Census is a template-matching non-parametric transform that depends on the relative ordering of pixel intensities. By ensuring sub-meter localization accuracy, their approach demonstrated performance on par with human-in-the-loop localization. A global localization approach for autonomous planetary rovers is also proposed in Geromichalos et al. [28], where the error drift produced by the onboard simultaneous localization and mapping (SLAM) algorithm is corrected by registering the generated local map onto a global orbital map using template-correlation.

Although template matching proved successful and reliable in the aforementioned cases, it is generally sensitive to lighting, viewpoint and in-plane rotation variations, necessitating correction steps. A phase-correlation-based pipeline for UAV global localization on HiRISE maps has been proposed by Wan et al. [23] to improve robustness to lighting variations throughout the Martian day. However, the method requires relatively large overlap between the template and

the map, imposing strict constraints on scale invariance. Additionally, template pre-processing is still needed to handle view-point and rotation changes.

B. HAND-CRAFTED FEATURE MATCHING

Feature keypoints are typically detected in high-contrast regions of an image, such as corners, edges or blobs. Hand-crafted feature-matching relies on manually designed detectors and descriptors to identify features with built-in invariance to scale, view-point, rotation or lighting.

Popular hand-crafted features such as SIFT [10], ORB [29], SURF [30] have been investigated for MbL in UAV applications. SIFT is one of the most widely used local hand-crafted descriptors that encodes the orientation and magnitude of image gradients around keypoints. SURF builds on SIFT by accelerating its feature computation, while ORB combines a FAST detector [31] with a BRIEF descriptor [32] to offer a fast and rotation-invariant alternative.

The performance of these features for UAV global localization was assessed in the work of Brockers et al. [12], which proposed an autonomous on-board pipeline to register simulated aerial images on Mars onto a HiRISE map. SIFT was demonstrated to achieve the highest localization accuracy, showing robustness to scale and view-point changes. However, its performance was negatively impacted by low sun elevation angles, where learning-based methods, such as SuperPoint [33], showed better performance.

C. LEARNING-BASED APPROACHES

Deep learning profoundly transformed the image matching problem, surpassing traditional computer vision methods in modern UAV visual localization tasks. By leveraging Convolutional Neural Networks (CNNs) and transformer-based architectures, learning-based methods can extract highly discriminative and robust feature representations, learning hierarchical features and pixel-wise contextual relationships directly from image data.

Several deep learning-based frameworks have been introduced to enhance image matching accuracy with greater robustness to variations in scale, illumination, and view-point. Among the popular self-supervised methods, SuperPoint [33] uses CNNs to jointly detect keypoints and generate corresponding descriptors. SuperGlue [34] refines SuperPoint feature matching by employing a graph neural network with attention mechanisms. LoFTR [19] introduced a Transformer-based [17] detector-free paradigm, which can leverage the global context provided by Transformers in a coarse-to-fine strategy to produce pixel-wise semi-dense correspondences in low-texture areas of the images. Robust and dense matching under challenging real-world variations is also tackled by RoMa [20], which utilizes features from the DINOv2 [18] vision foundation model, and DKM [21], which estimates a dense warp to pixel-wise matches.

Extensive research has explored learning-based strategies for Earth-based UAV global localization, where navigation images are registered to geo-referenced satellite imagery. The abundance of data fostered the adoption of deep-learning methods into existing pipelines to enhance robustness in challenging conditions. Surveys on deep learning for UAV localization [35], [36] and related applications in GNSS-denied scenarios [37] provide comprehensive reviews of this evolving field. In contrast, deep learning-based MbL on Mars remains largely unexplored, partially due to the limited volume of real aerial imagery, which constrains extensive model training. Recent efforts, such as JointLoc [38], have proposed vision-based UAV localization on Mars using SuperPoint and LightGlue [39], but primarily rely on purely synthetic datasets from artificial environments. In this work, we integrate LoFTR into our MbL pipeline for Ingenuity. To address the data scarcity issue, our training strategy leverages high-fidelity simulated datasets, generated from HiRISE Digital Terrain Models (DTMs) and ortho-projected images, combined with real Ingenuity imagery.

III. PROBLEM DESCRIPTION

We address the problem of global localization for UAVs in the Martian environment. For this work, we consider the task of Map-based Localization for Ingenuity flights as a blueprint for future UAV concepts for Mars exploration. We note that we focus specifically on MbL and consider the integration with VIO as part of a state estimator outside the scope this work.

Ingenuity is equipped with a nadir-pointing navigation camera that produces images I^{raw} at 640×480 resolution. Each image is undistorted using the CAHVOR [40] camera model and then orthoprojected on the map to get the navigation image I^{nav}, used in our MbL system. The reference map I^{map}, is an ortho-image with equirectangular projection created from HiRISE observations with 0.25m/pixel resolution. The size of the map is approximately $20km \times 20km$ and depicts the Jezero crater landing site for the Mars2020 mission.

During each flight, the on-board Visual Inertia Odometry (VIO) system produces relative position and orientation estimates [8]. The VIO estimate is propagated from the starting position, which is manually annotated before the flight, to produce an initial noisy global position, $\mathbf{t}^{vio} \in \mathbb{R}^2$, that serves as a position prior to identify a local map search area, $I^{map}(\mathbf{t}^{vio})$, in the MbL system.

Given a registration algorithm, $\mathcal{G}$, that produces correspondences between I^{nav} and $I^{map}(\mathbf{t}^{vio})$, our objective is to estimate the drift-free 2D global position:

$$\mathbf{t}^{nav} = \mathcal{H}\left(\mathcal{G}(I^{nav}, I^{map}(\mathbf{t}^{vio}))\right) \tag{1}$$

where $\mathcal{H}$ is a function representing an MbL pipeline that uses an affine transformation to align the navigation image to the local area of the map.

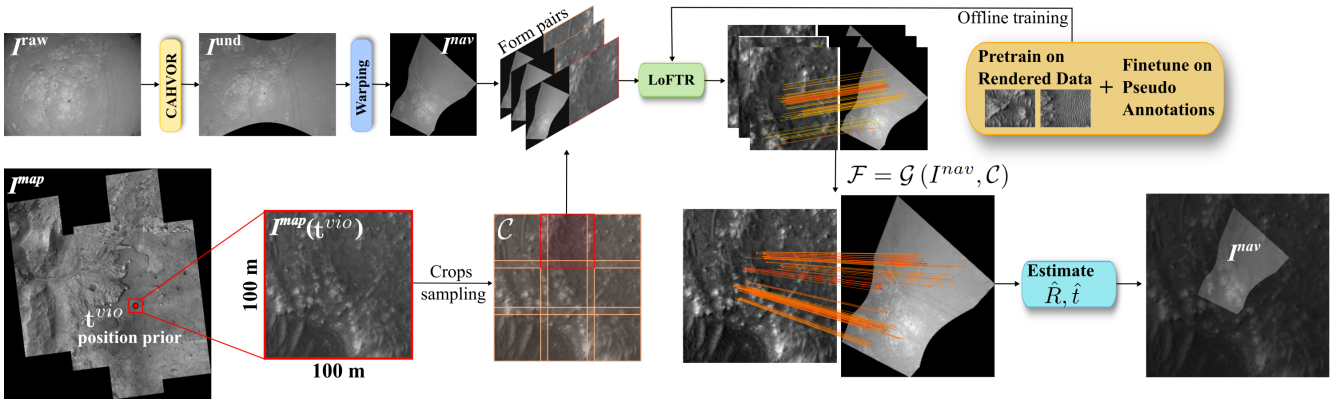

FIGURE 2. **Overview of the Map-based Localization pipeline. Using the position prior from VIO, $\mathbf{t}^{vio}$, we select a local search area $I^{map}(\mathbf{t}^{vio})$ on the map to register the orthoprojected Ingenuity navigation image, I^{nav}. This is obtained by first undistorting the raw image, I^{raw}, with a CAHVOR camera model and then warping the undistorted image, I^{und}, on the map. The registration is carried out by matching the navigation image on a set of crops, $\mathcal{C}$, of the search area using a fine-tuned LoFTR model. The model is pretrained using a large rendered dataset followed by fine-tuning on a small set of navigation images in a self-supervised manner. Finally, once a set of matches, $\mathcal{F}$, is established between the search area and I^{nav}, we estimate an affine transformation, $\hat{R}, \hat{t}$, that aligns the navigation image to the map.**

IV. MbL SYSTEM FOR UAVs ON MARS USING DEEP IMAGE MATCHING

We aim to define a registration algorithm, $\mathcal{G}$, and an MbL pipeline, $\mathcal{H}$, from Eq. 1 that can address the large scale difference present between the Ingenuity navigation images at low altitudes (8-12m) and the HiRISE map. Following the details about Ingenuity's camera specifications [41], at these altitudes the resolution would be between 0.028m/pixel and 0.042m/pixel. Since HiRISE resolution is 0.25m/pixel the scale factor is $\sim$9 at 8m altitude and $\sim$6 at 12m altitude. Due to this large scale difference, detailed terrain features that are visible in the navigation images are either absent or appear as coarse blobs in the map.

Registration methods using hand-crafted features, such as SIFT [10], frequently fail on this task as they depend on local appearance similarity to establish matches. Template-matching methods such as the Census transform [27] have shown some level of robustness to this problem when using larger templates that allow capturing high-level structural information, but they suffer in the presence of texture-less terrains. In contrast, the transformer-based LoFTR [19] model learns how to combine information in a global image context when composing pixel-wise feature representations. This allows the model to exploit discriminative features or structural patterns in a larger area and thus be more robust to texture-less terrain or large resolution difference.

Therefore, we use a fine-tuned LoFTR model as the registration algorithm in the MbL system. Prior to registration, the navigation image is orthoprojected into the map frame. The system then attempts to register it using LoFTR on a sequence of local map crops. The crops are dynamically selected given the initial global position estimate from VIO and the dimensions of the orthoprojected image. From this sequence, all matches are collected and filtered, before a 2D transformation is estimated that aligns the navigation image to the map. An overview of this process is shown in Fig. 2.

The subsections describing the different parts of our method are organized as follows. First, in Section A, we discuss background information regarding Transformers [17] and LoFTR [19] that contextualizes our choice of the registration algorithm. Then, in Section B, we present our approach for fine-tuning LoFTR on the Ingenuity flights, and finally in Section C, we describe the steps taken in our MbL pipeline.

A. PRELIMINARIES

1) OVERVIEW OF TRANSFORMERS

Originally developed for neural language processing problems, *Transformers* [17] have been increasingly and widely employed in computer vision tasks due to their simplicity and ability to learn meaningful associations within long sequences of data. At the core of these architectures lie the attention layers that learn contextual relationships between elements within an input data sequence (*self-attention*) or across different data sequences (*cross-attention*).

Given the input sequences F^i, $F^j \in \mathbb{R}^{N \times D}$ made of N vectors of dimension D, these are projected into distinct representation subspaces referred as Query, $Q = F^i W_Q \in \mathbb{R}^{N \times D_k}$, Key, $K = F^j W_K \in \mathbb{R}^{N \times D_k}$, and Value, $V = F^j W_V \in \mathbb{R}^{N \times D_v}$, using the learnable weight matrices W_Q, $W_K \in \mathbb{R}^{D \times D_k}$ and $W_V \in R^{D \times D_v}$. In the case of self-attention, $F^i = F^j$, while cross-attention applies otherwise.

The attention mechanism measures the relevance of each key vector in K to each query in Q with similarity scores computed from the dot-products $QK^T/\sqrt{D_k}$. Applying the softmax function yields the *attention weights* representing the contribution of each value vector in V to the resulting weighted sum. Thus, the attention function can be expressed as:

$$Att.(Q,\, K,\, V) = softmax\left(\frac{QK^T}{\sqrt{D_k}}\right) V \qquad (2)$$

Intuitively, the attention layer produces a new context vector for each query in Q that retrieves relevant information from all the input elements in relation to that query. In image matching, input sequences correspond to feature maps extracted from the images by CNN backbones. A Transformers-based approach like LoFTR offers the advantage of providing local features with context-rich and pixel-position-dependent representations that can effectively capture long-range dependencies between the pixels. This learned behavior enables dense feature matching in textureless or repetitive regions, where most detector-based methods often fail to produce consistent keypoints.

2) LOCAL FEATURE MATCHING WITH TRANSFORMERS (LoFTR)

LoFTR [19] produces semi-dense pixel-wise matches between two images I^A and I^B. The method leverages Transformers with self- and cross-attention layers to process feature maps from a ResNet-18 [42] backbone at two scales $\tilde{F}^A \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times d}$ and $\hat{F}^A \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times d}$, where H, W denote height and width of I^A and d is the feature dimension. Feature maps are produced from I^B with the same process.

The transformed feature maps can rely on representations capturing the image's global context to ensure more reliable matches than those produced solely by the CNN backbone. By following a coarse-to-fine approach, the transformed features are first matched at the coarse level to produce a set of predicted matches according to a confidence threshold and mutual-nearest-neighbor criteria. The coarse matches are then refined within local windows cropped from the higher resolution feature maps and a set of fine-level matches is generated with associated confidence scores.

Instead of directly applying the vanilla dot-product attention mechanism, LoFTR incorporates linear-attention layers [43] that exploit an Exponential Linear Unit (ELU) operator and the associativity property of matrix multiplication to bring the cost complexity from $O(N^2)$ down to $O(N)$, with N being the number of feature vectors composing each input sequence.

While any Transformer-based method would be suited as our registration algorithm, we choose LoFTR as a proof-of-concept model for two reasons. First, it has a good performance vs. computational complexity trade-off, and second, it is a modular approach that lends itself favorably to future development.

B. TRAINING LoFTR FOR INGENUITY FLIGHTS

One of the biggest hurdles in training deep learning-based models, is that a large-scale dataset with high-quality annotations is typically necessary, which is extremely time-consuming to obtain, especially when annotations are created with manual labor. For example, LoFTR [19] was trained on the MegaDepth [22] dataset that contains around 1M terrestrial in-the-wild image pairs. In order to avoid manually annotating the images, Structure-from-Motion (SfM) was used to retrieve the camera poses and project points between images to create ground-truth correspondences. In the case of fine-tuning an existing pretrained model, this requirement can be reduced to thousands or tens of thousands of training samples, which still requires a sizeable annotation effort.

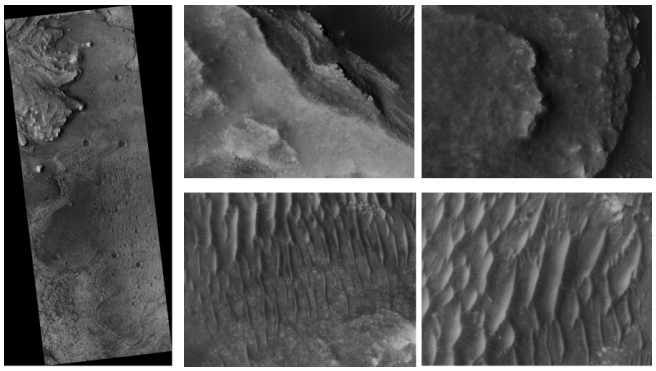

FIGURE 3. Examples from the rendered training set used to pretrain the model that show an orthographic map from HiRISE (left) and simulated navigation images (right).

In our case, we also have another fundamental restriction. There is a limited number of Ingenuity navigation images that can be used for training. At best, long flights have around 170 frames after frames with less than 8m of altitude are filtered out. The 8m threshold was selected because it was the minimum flying altitude of Ingenuity and to avoid frames during ascent and descent. Moreover, even if we had a large number of images, there are several challenges towards obtaining good quality ground-truth correspondences: 1) The VIO poses are too noisy to produce pixel-accurate projections between images, 2) SfM with classical hand-crafted features typically fails due to frequent low-texture terrain, and 3) the large scale difference between navigation images and the map makes manually finding corresponding features to annotate extremely difficult and time-consuming.

Therefore, we devise a strategy that involves two main steps: 1) Training a LoFTR model on a large scale dataset of Mars created in simulation, and 2) Fine-tuning that model on a much smaller number of Ingenuity navigation image-map pairs that are annotated with pseudo-labels from the off-the-shelf LoFTR model.

1) PRETRAINING WITH RENDERED DATA

For the first step, we follow the example of prior work [44] which uses the open-source simulation software, Blender, to generate a large-scale dataset of the Martian surface with the goal of addressing large illumination variation. We use Blender to render the publicly available[1] HiRISE map of the Jezero crater (14km $\times$ 6km) with custom camera properties and lighting.

We generated a dataset comprised of a rendered orthographic map and a large number of nadir-pointing simulated

[1] https://www.uahirise.org/dtm/ESP_045994_1985

navigation images up to 200m altitude using a perspective camera (see Fig. 3). To create a training set, we divide the map into a set of smaller map crops and pair them with navigation images that have at least 25% overlap. Each navigation image can be paired with more than one map crop. These pairs have known navigation camera poses and map crop coordinates, which are used to produce ground-truth correspondences for each training pair. Following this strategy, we randomly sampled observations across different combinations of three altitudes and 12 lighting conditions that resulted in approximately 150K navigation-map training pairs. These pairs were used to further train the off-the-shelf LoFTR model as an intermediate step to bridge the domain gap between its initial training on MegaDepth data and real navigation camera data..

We note that this intermediate training on the large simulated dataset is not meant to directly solve the registration problem for the Ingenuity navigation images, but rather for the model to obtain a suitable initial representation that can be rapidly fine-tuned. Even though the HiRISE rendered images visually differ from the Ingenuity navigation images in terms of resolution and brightness, we found experimentally that this intermediate training allows the model to learn tailored representations for HiRISE maps which are needed to register Ingenuity images and reduce the domain gap between Ingenuity and HiRISE.

2) SELF-SUPERVISION USING PSEUDO CORRESPONDENCES

The next step is to fine-tune the model on a small number of Ingenuity navigation images. Unlike the rendered data where ground-truth correspondences are known and can be used as supervision, we only have noisy estimates for most navigation image poses based on VIO. Instead of using these noisy estimates, we adopt a self-supervised approach by relying on the generalization ability of the off-the-shelf LoFTR model and obtain pseudo ground-truth correspondences between navigation images and the map from several Ingenuity flights. We wish to note that this approach is inspired by tools such as AnyLabeling [45] which uses foundation models to produce semi-automatic annotations.

Our procedure is as follows. First we run the model on navigation-map pairs to obtain initial sets of matches. For each example, we perform homography estimation with RANSAC and set the inlier reprojection error threshold to a strict value of 1 pixel. If the number of inliers exceeds a certain threshold (empirically set to 15) then we keep the homography matrix to be used for creating pseudo ground-truth during the training procedure. A visual inspection also ensures that incorrect homography estimations are excluded from training. This approach allows us to select 177 navigation images with good homography estimations. These are used to form 550 training pairs with map crops. Another 105 pairs are held out for validation. Using these pairs, we fine-tune the model directly on Ingenuity data. In practice, this process can be repeated multiple times, with every iteration using the fine-tuned model from the previous round in order to obtain the initial sets of matches. However, we found experimentally that for our particular domain a single iteration proved sufficient.

C. MbL PIPELINE

Our objective is to register a navigation image from Ingenuity and produce a position estimate with respect to an orbital map. To do so, the MbL pipeline $\mathcal{H}$ from Eq. 1 can be realized in two ways: 1) Orthoproject (warp) the undistorted navigation image, I^{und}, on the map using the onboard state-estimator rotation and altitude prior, followed up by matching and estimating a 2D affine transformation, and 2) Establish 2D-3D correspondences between I^{und} and the HiRISE DTM and solve a Perspective-n-Point (PnP) problem to get the pose. A drawback of option 1 is that in the absence of a depth image, the orthoprojection assumes a flat terrain and it is susceptible to noise in the VIO attitude estimate (pitch and roll). The max absolute error on these estimates has been shown as 3° for Ingenuity [8], with more recent work [46] showing further improvements.

Regardless, we adopt option 1 for the following reasons. First, the orthoprojection roughly aligns the navigation image to the map's rotation and size which simplifies parts of the MbL process and enables a more efficient matching process (e.g., choosing the proper size and number of the map crops, see Fig. 2). Second, a disadvantage of option 2 is that the HiRISE DTM has 4 times lower resolution in elevation measurement or *post* spacing (1 m/post [47]) than the texture map pixel resolution, which would have resulted in noisy 3D coordinates for our matches. This problem is exacerbated when considering MbL using orbital DTMs from CTX which have much lower resolution (20 m/post [48]). Therefore, option 1 lends itself more favorably towards future work using CTX. In addition, Ingenuity implements a terrain following algorithm enabled by frequent altitude measurements from the onboard LRF. Therefore, we are not concerned with estimating the vertical component (as there is no drift) and focus on the 2D position.

1) REFERENCE FRAME DEFINITION

We define three reference frames involved in our MbL pipeline, as illustrated in Fig. 4.

a) Map frame M

A Mars surface-fixed map frame is defined as a East-North-Up (ENU) coordinate system with origin on the map center. The horizontal xy plane aligns with the map projection plane, m, defined by the equirectangular projection used in the HiRISE map adopted for the Jezero crater site.

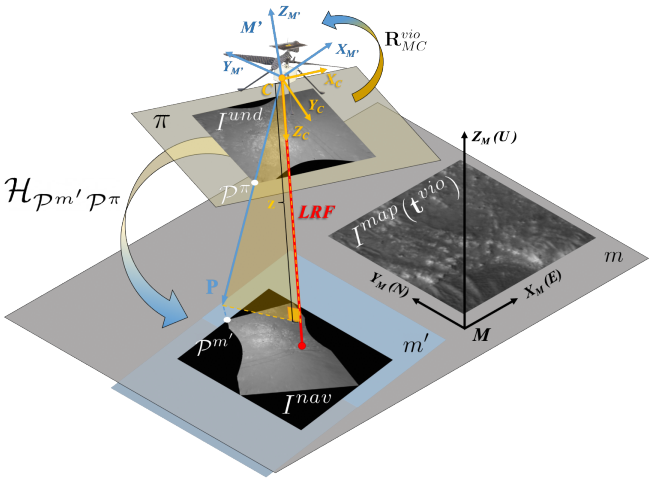

FIGURE 4. **Reference frames adopted in the MbL pipeline.**

b) Camera frame C

The Ingenuity camera frame is defined with its origin at the camera's optical center. The Z-axis points towards the scene along the optical axis, through the center of the image plane π. The X-axis points to the right along the image width, and the Y-axis completes the orthogonal set.

c) VIO-aligned frame M'

The pose prior estimate, $(\mathbf{R}_{MC}^{vio}, |\mathbf{t}_{MC}^{vio})$, from Ingenuity's onboard VIO provides the orientation and position of the camera with respect to the map, in terms of the rotation matrix $\mathbf{R}_{MC}^{vio} \in \mathbb{R}^{3\times3}$, and the translation vector, $\mathbf{t}_{MC}^{vio} \in \mathbb{R}^3$ in map coordinates. $\mathbf{R}_{MC}^{vio}$ is used to rotate the camera frame C onto a reference frame M' that is approximately aligned with M.

The angular misalignment between the map plane, m', estimated by the VIO (parallel to the $x_{m'}$-$y_{m'}$ axes) and the HiRISE map projection plane m, is affected by noise in the attitude estimate, however we note that VIO attitude estimate remains significantly more accurate than the position estimate $\mathbf{t}_{MC}^{vio}$. The latter is solely used to identify a local search area within the HiRISE map during the registration process. Specifically, its horizontal component, $\mathbf{t}^{vio} \in \mathbb{R}^2$, defines the center of the search region $I^{map}(\mathbf{t}^{vio})$.

2) NAVIGATION IMAGE PREPROCESSING

Given an observation from the Ingenuity navigation camera, the corresponding prior attitude estimate, $\mathbf{R}_{MC}^{vio}$, from the VIO, and the range measurement from the onboard LRF, the raw image I^{raw} is pre-processed in two steps. First, it is undistorted with a CAHVOR camera model [40] and second, it is warped onto the map projection plane *m'* estimated with the VIO attitude prior.

The raw navigation image, I^{raw}, is corrected for radial distortion using Ingenuity's CAHVOR camera model and an undistorted image, I^{und}, is produced along with the related intrinsics matrix $\mathbf{K}$. I^{und} is then warped to the plane m'. Under the assumption of nadir flights over flat terrain, we can approximate the z-coordinate, z, of the terrain points P in the camera frame C corresponding to the corners of I^{und}, using the LRF measurement (Fig. 4). The warping process first back-projects the four corner points in homogeneous pixel coordinates, $\mathcal{P}^{\pi}$ of I^{und}, from the image plane π to the camera frame C, and then transforms them in the frame M', using the rotation matrix $\mathbf{R}_{MC}^{vio}$:

$$P = z\mathbf{R}_{MC}^{vio}\mathbf{K}^{-1}\mathcal{P}^{\pi}. \tag{3}$$

Next we get the ortho-projected points $\mathcal{P}^{m'}$ on the plane m' by scaling the x and y coordinates of P using the map pixel size (0.25m). It is worth noting that any parallel plane to $x_{M'}$-$y_{M'}$ can be used due to the intrinsic independence of the orthographic projection from the the z-coordinate in M'. The four pairs of corner points on the undistorted image plane, $\mathcal{P}^{\pi}$, and on the VIO-aligned map plane $\mathcal{P}^{m'}$ are used to estimate the homography $\mathcal{H}_{\mathcal{P}^{m'}\mathcal{P}^{\pi}}$ between the two planes. Using the maximum and minimum values of the projected points $\mathcal{P}^{m'}$ we can determine the expected height h_{nav} and width w_{nav} of the projected navigation image. Finally, the estimated homography is applied to the undistorted image to produce the warped navigation image $I^{nav} = \mathcal{H}_{\mathcal{P}^{m'}\mathcal{P}^{\pi}}(I^{und})$ where $I^{nav} \in \mathbb{R}^{h_{nav}\times w_{nav}}$.

3) REGISTRATION AND POSITION ESTIMATION

The registration process of the navigation image involves applying the deep image matcher on a set of local map crops $\mathcal{C} = \{c_1, ..., c_N\}$ and producing a set of matches with the map $\mathcal{F} = \mathcal{G}(I^{nav}, \mathcal{C})$ that will be used to estimate Ingenuity's position.

The set $\mathcal{C}$ is dynamically determined for each navigation image by two factors. First, given the 2D noisy position estimate $\mathbf{t}^{vio}$ we can significantly narrow down the search region within the HiRISE map to a 100m $\times$ 100m local map area, $I^{map}(\mathbf{t}^{vio})$, centered at $\mathbf{t}^{vio}$. Even though the uncertainty of the VIO position estimate [8] and the short range of Ingenuity's flights would allow for a smaller search area, we decided on this conservative scenario that would account for drift in longer flights of a potential future mission.

The second factor is the size of the I^{nav} as estimated during the orthoprojection on the map. We divide the local map area into overlapping map crops where each $c_i \in \mathbb{R}^{h_{nav}\times w_{nav}}$ and with a minimum overlap of 64 pixels (empirically selected) with each neighboring crop. Each c_i is paired with I^{nav} and processed by the fine-tuned LoFTR to produce an initial set of matches. Using the confidence of the model for each match we keep only the top k=100, which is a good trade-off between keeping good matches and reducing the computational burden. After matching with all map crops, we pool together all matches and further refine

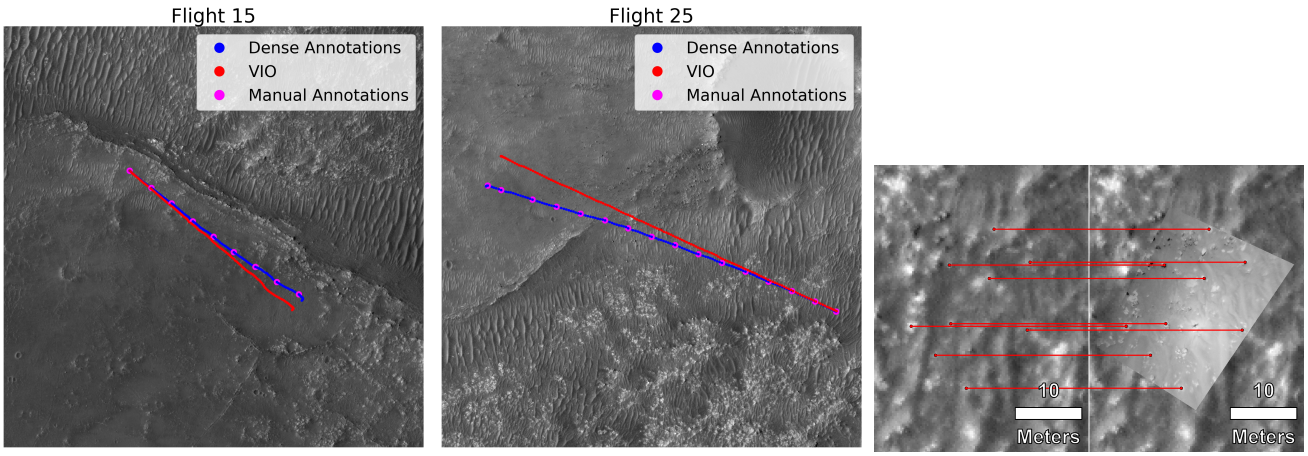

FIGURE 5. **Examples of sparse manual annotations (magenta), dense annotations (blue), and VIO positions (red) projected on map crops from HiRISE (left and middle). The dense annotations are interpolated between the manual ones. The crops are 800m $\times$ 800m. The drift from VIO is clearly noticeable with the landing site of each flight located at the direction of largest divergence with respect to the annotations. Examples of the manually placed tie-points during the annotation process (right).**

the set with a strict confidence threshold of 0.95 to obtain our final set of matches $\mathcal{F}$. We require this last step in order to eliminate bad matches from any c_i that do not actually have overlap with I^{nav}. The final step is to estimate an affine transformation $\hat{R}, \hat{t}$, using RANSAC to eliminate any outliers in $\mathcal{F}$, that aligns I^{nav} with the 100m $\times$ 100m local map area. In practice, the final 2D position estimate $\mathbf{t}^{nav}$ is obtained by applying the affine transformation on the pixel coordinate that corresponds to Ingenuity's projected location within I^{nav}.

V. EXPERIMENTAL EVALUATION

We evaluate our Map-based Localization pipeline on several Ingenuity flights. Besides using the VIO estimate as a prior to select the search region within HiRISE, the localization results reported are not part of a state estimator but rather independently estimated for each image using our MbL approach. For our experiments we use navigation images[2] from flights 9, 10, 11, 15, 20, 21, and 25. Only images taken with at least 8m altitude are considered. We chose these early Ingenuity flights due to their visual difficulty with low-texture terrain frequently containing repetitive patterns of dune fields. Details about these flights are reported in Table 1, while some of the flight trajectories over the HiRISE map can be seen in Fig. 5. More information on all Ingenuity flights can be found in the publicly available flight log[3].

For both LoFTR and the reported baselines, during position estimation we remove outliers by running RANSAC for 5000 iterations and setting the inlier reprojection error threshold to 1 pixel. Our main evaluation metric is the localization accuracy @5m, which is the percentage of queries with position error at or below 5m. The 5m threshold was chosen because of the global localization performance requirements for a Mars rotorcraft [3]. We also plot the Cumulative Distribution Function (CDF) of the localization accuracy up to 10m.

We designed experiments to showcase the viability of using a deep learning-based image registration model for the purpose of global localization of Ingenuity, and by extension for any future Mars rotorcraft with similar mission conditions. First, in subsection A, we compare directly to other baselines on localization accuracy to demonstrate the improvement achieved by using LoFTR and the potential for correcting the VIO drift. We also demonstrate the generalization ability of the model by comparing two variations of our model that differ in the number of flights used for fine-tuning. Second, in subsection B, we investigate the impact our intermediate training step using the rendered dataset of the Martian environment. In subsection C, we investigate the amount of fine-tuning training data required for domain adaptation. Subsection D provides insight into how the model learns to incorporate context by visualizing the learned attention weights. In subsection E, we conduct an analysis to determine a confidence metric over the quality of the localizations. Finally, in subsection F we investigate the ability of the model to predict match scores that correlate with accurate matches, while in subsection G we discuss the computational requirements of LoFTR.

During the experiments, our approach of training, first with the rendered dataset, and then fine-tuning on Ingenuity flights is referred to as *LoFTR-Fine*, with variations of the method defined in their respective experiment sections. Table 2 shows the datasets used to train each model. Any model with the suffix "-Single" was fine-tuned only on flight 9, and the suffix "-All" refers to the model trained on a small set of images from all flights. All fine-tuned models presented in our evaluations were trained only for 5 epochs on their respective training sets with short training times between 10-20 minutes on an NVIDIA RTX 3080 laptop

[2]https://mars.nasa.gov/mars2020/multimedia/raw-images/

[3]https://science.nasa.gov/mission/mars-2020-perseverance/ingenuity-mars-helicopter/

TABLE 1. Details regarding the Ingenuity flights used in our experiments. *LTST* refers to Local True Solar Time on Mars where 12:00 (noon) occurs when the Sun is highest in the sky. The *Nr Images* corresponds to the number of navigation images with at least 8-m altitude, while *Nr Training Images* refers to the number of images selected for creating pseudo ground-truth correspondences during fine-tuning of the model.

Flight	Sol	Horizontal Distance (m)	Start LTST	Max Altitude (m)	Nr Images	Nr Training Images
9	133	625	12:36	10	169	30
10	153	233	12:10	12	167	34
11	163	383	12:42	12	165	31
15	254	407	12:25	12	162	31
20	362	391	10:40	10	163	21
21	375	370	10:40	10	132	15
25	403	704	10:38	10	169	15

TABLE 2. Datasets used to train the models in our experiments. Rendered refers to the large-scale dataset we generated using a HiRISE map in Blender. The order of training is always MegaDepth → Rendered → Ingenuity.

	MegaDepth [22]	Rendered	Ingenuity
LoFTR-Scratch-Fine			✓
LoFTR-Pre	✓		
LoFTR-Pre-Fine	✓		✓
LoFTR-Fine	✓	✓	✓

GPU. We observed empirically that longer training times led to overfitting and performance degradation. For fine-tuning LoFTR, we follow the training procedure (loss definitions, loss weights etc.) as described in [19].

Finally, we plan to release the undistorted Ingenuity navigation images we used in our experiments along with our trained models upon publication.

1) INGENUITY FLIGHT POSITION ANNOTATION PROCESS

Currently there are no publicly available, reliable, ground-truth annotations for Ingenuity flights. While we produced pseudo correspondences to fine-tune LoFTR in Section IV.B, those are probably not reliable enough to produce positional ground-truth suitable for evaluating the MbL pipeline. Furthermore, manual annotations to a reference map require a careful and time-consuming process. In order to evaluate our approach, we produced a sparse set of manual annotations for the flights listed in Table 1, and then propagated the manual annotation results to the rest of the non-annotated images.

Images from Ingenuity's navigation camera were first geometrically corrected using the CAHVOR camera model to produce cropped, undistorted versions without the fisheye lens effect. A sparse subset of images taken at approximately 50m intervals was selected and imported into Geographic Information System (GIS) software, along with a geographically referenced HiRISE image basemap. Tie-points were manually assigned to visible surface features such as rocks, albedo variations, and intersecting ripple crests that are uniquely distinguishable in both the HiRISE and Ingenuity images. A perspective projection transform was then applied using each image's tie-points to warp it to the map coordinate system with an accuracy of ≤ 1 basemap pixel (25 cm).

Upon completing tie-points and transformations for each flight, the helicopter's shadow position was marked in each of the manually referenced images. Because the sun was not directly overhead, an offset was applied uniformly to all images per flight to denote the nadir position directly underneath the helicopter. The projection of the helicopter's shadow onto underlying terrain contains some additional uncertainty due to unresolved topographic variability, but is estimated to be less than 1 m. The coordinates of these nadir positions were recorded, along with the corresponding elevation values interpolated from the 1 m/pixel HiRISE DTM. Vehicle heading for each position was measured clockwise relative to north based on the orientations of shadows from Ingenuity's footpads and/or rectangular solar array with an accuracy of $\leq 2°$.

To create approximate dense image position annotations between sparse manual ones, we applied an affine transformation of the VIO trajectory coordinates in each segment, aligning them with the corrected manual annotations. Examples of the manual and the final dense annotations used to evaluate our approach are shown in Fig. 5.

A. COMPARISON TO BASELINES AND MODEL GENERALIZATION

In this first experiment, we compare both quantitatively and qualitatively to the most relevant image registration baselines for our MbL task that includes a template-matching approach, a classical hand-crafted feature, a learning-based approach, and the original LoFTR model:

- *Census* [27]: Census was successfully used in the recently introduced rover global localization system [26]. For our experiments, we adopt the global-to-local Census transform pipeline proposed in [49] that first matches the image to the entire map in order to get a coarse localization estimate, followed by matching local patches to refine the initial estimate.
- *SIFT and DenseSIFT* [10]: SIFT is a popular descriptor that has demonstrated robust performance in initial MbL studies for a future Mars rotorcraft [12]. We investigate its performance following the keypoint detector

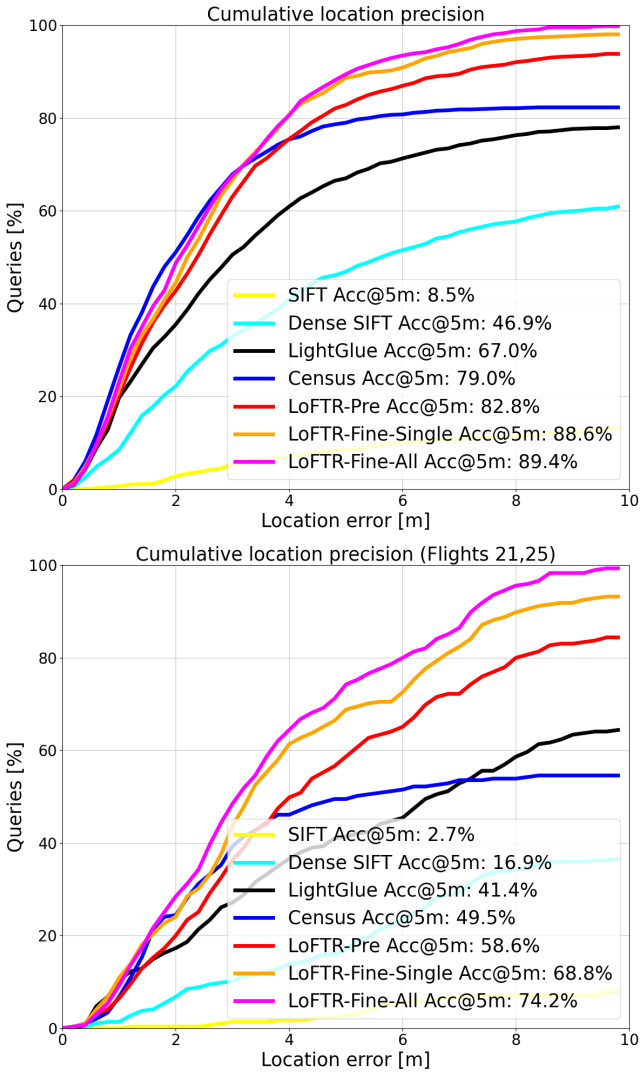

FIGURE 6. Localization accuracy results between baselines and our *LoFTR-Fine* variants on all flights (top) and on the most challenging flights of 21 and 25 (bottom). *LoFTR-Fine-Single* and *LoFTR-Fine-All* were fine-tuned only on navigation images from flight 9, and all flights respectively, while *LoFTR-Pre* is the off-the-shelf LoFTR [19] model.

by Lowe's algorithm [10] and via densely sampling SIFT descriptors over the image at 8-pixel intervals.

- *LoFTR-Pre* [19]: We compare to the original LoFTR model that was trained on the MegaDepth [22] dataset and has not been exposed to any data from the Martian environment.
- *LightGlue* [39]: A deep image matcher approach that uses Superpoint [33] to extract keypoints and also utilizes Transformers to aggregate image context. We evaluate LightGlue as a lightweight alternative to LoFTR. It employs an adaptive stopping mechanism to reduce inference time and it is also trained on the MegaDepth dataset.

TABLE 3. Our MbL pipeline using LoFTR-based registration has the potential for significant VIO drift correction. The *LoFTR-Fine-Single* error represents the mean of the frames with $\leq$ 5-m error over each flight to avoid including outliers from our independently estimated localization.

Flight	VIO Drift (m / %)	*LoFTR-Fine-Single* Error (m / %)
9	44.4 / 7.1	1.9 / 0.3
10	4.6 / 2.0	2.8 / 1.2
11	24.2 / 6.3	1.2 / 0.3
15	25.8 / 6.3	2.6 / 0.6
20	11.3 / 2.9	1.3 / 0.3
21	45.5 / 12.3	2.6 / 0.7
25	63.9 / 9.1	2.4 / 0.3

We use two variants of our model that were trained on the rendered dataset and then fine-tuned using the pseudo ground-truth correspondences. *LoFTR-Fine-All* is fine-tuned on a subset of images from all flights (177 navigation images that form 550 training pairs with map crops), while *LoFTR-Fine-Single* is fine-tuned only on images from flight 9 (30 navigation images, 61 training pairs). Our motivation for *LoFTR-Fine-Single* is to demonstrate the ability of the model to generalize to flights not used during training. This represents a more realistic scenario during a mission, where data from previous flights can be used to quickly adapt a model to be used in future flights on unseen terrain. We note that for the purpose of our evaluation, we tested all baselines on all navigation images regardless whether they were used during the self-supervision approach in Section IV.B. This only affects *LoFTR-Fine-All* and we found experimentally that the effect on the evaluation is minimal (0.3% higher when including all navigation images). This can be justified by the fact that we used only pseudo correspondences and only 15% of the 1127 navigation images from all flights were part of the self-supervised training with pseudo annotation.

The CDF of the localization accuracy over all flights is shown in Fig. 6 (top). *SIFT* fails to identify repeatable keypoints due to the very large scale difference between navigation image and map and barely carries out any successful registrations. *DenseSIFT*, which does not rely on a handcrafted keypoint detector, greatly outperforms SIFT but still falls short of the learning-based methods due to its dependency on local appearance similarity to establish matches.

While *Census* outperforms *LightGlue* and is on par with the *LoFTR-Fine* variants for accuracies roughly below 3m, its performance reaches a plateau relatively quickly with an Acc@5m of 79.0% and Acc@10m of 82.2%. On the other hand, *LoFTR-Fine-Single* and *LoFTR-Fine-All* have a much higher ceiling, achieving a 9.6% and 10.4% Acc@5m improvement over *Census*, and show an almost perfect Acc@10m of 98.2% and 99.8% respectively.

To further drive this point, we also show the CDF of the localization accuracy on the most challenging flights of 21 and 25 (based on the terrain and *Census* performance)

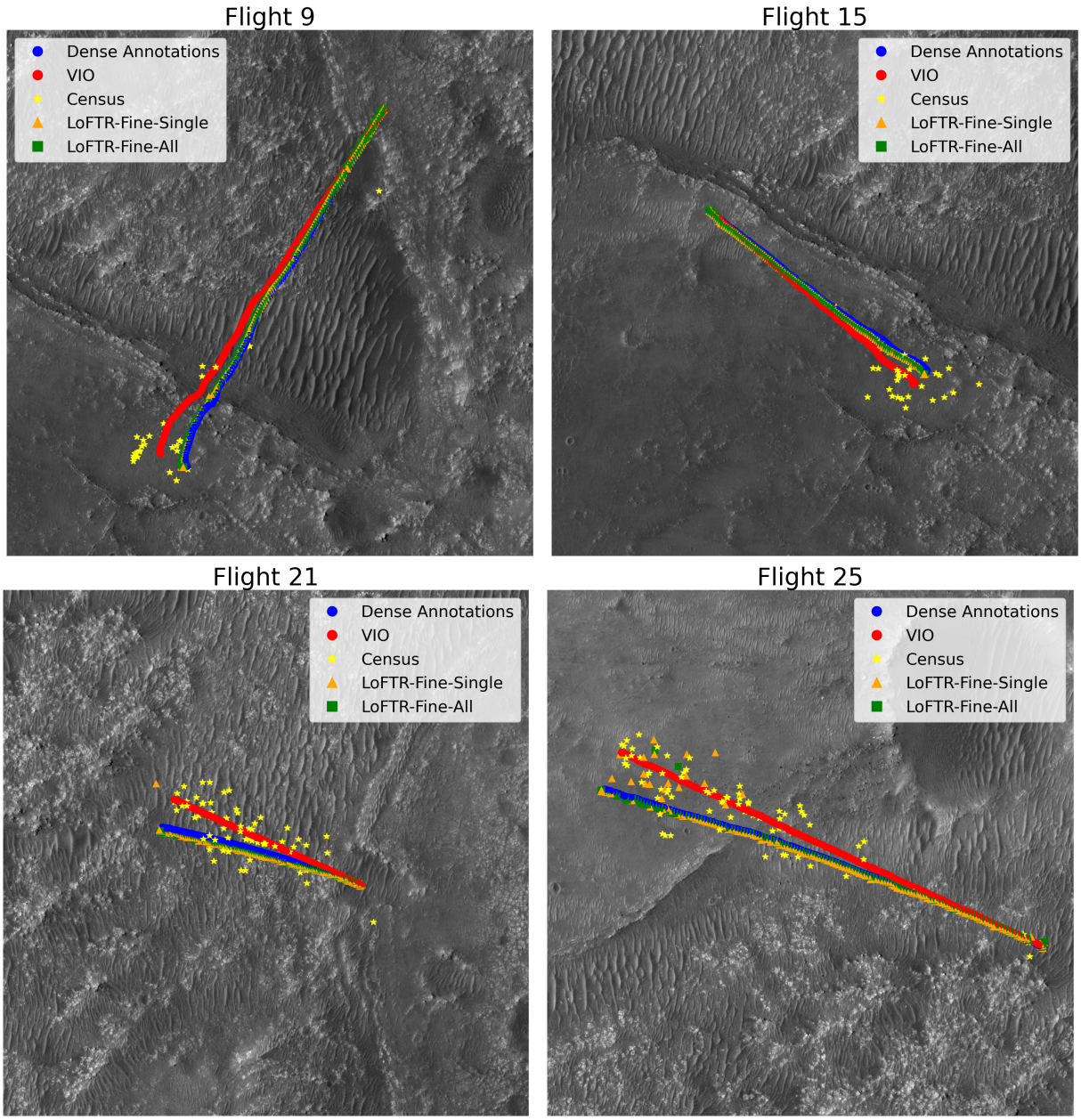

FIGURE 7. Reconstructed Ingenuity flights using localization estimates from *Census*, *LoFTR-Fine-Single*, and *LoFTR-Fine-All*. The latter is much more consistent and rarely estimates a position far outside of the path, *Census* clearly fails on challenging parts of the terrain, while *LoFTR-Fine-Single* only struggles on a portion of flight 25 where Ingenuity was flying above texture-less terrain. Note that *LoFTR-Fine-Single* was fine-tuned using only a small subset of navigation images from flight 9. The map crop dimensions are 800m $\times$ 800m.

in Fig. 6 (bottom). *Census* performance drops significantly on these flights with a 19.3% and 24.7% Acc@5m gap to *LoFTR-Fine-Single* and *LoFTR-Fine-All* respectively, while the two *LoFTR-Fine* variants maintain high Acc@10m of 93.2% and 99.3%. This suggests that *Census* can perform accurate registration when conditions are ideal (e.g., when terrain has discriminative textures), but it is less robust in more challenging cases frequently observed during flights.

The position estimations across entire flights for *Census*, *LoFTR-Fine-Single*, and *LoFTR-Fine-All* are plotted in Fig. 7. We can visually observe that *Census* frequently has wrong localizations outside the expected path of Ingenuity, usually in the presence of challenging terrain. In contrast, our two *LoFTR-Fine* variants are more consistently accurate in reconstructing Ingenuity's path. Finally, qualitative image matching examples are shown in Fig. 8.

1) CORRECTING THE VIO DRIFT

Since we evaluate our MbL pipeline independently of a state estimator, our estimations lack a temporal component and are thus not directly comparable to the VIO. Regardless, we provide some intuition of the potential VIO drift correction

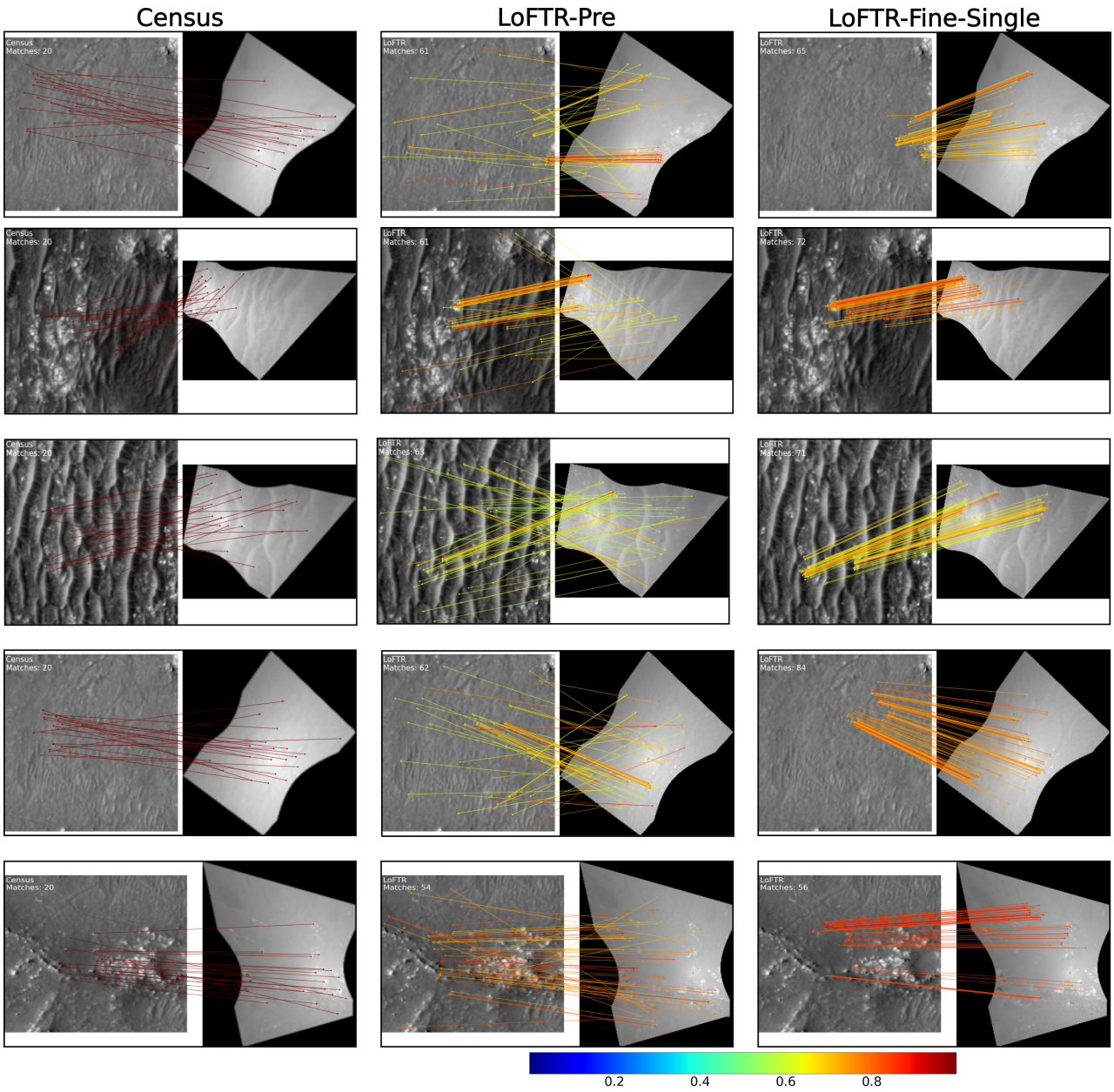

FIGURE 8. **Matching examples between the navigation image (right of every pair) and the map crop (left of every pair) during Ingenuity flights. From left to right, we show matches from *Census*, *LoFTR-Pre*, and *LoFTR-Fine-Single*. The color scheme shows the LoFTR model confidence for each match between 0 and 1 (see color bar), and the depicted matches are before geometric verification (e.g., RANSAC) takes place. Census does not return normalized confidence values so all matches are shown in the same color. *LoFTR-Fine-Single* is more consistent on finding accurate matches over a variety of terrains even though it was trained only over a single flight.**

that the LoFTR-based registration can achieve. Specifically, Table 3 shows the VIO drift error at the end of each flight, along with the mean position error across the flight from our *LoFTR-Fine-Single* model. We note that the mean error is estimated on frames with $\leq$5m position error as we make the assumption that outliers will be discarded by the state estimator. It is evident that our MbL system has the potential to significantly decrease the position error of Ingenuity caused by VIO drift, given frequent updates to the state estimator.

B. IMPACT OF THE PRETRAINED MODEL

In this section, we investigate the importance of our intermediate training step using the large rendered dataset. We compare three options fine-tuned on a single flight. *LoFTR-Fine-Single* follows our intermediate training strategy, *LoFTR-Pre-Fine-Single* uses the off-the-shelf LoFTR model trained on MegaDepth [22] with no intermediate training, and *LoFTR-Scratch-Fine-Single* initializes LoFTR with the backbone of the model using the original ResNet-18 [42] weights trained on ImageNet [50] and the rest of

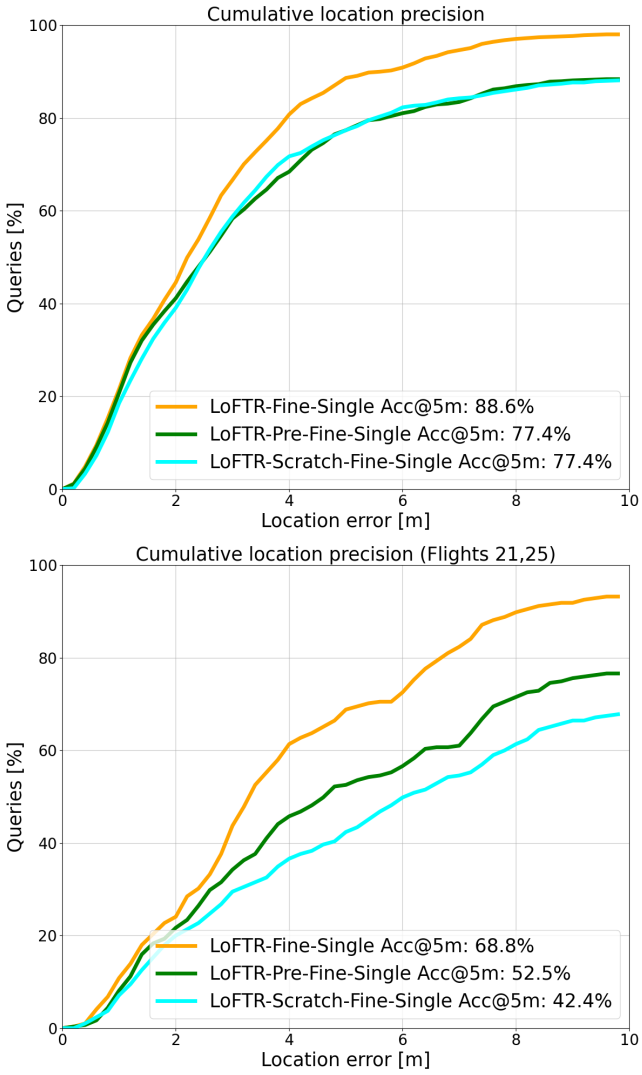

FIGURE 9. **Investigation over the impact of the pretrained model. We compare our *LoFTR-Fine-Single* that was pretrained using our rendered dataset to *LoFTR-Pre-Fine-Single* that used the off-the-shelf model, and to *LoFTR-Scratch-Fine-Single* that was trained from scratch. The top figure shows results on all flights while the bottom focuses on the most challenging flights 21 and 25. Our intermediate training on the rendered data is shown to have a clear advantage over fine-tuning the model directly on Ingenuity images, especially for flights with more challenging terrain.**

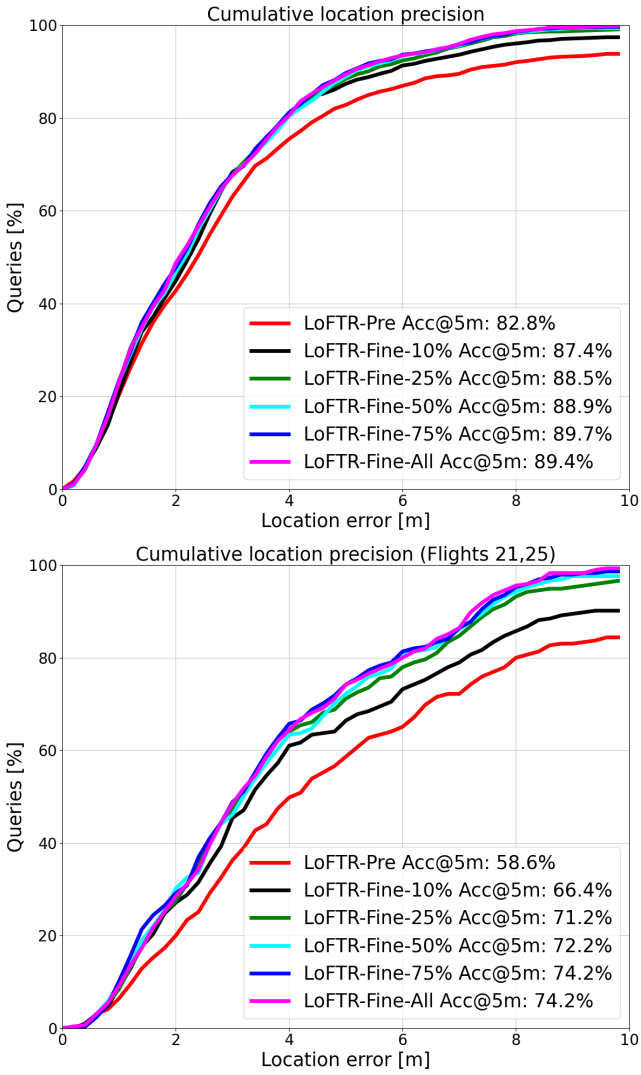

FIGURE 10. **Investigation over the effect of using different amounts of Ingenuity training data during fine-tuning. The top figure shows results on all flights, while the bottom focuses on the most challenging flights 21 and 25. The availability of small amounts of data (e.g., 10%) can still have impact over *LoFTR-Pre* with data availability having larger impact on the more challenging flights.**

the model layers initialized with random weights. Neither MegaDepth or the intermediate training is used for *LoFTR-Scratch-Fine-Single*.

The results are illustrated in Fig. 9. We notice a large performance gap of 11.2% on Acc@5m between *LoFTR-Fine-Single* and both *LoFTR-Pre-Fine-Single* and *LoFTR-Scratch-Fine-Single* baselines on all flights, and 16.3% and 26.4% gaps respectively for the more challenging flights 21 and 25. These gaps highlight the need for pretraining on a large relevant training set, especially when the target domain (in this case, the Ingenuity flights) has very limited data available for fine-tuning. Interestingly, the performance of *LoFTR-Pre-Fine-Single* drops compared to *LoFTR-Pre* (see Fig. 6) even though the former was trained with a few Ingenuity images. This indicates that in the absence of the intermediate training, it is difficult to learn an appropriate model directly on a handful of images with pseudo annotations.

Furthermore, *LoFTR-Pre-Fine-Single* and *LoFTR-Scratch-Fine-Single* show identical performance (on all flights) in terms of localization accuracy even though *LoFTR-Pre-Fine-Single* was pretrained on a large dataset. However, in practice the model trained from scratch produces more high confident outliers during matching that are filtered out during RANSAC. Indeed, the results on flights 21 and 25 illustrate

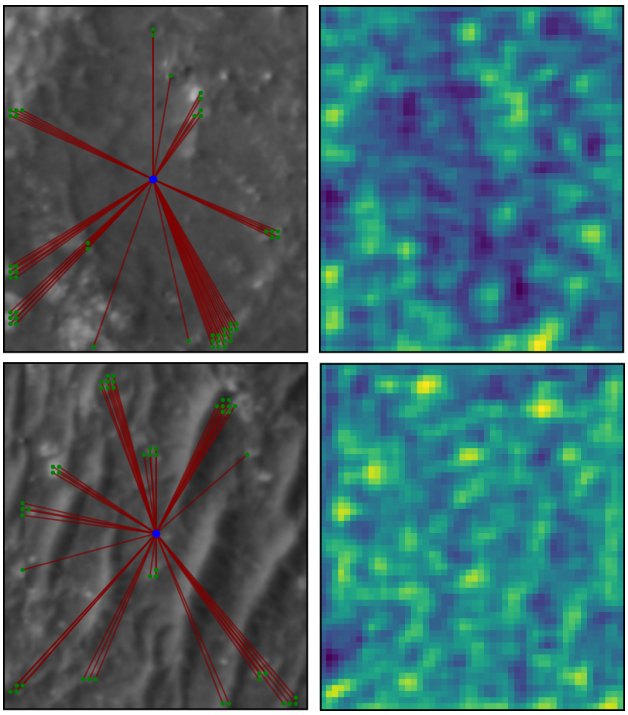

FIGURE 11. Each row shows a separate example of how Transformer blocks in *LoFTR-Fine-All* incorporate context from the image when estimating the feature representation for the center blue point. The heatmap of the attention weights is shown to the right, while the top 50 locations with the highest attention weights are shown to the left (green points). The model learns to focus on discriminative locations in the image and suppresses information from feature-less terrain (top) and repeated dune patterns (bottom).

that training from scratch is not ideal for challenging scenarios.

C. EVALUATION OVER THE AMOUNT OF TRAINING DATA

We are interested in quantifying the impact of data scarcity on model performance. How does the model performance degrade when fewer training samples with pseudo ground-truth correspondences are available for fine-tuning? In one plausible scenario, the model needs to be adapted very swiftly from a small number of navigation images. To investigate the adaptability of the model to a small fine-tuning set, we fine-tune the *LoFTR-Fine* model by randomly sampling subsets of 10%, 25%, 50%, and 75% of the 445 training pairs collected from all flights.

The results in Fig. 10 show that even when using only 10% of the data (44 training examples), there is a noticeable increase in performance from *LoFTR-Pre* of 4.6% for all flights and 7.8% for flights 21 and 25. This indicates that the intermediate training with the rendered dataset provides excellent initialization such that the model can adapt quickly with minimal amounts of training samples. It is also worth noting that the increase in data availability has larger impact on the localization accuracy of the more challenging flights since the model needs more examples to adapt to feature-less or repetitive terrains.

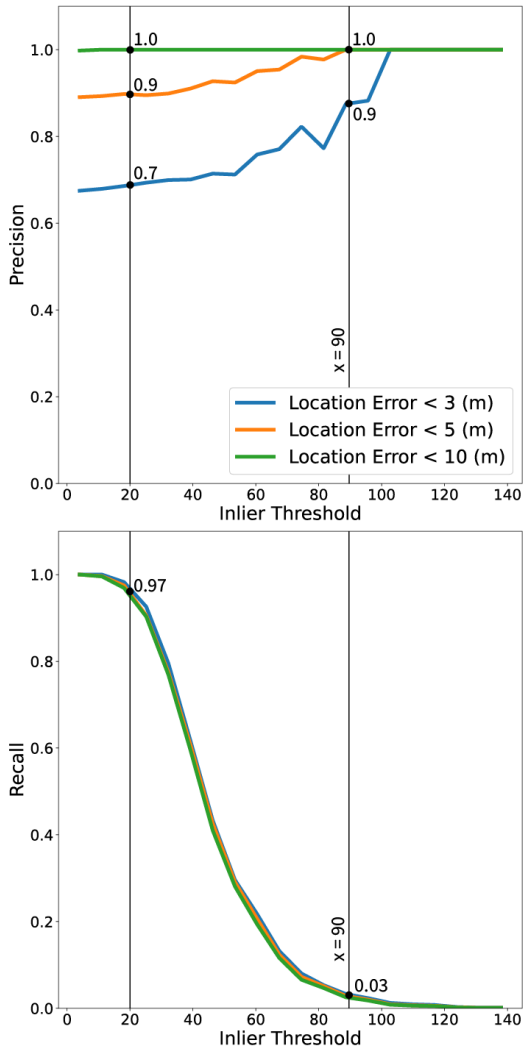

FIGURE 12. Inlier threshold is one way of selecting flights with high confidence in their localization estimate. The precision plot (top) shows the percentage of flights, where a localization was selected, using the inlier threshold, and it was within the error bound. The recall plot (bottom), shows the percentage of flights, where a good localization exists, and it was selected. Both plots show the results on a validation dataset. Vertical bars at $x = 20$ and $x = 90$ highlight specific values representing a low-precision/high-recall and a high-precision/low-recall regimes.

D. VISUALIZATION OF MODEL FEATURES

Due to the large complexity of deep learning-based models, it is difficult to interpret model output. Visualizing features from intermediate layers of the model [51] can provide useful information as to the learned representation. We visualize such features from our *LoFTR-Fine-All* model in Fig. 11 to offer insight into how the transformer blocks in LoFTR learn to incorporate information from the cropped map images when estimating the feature representation for a particular point. It is evident that the model focuses on more salient areas of the map while assigning less weight to texture-less regions.

E. CONFIDENCE ANALYSIS

Global localization estimates can only be used in flight if their accuracy falls within the mission's error tolerance; otherwise, they risk compromising the mission. However, the way MbL is integrated into flight operations influences the required balance between precision and recall. In this section, we explore how a secondary confidence metric, such as the number of inliers obtained when estimating the affine transformation in the final step of the MbL process, can help optimize this trade-off based on mission requirements.

One application of MbL is as a drop-in replacement for human-in-the-loop post-flight localization. In Ingenuity operations, global localization was performed post-flight due to the need for image downlink and manual annotation. Automating this process requires 100% precision, as errors would propagate to subsequent flights. However, as shown in Fig. 6 (top), our model achieves only 89.4% precision at a 5-meter error tolerance, falling short of this strict requirement.

To address this, a secondary confidence metric, which correlates with accuracy, can be employed to further refine the localizations until they satisfy the precision requirement. High-confidence estimates can be trusted by the flight system, while lower-confidence ones can be flagged for human review without jeopardizing the mission. For example, Fig. 12 illustrates how setting an inlier threshold impacts precision and recall. An inlier threshold of 90 ensures 100% precision at a 5-meter error tolerance, but only 3% of localizations meet this criterion, requiring human intervention 98% of the time. If a more lenient 10-meter tolerance is acceptable, no thresholding is needed, allowing 100% of autonomous localizations to be used.

MbL can also provide mid-flight global localization updates to correct VIO drift in real time. In this mode, the system prioritizes recall over precision, as frequent updates refine the position estimate through a Kalman filter. A higher number of measurements, even with some noise, ensures robustness. By adjusting the inlier threshold, we can control the trade-off between precision and recall. As illustrated in Fig. 12, setting the threshold to 20 results in 90% precision and 97% recall at a 5-meter error tolerance, making it suitable for mid-flight updates where noisy measurements are smoothed by the Kalman filter.

The number of inliers from the affine transformation estimation serves as a strong confidence metric because it directly reflects the geometric consistency between the matched navigation image and map features. A higher inlier count indicates that more correspondences align well under the estimated transformation, suggesting a more reliable localization estimate. Conversely, a low inlier count may signal mismatches, poor feature alignment, or insufficient visual overlap, making the estimate less trustworthy. However the number of inliers also depends on the number of features detected in the image, which could be low even if the geometric consistency is high.

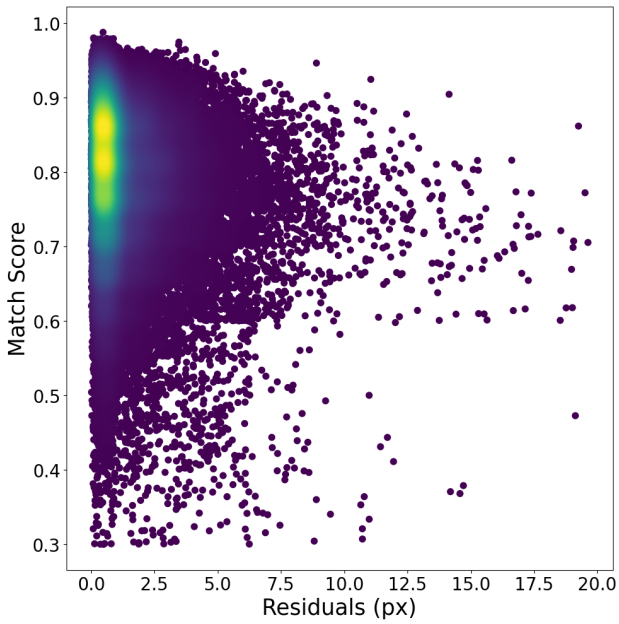

FIGURE 13. Match scores as a function of residuals produced by the *LoFTR-Fine-All* model for each match. The residuals correspond to the reprojection error in pixels after the RANSAC-based alignment process. High score values correlate well with low residuals, suggesting that the model is able to predict good matches with high confidence.

F. MODEL MATCH SCORE ANALYSIS

The LoFTR model learns to output a score for every predicted match that reflects the probability of soft mutual nearest neighbor matching. In this experiment, we investigate the correlation of this score with the actual quality of the matches between the Ingenuity navigation image and the map. The quality of each match is assessed based on its reprojection error in pixels (residual) after the RANSAC-based alignment process during MbL. In Fig. 13 we plot the match scores as a function of the residuals, where yellow values correspond to higher density of points. Each point corresponds to an individual match collected from all flights using our model *LoFTR-Fine-All*.

We observe that the highest density is situated at the intersection of high confidence values and low residuals (top left of the plot). Specifically, we find that out of all matches with ≥ 0.8 score, 66.8% of them have a residual $\leq$ 1px and 98.2% have a residual $\leq$ 5px. The corresponding numbers for matches with ≥ 0.9 score are 74.2% with residual $\leq$ 1px and 99.3% with $\leq$ 5px. This suggests that the model scores are calibrated and can predict accurate matches with high confidence. A few outliers where high scores correspond to high residuals and low scores correspond to low residuals can be attributed to texture-less terrains or terrains with repeated patterns. In both cases, the model was not able to compose a discriminative representation that can lead to either a wrong match, or a correct match with high ambiguity.

G. COMPUTATIONAL REQUIREMENTS

We report the LoFTR runtime as 122ms $\pm$ 3ms estimated over 100 trials on an NVIDIA RTX 3080 laptop GPU. This runtime accounts for a forward pass of the model that takes two images at 640×480 resolution as input and produces the set of matches and their scores. We choose to report the forward pass runtime because the overall MbL runtime depends on other external factors such as the size of the map search area and the size of the map crops. Additionally, we report the LoFTR runtime on the NVIDIA embedded hardware Jetson AGX Xavier as 606ms $\pm$ 11ms, and note that newer edge devices (e.g., Jetson AGX Orin) offer higher performance. The reported numbers are computed using the Python implementation without using any optimization frameworks (e.g., TensorRT) that can reduce inference runtime by multiple factors. We note that MbL does not need to run in real-time and its frequency depends on how often the VIO drift needs to be corrected as dictated by mission requirements.

VI. CONCLUSION

In this paper, we present a Map-based Localization system that overcomes the large resolution difference between the Ingenuity navigation images and the HiRISE map by incorporating a deep image matcher. Our main insight is that Transformer-based models are robust to this challenge by learning to integrate global image context. To resolve the issue of limited data available for training such a model, we bootstrap the learning process with a large rendered Martian dataset, followed by fine-tuning on a small set of Ingenuity images with pseudo annotations.

Our experimental results show that our MbL approach significantly outperforms Census transform-based template matching in localization accuracy, while the classical hand-crafted feature SIFT fails on this task. In addition, our model generalized well to unseen flights and rapidly adapts to this domain even when a minimal number of training images were available. Our deep image matcher model not only outperformed classical methods convincingly, but it did so with minimal dependency on in-domain training data which has always been a prohibiting factor for adopting deep learning models in space applications.

The success of our approach in a realistic mission scenario highlights the potential of deep learning for planetary navigation and localization. We believe this work will encourage broader adoption of deep learning in space applications and inspire further research into its viability for future missions.

FUTURE WORK

A useful direction of our work would be to investigate the performance of our approach using orbital map products created by CTX [7] instead of HiRISE. CTX maps pose a harder registration problem due to their much lower resolution (6m / pixel), but they provide approximately 99% coverage of Mars. In addition, future missions might conduct flights during different times of day. With the vast majority of orbital data being collected during a limited time range (afternoon), the registration algorithm has to be robust to different lighting conditions. Furthermore, we are currently integrating our MbL pipeline with VIO as part of a state estimator, in order to demonstrate the potential of our method for online drift correction. Finally, we plan to work on reducing the inference time on edge devices using techniques such as quantization and distillation.

REFERENCES

[1] J. Bapst, T. J. Parker, J. Balaram, T. Tzanetos, L. H. Matthies, C. D. Edwards, A. Freeman, S. Withrow-Maser, W. Johnson, E. Amador-French, J. L. Bishop, I. J. Daubar, C. M. Dundas, A. A. Fraeman, C. W. Hamilton, C. Hardgrove, B. Horgan, C. W. Leung, Y. Lin, A. Mittelholz, and B. P. Weiss, "Mars Science Helicopter: Compelling Science Enabled by an Aerial Platform," *Bulletin of the AAS*, vol. 53, mar 18 2021. https://baas.aas.org/pub/2021n4i361.

[2] I. R. Witte, D. L. Bekker, M. H. Chen, T. B. Criss, S. N. Jenkins, N. L. Mehta, C. A. Sawyer, J. A. Stipes, and J. R. Thomas, "No GPS? No problem! Exploring the dunes of titan with dragonfly using visual odometry," in *AIAA Scitech 2019 Forum*, p. 1177, 2019.

[3] J. L. Anderson, T. L. Brown, M. Cacan, G. Kubiak, A. Jasour, and N. Z. Rothenberger, "Lessons from ingenuity's climb up Jezero crater delta," in *2024 IEEE Aerospace Conference*, pp. 1–15, IEEE, 2024.

[4] NASA/JPL-Caltech/ASU/MSSS, "Mars report: The most extreme flights of nasa's ingenuity mars helicopter." Released on 2024-02-01. Accessed: 2025-07-21.

[5] H. F. Grip, D. Conway, J. Lam, N. Williams, M. P. Golombek, R. Brockers, M. Mischna, and M. R. Cacan, "Flying a helicopter on mars: How ingenuity's flights were planned, executed, and analyzed," in *2022 IEEE Aerospace Conference (AERO)*, pp. 1–17, 2022.

[6] A. S. McEwen, E. M. Eliason, J. W. Bergstrom, N. T. Bridges, C. J. Hansen, W. A. Delamere, J. A. Grant, V. C. Gulick, K. E. Herkenhoff, L. Keszthelyi, R. L. Kirk, M. T. Mellon, S. W. Squyres, N. Thomas, and C. M. Weitz, "Mars reconnaissance orbiter's high resolution imaging science experiment (HiRISE)," *Journal of Geophysical Research: Planets*, vol. 112, no. E5, 2007.

[7] M. C. Malin, J. F. Bell III, B. A. Cantor, M. A. Caplinger, W. M. Calvin, R. T. Clancy, K. S. Edgett, L. Edwards, R. M. Haberle, P. B. James, S. W. Lee, M. A. Ravine, P. C. Thomas, and M. J. Wolff, "Context camera investigation on board the mars reconnaissance orbiter," *Journal of Geophysical Research: Planets*, vol. 112, no. E5, 2007.

[8] D. S. Bayard, D. T. Conway, R. Brockers, J. H. Delaune, L. H. Matthies, H. F. Grip, G. B. Merewether, T. L. Brown, and A. M. San Martin, "Vision-based navigation for the nasa mars helicopter," in *AIAA Scitech 2019 Forum*, p. 1411, 2019.

[9] T.-H. Pham, W. Seto, S. Daftry, B. Ridge, J. Hansen, T. Thrush, M. Van der Merwe, G. Maggiolino, A. Brinkman, J. Mayo, *et al.*, "Rover relocalization for mars sample return by virtual template synthesis and matching," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 4009–4016, 2021.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[11] A. E. Johnson, Y. Cheng, N. Trawny, J. F. Montgomery, S. Schroeder, J. Chang, D. Clouse, S. Aaron, and S. Mohan, "Implementation of a map relative localization system for planetary landing," *Journal of Guidance, Control, and Dynamics*, vol. 46, no. 4, pp. 618–637, 2023.

[12] R. Brockers, P. Proença, J. Delaune, J. Todd, L. Matthies, T. Tzanetos, and J. B. Balaram, "On-board absolute localization based on orbital imagery for a future mars science helicopter," in *2022 IEEE Aerospace Conference (AERO)*, pp. 1–11, 2022.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[15] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7262–7272, 2021.

[16] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun, "Depth pro: Sharp monocular metric depth in less than a second," *arXiv preprint arXiv:2410.02073*, 2024.

[17] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[18] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research Journal*, pp. 1–31, 2024.

[19] 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*, pp. 8922–8931, 2021.

[20] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg, "Roma: Robust dense feature matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19790–19800, 2024.

[21] J. Edstedt, I. Athanasiadis, M. Wadenbäck, and M. Felsberg, "Dkm: Dense kernelized feature matching for geometry estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17765–17775, 2023.

[22] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2041–2050, 2018.

[23] X. Wan, J. Liu, H. Yan, and G. L. Morgan, "Illumination-invariant image matching for autonomous uav localisation based on optical sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 119, pp. 198–213, 2016.

[24] A. Ansar and L. Matthies, "Multi-modal image registration for localization in titan's atmosphere," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3349–3354, IEEE, 2009.

[25] J. Mustard, M. Adler, A. Allwood, D. Bass, D. Beaty, J. Bell, W. Brinckerhoff, M. Carr, D. Des Marais, B. Brake, *et al.*, "Report of the mars 2020 science definition team," *Mars Explor. Progr. Anal. Gr*, vol. 150, pp. 155–205, 2013.

[26] J. Nash, Q. Dwight, L. Saldyt, H. Wang, S. Myint, A. Ansar, and V. Verma, "Censible: A robust and practical global localization framework for planetary surface missions," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8642–8648, 2024.

[27] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Computer Vision—ECCV'94: Third European Conference on Computer Vision Stockholm, Sweden, May 2–6 1994 Proceedings, Volume II 3*, pp. 151–158, Springer, 1994.

[28] D. Geromichalos, M. Azkarate, E. Tsardoulias, L. Gerdes, L. Petrou, and C. Perez Del Pulgar, "Slam for autonomous planetary rovers with global localization," *Journal of Field Robotics*, vol. 37, no. 5, pp. 830–847, 2020.

[29] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, pp. 2564–2571, 2011.

[30] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision – ECCV 2006* (A. Leonardis, H. Bischof, and A. Pinz, eds.), (Berlin, Heidelberg), pp. 404–417, Springer Berlin Heidelberg, 2006.

[31] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Computer Vision – ECCV 2006* (A. Leonardis, H. Bischof, and A. Pinz, eds.), (Berlin, Heidelberg), pp. 430–443, Springer Berlin Heidelberg, 2006.

[32] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *Computer Vision – ECCV 2010* (K. Daniilidis, P. Maragos, and N. Paragios, eds.), (Berlin, Heidelberg), pp. 778–792, Springer Berlin Heidelberg, 2010.

[33] 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*, pp. 224–236, 2018.

[34] 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*, pp. 4938–4947, 2020.

[35] A. Couturier and M. A. Akhloufi, "A review on deep learning for uav absolute visual localization," *Drones*, vol. 8, no. 11, p. 622, 2024.

[36] Y. Xu, L. Pan, C. Du, J. Li, N. Jing, and J. Wu, "Vision-based uavs aerial image localization: A survey," *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 2018.

[37] O. Y. Al-Jarrah, A. S. Shatnawi, M. M. Shurman, O. A. Ramadan, and S. Muhaidat, "Exploring deep learning-based visual localization techniques for uavs in gps-denied environments," *IEEE Access*, vol. 12, pp. 113049–113071, 2024.

[38] X. Luo, X. Wan, Y. Gao, Y. Tian, W. Zhang, and L. Shu, "Jointloc: A real-time visual localization framework for planetary uavs based on joint relative and absolute pose estimation," *arXiv preprint arXiv:2405.07429*, 2024.

[39] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17627–17638, 2023.

[40] D. B. Gennery, "Generalized camera calibration including fish-eye lenses," *International Journal of Computer Vision*, vol. 68, pp. 239–266, 2006.

[41] J. Maki, M. P. Golombek, F. Ayoub, R. Deen, J. Delaune, P. Meras, T. Canham, J. Ravich, M. Cacan, N. Williams, *et al.*, "Ingenuity mars helicopter cameras: Description and results," in *56th Lunar and Planetary Science Conference*, p. 2685, 2025.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[43] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transformers with linear attention," in *International conference on machine learning*, pp. 5156–5165, PMLR, 2020.

[44] D. Pisanti, R. Hewitt, R. Brockers, and G. Georgakis, "Vision-based geo-localization of future mars rotorcraft in challenging illumination conditions," *arXiv preprint arXiv:2502.09795*, 2025.

[45] V. A. Nguyen, "Anylabeling - effortless data labeling with ai support"." https://github.com/vietanhdev/anylabeling, 2023.

[46] J. Delaune, D. S. Bayard, and R. Brockers, "Range-visual-inertial odometry: Scale observability without excitation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2421–2428, 2021.

[47] R. L. Kirk, E. Howington-Kraus, M. R. Rosiek, J. A. Anderson, B. A. Archinal, K. J. Becker, D. A. Cook, D. M. Galuszka, P. E. Geissler, T. M. Hare, I. M. Holmberg, L. P. Keszthelyi, B. L. Redding, W. A. Delamere, D. Gallagher, J. D. Chapel, E. M. Eliason, R. King, and A. S. McEwen, "Ultrahigh resolution topographic mapping of mars with mro HiRISE stereo images: Meter-scale slopes of candidate phoenix landing sites," *Journal of Geophysical Research: Planets*, vol. 113, no. E3, 2008.

[48] J. Laura, L. Adoram-Kershner, D. Meyer, B. Wheeler, K. Bauck, and R. Fergason, "Mars Reconnaissance Orbiter (MRO) Context Camera (CTX) orthoimage generated using Ames stereo pipeline derived digital terrain models." https://doi.org/10.5066/P9JKVWR3, 2023. U.S. Geological Survey.

[49] C. Basich, C. Mauceri, G. Kubiak, J. Delfa, A. Candela, P. Proenca, B. Ridge, and S. Chien, "Onboard autonomous health assessment and global localization for the mars helicopter: Towards multi-flight operations," *Proceedings of the International Symposium on Artificial Intelligence, Robotics and Automation in Space*, 2024.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.