Rover Relocalization for Mars Sample Return by Virtual Template Synthesis and Matching

Tu-Hoa Pham, William Seto, Shreyansh Daftry, Barry Ridge, Johanna Hansen,
Tristan Thrush, Mark Van der Merwe, Gerard Maggiolino, Alexander Brinkman,
John Mayo, Yang Cheng, Curtis Padgett, Eric Kulczycki and Renaud Detry
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA

Abstract— We consider the problem of rover relocalization in the context of the notional Mars Sample Return campaign. In this campaign, a rover (R1) needs to be capable of autonomously navigating and localizing itself within an area of approximately $50 \times 50\,\mathrm{m}$ using reference images collected years earlier by another rover (R0). We propose a visual localizer that exhibits robustness to the relatively barren terrain that we expect to find in relevant areas, and to large lighting and viewpoint differences between R0 and R1. The localizer synthesizes partial renderings of a mesh built from reference R0 images and matches those to R1 images. We evaluate our method on a dataset totaling 2160 images covering the range of expected environmental conditions (terrain, lighting, approach angle). Experimental results show the effectiveness of our approach. This work informs the Mars Sample Return campaign on the choice of a site where Perseverance (R0) will place a set of sample tubes for future retrieval by another rover (R1).

I. INTRODUCTION

The Mars 2020 *Perseverance* rover that launched in July will search for signs of ancient life on Mars by collecting samples from Martian rocks and soil using an arm-mounted drill. These sample will be stored in hermetically-sealed sample tubes and released at one or multiple sample cache *depots*, for possible recovery via a notional NASA-ESA follow-up mission that would land in 2028, Mars Sample Return (MSR) [25]. The mission would include a rover (the sample-fetching rover, SFR) and a rocket (the Mars ascent vehicle, MAV). SFR would drive to the sample cache depots, pick up the tubes and bring them back to the lander for transfer and launch to Mars orbit through the MAV. Finally, a probe would capture the container in orbit and bring it back to Earth for sample containment and analysis. This paper focuses on the sample retrieval phase of the campaign, in particular the problem of in-depot navigation for tube pickup.

While Mars rovers intended to survive the Martian winter have used radioisotope heating units or thermoelectric

generators, the base SFR design calls for neither to limit costs, and should complete its mission in a single season before shutting down forever. Accounting for a notional $10\,\mathrm{km}$ drive to the sample depot(s) and back to the MAV leaves only 30 sols for SFR to pick up 36 tubes. Due to limited Earth-Mars communication windows, *ground-in-the-loop* tube pickup takes a minimum 3-sol-per-tube pickup time. SFR thus needs to retrieve tube autonomously, with minimal guidance from operators on Earth.

Towards this, Perseverance will document depots by capturing images as it drops sample tubes. These images will be telemetered to Earth to reconstruct a *map* of the depot annotated with tube poses. SFR will retrieve sample tubes by *relocalizing* itself with respect to that map, years later. This is a difficult problem for multiple reasons. In contrast to relocalization on Earth, which often benefits from human-made objects and environments with distinctive visual structures, the surface of Mars mostly consists of desert-like environments with fewer salient features. The (changing) interaction of light with small rocks and other terrain features further complicates the problem. The difficulty of relocalization in changing environments is well-established on Earth [35]. Anecdotally, we observe the same challenges on Mars. We depict in Fig. 1 two stereo pairs that feature-based localization [20] failed to align despite being from the exact same viewpoint but under different lighting. In addition to lighting changes, relocalization must be robust to other environment changes such as dust deposition or accumulation of sand to form small drifts. In that sense, the nature and timeline of the mission permits dedicating special care to the crafting of the map, e.g., by selecting landmarks that are more likely persist over time.

In this paper, we propose a novel method for relocalization over changing environments by Virtual Template Synthesis and Matching (VTSM). Our work builds upon the state of the art in multiple areas of visual localization (see Section II) and offers the following contributions.

- A relocalization algorithm that synthesizes partial renderings of multiple points of interest on the map, as perceived from multiple virtual poses, and matches them to real observations across multiple modalities (sizes and filters) for changing environments (see Section III).
- A new dataset spanning the range of environmental conditions we expect to face on Mars, including 3 terrain

types, captured by 4 cameras along 60 viewpoints at 3 times of the day, totalling 2160 images annotated with ground-truth poses from a motion capture system (see Section IV). We release our dataset publicly to foster the research in this exciting problem for space exploration[1].

- An extensive performance and sensitivity analysis for our method, leading to recommendations on depot mapping and navigation strategies (see Section V).

We finally discuss challenges we encountered, current limitations and future extensions of our work (see Section VI).
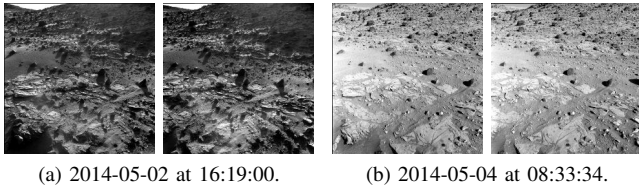

(a) 2014-05-02 at 16:19:00.   (b) 2014-05-04 at 08:33:34.

Fig. 1.   Stereo images captured on Mars from the same viewpoint at different times (available `https://pds-imaging.jpl.nasa.gov/`).

II. RELATED WORK

In this work, we consider the problem of *relocalization* of a rover with respect to a *map* built from images captured by another rover, years prior. Autonomous localization capabilities on Mars rovers are currently limited to motion estimation by visual-inertial odometry (VIO) [21], [18], and absolute orientation estimation using gravity and Sun position in the sky [1], [17]. Mars rovers do not implement onboard absolute position estimation. Instead, it is conducted on Earth, by aligning rover and orbital imagery [10], [6], [38] with meter-scale accuracy (Mars Reconnaissance Orbiter). While VIO could be used to estimate relative poses between rovers, it is also vulnerable to environment changes, e.g., lighting [31].

Localization techniques leveraging 3D structure [3] can mitigate the effects of lighting changes, e.g., by performing 3D registration using iterative closest point (ICP) [4] or fast point feature histograms (FPFH) [33]. However, those tend to converge towards local optima in the absence of an accurate pose prior, especially on the rather flat terrains that may serve for depot construction. We consider instead 3D (depth) reconstructed together with texture (color), e.g., through Structure from Motion (SfM) [36] or Simultaneous Localization and Mapping (SLAM) [5], which can accomodate maps built over multiple sessions under the same lighting [14] or different lighting over small pose changes [27]. Such 3D maps can be used for relocalization by correspondence search between features from 2D images to relocalize and a database of features associated to 3D map points [34]. The underlying matching schemes can also be accelerated to take advantage of multi-camera systems [13] but remain subject to similar limitations as 2D local feature matching under scene changes. Notably, [24] showed that image patches could be more robust to changing conditions than point features, though less accurate. While robustness can also be

attained by repeated traverses under different conditions [7], [30], depot construction on Mars would be single-shot. In a recent benchmark of visual localization in changing conditions [35], image retrieval techniques were shown to sometimes succeed at providing a coarse pose estimate when local feature matching would fail, e.g., by augmenting a mapping database with synthetic renderings [39]. 3D map rendering was also used for relocalization by minimizing distance metrics between real and synthetic observations, such as normalized information distance [28], normalized cross-correlation [19] and photometric error [26], showing some robustness moderate lighting changes. Changing shadows were further downweighted in the image alignment pipeline of [16] however large lighting changes remain a challenge. We build upon these works and extend relocalization capabilities by viewpoint synthesis to extensive changes in lighting and possibly scene geometry over the years.

Instead of relocalizing images across different lighting, shadow-invariant image transformations were developed in [9], [23] assuming infinitely-narrow camera sensor responses and illumination by a single Planckian source. While the former assumption can be relaxed [29], the latter may not always hold on Mars depending on the atmosphere radiance at capture time. Similar problems were also addressed recently using deep neural networks, e.g., to learn image relighting [2] and image representations that are robust to lighting changes [8], [41], [32]. While we believe such methods will result in future breakthroughs for robust relocalization on Earth, their applicability to space exploration remains restricted by the scarcity of data for training, low-compute for space-rated hardware, limited interpretability and concerns about generalization to events unseen during training. Still, we build our approach in such a way that it could accomodate further advances in either field.

III. METHOD

Our goal is to enable in-depot navigation for sample tube retrieval, by estimating the 6D pose of SFR with respect to a depot map built from images taken by Perseverance several years before. To do so, we propose a relocalization method that synthesizes partial renderings of the depot map from virtual viewpoints in the vicinity of its current pose estimate and compares those to actual observations. In the following, matrix variables are denoted in bold and scalar in italic.

A. Overview

Let $\mathcal{I}^{\text{M2020}} = (\mathbf{I}_{\text{L},i}^{\text{M2020}}, \mathbf{I}_{\text{R},i}^{\text{M2020}})_{i \in [1, N^{\text{M2020}}]}$ denote a set of N^{M2020} left and right stereo image pairs captured by Perseverance during depot construction on Mars and telemetered back to Earth. On Earth, we register all images $\mathcal{I}^{\text{M2020}}$ and build a map $\mathcal{M}$ equiped with a global frame $\mathcal{W}$, in which we express rover poses $^{\mathcal{W}}\mathbf{T}_{\text{M2020}}$ as well as tubes $^{\mathcal{W}}\mathbf{T}_{\mathcal{T}}$. In addition, we denote by $\widehat{\mathcal{M}}$ a subset of $\mathcal{M}$ in which areas likely to be affected by wind (e.g., sand) have been removed on Earth between the Mars 2020 and MSR missions.

Consider now the task of relocalizing SFR several years after Perseverance. While driving from landing site to depot,
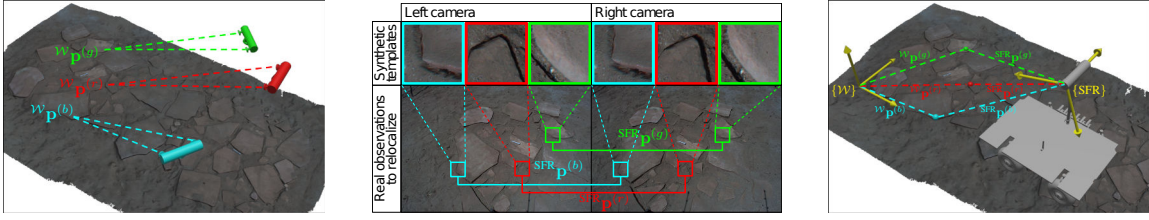
[1] `https://data.caltech.edu/records/1898`

---

[1] `https://data.caltech.edu/records/1898`

Fig. 2. We render points $^{\mathcal{W}}\mathbf{p}$ on the mesh built from Perseverance images from multiple virtual viewpoints (left). Synthetic templates are matched to real SFR images to compute points in camera frame $^{\text{SFR}}\mathbf{p}$ (middle). We align both point sets to estimate the pose $^{\mathcal{W}}\mathbf{T}_{\widehat{\text{SFR}}}$ (right, conceptual rover depicted).

SFR's position is monitored onboard through VIO, which induces drift growing with traversed distance and corrected by manually aligning surface and orbital imagery. When reaching the depot, relocalization with respect to $\mathcal{M}$ is performed by manually aligning SFR and Perseverance images instead. Due to mission time constraints, this is done only once. For in-depot navigation, we then assume an SFR pose estimate $^{\mathcal{W}}\mathbf{T}_{\widehat{\text{SFR}}}$ subject to VIO errors accumulated over drives post ground-in-the-loop initialization (e.g., $20\,\text{cm}$ for a few-meter drive, depending on specifications) and $1.5°$ uncertainty from onboard attitude estimation [1].

Pose alignment consists in computing a posterior pose given new camera images and a prior pose estimate $^{\mathcal{W}}\mathbf{T}_{\widehat{\text{SFR}}}$. We compute this alignment by simulating the SFR cameras from several viewpoint hypotheses $\mathcal{V}$ and iteratively match synthetic renderings $(\mathbf{I}_{\text{L}}^{\mathcal{V}}, \mathbf{I}_{\text{R}}^{\mathcal{V}})$ to real images $(\mathbf{I}_{\text{L}}^{\text{SFR}}, \mathbf{I}_{\text{R}}^{\text{SFR}})$. More precisely, in order to account for lighting and scene geometry changes between mapping and relocalization, we do not directly align full renderings but rather patches thereof, or *synthetic templates* $(\widehat{\mathbf{I}}_{\text{L}}^{\mathcal{V}}, \widehat{\mathbf{I}}_{\text{R}}^{\mathcal{V}})$, centered on a point $^{\mathcal{W}}\mathbf{p}$ and synthesized on the fly. This enables: 1) efficiently evaluating multiple viewpoint hypotheses without having to render full images; 2) prioritizing landmarks that are likely to remain visually similar over time, and conversely ignoring parts of the image that are likely to change; 3) facilitating pose search as a full image may contain cast shadows (edges that do not persist with lighting changes) but small patches may be uniformly lit. We match the synthetic templates to the real observations, resulting in $^{\text{SFR}}\mathbf{p}$, the point corresponding to $^{\mathcal{W}}\mathbf{p}$ in the SFR optical frame. We thus collect $N^{\mathcal{C}}$ correspondences $(^{\mathcal{W}}\mathbf{p}, ^{\text{SFR}}\mathbf{p})$ to update $^{\mathcal{W}}\mathbf{T}_{\widehat{\text{SFR}}}$ by least-squares transformation estimation, and further refine it over N^{iter} iterations. We summarize our method in Alg. 1 and Fig. 2 and discuss its components in the following.

B. Viewpoint Sampling

Pose search begins with an initial viewpoint estimate $^{\mathcal{W}}\mathbf{T}_{\widehat{\text{SFR}}}$, which is typically computed from the previous pose and the rover's VIO-derived motion since the last alignment. We account for uncertainty incurred by VIO by searching for a best-matching pose in the neighborhood of $^{\mathcal{W}}\mathbf{T}_{\widehat{\text{SFR}}}$. Rover localization is largely a planar problem, and for perfectly flat terrain our search should be constrained to a planar search. However, as we expect substantial relief through a depot, we conduct a full 6D search. We search the neighborhood of the initial pose by applying a perturbation transformation $^{\text{SFR}}\mathbf{T}_{\mathcal{V}}$

Algorithm 1 Virtual Template Synthesis and Matching

Precondition: $(\mathbf{I}_{\text{L}}^{\text{SFR}}, \mathbf{I}_{\text{R}}^{\text{SFR}})$ stereo pair captured by SFR, $\mathcal{M}$ depot map, $\widehat{\mathcal{M}}$ sampling mask, $^{\mathcal{W}}\mathbf{T}_{\widehat{\text{SFR}}}$ initial guess
1: **function** LOCALIZE$((\mathbf{I}_{\text{L}}^{\text{SFR}}, \mathbf{I}_{\text{R}}^{\text{SFR}}), \mathcal{M}, ^{\mathcal{W}}\mathbf{T}_{\widehat{\text{SFR}}})$
2:      **for** $i \leftarrow 1$ to N^{iter} **do**
3:          $\mathcal{C}_{\mathcal{W}/\text{SFR}} \leftarrow \{\}$ ▷ world/SFR correspondences
4:          **while** SIZE$(\mathcal{C}_{\mathcal{W}/\text{SFR}}) \neq N^{\mathcal{C}}$ **do**
5:              $^{\mathcal{W}}\mathbf{T}_{\mathcal{V}} \leftarrow$ RANDOMIZEVIEWPOINT$(^{\mathcal{W}}\mathbf{T}_{\widehat{\text{SFR}}})$
6:              $^{\mathcal{W}}\mathbf{p} \leftarrow$ SAMPLEPOINT$(\widehat{\mathcal{M}}, ^{\mathcal{W}}\mathbf{T}_{\mathcal{V}})$
7:              $(\widehat{\mathbf{I}}_{\text{L}}^{\mathcal{V}}, \widehat{\mathbf{I}}_{\text{R}}^{\mathcal{V}}) \leftarrow$ SYNTHESIZE$(\mathcal{M}, ^{\mathcal{W}}\mathbf{T}_{\mathcal{V}}, ^{\mathcal{W}}\mathbf{p})$
8:              $(u_{\text{L}}^{\text{SFR}}, v_{\text{L}}^{\text{SFR}}) \leftarrow$ MATCH$(\widehat{\mathbf{I}}_{\text{L}}^{\mathcal{V}}, \mathbf{I}_{\text{L}}^{\text{SFR}})$ ▷ Left
9:              $(u_{\text{R}}^{\text{SFR}}, v_{\text{R}}^{\text{SFR}}) \leftarrow$ MATCH$(\widehat{\mathbf{I}}_{\text{R}}^{\mathcal{V}}, \mathbf{I}_{\text{R}}^{\text{SFR}})$ ▷ Right
10:             **if** VALID$(u_{\text{L}}^{\text{SFR}}, v_{\text{L}}^{\text{SFR}}, u_{\text{R}}^{\text{SFR}}, v_{\text{R}}^{\text{SFR}})$ **then**
11:                $^{\text{SFR}}\mathbf{p} \leftarrow$ STEREO$(u_{\text{L}}^{\text{SFR}}, v_{\text{L}}^{\text{SFR}}, u_{\text{R}}^{\text{SFR}}, v_{\text{R}}^{\text{SFR}})$
12:                APPEND$(\mathcal{C}_{\mathcal{W}/\text{SFR}}, (^{\mathcal{W}}\mathbf{p}, ^{\text{SFR}}\mathbf{p}))$
13:             **end if**
14:          **end while**
15:          $^{\mathcal{W}}\mathbf{T}_{\widehat{\text{SFR}}} \leftarrow$ GETTRANSFORM$(\mathcal{C}_{\mathcal{W}/\text{SFR}})$
16:      **end for**
17:      **return** $^{\mathcal{W}}\mathbf{T}_{\widehat{\text{SFR}}}$
18: **end function**

---

characterized by translation and rotation search bounds $\widetilde{t}, \widetilde{r}$ (initially, $\widetilde{t} = \widetilde{t}_0 = 20\,\text{cm}$, $\widetilde{r} = \widetilde{r}_0 = 1.5°$). We build the rotational perturbation by randomly sampling a rotation axis from the unit sphere and a rotation angle from a uniform distribution on $[-\widetilde{r}, \widetilde{r}]$. Similarly, the perturbation's translational magnitude is sampled within $[-\widetilde{t}, \widetilde{t}]$ and its direction within the local surface tangent for the first iteration (planar search), then the 3D unit sphere for subsequent refinement.

C. Virtual Template Synthesis

In the following, we consider the depot map $\mathcal{M}$ as a textured polygon mesh. Using OpenGL, we build a rendering environment reproducing SFR rectified stereo calibration parameters, enabling the synthesis of images similar in appearance to what would be captured by the real rover at arbitrary viewpoints $^{\mathcal{W}}\mathbf{T}_{\mathcal{V}}$. To evaluate multiple viewpoint hypotheses efficiently, we do not render full stereo images (e.g., $5472 \times 3648 \approx 20\,\text{Mpixel}$) but instead square patches of side length ℓ centered on points of interest (e.g., $\ell = 256$, about $300\times$ smaller). The synthesis pipeline is as follows. Given a sampling mask $\widetilde{\mathcal{M}}$ of the depot areas that can reliably be used for relocalization (e.g., by keeping large rocks and filtering out sand), we randomly select a 3D point

$^{\mathcal{W}}\widetilde{\mathbf{p}}$ from the vertices constituting $\widehat{\mathcal{M}}$ visible from $^{\mathcal{W}}\mathbf{T}_{\mathcal{V}}$. When the scene geometry is not expected to change, we set $\widehat{\mathcal{M}} := \mathcal{M}$. We then project $^{\mathcal{W}}\widetilde{\mathbf{p}}$ into the left virtual camera frame, yielding 2D pixel coordinates $(u_{\mathrm{L}}^{\mathcal{V}}, v_{\mathrm{L}}^{\mathcal{V}})$. We thus synthesize a left template $\widehat{\mathbf{I}}_{\mathrm{L}}^{\mathcal{M}}$ centered on $(u_{\mathrm{L}}^{\mathcal{V}}, v_{\mathrm{L}}^{\mathcal{V}})$, together with an associated depth map $\widehat{\mathbf{D}}_{\mathrm{L}}^{\mathcal{M}}$. We use the full mesh $\mathcal{M}$ for rendering, which contains areas that may change over time but can still be used for matching. We estimate which pixels to keep from the synthetic template $\widehat{\mathbf{I}}_{\mathrm{L}}^{\mathcal{M}}$ in two ways. First, we render the depth map $\widehat{\mathbf{D}}_{\mathrm{R}}^{\widehat{\mathcal{M}}}$ associated to the sampling mask $\widehat{\mathcal{M}}$ and keep all pixels of $\widehat{\mathbf{I}}_{\mathrm{L}}^{\mathcal{M}}$ that have a depth value in $\widehat{\mathbf{D}}_{\mathrm{R}}^{\widehat{\mathcal{M}}}$ as they correspond to vertices of $\widehat{\mathcal{M}}$. Second, we identify pixels that may change locally (e.g., sand) but still serve as contrasting background for features in the foreground (e.g., rock edges). We define those as pixels of $\widehat{\mathbf{I}}_{\mathrm{L}}^{\mathcal{M}}$ that do *not* have a depth value in $\widehat{\mathbf{D}}_{\mathrm{L}}^{\widehat{\mathcal{M}}}$ but are beyond edges marking depth discontinuity in $\widehat{\mathbf{D}}_{\mathrm{L}}^{\mathcal{M}}$.

We synthesize the right template $\widehat{\mathbf{I}}_{\mathrm{R}}^{\mathcal{V}}$ by similarly generating color and depth patches from a virtual *right* camera.

D. Synthetic-to-Real Template Matching

We now search for $^{\mathrm{SFR}}\mathbf{p}$, the point corresponding to $^{\mathcal{W}}\mathbf{p}$ in the SFR camera frame, by separately searching for left and right templates $\widehat{\mathbf{I}}_{\mathrm{L}}^{\mathcal{V}}, \widehat{\mathbf{I}}_{\mathrm{L}}^{\mathcal{V}}$ in the real SFR images $\mathbf{I}_{\mathrm{L}}^{\mathrm{SFR}}, \mathbf{I}_{\mathrm{R}}^{\mathrm{SFR}}$, yielding SFR pixel coordinates $(u_{\mathrm{L}}^{\mathrm{SFR}}, v_{\mathrm{L}}^{\mathrm{SFR}}), (u_{\mathrm{R}}^{\mathrm{SFR}}, v_{\mathrm{R}}^{\mathrm{SFR}})$ of maximum normalized cross-correlation. We make use of the epipolar constraint between rectified stereo images to reject matches such that the vertical difference $|u_{\mathrm{L}}^{\mathrm{SFR}} - u_{\mathrm{R}}^{\mathrm{SFR}}|$ exceeds a chosen threshold ϵ_u. If $|u_{\mathrm{L}}^{\mathrm{SFR}} - u_{\mathrm{R}}^{\mathrm{SFR}}| \leq \epsilon_u$, we calculate $^{\mathrm{SFR}}\mathbf{p}$ using the average vertical coordinate $0.5 * (u_{\mathrm{L}}^{\mathrm{SFR}} + u_{\mathrm{R}}^{\mathrm{SFR}})$ and horizontal disparity $(v_{\mathrm{L}}^{\mathrm{SFR}} - v_{\mathrm{R}}^{\mathrm{SFR}})$.

We exploit two strategies to facilitate synthetic-to-real matching. First, we search for matches using different variations of synthetic templates $(\widehat{\mathbf{I}}_{\mathrm{L}}^{\mathcal{V}}, \widehat{\mathbf{I}}_{\mathrm{R}}^{\mathcal{V}})$: size (e.g., $\ell/2$-length sub-template) and derivative order (e.g., direct grayscale, or processed through Sobel, Laplacian operators). Smaller templates are easier to match across larger viewpoint differences at the cost of more false positives to filter out, while differentiating templates partially mitigates the effects of lighting differences. Second, rather than searching for templates $(\widehat{\mathbf{I}}_{\mathrm{L}}^{\mathcal{V}}, \widehat{\mathbf{I}}_{\mathrm{R}}^{\mathcal{V}})$ in the full SFR images $(\mathbf{I}_{\mathrm{L}}^{\mathrm{SFR}}, \mathbf{I}_{\mathrm{R}}^{\mathrm{SFR}})$, we compute bounds on their possible pixel coordinates based on the pose uncertainty $\widetilde{t}, \widetilde{r}$ and perform the template search on these sub-images (e.g., 800×800 patch within 5472×3648).

E. Pose Update

We repeat the steps described in Sections III-B to III-D until reaching a target number $N^{\mathcal{C}}$ of world-camera point correspondence candidates $(^{\mathcal{W}}\mathbf{p}, {}^{\mathrm{SFR}}\mathbf{p})$ and estimate a transformation using the Umeyama algorithm [40] and RANSAC for outlier rejection [11] within the Point Cloud Library [33]. If successful, we set the resulting maximum-inlier transformation as new pose estimate $^{\mathcal{W}}\mathbf{T}_{\widehat{\mathrm{SFR}}}$ and repeat the process for N^{iter} iterations or until the pose update converges within a chosen threshold (e.g., $1\,\mathrm{mm}$). We propose additional mechanisms to facilitate the pose search.

When a transformation cannot be estimated from the correspondence candidates collected at this iteration:

- STALL: get new correspondences from the same pose
- RESEED: apply a random perturbation to the current pose before collecting new correspondences

The STALL procedure is well suited when we already have at least one successful iteration, i.e., the current pose estimate is already close to the real pose. In contrast, RESEED lets us evaluate multiple guesses within a potentially large initial uncertainty range (e.g., $50\,\mathrm{cm}$) while maintaining smaller synthetic viewpoint variations (e.g., $20\,\mathrm{cm}$) for local search.

When a transformation is successfully estimated:

- ANNEAL: decrease synthetic viewpoint randomization following $\widetilde{t} := \gamma\widetilde{t}, \widetilde{r} := \gamma\widetilde{r}$, with $\gamma \in [0, 1]$
- DISTRIBUTE: randomize synthetic viewpoints around multiple poses rather than the current estimate only
- REUSE: carry over a set number of correspondences (inliers) throughout successful iterations

The γ parameter facilitates convergence by reducing synthetic viewpoint randomization over time, e.g., stop randomizing ($\gamma = 0$), halve every iteration ($\gamma = 0.5$), keep constant ($\gamma = 1$). In DISTRIBUTE, rather than focusing viewpoint synthesis around the maximum-inlier transformation from RANSAC, we spread it around the best pose candidates with frequency weighed by their inlier count (e.g., 3 candidates with 10, 15, 25 inliers, would have a 20, 30, 50 % pick rate, respectively) to avoid local minima. In REUSE, we do not restart the world-camera point correspondence search from scratch every time but instead keep a set number inliers from the past iteration (e.g., 50 %). This helps stabilize the pose search and avoid large variations across iterations. Finally, we limit the number of times STALL and RESEED can be performed consecutively and return a failure code if reached.

IV. DATASET

(a) Flagstone testbed. (b) Camera pose grid around tubes $\mathcal{T}_A$, $\mathcal{T}_B$.

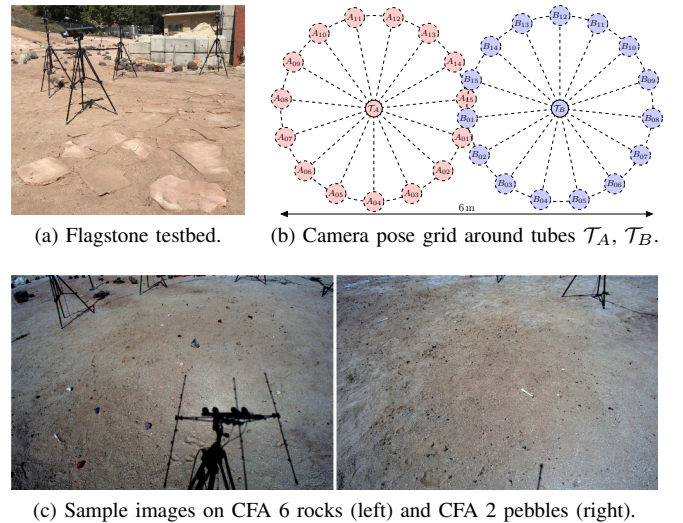(c) Sample images on CFA 6 rocks (left) and CFA 2 pebbles (right).

Fig. 3. Outdoor dataset: (a) testbed, (b) capture grid, (c) sample images.

To evaluate the performance of our proposed method towards mission planning, we collect an outdoor dataset

capturing representative conditions we expect to face on Mars. In a preliminary study [31], we showed on an indoor dataset that lighting changes and lack of consistent features throughout were a major challenge for rover localization. We investigate these problems further by collecting a new dataset, outdoor, enabling: 1) natural light and shadows from the sun that change continuously; 2) diverse, representative terrains, based on discussions with Mars geologists on where sample tubes could be dropped; 3) a large workspace to evaluate different depot imaging strategies as well as in-depot navigation during tube recovery (e.g., how far SFR can deviate from areas imaged by Perseverance).

We construct a camera acquisition setup for outdoor use consisting of four FLIR BlackFly S cameras (5472×3648, color, $77°$ field of view) arranged as two stereo pairs of baseline 20 and $40\,\text{cm}$, representative of the Perseverance rover's optics [22]. The cameras are covered by an aluminum plate serving as heat shield for extended use under sunlight and rigidly linked to a frame carrying motion capture (MoCap) markers for pose ground-truthing using 10 Vicon T-160 cameras (see Fig. 3a). The dataset captures the following:

- 3 terrain types (see Figs. 3a and 3c):
  - "Flagstone": broken stone slabs covered with a thin layer of sand similarly to fractured bedrock on Mars.
  - "CFA6": a rock distribution of *cumulative fractional area* (CFA – a measure of rock density [15]) equal to $6\,\%$, the smallest rocks still visible from orbit to guide the choice of depot location. Rocks encountered in practice would only be this big or smaller.
  - "CFA2": small rocks only visible from surface imagery (not orbit), here pebbles on dust and sand.
- 3 image capture times: "am" (9 am to 10 am), "nn" (noon to 1 pm), "pm" (3 pm to 4 pm)
- 2 sample tubes with variable visibility: unoccluded or in a crack between slabs (flagstone), unoccluded or $25\,\%$ covered by sand (CFA 2), 50 or $75\,\%$ occluded (CFA6)
- 15 camera tripod positions along two circles, each centered on a sample tube (30 stops total, see Fig. 3b)
- 2 camera orientations at each stop (look at each tube)
- 4 cameras: 2 stereo pairs of baseline 20 and $40\,\text{cm}$

Overall, our outdoor dataset comprises 540 capture configurations, totalling 2160 images collected over the course of 3 days, annotated with reference poses from MoCap.

V. Experiments

In this Section, we examine the performance of our method on different types of terrains and lighting conditions, then assess different depot mapping and navigation strategies to make recommendations for Mars Sample Return planning.

A. Relocalization Performance

The first step consists in building a map of the depot. For each terrain and capture time (9 combinations total), we build a depot map from 60 viewpoints captured with the $40\,\text{cm}$-baseline stereo cameras, representative of those Perseverance would use for depot imaging. Using the Agisoft

Metashape software, we generate a textured mesh $\mathcal{M}$ of the full scene. For Flagstone and CFA6 (possible scene perturbations, e.g., from Martian wind), we manually process $\mathcal{M}$ with the Blender 3D graphics software to only keep rocks in the sampling mask $\widehat{\mathcal{M}}$. For CFA2 (undisturbed scene), we keep the full mesh and set $\widehat{\mathcal{M}} := \mathcal{M}$. We depict the resulting meshes and masks in Figs. 4b to 4d. We then use this map to relocalize images taken at different times of the day, see Fig. 4a. Viewpoints to relocalize (green arrows) overlap with mapping images (blue) when they are from the same capture time, otherwise vary slightly over different capture sessions.

We evaluate our method by starting away from the ground-truth pose $^{\mathcal{W}}\mathbf{T}_{\text{SFR}}$, applying a random transformation of 10-to-20 cm translation and $1.5°$ rotation components (see Section III-A). We run VTSM using the resulting perturbed pose $^{\mathcal{W}}\mathbf{T}_{\widehat{\text{SFR}}}$ as initial guess, $\widetilde{t}_0 = 20\,\text{cm}$ and $\widetilde{r}_0 = 1.5°$ initial search bounds with decay parameter $\gamma = 0.5$, virtual template sizes 128 and 256, and epipolar constraint threshold $\epsilon_u = 8\,\text{pixels}$. A run is considered successful if $N^{\mathcal{C}} = 100$ correspondences are found and the pose is successfully updated for $N^{\text{iter}} = 5$ iterations. The transformation estimation error is defined as the difference between ground-truth pose and final estimate, $^{\text{SFR}}\mathbf{T}_{\widehat{\text{SFR}}} = (^{\mathcal{W}}\mathbf{T}_{\text{SFR}})^{-1} \cdot {}^{\mathcal{W}}\mathbf{T}_{\widehat{\text{SFR}}}$.

We report in Fig. 4e the VTSM relocalization success rate and average error on each terrain type, sorted by time difference between mapping and relocalization: $0\,\text{h}$ (same time for both), $3\,\text{h}$ (e.g., noon relocalization vs morning map), $6\,\text{h}$ (e.g., afternoon relocalization vs morning map). We also depict Flagstone detailed results in Fig. 4f as 2D plots where each point's coordinates represent the distance (linear and angular) between the viewpoint to relocalize and the nearest viewpoint used to map the depot, and its color the relocalization accuracy. As a baseline to our method, we also report these metrics when performing localization by matching local features between the same relocalization and nearest mapping viewpoints. We did so using the LIBVISO2 package [12], modified to use SIFT features for better robustness in exchange for longer computation time.

We observe the following. First, while SIFT-based localization is generally more accurate than VTSM when it successfully estimates a transformation, this success rate decreases significantly with lighting changes across all terrains. Notably, it completely drops to $0\,\%$ with $6\,\text{h}$ of natural sunlight difference while VTSM achieves $100\,\%$ across all configurations. We note that relocalization errors are largest on CFA6, possibly due to the small size of rock features together with their sparsity in the sampling mask (see Fig. 4c), suggesting that it may be preferrable to build the depot on large-enough rocks if the scene is expected to be affected by Martian wind (Flagstone), or to find areas where sand will remain undisturbed (CFA2). Second, Fig. 4f illustrates that VTSM is generally successful throughout the assessed range of $40\,\text{cm}$ and $20°$ between test and mapping viewpoints. This prompts us to consider larger ranges in the next sections. Finally, we report an average run time of $3\,\text{min}$ $55\,\text{s}$ on a $4.5\,\text{GHz}$ CPU-only, single-threaded implementation compatible with space-rated hardware.

For the sake of completeness, we considered alternative methods that may *not* be compatible with space-rated hardware, such as neural-network-based features [37], that we observed generally performing worse than SIFT on Mars-like terrains under lighting changes [31]. We also implemented the image transform of [9] as a preprocessing step and observed that the added noise, as reported by [29], particularly hindered subsequent local feature matching on our relatively barren terrains. While this may be alleviated with additional training, it remains unclear whether the single Planckian source assumption would hold at tube pickup time on Mars. Finally, we trained the benchmark-leading [35] image retrieval technique of [39] on our dataset. While image retrieval itself aims at returning *the nearest pose in a predefined database* rather than the actual rover pose itself, we could envision using such a system to seed VTSM with a coarse estimate for further refinement. However, we again observed the method to fail at retrieving such a pose, which may be due to the DenseVLAD features employed being derived from SIFT, therefore subject to similar limitations.

B. Parameter Sensitivity Analysis

While the $40\,\mathrm{cm}$-baseline stereo cameras are representative of those Perseverance will use for depot mapping, the SFR design may be constrained by further size and payload requirements. Keeping the depot maps built from the $40\,\mathrm{cm}$-baseline stereo images, we now relocalize images taken by other cameras in a $20\,\mathrm{cm}$-baseline configuration. We report the resulting errors on all three terrains over $6\,\mathrm{h}$ time difference between mapping and relocalization in Fig. 5a. We observe that while the smaller stereo baseline results in higher relocalization errors, these are expected as the theoretical depth uncertainty itself also increases by 6.5-to-$26.0\,\mathrm{mm}$ for a $1\,\mathrm{pixel}$ disparity uncertainty 3-to-$6\,\mathrm{m}$ ahead of the camera. As success rates remain similar, we infer here that VTSM itself is robust to camera changes, with its accuracy contingent on that permitted by the chosen setup.

We now consider an alternative scenario where the rover's position uncertainty suddenly grows beyond the previous $20\,\mathrm{cm}$ to $50\,\mathrm{cm}$ (e.g., slippage). To address this, one possibility is to simply increase the randomization range $\widetilde{t}_0$ when generating synthetic viewpoints to $50\,\mathrm{cm}$ around the current pose guess $^{\mathcal{W}}\mathbf{T}_{\widehat{\mathrm{SFR}}}$. However, the increased search range also requires more attempts to sample poses closer to the real one, while also generating more false positives. Furthermore, template matching takes longer as the increased uncertainty only lets us restrict the search to about 2000×2000 patches within the real images instead of 800×800 as described in Section III-D. Instead, we propose to keep VTSM virtual viewpoint randomization at $20\,\mathrm{cm}$, but around *multiple pose seeds* randomly sampled within the $50\,\mathrm{cm}$ uncertainty range. We choose to run one VTSM iteration over 100 such pose seeds and select the one resulting in the maximum number of correspondence inliers to run the rest of the algorithm on. Adding these results to Fig. 5a, we observe that success rates remain at $100\,\%$ and that relocalization errors are slightly lower, which may stem from the search over multiple pose
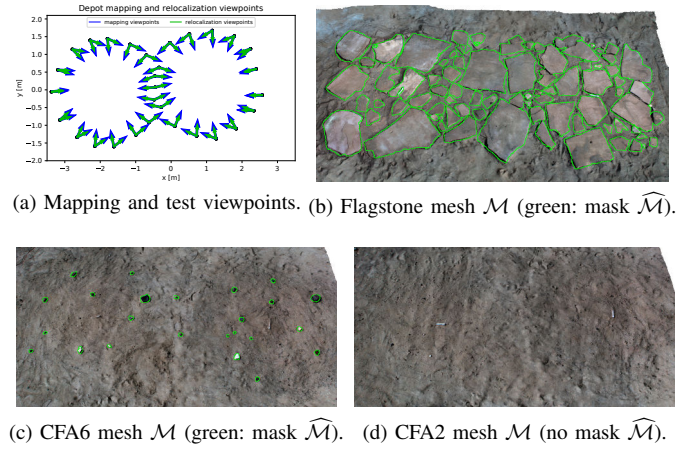

(a) Mapping and test viewpoints.

(b) Flagstone mesh $\mathcal{M}$ (green: mask $\widehat{\mathcal{M}}$).

(c) CFA6 mesh $\mathcal{M}$ (green: mask $\widehat{\mathcal{M}}$).

(d) CFA2 mesh $\mathcal{M}$ (no mask $\widehat{\mathcal{M}}$).

	Terrain	Flagstone			CFA6			CFA2		
	Time diff. [h]	0	3	6	0	3	6	0	3	6
SIFT	Success [%]	98.9	66.7	0.0	100.0	70.7	0.0	98.9	52.9	0.0
	Error [mm]	0.8	22.8	N/A	0.4	6.5	N/A	0.6	10.7	N/A
VTSM	Success [%]	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Error (init) [mm]	148.1	152.3	149.2	149.7	147.9	150.0	149.3	150.5	148.0
	Error (end) [mm]	51.2	56.0	50.2	65.4	75.9	98.5	51.1	52.8	76.6

(e) Localization success rate and average error on all terrain-time differences.
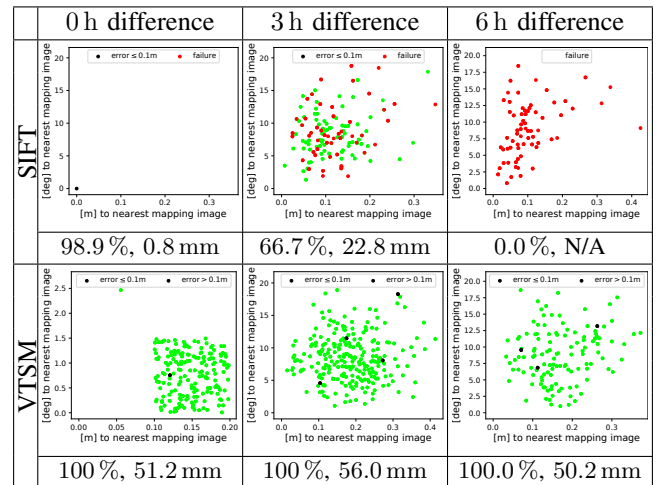
(f) Transformation estimation success rate and average error on flagstone depot across 0, 3, $6\,\mathrm{h}$ capture time and lighting differences. Note: SIFT-$0\,\mathrm{h}$ amounts to estimating the (zero) transformation between the same images and VTSM-$0\,\mathrm{h}$ illustrates the effects of 10-to-$20\,\mathrm{cm}$, 0-to-$1.5°$ pose randomization alone. 3 and $6\,\mathrm{h}$ plots include pose variations across different depot traverses.

Fig. 4. VTSM evaluation on all terrain-time differences.

seeds within a $50\,\mathrm{cm}$ range being a better initializer than relocalizing from a single one within $20\,\mathrm{cm}$. Fig. 5b depicts real and synthetic images from VTSM showing successful matching on all terrains despite strong lighting changes between mapping and relocalization. We report an average $1\,\mathrm{h}\ 4\,\mathrm{min}$ run time per relocalization attempt with $50\,\mathrm{cm}$ uncertainty, including $1\,\mathrm{h}$ to evaluate the 100 pose seeds, which could be improved using other search schemes or early stop criteria (e.g., minimum inlier ratio).

C. Depot Imaging Strategy and Relocalization Range

Finally, we consider the case where SFR deviates from the path Perseverance took when imaging the depot. We

	Terrain-time	Flagstone-6 h			CFA6-6 h			CFA2-6 h		
	Configuration	**ref**	**bl**	**50cm**	**ref**	**bl**	**50cm**	**ref**	**bl**	**50cm**
VTSM	Success [%]	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
VTSM	Error (init) [mm]	149.2	146.2	363.8	150.0	153.7	379.2	148.0	146.5	377.1
VTSM	Error (end) [mm]	50.2	58.2	49.3	98.5	123.0	83.5	76.6	84.6	62.2

(a) VTSM results on different configurations. Reference results **ref**: 40 cm relocalization baseline, 10-to-20 cm initial randomization. Two variations: **bl**: 20 cm relocalization baseline, **50cm**: 25-to-50 cm initial randomization.

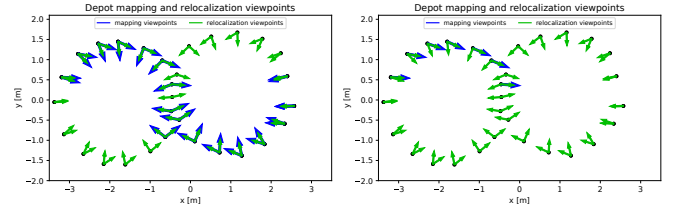|	Synthetic, random	Real, relocalize	Synthetic, final
Flagstone			
CFA6			
CFA2			
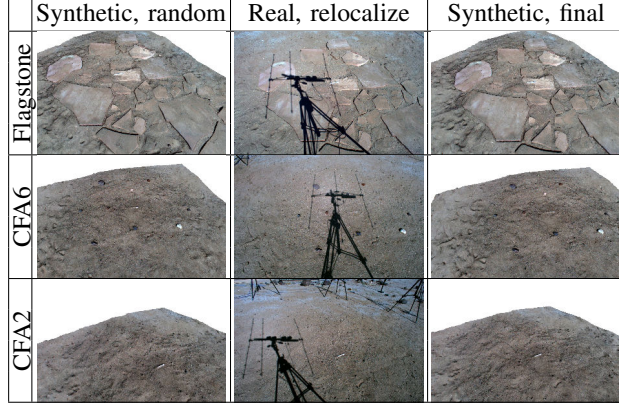
(b) Afternoon relocalization vs morning map with 50 cm randomization. Note the real image strong shadows, which could occur on Mars (rover shadow).

Fig. 5.   Results on alternative relocalization configurations.

(a) Alternative imaging trajectories: wave (left) and forward (right).

(b) Forward trajectory with step size 6 (1.9 m step). Red arrows indicate mapping images that were not successfully aligned with the others (blue arrows) during mesh reconstruction. Left: Flagstone and CFA6, right: CFA2.

Flagstone - 6 h	CFA6 - 6 h	CFA2 - 6 h
100.0 %, 104.2 mm	100.0 %, 121.6 mm	77.6 %, 84.4 mm

(c) Transformation estimation success rate and average error on forward trajectory with step size 6. All maps are effectively built from a single stereo pair. Failure cases for CFA2 can be attributed to half the mesh missing.

	Trajectory	Wave				Forward			
	Step size [m]	0.4	0.8	1.5	2.0	0.4	0.8	1.2	1.9
F.stone	Range [m]	3.0	3.0	3.0	6.0	3.0	3.0	5.0	6.0
F.stone	Success [%]	100.0	100.0	100.0	86.2	100.0	100.0	100.0	100.0
F.stone	Error [mm]	56.9	50.8	65.9	248.2	80.4	84.9	116.7	104.2
CFA6	Range [m]	3.0	3.0	3.5	6.0	4.0	4.0	4.0	6.0
CFA6	Success [%]	100.0	100.0	100.0	100.0	100.0	100.0	98.3	100.0
CFA6	Error [mm]	85.2	83.9	122.4	119.2	75.6	110.9	139.3	121.6
CFA2	Range [m]	2.5	3.0	4.5	6.0	3.0	3.0	5.0	3.5
CFA2	Success [%]	100.0	100.0	100.0	77.6	100.0	100.0	100.0	77.6
CFA2	Error [mm]	60.9	77.3	189.1	312.8	74.3	75.1	94.4	84.4

(d) VTSM results on wave and forward imaging trajectories. We sub-sample each with four step sizes, skipping viewpoints in between. "Range" denotes the maximum distance between the resulting mapping and relocalization images.

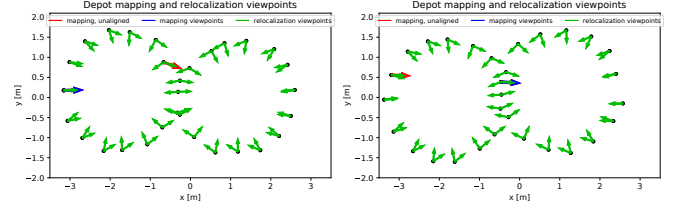Fig. 6.   Results on different mapping step sizes and relocalization ranges.

do so in two ways. First, we consider two alternative depot imaging trajectories for Perseverance (see Fig. 6a): a *wave* trajectory where the rover navigates between two consecutive sample tubes imaging both, and a *forward* trajectory where the rover only looks ahead while driving. This lets us evaluate relocalization from viewpoints further from the mapping set. Second, we sub-sample each depot imaging set by only using viewpoints every x m, which lets us evaluate Perseverance imaging density requirements to enable SFR relocalization. We observe the following. First, our meshing software starts failing to align input images when the imaging step size exceeds 1 m due to insufficient overlap between viewpoints. We depict in Fig. 6b the forward path with step size 1.9 m, leaving only the first and last viewpoints of the trajectory as mapping images, which the software could not be align. We thus obtained a mesh from only the first viewpoint's stereo pair for Flagstone and CFA6, and from the final viewpoint for CFA2. The latter being in the middle of the depot resulted in only half of it being 3D-modeled. While this issue may be mitigated using further image alignment techniques, we are also interested in evaluating our method against less accurate meshes. Fig. 6c illustrates that VTSM relocalization is still successful up to the 6 m away from the nearest mapping image on Flagstone and CFA6, with errors mostly below 10 cm up to 3 m away. Failure cases on CFA2 appear, expectedly, when attempting to relocalize images of the depot half that could not be mapped. We report results on all configurations in Fig. 6d, observing that relocalization accuracy drops after 1 m depot imaging step in most cases.
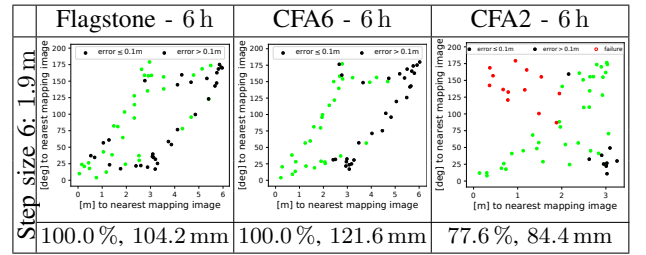
VI. DISCUSSION

The problem of retrieving sample tubes on Mars years after they have been dropped by another rover is a difficult task due to the unknown of how depots may change over time. In this paper, we presented a complete relocalization pipeline matching partial renderings of a depot map over multiple virtual viewpoints to real images. Our approach estimated poses with 100 % success rate across all terrains and lighting differences when local feature matching would completely fail, with average error below 10 cm in both nominal and extended conditions. Further analysis permitted by our large-scale dataset showed that our method maintained similar performance for at least 3 m away from poses imaged by Perseverance, with relocalization failing only when depot mapping itself fails. Based on results across different experimental conditions, we recommend that depots are

constructed on fractured bedrock on Mars akin to flagstone on Earth, and imaged by Perseverance from viewpoints no further than $1\,\mathrm{m}$ apart. Sparse, or even no rocks can be considered if the effects of Martian wind can be deemed negligible from surface or orbital imagery.

Our work lends itself to multiple development opportunities. First, the viewpoint randomization process could appropriately be implemented as a particle filter that updates search parameters rather than following a fixed schedule. We expect this would improve relocalization accuracy and computational efficiency towards being used onboard the rover. Machine learning techniques could also be used to identify salient points on the map that are most likely to yield successful matches, as done manually in [19]. As a longer-term development, the synthetic matching pipeline could be applied to image modalities other than direct pixel intensity and its derivatives. We could, for example, convert both synthetic and real images to lighting-invariant representations using recent neural-network-based techniques [41], or synthetically relight the depot map on the fly to reproduce Mars lighting conditions at SFR relocalization time.

References

[1] K. S. Ali, C. A. Vanelli, J. J. Biesiadecki, M. W. Maimone, Y. Cheng, A. M. San Martin, and J. W. Alexander. Attitude and position estimation on the mars exploration rovers. In *IEEE Int. Conf. Syst. Man Cybern.*, 2005.

[2] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool. Night-to-day image translation for retrieval-based localization. In *IEEE Int. Conf. Robot. Autom.*, 2019.

[3] J. N. Bakambu, P. Allard, and E. Dupuis. 3d terrain modeling for rover localization and navigation. In *IEEE Can. Conf. Comput. Robot Vision*, 2006.

[4] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1992.

[5] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Trans. Rob.*, 2016.

[6] P. J. Carle, P. T. Furgale, and T. D. Barfoot. Long-range rover localization by matching lidar scans to orbital elevation maps. *J. Field Rob.*, 2010.

[7] W. Churchill and P. Newman. Experience-based navigation for long-term localisation. *Int. J. Rob. Res.*, 2013.

[8] L. Clement, M. Gridseth, J. Tomasi, and J. Kelly. Learning matchable image transformations for long-term metric visual localization. *IEEE Rob. Autom. Lett.*, 2020.

[9] P. Corke, R. Paul, W. Churchill, and P. Newman. Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation. In *IEEE/RSJ Int. Conf. Intell. Rob. Syst.*, 2013.

[10] F. Cozman, E. Krotkov, and C. Guestrin. Outdoor visual position estimation for planetary rovers. *Auton. Robots*, 2000.

[11] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 1981.

[12] A. Geiger, J. Ziegler, and C. Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *IEEE Intell. Veh. Symp.*, 2011.

[13] M. Geppert, P. Liu, Z. Cui, M. Pollefeys, and T. Sattler. Efficient 2d-3d matching for multi-camera visual localization. In *IEEE Int. Conf. Robot. Autom.*, 2019.

[14] R. Giubilato, M. Vayugundla, M. J. Schuster, W. Stürzl, A. Wedler, R. Triebel, and S. Debei. Relocalization with submaps: Multi-session mapping for planetary rovers equipped with stereo cameras. *IEEE Rob. Autom. Lett.*, 2020.

[15] M. Golombek and D. Rapp. Size-frequency distributions of rocks on mars and earth analog sites: Implications for future landed missions. *J. Geophys. Res.: Planets*, 1997.

[16] M. Gridseth and T. Barfoot. Towards direct localization for visual teach and repeat. In *IEEE Conf. Comput. Robot Vision*, 2019.

[17] A. Lambert, P. Furgale, T. D. Barfoot, and J. Enright. Visual odometry aided by a sun sensor and inclinometer. In *IEEE Aerosp. Conf.*, 2011.

[18] R. Li, K. Di, A. B. Howard, L. Matthies, J. Wang, and S. Agarwal. Rock modeling and matching for autonomous long-range mars rover localization. *J. Field Rob.*, 2007.

[19] D. A. Lorenz, R. Olds, A. May, C. Mario, M. E. Perry, E. E. Palmer, and M. Daly. Lessons learned from osiris-rex autonomous navigation using natural feature tracking. In *IEEE Aerosp. Conf.*, 2017.

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 2004.

[21] M. Maimone, Y. Cheng, and L. Matthies. Two years of visual odometry on the mars exploration rovers. *J. Field Rob.*, 2007.

[22] J. Maki, C. McKinney, R. Willson, R. Sellar, D. Copley-Woods, M. Valvo, T. Goodsall, J. McGuire, K. Singh, T. Litwin, et al. The mars 2020 rover engineering cameras. In *Lunar Planet. Sci. Conf.*, 2020.

[23] C. McManus, W. Churchill, W. Maddern, A. D. Stewart, and P. Newman. Shady dealings: Robust, long-term visual localisation using illumination invariance. In *IEEE Int. Conf. Robot. Autom.*, 2014.

[24] C. McManus, B. Upcroft, and P. Newman. Scene signatures: Localised and point-less features for localisation. In *Rob.: Sci. Syst.*, 2014.

[25] B. K. Muirhead, A. K. Nicholas, J. Umland, O. Sutherland, and S. Vijendran. Mars sample return campaign concept status. *Acta Astronaut.*, 2020.

[26] K. Ok, W. N. Greene, and N. Roy. Simultaneous tracking and rendering: Real-time monocular localization for mavs. In *IEEE Int. Conf. Robot. Autom.*, 2016.

[27] S. Park, T. Schöps, and M. Pollefeys. Illumination change robustness in direct visual slam. In *IEEE Int. Conf. Robot. Autom.*, 2017.

[28] G. Pascoe, W. Maddern, A. D. Stewart, and P. Newman. Farlap: Fast robust localisation using appearance priors. In *IEEE Int. Conf. Robot. Autom.*, 2015.

[29] M. Paton, K. MacTavish, C. J. Ostafew, and T. D. Barfoot. It's not easy seeing green: Lighting-resistant stereo visual teach & repeat using color-constant images. In *IEEE Int. Conf. Robot. Autom.*, 2015.

[30] M. Paton, K. MacTavish, M. Warren, and T. D. Barfoot. Bridging the appearance gap: Multi-experience localization for long-term visual teach and repeat. In *IEEE/RSJ Int. Conf. Intell. Rob. Syst.* IEEE, 2016.

[31] T.-H. Pham, W. Seto, S. Daftry, A. Brinkman, J. Mayo, Y. Cheng, C. Padgett, E. Kulczycki, and R. Detry. Rover localization for tube pickup: Dataset, methods and validation for mars sample return planning. In *IEEE Aerosp. Conf.*, 2020.

[32] N. Piasco, D. Sidibé, V. Gouet-Brunet, and C. Demonceaux. Learning scene geometry for visual localization in challenging conditions. In *IEEE Int. Conf. Robot. Autom.*, 2019.

[33] R. B. Rusu and S. Cousins. 3d is here: Point cloud library (pcl). In *IEEE Int. Conf. Robot. Autom.*, 2011.

[34] T. Sattler, B. Leibe, and L. Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.

[35] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *IEEE Conf. Comput. Vision Pattern Recognit.*, 2018.

[36] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *IEEE Conf. Comput. Vision Pattern Recognit.*, 2016.

[37] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 118–126, 2015.

[38] Y. Tao, J.-P. Muller, and W. Poole. Automated localisation of mars rovers using co-registered hirise-ctx-hrsc orthorectified images and wide baseline navcam orthorectified mosaics. *Icarus*, 2016.

[39] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *IEEE Conf. Comput. Vision Pattern Recognit.*, 2015.

[40] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1991.

[41] L. von Stumberg, P. Wenzel, Q. Khan, and D. Cremers. Gn-net: The gauss-newton loss for multi-weather relocalization. *IEEE Rob. Autom. Lett.*, 2020.