2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

BEVIO: Efficient Bird's-Eye-View based Sparse-Update Visual-Inertial Odometry for Lunar Day-Night Navigation

Mohit Singh[1,2,*], Shehryar Khattak[1], Ashish Goel[1], Michael Paton[1], Kostas Alexis[2], and Issa A. Nesnas[1]

***Abstract*— Visual–Inertial Odometry (VIO) provides smooth, high-rate state estimates and has been widely used for robotic navigation in both terrestrial and planetary applications. However, its performance is typically dependent on the frequency of visual updates, which is a challenge for planetary rovers operating under extreme resource constraints and low frame rates. This work investigates enabling reliable VIO with very sparse visual updates for lunar rover applications, addressing both day and night-time operations where feature associations become especially difficult under self-illumination conditions. We propose a Bird's Eye View (BEV)–based image matching scheme that remains robust to larger inter-frame motions and more reliable feature matching despite significant visual appearance changes. We extensively evaluate our proposed approach, BEVIO, through high-fidelity photorealistic lunar and real-time robotic experiments conducted using a half-scale lunar rover, in a long-term day–night deployment at Plaster City, CA, USA. The results demonstrate that our method enables reliable day and nighttime self-illuminated traverses at visual update rates as low as 0.25 Hz, underscoring its suitability for navigation on power- and compute-limited lunar rovers.**

I. INTRODUCTION

Over the past five decades, robotic space explorers have enabled humankind to develop a rich understanding of our solar system. Future mission concepts continue to push the boundaries of exploration. One such concept is the Endurance mission [1, 2], which plans to explore the South Pole-Aitken (SPA) Basin of Earth's Moon and traverse over 2000 km across four years. The Radioisotope Thermoelectric Generator (RTG) powered version of the rover targets a maximum speed of 1 km/hour ($\approx 0.28\,\mathrm{m/s}$) and an ability to traverse during both the lunar day and night. The solar-powered version can only traverse during the day due to a lack of energy for anything but surviving the extreme cold of the lunar nights. This mobility represents a substantial increase over past missions, such as Mars 2020 Perseverance, which traversed at a top speed of $0.042\,\mathrm{m/s}$. Endurance is being designed to operate at higher traversal speeds with a higher degree of autonomy. Additional challenges arise in dark regions during nighttime missions or during daytime in dark shadows (e.g., from craters), where the rover needs to self-illuminate the environment in its proximity for sustained perception. Such a rover is limited by both onboard computing and energy; to this end, it is important to develop and study methods that can enable long-term, reliable performance while minimizing the computing and energy resources.

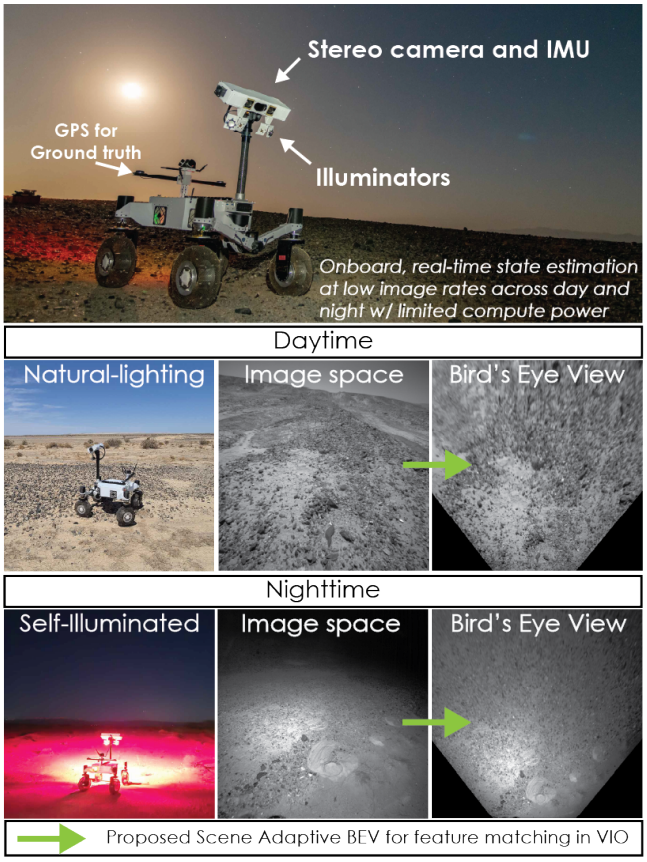


Fig. 1: Half-scale ERNEST rover, deployed in planetary-like desert terrain, navigating with onboard real-time state estimation. The challenges include fast traversal in day and nighttime operation, using constrained compute resources. Proposed solution employs a scene adaptive Bird's Eye View perspective to match features, enabling Visual-Inertial Odometry at lower image frame rates than the baseline approach, thus reducing the computational requirements.

We are investigating state estimation for the Endurance by utilizing a half-scaled rover ERNEST (Exploration Rover for Navigating Extreme Terrain) [2], illustrated in Fig. 1. It fuses visual and inertial sensors for state estimation. Hence, it is important to study the characteristics of Visual-Inertial Odometry (VIO) with a focus on low frame rates, thus reducing the need for frequent image capturing and onboard LED strobing, minimizing computational and energy requirements.

*Corresponding author: mohit.singh@ntnu.no
[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA.
[2]Autonomous Robots Lab at the Norwegian University of Science and Technology, Trondheim, Norway
The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004).

We propose a scene-adaptive, Bird's-Eye View (BEV)-perspective, feature-matching framework to increase inlier feature matches across a larger baseline between subsequent camera images. Our proposed method integrates into the feature-matching frontend module of the xVIO [3]. We conduct detailed experiments in simulation and in the field using a half-scale rover.

We find that by employing BEV, reliable VIO can be achieved at frame rates as low as $0.25\,\mathrm{Hz}$. In spatial terms, this corresponds to $\approx 1\,\mathrm{m}$ distance traveled between two consecutive images. To emphasize the challenge in the problem we tackle, it is important to understand that in missions like the Mars 2020 Perseverance, where the operation is limited to daytime, the image updates were performed every $1\,\mathrm{m}$ due to computational limitations. For fair comparison, the scale-compensated baseline between images for the Perseverance rover is approximately $0.5\,\mathrm{m}$, accounting for the half-scale prototype ERNEST, with the camera mounted at roughly half the height of the Perseverance and Endurance rovers. Our findings indicate a promising two-fold improvement in the baseline between images over the scale-compensated baseline for the Perseverance rover, while operating across day and night-time, with up to six times higher average speeds.

Further, to compare against earth-based flying systems with VIO, consider an aerial robot with a camera operating at $20\,\mathrm{Hz}$ and flying at $20\,\mathrm{m/s}$, resulting in an image-to-image baseline of $1\,\mathrm{m}$. In this case, the flying system benefits from Inertial Measurement Unit (IMU) biases evolving unconstrained for only $0.05\,\mathrm{s}$. By contrast, in the problem we address with rover traversing at average speeds of $0.25\,\mathrm{m/s}$, the IMU biases evolve unconstrained for approximately $4.0\,\mathrm{s}$, which is significantly longer.

The remainder of this paper is organized as follows: Section II reviews related work, Section III presents our scene-adaptive BEV method, Section IV describes VIO integration, and Section V describes the simulation and field experiments, while Section VI draws conclusions.

II. RELATED WORK

Vision-based state estimation, and VIO in particular, have evolved significantly over the past two decades. Seminal contributions to VIO can be broadly categorized into loosely-coupled methods (fusing independent visual and inertial estimates) [4], tightly-coupled methods (joint optimization of both modalities) [5], filter-based frameworks (e.g., Extended Kalman Filters) [6], and optimization-based systems (e.g., sliding-window bundle adjustment) [7]. On earth-based systems, often the choice of VIO can be across this broad spectrum, while for space applications, the early solutions have been loosely-coupled Visual Odometry (VO) with IMU or wheel odometry, while recent solutions are also based on bundle-adjustment, and tightly-coupled filter-based methods.

For earth-based systems, the Multi-State Constraint Kalman Filter (MSCKF) [6] is a foundational filter-based method that encodes feature constraints in the null space of the feature Jacobian, decoupling updates from feature count and enabling accurate, efficient VIO. Building on this, OpenVINS [8] further refines the MSCKF framework. In the optimization family, OKVIS [5] introduced keyframe-based nonlinear optimization with marginalization, improving robustness. Meanwhile, ROVIO [9] employs an iterated EKF with a semi-direct, patch-based photometric error and a robocentric formulation, improving initialization robustness and performance in textureless environments.

These methods have been extensively tested on Earth under various conditions, including dynamic motions and varying illumination. VIO is applied across a wide range of domains, including robotics, where it has been deployed on aerial, ground, and underwater robots [10]. Naturally, for space applications, the vision-based state-estimation method shares the same fundamental components, while mission-specific requirements like computational efficiency and energy constraints further tailor the solutions.

Among one of the early works, the Mars Exploration Rovers (MER) [11] use state estimation from wheel odometry and frame-to-frame feature tracking based VO. They detect Harris corners and perform stereo matching to obtain the three-dimensional location and corresponding covariances. The rover then moves, and the past feature locations are re-projected into the image frame. A correlation-based search is used to match the features. The estimation is performed in two steps: first, with a least-square fit with RANSAC, followed by Maximum likelihood estimation. Towards lunar applications, some works proposed a downward-facing stereo camera with self-illumination [12] where they use a downwards facing stereo camera pair with self-illumination, tracking features to perform state-estimation. While other works proposed an image-enhancement method over a stereo visual odometry with Harris corner detection and feature tracking for motion estimation [13].

In more recent developments, the Mars Helicopter [14] used an Extended Kalman Filter (EKF)-based VIO. The full sensor stack includes a downwards-facing camera, a single-beam range-finder, and an IMU. The method uses FAST [15] features and KLT [16] tracking with gyro-derotation, followed by RANSAC outlier rejection. The Yutu-2 lunar rover state estimation is based on SURF [17] features matching and manually selected features for stereo bundle adjustment based on IMU-assisted VO. A recent work, LuVo [18] utilized a BEV perspective for enhancing feature matching across large translations between images. Their solution proposes using learning-based feature matching with a remote, earth-based server for Graphics Processing Unit (GPU) computations, whereas the proposed method runs onboard the rover. Further, they obtain the BEV perspective by assuming flat terrain and using rover attitude to compute the warping.

Most of the prior methods targeting space deployment have been developed for slow-moving rovers (e.g. $0.04\,\mathrm{m/s}$ for the Mars 2020 Perseverance rover) or flying robots, which can be fast but function with high overlap between subsequent images due to higher altitude from the ground. To this end, we tackle the problem of a fast-moving planetary rover ($\approx 0.25\,\mathrm{cm/s}$) on challenging terrain for long-

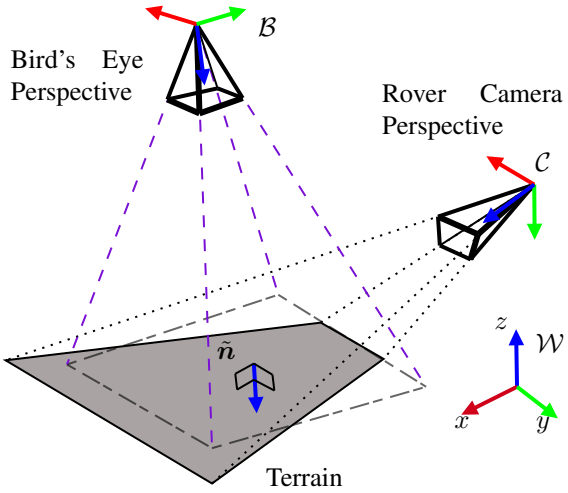


Fig. 2: Illustration of the perspectives of the Image Space and Bird's Eye View, alongside the coordinate frames used in the proposed formulation.

term traversal across day and night with an onboard real-time solution. Our approach uses real-time three-dimensional terrain information from dense stereo to identify the plane normal for the BEV projection. While feature matching in BEV enables us to reduce the image Frames Per Second (FPS), our VIO integrated solution outputs high-rate state estimates, necessary for fast lunar rover navigation.

III. METHOD

In this section, we describe the insight behind BEV and how we obtain the scene-adaptive BEV perspective, given the image from the rover's navigation camera and the stereo depth point cloud. We also discuss feature detection and matching after obtaining the images in the BEV perspective. The derivations in this section stem from the two-dimensional projective geometry described in [19].

The key insight is that in BEV perspective, the scene appearance remains consistent between frames; features primarily translate laterally in the image plane rather than undergoing large appearance changes. This greatly reduces descriptor mismatch compared to the Image Space (IS), where inter-frame appearance change is significantly larger.

A. Notations

- $\mathcal{C}$: Rover left stereo navigation camera coordinate frame, with the z-axis along optical axis.
- $\mathcal{I}$: Rover IMU coordinate frame.
- $\mathcal{B}$: Virtual BEV camera coordinate frame, with the z-axis perpendicular to the ground plane.
- $\mathcal{W}$: World coordinate frame, with the z-axis parallel to the gravity vector and pointing upwards.
- (H, W): Height and width of the camera image.
- $(H_{\text{BEV}}, W_{\text{BEV}})$: Height and width of the BEV image.
- Vectors are denoted by boldface symbols.

B. Bird's-Eye-View transformation

In our formulation, the rover navigation camera observes the scene (Fig. 2). We assume that the scene is approximately flat and a plane (parametrized by the normal unit vector $\tilde{\boldsymbol{n}}$) can be fit to this terrain. We compute a homography transformation M, defined as:

$$M = K R_{\mathcal{B},\mathcal{C}} K^{-1}, \tag{1}$$

where K is the camera intrinsic matrix and $R_{\mathcal{B},\mathcal{C}} \in SO(3)$ denotes the rotation matrix from $\mathcal{C}$ to $\mathcal{B}$. The intrinsics K are obtained through camera calibration, and we assume distortion-free, rectified input images.

The rotation matrix $R_{\mathcal{B},\mathcal{C}}$ is derived using the plane normal $\tilde{\boldsymbol{n}}$ and the axes of $\mathcal{C}$. Let $\tilde{\boldsymbol{x}}_{\mathcal{C}}$ and $\tilde{\boldsymbol{z}}_{\mathcal{C}}$ denote the x- and z-axes of frame $\mathcal{C}$, then $R_{\mathcal{B},\mathcal{C}}$ can be given by:

$$R_{\mathcal{B},\mathcal{C}} = \left[\left(\tilde{\boldsymbol{n}} \times \tilde{\boldsymbol{z}}_{\mathcal{C}}\right) \times \tilde{\boldsymbol{x}}_{\mathcal{C}} \quad \tilde{\boldsymbol{n}} \times \tilde{\boldsymbol{z}}_{\mathcal{C}} \quad \tilde{\boldsymbol{n}} \right]. \tag{2}$$

To control the resolution of the BEV image, we apply a scaling operation with matrix

$$S = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where s is the scaling factor. Furthermore, a $2D$ translation is applied in image coordinates using a translation matrix T to align the midpoint of the lower edge of the input image $(H, W/2)$ with the bottom of the BEV frame $(H_{BEV}, W_{BEV}/2)$. Scaling is a tunable heuristic; it ensures fine resolution for nearby points. In our case, we set $s = 0.8$. Translation minimizes unmapped regions manifesting as pixels with 0 intensity. The complete homography matrix M^* is then given by

$$M^* = TSM. \tag{4}$$

C. Plane normal estimation

We estimate the ground plane normal by fitting a plane to the 3D points obtained from the stereo depth image $D \in \mathbb{R}^{H \times W}$ using least squares with RANSAC. To reduce computation, we downsample the depth image to $(H/g, W/g)$ points, where g is the downsampling factor. The selected points are the centers of non-overlapping patches of size (g, g). Using the pixel coordinates, depth values, and K, we obtain the corresponding 3D points in $\mathcal{C}$, denoted as $\mathbf{P}_{\mathcal{C}}$, and estimate the plane normal $\tilde{\boldsymbol{n}}_{\mathcal{C}}$.

We transform $\tilde{\boldsymbol{n}}_{\mathcal{C}}$ into the world frame:

$$\tilde{\boldsymbol{n}}_{\mathcal{W}} = R_{\mathcal{W},\mathcal{C}}\, \tilde{\boldsymbol{n}}_{\mathcal{C}}, \tag{5}$$

where $R_{\mathcal{W},\mathcal{C}}$ is obtained from the rover's orientation estimate $\mathbf{q}_{\mathcal{W},\mathcal{C}}$ provided by the state-estimation filter (described in Section IV). To smoothen the computed normal in the world frame, we maintain a sliding window of normals $\mathbb{N} = \{\tilde{\boldsymbol{n}}_{\mathcal{W},i} \mid t - m \leq i \leq t\}$ and compute their moving average $\tilde{\boldsymbol{n}}^*_{\mathcal{W}}$. The averaged normal is then transformed back into $\mathcal{C}$ as follows:

$$\tilde{\boldsymbol{n}}^*_{\mathcal{C}} = R^T_{\mathcal{W},\mathcal{C}} \tilde{\boldsymbol{n}}^*_{\mathcal{W}}. \tag{6}$$

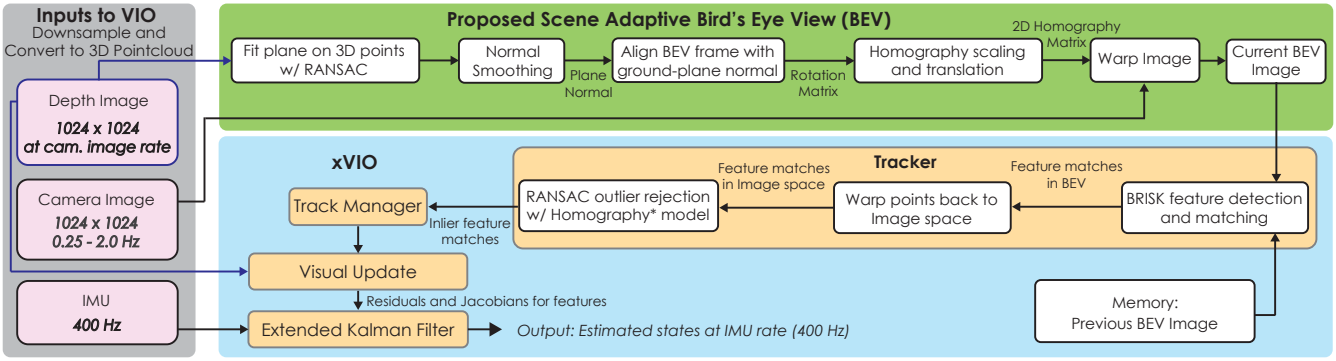


Fig. 3: Overview for the proposed method and its integration into VIO. The blocks in white color illustrate the parts of the proposed approach, and blocks in yellow color illustrate, modules of the baseline method xVIO[3].

D. Feature matching in BEV

Given input images, I_{t-1} and I_t (each of dimensions $H \times W$) from the navigation camera at times $t-1$ and t respectively. The BEV images are obtained by warping the camera images using the homography at the corresponding times:

$$I_{t,\text{BEV}} = \phi\left(I_t, M_t^*\right), \tag{7a}$$

$$I_{t-1,\text{BEV}} = \phi\left(I_{t-1}, M_{t-1}^*\right), \tag{7b}$$

where ϕ is the warping function defined from a pixel coordinate $\boldsymbol{p}_{\text{IS}}$ in IS image to pixel coordinate in $\boldsymbol{p}_{\text{BEV}}$ in BEV image and is given by

$$\boldsymbol{p}_{\text{BEV}} = M\ \boldsymbol{p}_{\text{IS}} \tag{8}$$

for a given homography matrix M. The unmapped regions are assigned zero intensity. The warped BEV images are of dimensions of $(H_{\text{BEV}} \times W_{\text{BEV}})$. To avoid spurious detections, feature extraction is masked to exclude boundaries between valid and unmapped areas.

In our formulation, we employ BRISK [20] features and descriptors. This choice is motivated by its robustness to illumination variations across images captured from large baselines, especially compared to optical-flow based feature tracking using KLT[16] (used in xVIO [3]), which assumes constant intensity. KLT is suitable for high FPS image streams. Additionally, BRISK was chosen due to the computational efficiency of its binary descriptors over floating-point descriptors [17, 21]. We detect BRISK features on $I_{t,\text{BEV}}$ and $I_{t-1,\text{BEV}}$ and match them with normalized Hamming distance as described in [20].

The matched features contain outliers, and to reject these, we use a homography model with RANSAC. During this step, the estimated homography is also decomposed to recover the supporting plane normal, which is compared against the ideal flat-terrain normal for consistency verification. This yields a robust set of matched visual features. Although the feature detection and matching operations are performed in the BEV space, we transform the points back to IS using the inverse of the homography matrices corresponding to each timestep. This enables us to detect and match features in BEV, without changing the downstream formulation of VIO. Fig. 3 showcases the overall framework.

IV. INTEGRATION INTO VISUAL-INERTIAL ODOMETRY

We use xVIO [3] as the underlying VIO framework. The input sensor modalities include the camera images, the IMU measurements, and the range measurements. In our case, the image is from the left camera of the stereo pair, and the range measurements come from the stereo matching depth image. The IMU measurements are from a MEMS IMU mounted rigidly with respect to the cameras. We use the same notations for coordinate frame from Section III-A. The state vector in xVIO is defined as

$$\mathbf{s} = \begin{bmatrix}\mathbf{s}_I^T & \mathbf{s}_V^T\end{bmatrix}^T, \tag{9}$$

where $\mathbf{s}_I$ denotes the inertial state associated with the IMU, and $\mathbf{s}_V$ represents the visual state corresponding to tracked features.

The inertial state $\mathbf{s}_I \in \mathbb{R}^{16}$ comprises the position, velocity, orientation, gyroscope biases, and accelerometer biases of the IMU, and can be written as

$$\mathbf{s}_I = \begin{bmatrix}\mathbf{p}^T & \mathbf{v}^T & \mathbf{q}^T & \mathbf{b}_g^T & \mathbf{b}_a^T\end{bmatrix}^T. \tag{10}$$

The error-state is partitioned in the same manner (inertial and visual components), which enables EKF linearization and covariance propagation.

Internally, xVIO is structured into three main modules: *tracker*, *track manager*, and *visual update*.

- **Tracker:** Implements frame-to-frame feature tracking, serving as the visual front-end. Our proposed BEV-based feature matching is integrated within this module.
- **Track manager:** Maintain keyframes, and triggers their creation based on the number of tracked features output by the tracker.
- **Visual update:** Handles computation of Jacobians and residuals for EKF update associated with features.

Given the orientation of the IMU in the world frame, $\mathbf{q}$, and the camera-to-IMU transformation, $\mathbf{q}_{\mathcal{C},\mathcal{I}}$, we can compute the orientation of the camera in the world frame, $\mathbf{q}_{\mathcal{W},\mathcal{C}}$, which is then used for plane normal estimation (see Section III-C). Specifically, we use the predicted orientation, $\hat{\mathbf{q}}$, obtained from the filter prediction step, as it is required prior to the image measurement update. For further details on xVIO, the reader is referred to [3].

V. EXPERIMENTS

This section presents a comprehensive evaluation of our proposed method against the baseline VIO approach, using both qualitative and quantitative analyses. Here, baseline VIO refers to a modified version of xVIO [3] with a BRISK feature matching front-end instead of KLT [16], and a stereo depth image for range measurements instead of a single laser range-finder. First, we assess the performance gains achieved by employing the BEV perspective compared to the IS used in the baseline for feature matching. Second, we assess the possible reduction in image FPS (i.e., to enable feature matching at larger baselines) achieved by utilizing BEV in the feature matching front-end of VIO. Prior to the evaluation, we present the experimental setup, commencing with the rover platform employed in this study.

A. ERNEST rover prototype

Our studies were conducted using data collected from simulated and real-world environments, with a half-scale Endurance rover prototype named "ERNEST" serving as the robot platform. ERNEST is a four-wheeled surface mobility rover with a suspension that can be configured to be actively controlled or passive. The active suspension allows the rover to navigate more extreme terrains and slopes up to $30°$. For the Endurance mission prototype testing, the rover was configured to use its passive suspension given the anticipated terrain. In our experiments, the rover maintains an average velocity of $0.24\,\mathrm{m/s}$.

The state-estimation sensor suite includes a synchronized global shutter stereo camera pair and a VectorNav VN100 MEMS IMU, both mounted on a fixed mast at the front of the rover. The stereo camera pair is pitched down by $30°$, and the Field of View (FOV) for each camera is $90°\times90°$. The rover's autonomy stack runs on the onboard computer, with high-level commands provided from a ground control station. The camera is triggered at $2\,\mathrm{Hz}$, while the IMU is sampled at $400\,\mathrm{Hz}$. To emulate a low frame rate image stream, we developed a pre-processing script that skips a specified number of images to achieve the desired image FPS. The cameras were calibrated using the Kalibr toolbox. The cameras outputs grayscale images of dimension 1024×1024 pixels. Fig. 1 shows the ERNEST rover in the field during the data collection in daytime and nighttime scenes. For nighttime perception, the rover uses a set of LED illuminators positioned below the stereo cameras on the mast. The LEDs irradiate red light for their highest optical response on camera pixels.

B. Real-world experiment setup

To analyze the state estimation, we collected extended data in the Plaster City desert region in California, USA. The scenes in the data qualitatively mimic a planetary environment. The navigation system continuously generates hazard-free trajectories for the rover to follow. During the traversal, the rover covers paths of over hundreds of meters in both day and nighttime scenarios. Over its course, the rover traverses numerous regions, spanning sandy, rocky, and uneven terrain. Table I shows the lengths of the different trajectories used in this evaluation. Fig. 4 shows the images from the rover left camera from the real-world day and nighttime, and for the nighttime simulated environment. The trajectories corresponding to real-world data from nighttime are *Night-1* and *Night-2*, while for daytime scenes are *Day-1* and *Day-2*.

TABLE I: Trajectory lengths (m) for different runs.

Night-1	Night-2	Night-3	Day-1	Day-2
115 m	80 m	100 m	130 m	115 m

C. Lunar simulation setup

To conduct our simulation analysis, we used the DARTS[22] lunar simulator, which provides high-fidelity physics and image simulations for multiple lunar scenarios. This environment models both the complex terra-mechanics of traversing rocky terrains as well as the illumination effects from the sun and the rover's lights. The simulation uses an identical model of the Ernest rover, and the software stack is configured to function identically in both the simulator and real-world scenarios. The trajectory corresponding to simulated data is *Night-3*.

(a) Day (b) Night (c) Night simulated

Fig. 4: Images from the rover left camera from the real-world daytime, real-world nighttime, and simulated nighttime environment.

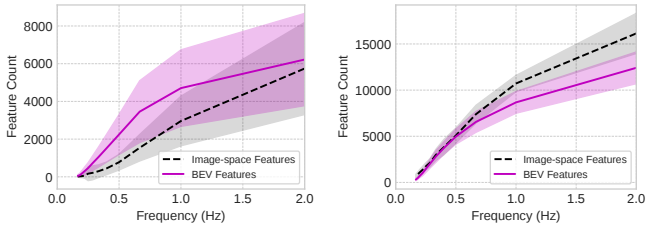
D. Effectiveness of Bird's Eye View

Prior to evaluating the VIO performance, it is crucial to analyze the impact of reduced image frame-rate on feature matching using controlled study, independent from VIO. To this end, we evaluate the number of detected features, the ratio of inliers to total matched features, and the histogram of feature locations across varying frame rates.

Given a sequential image dataset, we select an image at time t and match its features with another image captured at time $t+0.5k$ (factor of 0.5 is used, given the recorded images are at 2 FPS). By increasing the value of k, we effectively simulate a decreasing image frame rate. To obtain a statistical evaluation, $t \in \{t_1, t_2, \ldots, t_n\}$ with $t_1 < t_2 < \cdots < t_n$.

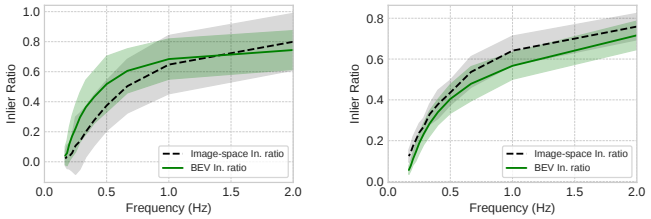
Fig. 5 illustrates how the feature count varies with the image FPS. Notably, as the frequency decreases, a greater number of features are successfully matched using BEV perspective compared to IS at nighttime. We observe a similar trend in the inlier ratio in Fig. 6, where the ratio is consistently higher for the BEV approach at lower frame rates than for the IS for nighttime. We also observe that for daytime scenes, both "number of feature matched" and

"inlier ratio" are better for baseline; however, these matched features are distributed in the top part of the image, i.e., the far-off region of the scene. Due to the large distance of feature points from the rover, these matches are less informative for the rover's ego-motion, as they have lower parallax compared to nearby features, and they exhibit higher stereo depth estimation noise. The latter rationale is further elaborated in the next sub-section.



(a) Nighttime real-world. (b) Daytime real-world.

Fig. 5: Comparison of the number of features matched in IS vs. BEV in nighttime and daytime scenarios in real-world environment. The error bounds indicates maximum and minimum values.



(a) Nighttime real-world. (b) Daytime real-world.

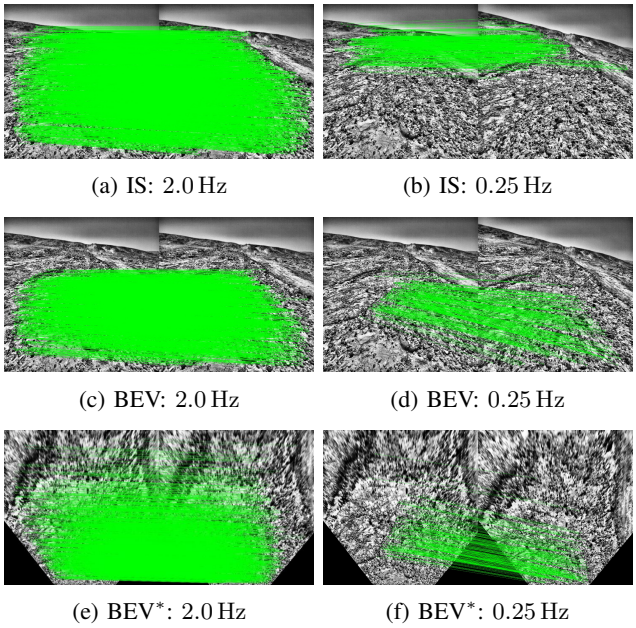
Fig. 6: Comparison of the inlier ratio in IS vs. BEV in nighttime and daytime scenarios in real-world environment. The error bounds indicate maximum and minimum values.

E. Distribution of the matched features

In our specific scenario, where the majority of the scene consists of a two-dimensional plane, a distribution in the lower region of the image indicates that the matched features are in the immediate vicinity of the rover. Conversely, a distribution in the upper region of the image denotes far-off features. Given this insight, we first showcase the effect of the matched features in BEV and IS in day and nighttime scenes at $2.0\,\mathrm{Hz}$ and $0.25\,\mathrm{Hz}$.

Fig. 7 shows the daytime case, where we observe that in the first row the matched features at $2.0\,\mathrm{Hz}$ have large coverage over the image, whereas at $0.25\,\mathrm{Hz}$ the number of features reduces and are distributed towards far-off points in the top region of the image. In the second row, we observe that BEV at $0.25\,\mathrm{Hz}$ matches features much closer to the rover. The last row shows the matched feature in the BEV perspective, visualized in BEV. Fig. 8 shows a similar comparison for night time, where at $0.25\,\mathrm{Hz}$ IS matches far lesser features compared to BEV for the same FPS. Lastly, we see that at nighttime, BEV matches a higher number of features, closer to the rover.

To assess the distributions of matched features at more intervals of time-periods between images, we overlay the smoothed histogram curves of the features along the Y axis in the image coordinates to compare the distributions. This can be seen in Fig. 9. For features from nearby scenes (i.e., points in the lower region of the image), the distribution flattens much earlier in IS than in BEV for both day and nighttime scenes. Specifically, in the case of IS, the distributions in the lower region of the image flattens above $2.0\,\mathrm{s}$ for both day and nighttime scenes, while BEV delays the flattening to over $4.0\,\mathrm{s}$. Showcasing its ability to match reliable features across larger baselines. For day-time (Fig. 9b), the IS distribution shifts to the upper region of the plot with increasing time-period, indicating that less informative far-off features are being matched. For nighttime scenes, (Fig. 9c) BEV matches substantially more features, than IS (Fig. 9d). Lastly, due to limited self-illumination around the rover, far-off features are not matched in nighttime. This demonstrates the effectiveness of the BEV approach for matching visual features in planetary scenarios during both day and night.



(a) IS: $2.0\,\mathrm{Hz}$ (b) IS: $0.25\,\mathrm{Hz}$

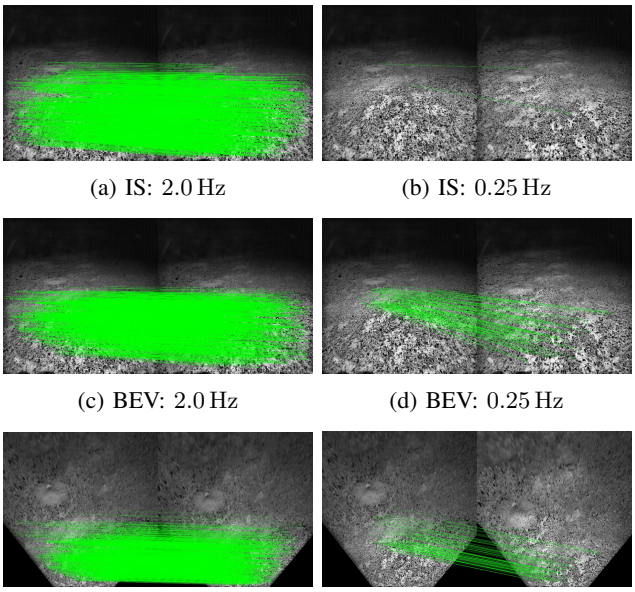
(c) BEV: $2.0\,\mathrm{Hz}$ (d) BEV: $0.25\,\mathrm{Hz}$

(e) BEV*: $2.0\,\mathrm{Hz}$ (f) BEV*: $0.25\,\mathrm{Hz}$

Fig. 7: Day-time feature matching comparison between feature matching in IS and BEV between images at $2.0\,\mathrm{Hz}$ and $0.25\,\mathrm{Hz}$. Here, BEV denotes feature matched in BEV perspective and visualized in IS for comparability, and BEV* denotes that the features are matched in BEV and visualized in BEV perspective.

F. Analysis of VIO

We present an ablation study on odometry performance across the five distinct trajectories, evaluating VIO performance with Relative Position Error (RPE) (given subsequent trajectory fragments lengths of $10\,\mathrm{m}$) against ground-truth from GPS. We progressively reduce the frame rate of images by skipping the images from the dataset in both day and nighttime conditions, using real and simulated data. The numerical analysis of the VIO for different scenarios is compiled in Table II, which shows the Root Mean Square Error (RMSE) of the RPE. A common failure mode is a decrease in the number of matched features matched to



(a) IS: 2.0 Hz (b) IS: 0.25 Hz

(c) BEV: 2.0 Hz (d) BEV: 0.25 Hz

(e) BEV*: 2.0 Hz (f) BEV*: 0.25 Hz

Fig. 8: Nighttime feature matching comparison between feature matching in IS and BEV between images at 2.0 Hz and 0.25 Hz. Here, BEV denotes feature matched in BEV perspective and visualized in IS for comparability, and BEV* denotes that the features are matched in BEV and visualized in BEV perspective.

TABLE II: Comparison of RMSE of RPE in meters for VIO with BEV and IS

Time-period (s)		0.5	1.0	2.0	2.5	3.0	3.5	4.0
Frequency (Hz)		2.0	1.0	0.5	0.4	0.33	0.28	0.25
Night-1 (real-world)	BEV	0.263	0.273	0.271	0.276	0.341	-	-
	IS	0.282	0.274	0.274	-	-	-	-
Night-2 (real-world)	BEV	0.129	0.127	0.123	0.122	0.126	0.126	0.213
	IS	0.124	0.124	0.126	0.126	-	-	-
Night-3 (simulated)	BEV	0.281	0.300	0.304	0.303	0.310	-	-
	IS	0.285	0.309	0.299	0.370	-	-	-
Day-1 (real-world)	BEV	0.282	0.268	0.252	0.273	0.266	-	-
	IS	0.322	0.261	0.257	0.240	-	-	-
Day-2 (real-world)	BEV	0.172	0.172	0.172	0.172	0.178	0.201	0.230
	IS	0.175	0.174	0.175	0.179	0.188	0.199	-

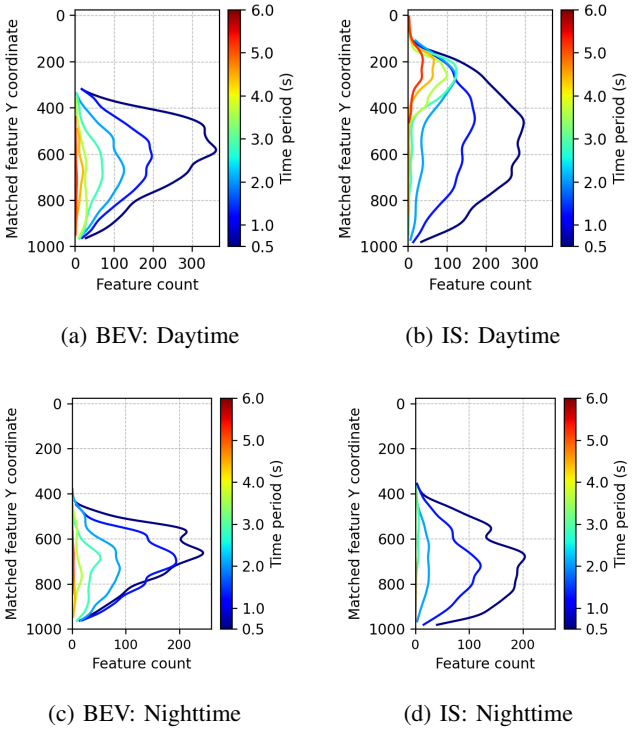
as low as 0, which causes large errors in the subsequent updates of the Kalman Filter, eventually causing the filter to diverge. We report the minimum frequency at which the VIO runs successfully in Table III. Based on this evaluation, we showcase the envelope of successful performance of the proposed BEV vs. the baseline IS in Fig. 10.

G. Scene adaptive BEV, Fixed BEV, and IS

We compare three cases of running VIO: (i) the proposed scene-adaptive homography estimation for BEV, (ii) a fixed ideal homography for BEV, and (iii) the baseline using IS. Figure 11 presents the qualitative results for each case.

The experiments were conducted at an image update frequency of 0.33 Hz, with the rover first traversing uneven terrain that induced large variations in attitude, followed by a flat region. The attitude variations are visible between the first and second rows of images, while the transition to flat terrain occurs between the second and third rows.

The proposed scene-adaptive BEV (first column) consistently maintains reliable feature matches across both conditions. In contrast, case (ii) struggles under large attitude changes, while case (iii) fails to maintain matches in the flat region due to significant perspective changes.



(a) BEV: Daytime (b) IS: Daytime

(c) BEV: Nighttime (d) IS: Nighttime

Fig. 9: Distribution of the feature locations along Y axis.

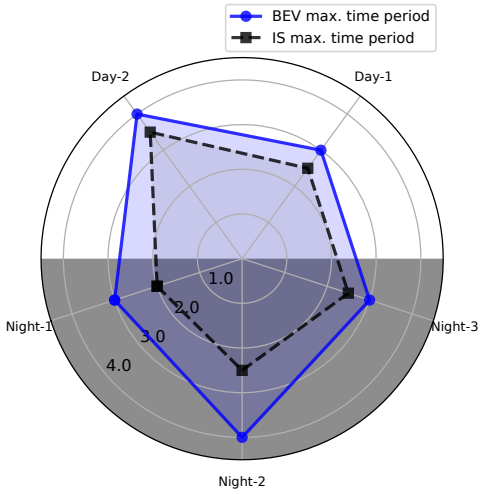


Fig. 10: Envelope of maximum time period between image measurements where the VIO method runs successfully

VI. CONCLUSIONS

In this work, we tackle the problem of visual-inertial state estimation for a lunar rover, in the context of the Endurance mission concept. Such a system faces challenges due to constrained computational and energy resources; it is tasked to navigate across day and nighttime scenes, operating under solar or self-illumination. We proposed a solution based on scene adaptive BEV that enables increasing the temporal

TABLE III: Minimum reliable image FPS for BEV and IS.

Condition	BEV (Hz)	IS (Hz)	(%)
Night-1	0.33	0.50	66.7
Night-2	0.25	0.40	62.5
Night-3	0.33	0.40	82.5
Day-1	0.33	0.40	82.5
Day-2	0.25	0.28	89.3
Overall Mean			**76.6**
Nighttime Mean			**70.3**
Daytime Mean			**85.9**

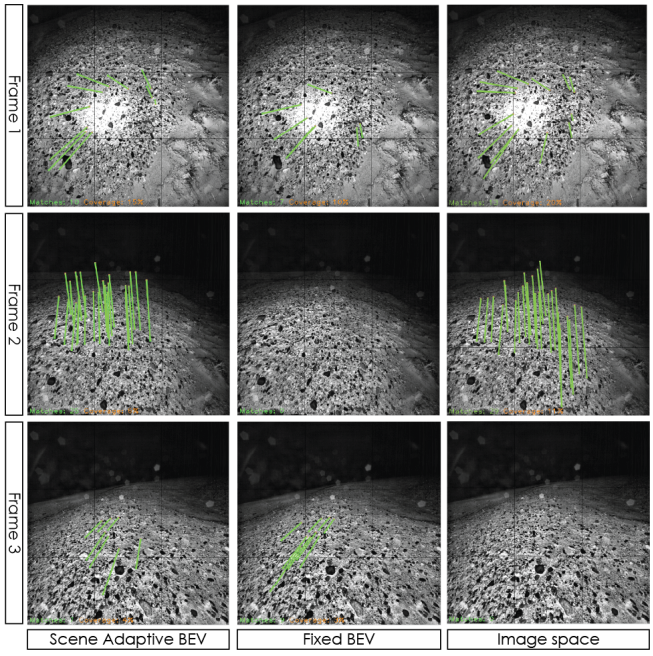


Fig. 11: Comparison of feature matching performance over multiple subsequent images. The columns represent different configurations of the VIO, while the time increases from top to bottom.

and spatial sparsity of the image updates (i.e., the image FPS). The perspective equalizing property of BEV explains its ability to maintain reliable matches at larger inter-frame baselines, enabling sparser visual updates. The rover navigation benefits from increased sparsity in the visual updates, both by lowering computational and energy requirements associated with computing, camera exposure, and strobing LEDs in dark scenes. We compare our approach against a baseline VIO method that uses feature matching in IS. In our quantitative evaluations from data collected across day and night on a half-scale lunar rover prototype in a planetary-like environment, we demonstrate that the proposed method enables VIO at a minimum update rate of $0.25\,\mathrm{Hz}$, which is 62.5% of the minimum update rate of the baseline method. Overall, we demonstrate that our approach operates in a larger envelope of sparse visual updates.

Future work may focus on developing a deeper understanding of the proposed BEV perspective and its influence on VIO. In particular, it is necessary to investigate whether feature matching in BEV alters the pixel measurement noise characteristics, and to determine if commonly assumed constant noise model remains valid. Furthermore, the impact of feature matching on objects that deviate from the estimated plane fit in the 3D point cloud warrants further investigation.

REFERENCES

[1] J. T. Keane, "Endurance: Lunar South Pole–Aitken Basin Traverse and Sample Return Rover."

[2] J. D. Baker, H. W. Stone, J. O. Elliott, J. T. Keane, R. P. Kornfeld, H. D. Nayar, and I. A. Nesnas, "The endurance mission progress," in *2025 IEEE Aerospace Conference*, 2025, pp. 1–14.

[3] J. Delaune, D. S. Bayard, and R. Brockers, "xvio: A range-visual-inertial odometry framework," *arXiv preprint arXiv:2010.06677*, 2020.

[4] S. Weiss, M. W. Achtelik, S. Lynen, M. Chli, and R. Siegwart, "Real-time onboard visual-inertial state estimation and self-calibration of mavs in unknown environments," in *2012 IEEE International Conference on Robotics and Automation*, 2012, pp. 957–964.

[5] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[6] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," in *Proceedings 2007 IEEE international conference on robotics and automation*. IEEE, 2007, pp. 3565–3572.

[7] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE transactions on robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[8] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "Openvins: A research platform for visual-inertial estimation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 4666–4672.

[9] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct ekf-based approach," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 298–304.

[10] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. Leonard, "Past, present, and future of simultaneous localization and mapping: Towards the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, p. 1309–1332, 2016.

[11] M. Maimone, Y. Cheng, and L. Matthies, "Two years of visual odometry on the mars exploration rovers," *Journal of Field Robotics*, vol. 24, no. 3, pp. 169–186, 2007.

[12] M. Wagner, D. Wettergreen, and P. Iles, "Visual odometry for the lunar analogue rover "artemis"," in *ISAIRAS*, 2012.

[13] L. Li, J. Lian, L. Guo, and R. Wang, "Visual odometry for planetary exploration rovers in sandy terrains," *International Journal of Advanced Robotic Systems*, vol. 10, no. 5, p. 234, 2013.

[14] 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*, 2019, p. 1411.

[15] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European conference on computer vision*. Springer, 2006, pp. 430–443.

[16] C. Tomasi and T. Kanade, "Detection and tracking of point," *Int J Comput Vis*, vol. 9, no. 137-154, p. 3, 1991.

[17] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[18] R. Soussan, J. McCaffery, S. McMichael, and M. Deans, "Luvo: Lunar visual odometry using homography-based image feature matching," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 13 428–13 435.

[19] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[20] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *2011 International Conference on Computer Vision*, 2011, pp. 2548–2555.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[22] J. Garcia-Bonilla, C. Leake, A. Elmquist, T. D. Hasseler, V. Steyert, A. Gaut, and A. Jain, "Dshell-darts: A reusability-focused multi-mission aerospace and robotics simulation toolkit," in *2025 IEEE Aerospace Conference*, 2025, pp. 1–13.