

PRISM: Multimodal Terrain Mapping for Rover Navigation in Unstructured Environments

R. Castilla-Arquillo^{1,2}, C. J. Pérez-del-Pulgar², L. Gerdes², A. Garcia-Cerezo², and M. Olivares-Mendez¹

Abstract—Robotic navigation in unstructured environments requires robust situational awareness to safely traverse hazards such as steep slopes and rocky terrain. To address this challenge, perception systems increasingly rely on multimodal sensor fusion. Specifically, integrating thermal imagery with standard optical and depth sensors enhances terrain differentiation, directly improving the reliability of mapping algorithms. This paper presents PRISM, a multimodal perception system for terrain mapping in unstructured settings. PRISM leverages a custom sensor suite to capture aligned RGB, depth, and thermal (RGB-D-T) imagery. At its core is OmniUnet, a novel vision transformer-based network specifically designed for multimodal semantic terrain segmentation. We validated the proposed system using two newly annotated datasets (BASEPROD and LAENTIEC) and demonstrate its real-world applicability through physical field experiments. Deployed on a resource-constrained embedded computer, PRISM efficiently generates traversability maps that directly enable autonomous navigation via a rover’s Guidance, Navigation, and Control (GNC) subsystem.

I. INTRODUCTION

Advancements in rover autonomy, driven by the Guidance, Navigation, and Control (GNC) system, have enabled reliable operation in unstructured and challenging terrains where direct human control is limited by communication constraints or environmental complexity [1]. Such environments often feature hazards like steep slopes, dense vegetation, and uneven ground, which can lead to slippage or entrapment [2], [3]. To address these risks, the navigation module of the GNC system relies on terrain mapping generated from onboard sensor data.

The terrain mapping process consists of two main stages: terrain segmentation and traversability analysis [4]. In the first stage, segmentation identifies distinct terrain regions and obstacles based on surface features. This is commonly handled by deep learning models for image segmentation, such as U-Net [5], Mask R-CNN [6], and YOLO [7]. More recent approaches incorporate Vision Transformers (ViT) [8] and hybrid architectures, such as Swin-Unet [9], which have demonstrated strong performance in complex environments [10], [11]. Most of these segmentation models, however, are designed for a single input modality, which limits their direct applicability when perception must draw on multiple complementary sensors simultaneously.

During the second stage of terrain mapping, traversability analysis utilizes segmented data to produce a map assessing navigational difficulty. Traversability is generally derived from geometric features, such as slope and roughness extracted from elevation data, or from machine learning models that predict rover behavior based on kinematics, energy use, and slip risk [12]–[14]. Because the Navigation module relies on these maps for path planning, the accuracy of these foundational stages dictates the overall safety of autonomous operations.

In the context of multimodal terrain mapping, the specific sensors a perception system fuses depend on the target operating environment. Within structured environments, current research combines RGB with stereoscopic images to segment small obstacles [15]–[17] and on-road features [18], while pairing thermal imagery with RGB data has gained traction for on-road elements [19]–[22]. Such fusion overcomes the limitations of any single sensor in complex scenes [23], a principle reflected in rover sensor suites: the ScienceCam on the Light Weight Rover Unit (LRU) combines grayscale, RGB, hyperspectral, and thermal imaging [24], and Mars rovers such as Curiosity [25] and Perseverance [26] carry similarly diverse suites for navigation and science. Perception in unstructured terrain, by contrast, still relies predominantly on fusing RGB and LiDAR [27]–[29], and the integration of thermal imaging remains comparatively underexplored [30], [31].

Addressing this gap is important, as thermal imaging offers unique insights in unstructured terrain. Surface thermal properties, such as thermal inertia, relate directly to traversability: under solar heating, granular sandy soils reach higher surface temperatures than more compacted soils, providing a cue for distinguishing terrain types that is difficult to recover from RGB data alone [30]. The cue is even stronger on Mars, where low atmospheric pressure heightens thermal contrasts and supports the estimation of rover slippage [32], [33]. Exploiting it, however, requires two integrated elements: first, an RGB-D-T sensing system that captures visible light (RGB), depth (D), and thermal (T) modalities; and second, a model capable of fusing these heterogeneous channels within a single architecture [34].

Several multimodal datasets have been proposed to support navigation in unstructured terrain. RUGD [35] pairs a monocular camera with LiDAR on a mobile robot to record diverse terrain classes, and RELLIS-3D [36] provides synchronized stereo vision and LiDAR data with semantic terrain annotations across varied outdoor settings. The CAVS dataset [37] contributes paired image and LiDAR data capturing fluctuations in lighting, weather, and vegetation density, while BASEPROD

¹R. Castilla-Arquillo and M. Olivares-Mendez are with the Space Robotics (SpaceR) Research Group, SnT, University of Luxembourg, Luxembourg, Luxembourg (e-mail:raul.castilla@uni.lu).

²R. Castilla-Arquillo, C. J. Pérez-del-Pulgar, L. Gerdes, and A. García Cerezo are with the Department of Automation and Systems Engineering, Universidad de Málaga, Andalucía Tech, 29070 Málaga, Spain (carlosperez@uma.es).

[38] offers RGB, depth, and thermal data collected by a rover in the Martian-like Bardenas semi-desert of Spain, matching the modality requirements of this work. Few systems, however, integrate thermal sensing into a multimodal segmentation pipeline and validate the resulting traversability map on a rover under real field conditions. Closing this gap, from offline segmentation to a deployed, field-tested mapping system, is the focus of this work.

Building on this, we present PRISM (Perception for RGB, Infrared, and Spatial Mapping), a multimodal terrain mapping system designed to integrate with the Navigation module of the GNC architecture for autonomous rover operations in unstructured environments. The system leverages RGB-D-T data to produce traversability maps that account for rover constraints. To process this data, we implement our previously introduced OmniUnet architecture [34], a deep learning model that combines the strengths of the Omnivore and U-Net frameworks. To support this system, we developed a multimodal sensor housing to mount multiple cameras on a robotic platform, the Rover Autonomy Testbed (RAT). We also contribute two publicly available labeled multimodal datasets: a labeled subset of the BASEPROD dataset [38], collected in the Bardenas semi-desert, and a new dataset gathered during rover traverses at the University of Málaga’s Area of Experimentation in New Technologies for Emergencies (LAENTIEC). Both span a variety of terrain types, such as dirt, grass, and gravel. We train and validate the terrain segmentation component, powered by OmniUnet, on both the BASEPROD and LAENTIEC datasets, and we validate the PRISM system through field experiments with the RAT platform in the unstructured terrain of LAENTIEC. The implementation was carried out in ROS2 and deployed on an onboard embedded computer. The source code and datasets are publicly available on GitHub¹.

II. THE GUIDANCE, NAVIGATION, AND CONTROL SYSTEM

The Guidance, Navigation, and Control (GNC) system employed in this work enables autonomous rover operation in complex, unstructured environments by fusing data from multiple sources. Inspired by architectures used in planetary exploration missions, it combines drone-based photogrammetry, which simulates the multispectral satellite data used in those missions, with the rover’s onboard sensor data. These include exteroceptive sensors, such as cameras and LiDAR, for environmental perception, and proprioceptive sensors, such as wheel encoders and inertial units, for tracking the rover’s motion.

When an operator assigns a scientific target or point of interest as a goal, the GNC system processes this objective to plan a traversable path based on the current terrain. It then commands the rover’s actuators to follow the generated trajectory, ensuring accurate and safe motion toward the target. As illustrated in Figure 1, the architecture comprises five main subsystems that collaborate in real time to generate terrain maps, estimate position, plan safe paths, and control movement. This section provides a brief overview of each

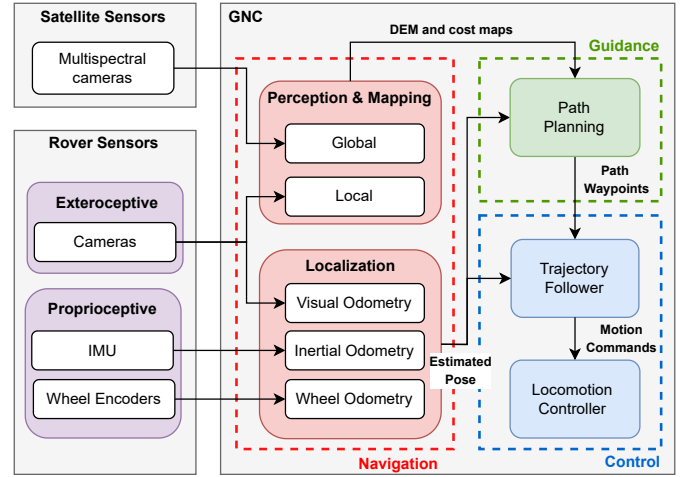


Fig. 1: Architecture of the Guidance, Navigation, and Control (GNC) system. The Path Planning module receives reference inputs from the operator, while the Locomotion Controller generates actuator command signals.

component, details the specific algorithms used, and provides access to our open-source implementations.

Perception & Mapping (Navigation): Responsible for building global and local maps to be used for path planning [39]. The global map, obtained from drone photogrammetry, provides a broad, low-resolution view of the terrain and supports route planning over large areas. The local map, continuously updated during traversal using exteroceptive sensors, captures nearby obstacles and fine-scale terrain features not visible in the global data. These maps are generally combined into Digital Elevation Maps (DEMs) and traversability maps [4]. A DEM represents the environment as a two-dimensional grid, with each cell storing the elevation at that location. Traversability maps, on the other hand, assign a traversal cost to each cell based on terrain characteristics such as slope or roughness. Both maps are sent to the Path Planning subsystem to support the computation of safe and efficient paths. In this work, the mapping pipeline is driven by our custom global and local DEM layers, which are based on the algorithms presented in [40]. The implementations for both layers are available as open-source repositories on GitHub².

Localization (Navigation): Estimates the rover’s pose, including its position, orientation, and velocity, by combining data from exteroceptive and proprioceptive sensors [41]. Wheel odometry tracks movement using encoder data, while visual odometry relies on camera images to correct for slippage and improve accuracy. Inertial odometry, derived from the Inertial Measurement Unit (IMU), which records acceleration and angular velocity, improves pose estimation by providing reliable motion data during rapid movements or in visually degraded environments. Localization can be integrated with mapping through Simultaneous Localization and Mapping (SLAM), a process in which the system incrementally constructs a map of the environment while simultaneously improving its

¹<https://github.com/spaceuma/MultimodalNavigation>

²Global DEM: https://github.com/spaceuma/nav2_global_dem_layer; Local DEM: https://github.com/spaceuma/nav2_local_dem_layer

estimate of the rover's pose within that map [42]. The final pose estimate is shared with the Path Planning and Trajectory Follower subsystems. State estimation is achieved using a GNSS/IMU localization node, which is available as an open-source repository³.

Path Planning (Guidance): Determines a route to a target location, which may be manually assigned by an operator or selected autonomously based on scientific interest or mission objectives. It takes into account the rover's current pose and the terrain information provided by the perception and mapping subsystem. Using this data, it generates a sequence of waypoints that avoid obstacles and account for terrain traversability. The planned path aims to minimize risk while optimizing travel distance or energy consumption, depending on the mission constraints [43]. Once computed, the waypoint sequence is passed to the Trajectory Follower for execution. For trajectory generation, we deployed a Fast Marching Method (FMM) planner [44] adapted for unstructured terrain, with the source code publicly accessible on GitHub⁴.

Trajectory Follower (Control): Converts the planned path into motion commands in real time. It dynamically adjusts the rover's speed, heading, and orientation to follow the trajectory accurately while responding to terrain conditions and deviations. One of the most widely used algorithms for this task is the Pure Pursuit algorithm [45], which calculates steering commands based on geometric criteria. Key parameters include the look-ahead distance, which defines how far ahead along the path the algorithm targets, and the width of the safety corridor, which sets the maximum allowable deviation from the trajectory before the path is considered unrecoverable. The resulting motion commands are sent to the Locomotion Controller for execution. Specifically, our framework implements Conservative Pure Pursuit Waypoint Navigation with a safety corridor [46], which is available online⁵.

Locomotion Control (Control): Executes movement commands through the rover's actuation system, whether using wheels, tracks, or legs. It compensates for terrain disturbances, enforces actuator limits, and applies control strategies such as Proportional-Integral-Derivative (PID) [47] control or Model Predictive Control (MPC) [48] to ensure stable and accurate locomotion. In wheeled systems, it may also coordinate differential or skid steering and manage suspension dynamics to maintain stability and traction across challenging terrain.

Together, these five subsystems form an integrated GNC architecture that enables robust and adaptive navigation in unstructured environments. Its modular design allows for scalability and adaptability across different robotic platforms and mission scenarios. This foundational structure serves as the basis for integrating advanced perception systems, such as the multimodal terrain-mapping module in this work.

III. MULTIMODAL TERRAIN MAPPING SYSTEM

The proposed multimodal terrain mapping system, PRISM, is integrated into the Perception & Mapping subsystem, which

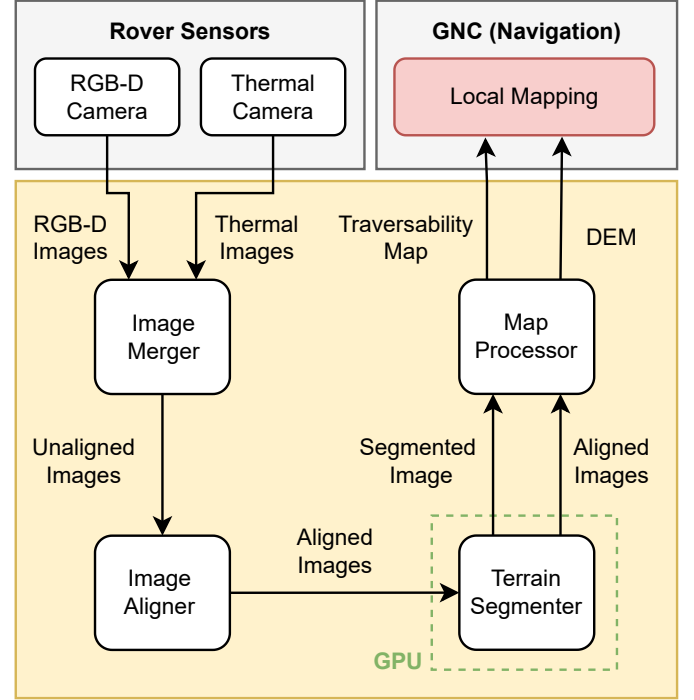


Fig. 2: Architecture of the Perception for RGB, Infrared, and Spatial Mapping (PRISM) system.

generates local maps using data from the rover's exteroceptive sensors. These maps provide elevation and traversability information and are forwarded to the path-planning subsystem to ensure safe navigation. As illustrated in Figure 2, PRISM consists of four interconnected components that operate concurrently: the image merger, image aligner, terrain segmenter, and map processor.

The process begins with the image merger, which integrates the RGB-D and thermal camera images based on their timestamps. These temporally synchronized but spatially unaligned images are then processed by the image aligner, which leverages depth data from the stereo camera to achieve precise spatial alignment. Once aligned, the images are passed to the terrain segmenter, which operates on an embedded Graphics Processing Unit (GPU) to generate segmentation masks. These masks, along with the aligned images, are then used by the map processor to construct a Digital Elevation Map (DEM) and a traversability map that incorporate depth and segmentation data, respectively. Finally, both maps are integrated into a multilayered DEM forming the rover's local map. This map is then sent to the rover's GNC system to enhance its ability to navigate unstructured terrain.

A. Image Merger

The process begins by retrieving the latest data from the sensor stacks to initialize a unified multi-channel matrix, M_{out} , which is flattened into a one-dimensional array for efficient serialization. The data is mapped into five channels: channels 0 through 2 contain the RGB data, channel 3 holds the depth information, and channel 4 stores the thermal data. Any out-of-bounds thermal pixels are set to zero. Finally, the

³https://github.com/spaceuma/nav2_gnss_imu_localization_node

⁴https://github.com/spaceuma/nav2_fmm_multilayered_planner

⁵https://github.com/spaceuma/nav2_conservative_pursuit

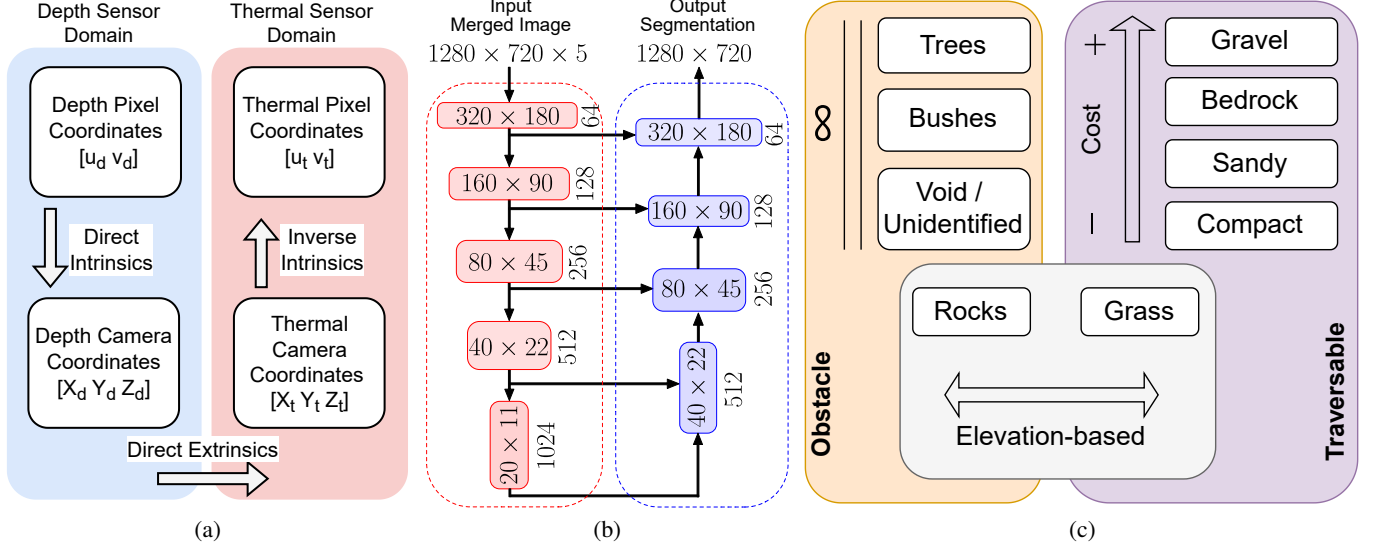


Fig. 3: Core pipeline stages of the PRISM system: (a) thermal-to-depth alignment (Image Aligner); (b) OmniUnet-based terrain segmentation (Terrain Segmenter); and (c) terrain- and elevation-based traversability mapping (Map Processor).

rover’s current pose and GPS coordinates are appended to the payload, providing spatial context before it is published to the navigation stack.

B. Image Aligner

Although the stereo camera inherently outputs registered RGB and depth images, the thermal camera requires a separate alignment step to match this coordinate frame. This registration relies on the Pinhole Camera Model (PMC), which mathematically models the projection of a 3D scene onto a 2D image plane. The model uses extrinsic parameters to define the camera’s spatial position and orientation, along with intrinsic parameters that specify the focal length and optical center. As shown in Fig. 3a, the proposed method aligns the thermal image by adapting standard stereo depth registration techniques. The algorithm initializes by storing the incoming depth and thermal frames as matrices and creating an empty aligned thermal matrix.

To populate this matrix, the system processes every pixel in the depth sensor domain with coordinates $[u_d, v_d]$ through a specific sequence of geometric transformations. First, it applies the depth camera’s intrinsic parameters to back-project the 2D pixel into 3D depth camera coordinates $[X_d, Y_d, Z_d]$. Next, extrinsic parameters are applied to transform this 3D point into the thermal camera’s coordinate space, yielding $[X_t, Y_t, Z_t]$. Finally, the thermal camera’s intrinsic parameters are used to project this 3D point onto the 2D thermal sensor domain, calculating the corresponding thermal pixel coordinates $[u_t, v_t]$. If these projected coordinates fall within the boundaries of the thermal image frame F_T , the corresponding temperature value is assigned to that pixel location in A_T . Out-of-bounds pixels are ignored.

C. OmniUnet terrain segmenter

The terrain segmenter receives the aligned multimodal images and generates a semantic segmentation mask of the rover’s surroundings. In addition to producing this segmented image, the module concurrently forwards the aligned multimodal images to the map processor for further integration. This component is powered by the OmniUnet architecture [34], as depicted in Fig. 3b, which is a transformer-based neural network designed to process combined RGB, depth, and thermal imagery. The model uses an Omnivore backbone [49] with shifted-window attention mechanisms to extract features across different sensory modalities, alongside a U-Net decoding strategy to produce the final terrain classifications. As illustrated in the diagram, the red blocks represent the Swin Transformer blocks responsible for feature extraction, while the blue blocks denote the convolutional layers with Gaussian Error Linear Unit (GELU) activations used during the decoding stage. The numbers inside these blocks indicate the spatial dimensions of the feature maps, while the numbers outside denote the channel dimensions at each respective stage. To ensure the system remains suitable for autonomous field operations, the segmenter is deployed on an onboard embedded GPU. This hardware configuration enables the network to perform efficient inference when classifying complex, unstructured terrain, such as sand, bedrock, and compact soil.

D. Map processor

The map processor performs two primary functions. First, it generates a multilayered DEM that combines RGB, thermal, elevation, and segmentation data. Second, it creates a traversability map using the terrain classifications from the terrain segmenter and the elevation data from the DEM. The multilayered DEM is generated following the procedure outlined in Algorithm 1. The system first collects depth and

Algorithm 1 Computation of elevation (M) and segmentation (S) maps.

Initialization:

```

1:  $res \leftarrow \text{obtainMapResolution}()$ 
2:  $pos \leftarrow \text{getCameraPose}()$ 
3:  $F_D \leftarrow \text{obtainDepthFrame}()$ 
4:  $F_S \leftarrow \text{obtainAlignedSegmentationFrame}()$ 
5:  $P_w \leftarrow \text{initializeWorldPointsArray}(F_D)$ 
6:  $P_s \leftarrow \text{initializeSegmentationPointsArray}(F_S)$ 
7:  $n \leftarrow 0$ 

```

Creating arrays with world points and seg. values:

```

8: for  $[u_d, v_d]$  in  $F_D$  do
9:    $[X_w, Y_w, Z_w] \leftarrow \text{getWorldCoord}(F_D, [u_d, v_d], pos)$ 
10:   $P_w(n) \leftarrow [X_w, Y_w, Z_w]$ 
11:   $F_s(n) \leftarrow P_s([u_d, v_d])$ 
12:   $n \leftarrow n + 1$ 
13: end for

```

Setting map bounds:

```

14:  $x_{offset} \leftarrow \min(P_w[X_w])$ 
15:  $M_{width} \leftarrow (\max(P_w[X_w]) + x_{offset}) \cdot res$ 
16:  $M_{height} \leftarrow \max(P_w[Y_w]) \cdot res$ 
17:  $M \leftarrow \text{initializeMatrix}([M_{width}, M_{height}])$ 
18:  $S \leftarrow \text{initializeMatrix}([M_{width}, M_{height}])$ 

```

Generating elevation and segmentation maps:

```

19: for  $n$  in  $P_w$  do
20:   $u_m \leftarrow (P_w(n)[X_w] + x_{offset}) \cdot res$ 
21:   $v_m \leftarrow P_w(n)[Y_w] \cdot res$ 
22:   $M([u_m, v_m]) \leftarrow P_w(n)[Z_w]$ 
23:   $S([u_m, v_m]) \leftarrow P_s(n)$ 
24: end for

```

segmentation frames, along with the depth camera's orientation. It then processes the depth frame iteratively, converting each depth value into world coordinates using the camera's intrinsic parameters. This step yields two main arrays: one containing the spatial coordinates and the other containing the terrain classes. The system then evaluates the coordinates to determine the appropriate map dimensions for a specific resolution. Finally, it initializes the elevation and segmentation matrices and populates them with the aligned data.

To construct the traversability map, the system evaluates both the predicted terrain class and the corresponding ground elevation. Figure 3c depicts this process and provides examples of different terrain categories. The map classifies the environment into two main groups based on navigation difficulty.

Obstacle: These areas receive an infinite cost regardless of the depth camera measurements. Certain terrains are classified as obstacles if their geometry is too steep or rough for the rover to cross safely. As a safety measure, any object not recognized by the terrain segmenter is automatically treated as an obstacle.

Traversable: These areas are assigned varying costs based on how easily the rover can navigate them. Specific terrain

types receive higher costs depending on their height and surface properties.

The costs generated from this process must be configured to meet the requirements of the GNC navigation stack. Once established, the traversability map can be further improved by incorporating data from other rover sensors. This combined approach is particularly useful for ambiguous areas such as grass and rocks, where relying solely on elevation or visual data can lead to classification errors. For example, if the system relies exclusively on elevation data, it might interpret a patch of tall grass as a solid obstacle. However, by including the neural network prediction, the system correctly identifies the grass and allows the rover to pass through. Conversely, the segmenter might identify a rock and flag it as an obstacle. By checking the elevation data, the system can determine if the rock is actually low enough for the rover to drive over safely.

IV. IMPLEMENTATION AND EXPERIMENTS

To validate the proposed PRISM architecture, this section details the complete system implementation and the experimental procedures. First, we describe the hardware configuration of the robotic platform and the custom multimodal sensor suite. Next, we introduce the datasets used for training and evaluation, followed by the specific training parameters used for the OmniUnet model. Finally, we present results from field experiments conducted in an unstructured environment to demonstrate the system's real-world applicability and performance.

A. Hardware

The RAT rover (Figure 4a) was used to perform the experiments. Its locomotion system features a Double-Bogie configuration with a passive longitudinal differential and a $4 \times 4 \times 4$ layout, incorporating four wheels, each with independent driving and steering capabilities [50]. Additionally, it includes an Xsens MTi-670 IMU and an Emlid Reach M+ GNSS receiver.

Multimodal images were generated by combining data from a Realsense D435i stereo camera and an Optris PI-640i thermal camera. The Realsense RGB sensor produces images with dimensions up to 1920×1080 pixels and a Field of View (FOV) of $69^\circ \times 42^\circ$. It provides stereoscopic depth images at a resolution of 1280×720 , with a FOV of $87^\circ \times 58^\circ$ and depth accuracy of less than 2% at a distance of 2 m. The Optris PI-640i thermal camera is based on uncooled microbolometer technology with a resolution of 640×480 pixels. It operates in the Long-Wave Infrared (LWIR) spectral range from $8 \mu\text{m}$ to $14 \mu\text{m}$ and can measure temperatures ranging from -20°C to 900°C with a thermal sensitivity of 0.04°C . Additionally, it features a germanium optic with a $60^\circ \times 45^\circ$ FOV.

A custom 3D-printed housing was designed to enclose both cameras (Fig. 4b), ensuring that the depth plane of the Realsense camera was vertically aligned with the optical plane of the thermal camera. This alignment enabled accurate projection of thermal data into the 3D world. The multimodal sensor was mounted on the platform at a height of 65 cm with

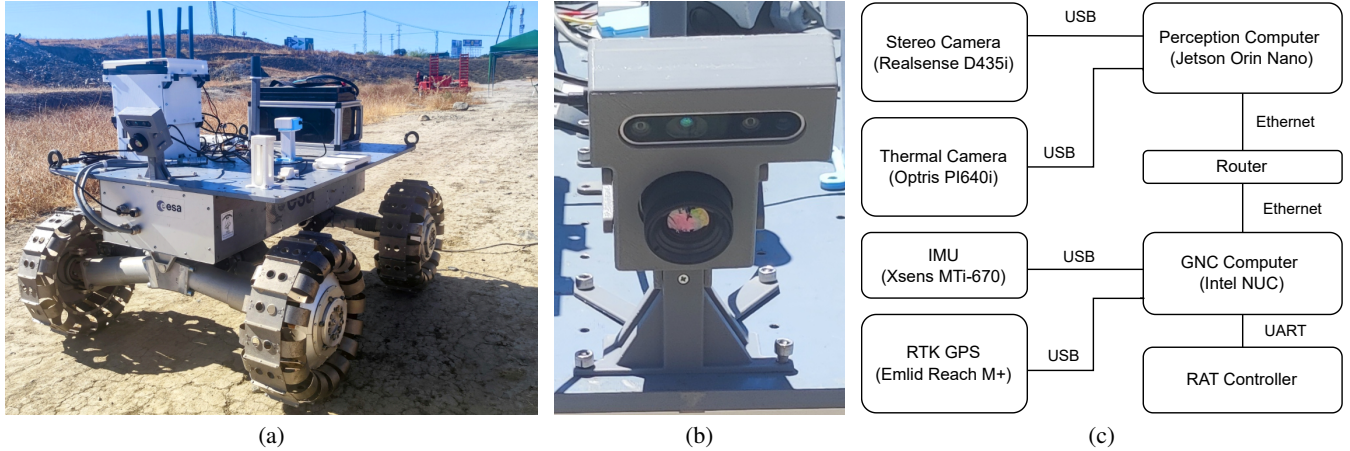


Fig. 4: Hardware configuration: (a) RAT platform during field operations, (b) 3D-printed multimodal sensor housing, and (c) block diagram of the system’s hardware architecture.

a 20° inclination to capture surface images as comprehensively as possible.

The RAT was equipped with two onboard computers linked via an Ethernet router; a diagram is depicted in Fig. 4c. One computer was dedicated to the rover’s GNC and the other handled perception. The GNC computer was an Intel NUC with an Intel Core i5 processor running at 1.8 GHz and 16 GB of RAM. This unit received data from the IMU and RTK GPS via USB connections and sent commands to the RAT controller over UART. The GNC architecture was programmed in ROS 2 and used the NAV2 [51] stack to execute a goal command, following a trajectory computed with the fast marching algorithm [52]. The perception computer was a Jetson Orin Nano featuring an ARM processor operating at 1.5 GHz, 8 GB of RAM, and an AI performance of 40 TOPS. It connected directly to the stereo and thermal cameras via USB. This computer was chosen over a high-performance GPU due to its low power consumption of 15 W and its limited resource architecture.

B. Terrain Segmenter Training

Because there are currently no publicly available multimodal datasets of the Martian surface, multimodal images of unstructured terrestrial environments were utilized to train and evaluate the proposed system. We contribute two publicly available labeled multimodal datasets for this purpose. To ensure complete consistency across all our data and field experiments, we used the same multimodal sensors and calibration parameters as those originally used to capture the BASEPROD dataset. The generated segmentation masks are single-channel images matching the resolution of the captured data, with unique integers representing different terrain classes. RGB images are provided in PNG format, whereas depth and radiometric thermal data are stored as CSV files.

The RUGD Dataset: The network was initially trained and evaluated using the RUGD dataset to establish a performance baseline. This dataset contains approximately 7,500 labeled

RGB images of unstructured outdoor scenes annotated across 25 semantic classes [35].

The Bardenas Dataset: The first multimodal dataset builds upon the BASEPROD dataset [38], which was collected during traverses with the MaRTA rover [39] in the Bardenas Reales semi-desert located in Cabanillas, Navarra, Spain. This environment features diverse terrains and geological obstacles, such as rocks and exposed bedrock, that closely mimic Martian surface characteristics. While the original BASEPROD dataset contains ca. 36,000 raw multimodal images, it does not provide semantic annotations. To address this, we contribute a newly labeled subset of this data. We manually labeled 950 representative images for training and 190 for validation. The annotations encompass eight classes: void, compact, grass, bedrock, sandy, gravel, rock, and bush. We have published these labeled segmentation masks on Zenodo⁶ [53] to support future research.

The LAENTIEC Dataset: The second multimodal dataset was generated using the RAT rover at LAENTIEC, an unstructured 90 000 m² testing facility at the University of Malaga, Spain. LAENTIEC provides varying elevations and diverse terrain types including dirt, gravel, and sand. Captured in spring, this dataset features denser green vegetation than the Bardenas dataset, though the site typically dries out in summer. The LAENTIEC dataset [54] consists of 310 multimodal images, with 62 reserved for validation. It classifies six semantic classes: void, compact, grass, gravel, sky, and tree/bush. This newly collected dataset has also been made publicly available on Zenodo⁷.

The training was conducted on an Nvidia DGX Station using a GPU Tesla V100 DGXS with 32 GB of V-RAM. The network was trained for 50 epochs for all modalities with a batch size of 16 images and a learning rate of 2e−5. The optimization function was a combined loss comprising Dice loss and cross-entropy loss. The datasets were split into 80/20 training/validation sets.

⁶<https://doi.org/10.5281/zenodo.15496884>

⁷<https://doi.org/10.5281/zenodo.8032971>

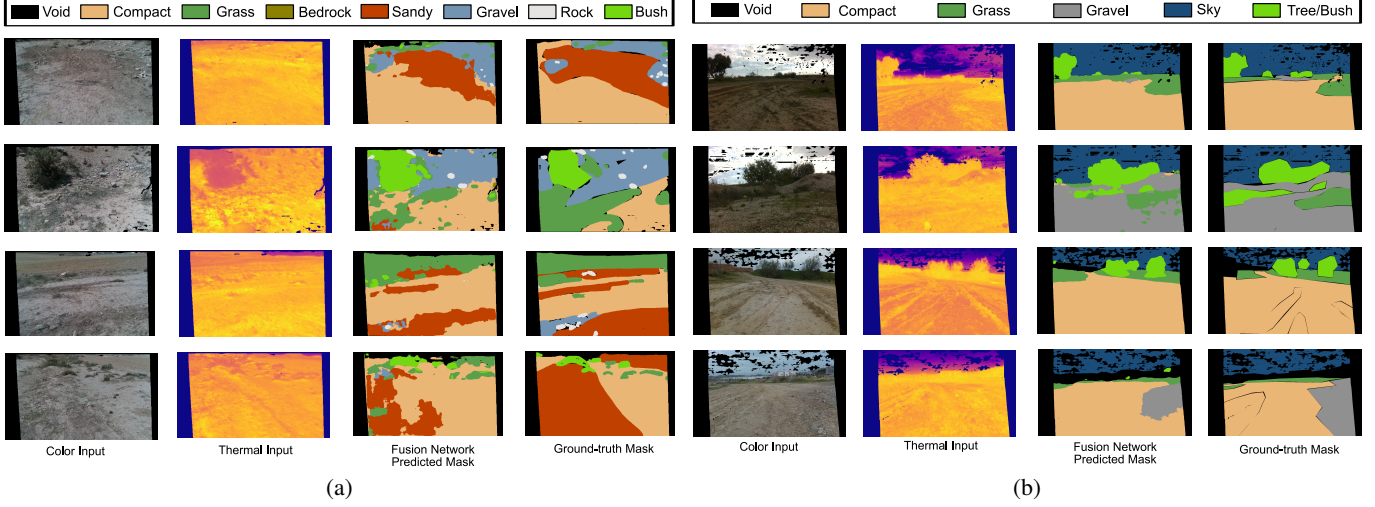


Fig. 5: Qualitative comparison of OmniUnet multimodal segmentation. Each subfigure presents the RGB and thermal inputs alongside the predicted and ground-truth segmentation masks for (a) the Mars-analog Bardenas dataset and (b) the LAENTIEC dataset.

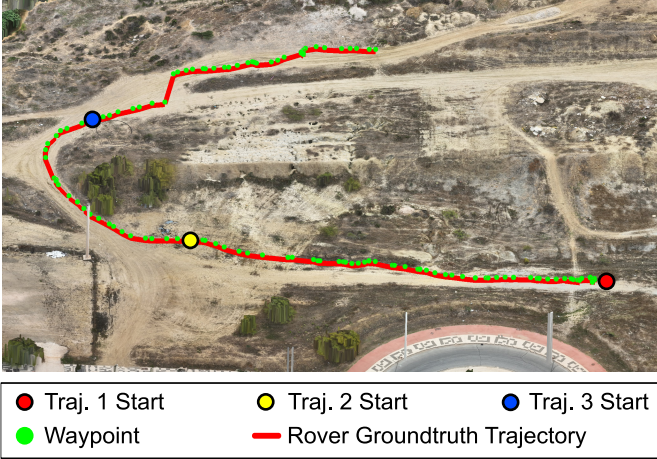


Fig. 6: Trajectories followed by the RAT platform during field tests at the LAENTIEC facility, detailing both autonomous and teleoperated segments.

Due to the limited number of images in the LAENTIEC dataset, OmniUnet was initially pre-trained on RGB-D-T images from the Bardenas dataset. The learned weights from this pre-training were then used to train the LAENTIEC dataset. During this stage, the first two network layers were frozen, and training was performed for 20 epochs using LAENTIEC images. Table I presents the per-class pixel accuracy, as well as overall metrics, including total pixel accuracy (excluding the void class for more realistic results), mean class pixel accuracy, and mean class intersection over union. Additionally, Figures 5a and 5b show the segmentation masks produced by the network with both the Bardenas and LAENTIEC datasets alongside their corresponding color and thermal inputs.

Regarding the quantitative results, the network achieved a baseline pixel accuracy of 92.58% on the RUGD dataset. For the multimodal inputs, both OmniUnet models (trained

TABLE I: Dataset statistics and metrics (%) of the OmniUnet network evaluated on the RUGD, Bardenas, and LAENTIEC datasets. Classes with dashes (-) were not present or not evaluated per class.

Metric/Class	RUGD	Bardenas	LAENTIEC
<i>Dataset Statistics</i>			
Total Images	7,435	1,140	310
Training	5,948	950	248
Validation	1,487	190	62
Number of Classes	25	8	6
<i>Class Pixel Accuracy</i>			
Void	-	90.48	94.16
Compact	-	77.44	78.68
Grass	-	39.04	82.72
Bedrock	-	28.30	-
Sandy	-	26.57	-
Gravel	-	51.83	27.04
Rock	-	18.40	-
Bush/Tree	-	40.68	41.20
Sky	-	-	94.25
<i>Average Metrics</i>			
Total PA	92.58	80.37	92.85
Mean PA	39.72	46.59	69.68
Mean IoU	34.90	38.77	63.13

on Bardenas and LAENTIEC images) achieved a correct pixel classification rate exceeding 80%. In the Bardenas dataset, the model demonstrated strong performance in identifying surfaces relevant to safe navigation, successfully detecting compact terrain (the safest to traverse) at 77.44% and gravel at 51.83%. For obstacle detection, the model identified bushes 40.68% of the time; however, rock detection was more difficult, dropping to 18.40% likely due to the variability in rock thermal signatures.

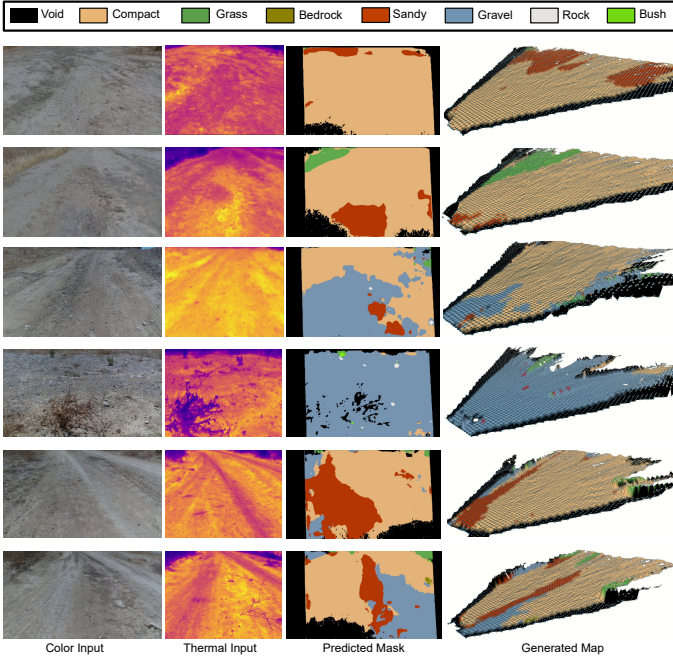


Fig. 7: Visual outputs from the PRISM system during field operations: (a) input RGB image, (b) input thermal image, (c) predicted terrain segmentation mask, and (d) the final 3D traversability map utilized for navigation.

TABLE II: Performance metrics and time distribution for the three field test trajectories executed by the RAT rover.

Traj.	Dist. (m)	Gen. maps	Stops	Test dur. (s)	Motion (s)	Static (s)
1	71	91	13	1881	806	1075
2	67	52	0	1084	1084	0
3	53	68	9	1433	633	800

C. Experiments

We evaluated the proposed PRISM architecture through a series of field tests with the RAT rover at UMA’s LAENTIEC facility. Developed in ROS 2 and accessible on GitHub⁸, the system constructs traversability maps by integrating DEM data with OmniUnet semantic segmentations. For these experiments, we selected model weights trained on the Bardenas dataset because it closely resembles the summer conditions presented at the test site. Operating at peak CPU capacity, the perception hardware completed the full map generation cycle in approximately 20s from image capture, while the Terrain Segmenter alone averaged 673ms per prediction. Ultimately, the setup successfully projected traversability maps up to 10m ahead of the rover at a resolution of 0.1 m.

The experimental route comprised three distinct segments (Figure 6) executed under two operational modes. Trajectories 1 and 3 were navigated autonomously for a combined distance of 124m. In this mode, an initial path was computed via the NAV2 interface using a global DEM acquired by a drone. Every 5m, the rover paused (indicated by green markers in Figure 6) to allow the GNC computer to receive an updated

local traversability map, integrate it into the global map, and replan the route. Conversely, Trajectory 2 involved a 67m teleoperated transition between map zones using a gamepad, during which local maps continued being generated. Table II summarizes the key metrics from these runs, and Figure 7 presents the resulting PRISM traversability maps. A video recording of the field tests is publicly accessible⁹.

V. CONCLUSIONS AND FUTURE WORK

This study introduced PRISM, a multimodal perception system for terrain segmentation and traversability mapping using RGB-D-T data. At its core, we implemented OmniUnet, a novel transformer-based architecture validated on two new labeled datasets from the Bardenas semi-desert and LAENTIEC facility. OmniUnet demonstrated robust performance, achieving total pixel accuracy exceeding 80 % on the Bardenas and reaching 92.85 % on the LAENTIEC dataset. The integration of thermal imagery proved essential for distinguishing surfaces with similar visual appearances but differing thermal inertia, such as compact and sandy soils. We successfully validated the PRISM pipeline through field experiments with the RAT platform, demonstrating efficient, autonomous navigation on resource-constrained hardware. Future work will enhance segmentation robustness by integrating specialized obstacle detection modules and adopting thermal stereo imaging to ensure reliable depth estimation independent of ambient lighting conditions.

ACKNOWLEDGMENTS

This work was supported by the Spanish national government project no. PID2024-160373OB-C21, entitled “Hybrid Perception and Navigation for Planetary Exploration”. Additionally, field testing was partially supported by the European Space Agency under activity no. 4000140043/22/NL/GLC/ces.

REFERENCES

- [1] M. B. Quadrelli, L. J. Wood, J. E. Riedel, M. C. McHenry, M. Aung, L. A. Cangahuala, R. A. Volpe, P. M. Beauchamp, and J. A. Cutts, “Guidance, navigation, and control technology assessment for future planetary science missions,” *Journal of Guidance, Control, and Dynamics*, vol. 38, no. 7, pp. 1165–1186, 2015.
- [2] L. Wijayathunga, A. Rassau, and D. Chai, “Challenges and solutions for autonomous ground robot scene understanding and navigation in unstructured outdoor environments: A review,” *Applied Sciences*, vol. 13, no. 17, 2023.
- [3] R. Gonzalez and K. Iagnemma, “Slippage estimation and compensation for planetary exploration rovers. State of the art and future challenges,” *Journal of Field Robotics*, vol. 35, no. 4, pp. 564–577, 2018.
- [4] P. Borges, T. Peynot, S. Liang, B. Arain, M. Wildie, M. Minareci, S. Lichman, G. Samvedi, I. Sa, N. Hudson *et al.*, “A survey on terrain traversability analysis for autonomous ground vehicles: Methods, sensors, and challenges,” *Field Robot.*, vol. 2, no. 1, pp. 1567–1627, 2022.
- [5] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.
- [6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [7] T. Diwan, G. Anirudh, and J. V. Tembhurne, “Object detection using yolo: Challenges, architectural successors, datasets and applications,” *multimedia Tools and Applications*, vol. 82, no. 6, pp. 9243–9275, 2023.

⁸<https://github.com/spaceuma/MultimodalNavigation>

⁹<https://youtu.be/6vFK4U6ZmFI>

- [8] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Computer Vision (ICCV)*, 2021.
- [9] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*, 2022, pp. 205–218.
- [10] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [11] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.
- [12] R. Zhou, L. Ding, H. Gao, W. Feng, Z. Deng, and N. Li, "Mapping for planetary rovers from terramechanics perspective," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1869–1874.
- [13] M. Ono, B. Rothrock, K. Otsu, S. Higa, Y. Iwashita, A. Didier, T. Islam, C. Laporte, V. Sun, K. Stack *et al.*, "Maars: Machine learning-based analytics for automated rover systems," in *2020 IEEE aerospace conference*, 2020, pp. 1–17.
- [14] D. Gaines, G. Doran, M. Paton, B. Rothrock, J. Russino, R. Mackey, R. Anderson, R. Francis, C. Joswig, H. Justice *et al.*, "Self-reliant rovers for increased mission productivity," *Journal of Field Robotics*, vol. 37, no. 7, pp. 1171–1196, 2020.
- [15] K. Gupta, S. A. Javed, V. Gandhi, and K. M. Krishna, "Mergenet: A deep net architecture for small obstacle discovery," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 5856–5862.
- [16] M. Hua, Y. Nan, and S. Lian, "Small obstacle avoidance based on RGB-D semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 1–9.
- [17] L. Sun, K. Yang, X. Hu, W. Hu, and K. Wang, "Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images," *IEEE robotics and automation letters*, vol. 5, no. 4, pp. 5558–5565, 2020.
- [18] Q. Zhao, Y. Wan, J. Xu, and L. Fang, "Cross-modal attention fusion network for RGB-D semantic segmentation," *Neurocomputing*, vol. 548, p. 126389, 2023.
- [19] J. Vertens, J. Zürn, and W. Burgard, "Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 8461–8468.
- [20] B. Wang, H. Zhao, and Y. Zhuang, "DaCFN: divide-and-conquer fusion network for RGB-T object detection," *International Journal of Machine Learning and Cybernetics*, pp. 1–14, 2023.
- [21] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," *IEEE Transactions on Intelligent Transportation Systems*, 2023.
- [22] X. Liu, J. Wang, and J. Li, "URTSegNet: A real-time segmentation network of unstructured road at night based on thermal infrared images for autonomous robot system," *Control Engineering Practice*, vol. 137, p. 105560, 2023.
- [23] Y. Zhang, A. Carballo, H. Yang, and K. Takeda, "Perception and sensing for autonomous vehicles under adverse weather conditions: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 146–177, 2023.
- [24] A. Wedler, M. Vayugundla, H. Lehner, P. Lehner, M. J. Schuster, S. G. Brunner, W. Stürzl, A. Dömel, H. Gmeiner, B. Vodermayr *et al.*, "First results of the ROBEX analogue mission campaign: Robotic deployment of seismic networks for future lunar missions," in *International Astronautical Congress (IAC)*, vol. 68, 2017.
- [25] J. P. Grotzinger, J. Crisp, A. R. Vasavada, R. C. Anderson, C. J. Baker, R. Barry, D. F. Blake, P. Conrad, K. S. Edgett, B. Ferdowski *et al.*, "Mars Science Laboratory mission and science investigation," *Space science reviews*, vol. 170, pp. 5–56, 2012.
- [26] J. Bell, J. Maki, G. Mehall, M. Ravine, M. Caplinger, Z. Bailey, S. Brylow, J. Schaffner, K. Kinch, M. Madsen *et al.*, "The Mars 2020 Perseverance rover mast camera zoom (Mascam-Z) multispectral, stereoscopic imaging investigation," *Space science reviews*, vol. 217, pp. 1–40, 2021.
- [27] T. H. Y. Leung, D. Ignatyev, and A. Zolotas, "Hybrid terrain traversability analysis in off-road environments," in *2022 8th International Conference on Automation, Robotics and Applications (ICARA)*, 2022, pp. 50–56.
- [28] C. Zhong, B. Li, and T. Wu, "Off-road drivable area detection: A learning-based approach exploiting lidar reflection texture information," *Remote Sensing*, vol. 15, no. 1, p. 27, 2022.
- [29] O. Kim, J. Seo, S. Ahn, and C. H. Kim, "Ufo: Uncertainty aware lidar-image fusion for off-road semantic terrain map estimation," *arXiv preprint arXiv:2403.02642*, 2024.
- [30] R. Gonzalez, A. Lopez, and K. Iagnemma, "Thermal vision, moisture content, and vegetation in the context of off-road mobile robots," *Journal of Terramechanics*, vol. 70, pp. 35–48, 2017.
- [31] Y. Iwashita, K. Nakashima, J. Gatto, S. Higa, A. Stoica, N. Khoo, and R. Kurazume, "Virtual IR sensing for planetary rovers: Improved terrain classification and thermal inertia estimation," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6302–6309, 2020.
- [32] C. Cunningham, I. A. Nesnas, and W. L. Whittaker, "Improving slip prediction on Mars using thermal inertia measurements," *Autonomous Robots*, vol. 43, pp. 503–521, 2019.
- [33] R. Castilla-Arquillo, A. Mandow, C. J. Pérez-del Pulgar, C. Álvarez-Llamas, J. M. Vellido, and J. Laserna, "Thermal imagery for rover soil assessment using a multipurpose environmental chamber under simulated Mars conditions," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–12, 2024.
- [34] R. Castilla-Arquillo, C. J. Pérez-Del-Pulgar, L. Gerdes, A. García-Cerezo, and M. Olivares-Mendez, "OmniUnet: A multimodal network for unstructured terrain segmentation on planetary rovers using RGB, depth, and thermal imagery," in *2025 International Conference on Space Robotics (iSpaRo)*, 2025, pp. 161–167.
- [35] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "A RUGD dataset for autonomous navigation and visual perception in unstructured outdoor environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 5000–5007.
- [36] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "Relis-3d dataset: Data, benchmarks and analysis," in *2021 IEEE international conference on robotics and automation (ICRA)*, 2021, pp. 1110–1116.
- [37] S. Sharma, L. Dabbiru, T. Hannis, G. Mason, D. W. Carruth, M. Doude, C. Goodin, C. Hudson, S. Ozier, J. E. Ball *et al.*, "Cat: Caves traversability dataset for off-road autonomous driving," *IEEE Access*, vol. 10, pp. 24 759–24 768, 2022.
- [38] L. Gerdes, T. Wiese, R. Castilla Arquillo, L. Bielenberg, M. Azkarate, H. Leblond, F. Wiltig, J. Ortega Cortés, A. Bernal, S. Palanco *et al.*, "BASEPROD: The Bardenas semi-desert planetary rover dataset," *Scientific Data*, vol. 11, no. 1, p. 1054, 2024.
- [39] M. Azkarate, L. Gerdes, T. Wiese, M. Zwick, M. Pagnamenta, J. Hidalgo-Carrio, P. Poulakis, and C. J. Perez-del Pulgar, "Design, Testing, and Evolution of Mars Rover Testbeds: European Space Agency Planetary Exploration," *IEEE Robotics and Automation Magazine*, vol. 29, no. 3, pp. 10–23, sep 2022.
- [40] G. J. Paz-Delgado, M. Azkarate, J. R. Sánchez-Ibáñez, C. J. Pérez-del Pulgar, L. Gerdes, and A. J. García-Cerezo, "Improving autonomous rover guidance in round-trip missions using a dynamic cost map," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 7014–7019.
- [41] M. Zarei and R. Chhabra, "Advancements in autonomous mobility of planetary wheeled mobile robots: A review," *Frontiers in Space Technologies*, vol. 3, p. 1080291, 2022.
- [42] C. Stachniss, J. J. Leonard, and S. Thrun, "Simultaneous localization and mapping," in *Springer handbook of robotics*, 2016, pp. 1153–1176.
- [43] J. R. Sánchez-Ibáñez, C. J. Pérez-del Pulgar, and A. García-Cerezo, "Path planning for autonomous mobile robots: A review," *Sensors*, vol. 21, no. 23, p. 7898, 2021.
- [44] J. R. Sánchez-Ibáñez, C. J. Perez-del Pulgar, M. Azkarate, L. Gerdes, and A. García-Cerezo, "Dynamic path planning for reconfigurable rovers using a multi-layered grid," *Engineering Applications of Artificial Intelligence*, vol. 86, pp. 32–42, 2019.
- [45] R. S. Wallace, A. Stentz, C. E. Thorpe, H. P. Moravec, W. Whittaker, T. Kanade *et al.*, "First results in robot road-following," in *IJCAI*, vol. 2, 1985, pp. 1089–1095.
- [46] L. Gerdes, M. Azkarate, J. Sánchez-Ibáñez, L. Joudrier, and C. Perez-del Pulgar, "Efficient autonomous navigation for planetary rovers with limited resources," *Journal of Field Robotics*, vol. 37, no. 7, pp. 1153–1170, 2020.
- [47] J. Kim and J. Lee, "A kinematic-based rough terrain control for traction and energy saving of an exploration rover," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 3595–3600.

- [48] J. He, Y. Sun, L. Yang, and F. Gao, "Model predictive control of a novel wheeled-legged planetary rover for trajectory tracking," *Sensors*, vol. 22, no. 11, p. 4164, 2022.
- [49] R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 102–16 112.
- [50] A. Medina, L. Mollinedo, K. Kapellos, C. Crespo, and P. Poulakis, "Design and realization of a rover autonomy testbed," in *ESA Workshop on Advanced Space Technologies for Robotics and Automation (ASTRA)*, 2015.
- [51] S. Macenski, F. Martin, R. White, and J. Ginés Clavero, "The marathon 2: A navigation system," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.
- [52] J. Sánchez-Ibáñez, C. Pérez-del Pulgar, M. Azkarate, L. Gerdes, and A. García-Cerezo, "Dynamic path planning for reconfigurable rovers using a multi-layered grid," *Engineering Applications of Artificial Intelligence*, vol. 86, pp. 32–42, 2019.
- [53] R. Castilla Arquillo, C. J. del Pulgar Mancebo, L. Gerdes, A. García-Cerezo, and A. Martínez Petersen, "Fusion of Color, Depth, and Thermal Modalities for Rover Navigation in Unstructured Environments," may 2025. [Online]. Available: <https://doi.org/10.5281/zenodo.15496884>
- [54] I. Naranjo-Ortega, R. Castilla-Arquillo, and C. J. P. del Pulgar-Mancebo, "RGBDT dataset LAENTIEC experimental area," May 2023. [Online]. Available: <https://doi.org/10.5281/zenodo.8032971>