Semantic-Aware Image Matching for Large-Scale 3D Reconstruction of the Martian Surface from Rover Images

Zhaojin Li, Bo Wu*

***Abstract*— High-resolution images of the Martian surface, collected by cameras onboard rovers, offer unique insights unavailable from satellite images and are crucial for rover navigation and geological study on the Martian surface. However, the large variations in spatial resolution and viewpoint across images acquired from different rover stations, exacerbated by the textureless nature of the Martian surface, pose significant challenges for effective 3D surface reconstruction, a fundamental task in planetary topographic mapping. Thus, this paper proposes a deep learning-based approach to enable robust image matching constructed from multi-level semantic cues, for large-scale 3D reconstruction from rover images. First, a Siamese transformer-based neural network is used to perform semantic segmentation of the rover images, thereby extracting multi-level semantic cues. Second, feature matching is performed using rover images collected from different stations, in which these semantic cues are integrated to enhance feature descriptor construction, contextual aggregation, and outlier removal, yielding robust cross-station matches. These matches facilitate the bundle adjustment to link cross-station rover images accurately. Third, in the dense matching of rover images, frequency-domain matching is proposed and embedded with semantic cues to improve matching reliability and preserve surface discontinuities. Lastly, the disparity maps generated from the matching results are used to derive the 3D point clouds, which are then meshed to generate 3D surface models. Experiments are conducted on two image datasets of typical Martian scenes collected by the Zhurong rover to evaluate the performance of the proposed method. The results indicate that image residuals of around 1.5 pixels on average are achieved for the bundle adjustment of cross-station images using the matched feature points, and the final generated 3D models exhibit an accuracy better than 0.5 m. Compared with the cutting-edge commercial software, the generated 3D models from our method exhibit superior quality in terms of both accuracy and coverage, highlighting the effectiveness of the semantic-aware image matching algorithm.**

***Index Terms*— Image matching; Semantic; Mars; Rover; 3D Reconstruction**

I. INTRODUCTION

THREE-DIMENSIONAL (3D) mapping of the Martian surface is of great importance for Mars exploration missions [1], topographic analysis [2], and geomorphologic and geological characterization [3], [4]. Various studies have been conducted using satellite images [5], [6], [7] to generate the digital elevation model (DEM) covering large areas. Despite advances in photoclinometry [8], [9] and the deep-learning [10], [11], which can enhance DEMs to recover pixel-wise details, the retrieved 3D information is still limited by the original resolution of the images. While sub-pixel super-resolution techniques [12] offer potential for resolution enhancement, training a robust super-resolution network for 3D mapping is still challenging. In contrast, the rover image captured from close range could reveal subtle details of the planetary surface surface [13], [14]. Hence, there is growing interest in 3D reconstruction from rover images [15] to achieve improved large-scale 3D models.

Rover images are typically captured at waypoints along the traverse, and reconstructing 3D surfaces from close-range images (e.g., within 15 m) captured at a single station is straightforward [16]. However, the limited baseline between the cameras introduces significant uncertainty in recovering 3D information of distant regions through photogrammetry. This challenge is further compounded by the drift in the onboard inertial measurement unit (IMU) system [15]. Moreover, image resolution decreases as the viewing distance increases, leading to poor-quality reconstructions of distant regions. A pioneering work is thus proposed to integrate rover images with satellite imagery, achieving sub-pixel accuracy in mapping while simultaneously localizing the rover in global coordinate [17]. While other study [18] focuses on using rover images solely, to deal with the significant variations in the appearance of the same landforms owing to changes in viewpoint and resolution, which hinders feature matching and subsequently impedes the integrated bundle adjustment. Recent studies have demonstrated the benefits of introducing initial position and pose parameters into the matching pipeline [19], [20], [21] to successfully match challenging image pairs. However, unlike Earth, where the Global Navigation Satellite System (GNSS) is available, the original position and pose parameters of rover images on Mars may diverge from the actual values, limiting the effectiveness of the matching algorithm. The textureless nature of the Martian surface further complicates the matching process, given the challenges in constructing distinctive feature descriptors. Recent breakthroughs in deep learning have facilitated the development of more distinctive and robust

Manuscript submitted on September 9th, 2024. This work was supported by grants from the Research Grants Council of Hong Kong (Project No: PolyU 15215822, Project No: PolyU 15210520, RIF Project No: R5043-19, CRF Project No: C7004-21GF). The authors thank all individuals who worked on the dataset archives to make them publicly available.

Z. Li, and B. Wu are with the Planetary Remote Sensing Laboratory, Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (bo.wu@polyu.edu.hk)

feature descriptors [22], [23], [24], [25] and end-to-end matching algorithms based on neural networks. Although promising results have been achieved on many common datasets, these learning-based algorithms still struggle to establish connections between textureless cross-station rover images owing to the aforementioned challenges.

Furthermore, dense matching is also hindered by the textureless planetary surface. Dense matching is typically formulated as a global optimization problem, with two key components, i.e., the data and smoothness costs [26]. The data cost, which measures the similarity between two small patches from the image pair based on grayscale intensities, is likely to be confused by the featureless and homogeneous planetary surface and further complicated by variations in resolution [27]. Moreover, the appearance of landforms varies significantly, rendering a uniform solution suitable for one landform unsuitable for others. The smoothness cost explicitly assumes that the disparity between neighboring pixels is minimal, which is not the case in areas with occlusion and sharp features where discontinuities occur [28]. To address this smoothness problem, edge-aware [29], [30] and texture-aware dense image matching [31] methods have been proposed, incorporating the edges detected by the Canny operator. However, the Canny operator may yield fragmented and noisy results in natural scenarios.

Recently, deep learning has gained significant attention for facilitating image matching for 3D reconstruction by automatically extracting multi-level robust features. These features enable semantic segmentation, allowing image matching algorithms to adapt to semantic classes thus becoming semantic-aware. This capability significantly improves matching performance in challenging scenarios [32], [33], especially for textureless regions, where traditional methods often fail due to the lack of distinct visual features. In light of this, we propose an innovative semantic-aware image matching method designed to achieve large-scale 3D reconstruction from rover images. In our work, semantic refers to both contextual information (e.g., classifying pixels as rocks, sand, or rover) and high-dimensional feature representations extracted by the segmentation network. The key contributions of this work are as follows:

(1) To retrieve tie-points for cross-station rover images, a semantic-aware feature matching approach is proposed, which extends the SuperGlue paradigm by integrating multiple levels of semantic cues derived from a transformer-based segmentation neural network. The features calculated from the backbone, decoded head, and the Softmax operator are successively used in the initial descriptor construction, contextual attention aggregation, and outlier removal throughout the entire matching workflow;

(2) To improve the quality and accuracy of the dense image matching, the semantic cues are utilized as a complementary cue to phase correlation in calculating the cost volume. Furthermore, the edges extracted from semantic segments are exploited to adaptively adjust penalties and constrain cost propagation, thereby preserving the actual shapes and discontinuities;

(3) To extract the semantic cues, a Siamese transformer-based segmentation neural network is devised to fuse multi-scale features. Using this Siamese architecture, the network can be trained in a semi-supervised manner, thereby mitigating the need for annotated data.

The remainder of this article is organized as follows. Section 2 provides a brief overview of the related work. Section 3 outlines the proposed semantic-aware image matching pipeline, including the Siamese semantic segmentation, semantic-aware feature matching, and semantic-aware dense matching. Section 4 presents the results of experiments based on two representative datasets collected by the Chinese Zhurong rover, and evaluates the performance of the proposed methods. Finally, the concluding remarks are summarized in Section 5.

II. Related Work

Photogrammetric 3D reconstruction from rover images typically involves two consecutive steps, namely, structure-from-motion (SfM) and multi-view stereo (MVS) [34]. While SfM refines the initial exterior orientation (EO) to eliminate possible inconsistency and integrate cross-station images, MVS retrieves pixel-wise correspondences from the images to generate 3D point clouds.

SfM first retrieves tie-points and then establishes constraint functions based on these matches, enabling the simultaneous calculation of the actual positions of the images and the 3D coordinates of these points. The scale-invariant feature transform (SIFT) algorithm algorithm [35] is a widely used feature matching technique, which considers grayscale descriptors and dominant orientations to achieve rotation and scale invariance. This algorithm has been extended to affine invariance [36] through simulated image generation based on possible camera positions and rotations. Recent studies have endeavored to further refine the feature matching between images exhibiting more common motion in the 3D world by incorporating the original pose and position parameters to align the views, thereby guiding SIFT-based feature matching [19], [20].

With the emergence of deep learning, end-to-end learning-based algorithms such as SuperGlue have demonstrated significant superiority over traditional SIFT algorithms [21], [22], [37]. Instead of manually designing features, a convolutional backbone is used to extract higher-level and more comprehensive feature representations, which are then embedded with positional and contextual information. The merit of the convolutional backbone has been further substantiated by Zhong et al., [18], which succeeds in connecting rover images captured by China's Yutu-2 rover using the learned features. Another promising semantic-aware strategy has also benefited from the advancements in deep-learning-based semantic segmentation. These algorithms attempt to delimit the search space for subsequent feature detection and matching [38]. Although satisfactory results have been obtained in certain scenarios, much of the valuable

information generated during the segmentation process is overlooked, limiting the potential of such techniques.

The same feature description issue is also encountered in the dense matching phase. Unlike feature matching, dense matching typically involves patch-based similarity measurements, and the corresponding approaches can be categorized into spatial and frequency domains. Representative spatial domain algorithms include those based on Census and mutual information [26]. Census, which encodes grayscale information to accelerate the overall workflow, has been widely used in various applications. In contrast, frequency-based algorithms aim to enhance the expressibility of textureless surfaces [39], [40]. Furthermore, complementary approaches have been developed by leveraging semantic labels to enhance domain similarity measurements for better accuracy [1. Apart from the similarity measurement, discontinuity preservation is another primary concern in dense matching. Rothermel et al., [29] have already introduced edge detection into the dense matching pipeline to adapt the parameters accordingly. The traditional Canny edge detector, while effective in certain scenarios, encounters challenges in scenarios involving barren planetary surfaces. In contrast, semantic edges, derived from high-level semantic features, can provide more comprehensive and meaningful representations of image structures. Pioneering research on urban scenes has used semantic edges [42] to confine disparity computations to these crucial edges. This concept was later extended to harness the disparity calculated from the semantic edges (i.e., rooftops) as constraints for refining the disparities in building structures [30]. Similarly, on planetary surfaces, semantic edges help clarify the boundaries of discontinuities where occlusion occurs, providing critical cues for accurate 3D reconstruction.

In summary, advancements in deep learning have enabled the retrieval of semantic information at multiple levels. Hence, the incorporation of semantic information to facilitate image matching has gained prominence in both sparse image matching and dense matching phases. However, in-depth investigations into harnessing multi-level semantic information to enrich gray-scale information for image matching on textureless Martian surfaces remain scarce. It is thus desirable to develop a more elegant and efficient method to integrate semantic information into the photogrammetric process to achieve optimal 3D reconstruction results.

III. Semantic-Aware Image Matching for Large-Scale 3D Reconstruction

A. Overview of the Approach

The photogrammetric process for generating 3D models from rover involves four consecutive stages: feature matching, bundle adjustment, dense matching, and point cloud generation. Specifically, bundle adjustment ensures the consistency of images both across and within stations, based on the tie-points extracted through sparse image matching. Dense matching aims to match images on a pixel-by-pixel basis to produce disparity images, which are then used for subsequent point cloud generation. However, the textureless Martian surface and significant variations in resolution and viewpoint hinder both sparse and dense matching.

In this paper, we surmount the aforementioned challenges by leveraging semantic information. The overview of the proposed semantic-aware image matching algorithm is illustrated in Fig. 1. Beginning with images captured from different stations, the Siamese transformer-based neural network is employed to extract multi-level semantic cues. Regarding tie-point matching, keypoints are detected using SuperPoint [37], and initially described by the semantic features derived from the segmentation network. These descriptors are then augmented with contextual information derived from distinct keypoints considering the semantic segments through the attention mechanism. These keypoints are matched through soft partial assignment to establish one-to-one correspondences [43], and are filtered by measuring the distance of the semantic probabilities. Subsequently, the matches are aggregated into tie-point tracks, followed by integrated bundle adjustment to connect cross-station images by inferring consistent EOs. Based on the refined EOs, dense matching is conducted for each stereo image pair to compute the disparity images. Within the proposed framework, the images are initially transformed into the frequency domain for phase correlation calculation, and successively incorporated with the semantic probabilities for similarity measurement. The semantic boundaries, extracted from the segments, are utilized to adapt the parameters and preserve the discontinuities between the landforms. From the disparity images, point clouds are calculated through space intersection, and thereafter interpolated to produce 3-D models, which are then textured with the corresponding imagery.

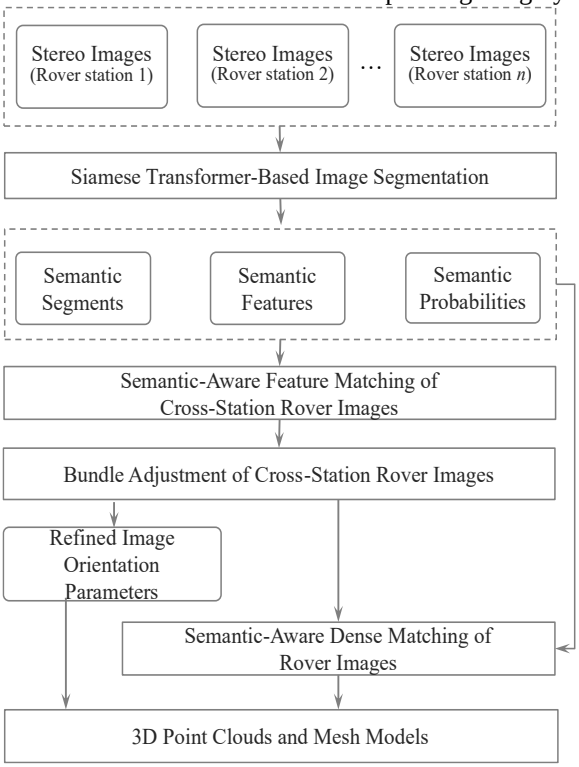


Fig. 1. Overview of the proposed approach.

B. Transformer-Based Semantic Segmentation of Images

To automatically segment the large number of images into semantic classes and extract multi-level semantic cues for subsequent processing, a neural network for semantic segmentation is the prerequisite. The cutting-edge Swin-Transformer framework [44], which combines the strengths of both convolutional and transformer-based networks, is exploited as the backbone of the segmentation network. Consistent with our previous work [45], the network is designed in a Siamese architecture, taking two overlapping images as the input, as illustrated in Fig. 2.

For each branch in the network, we employ Swin-T [44] as the backbone and utilize the UperNet framework [46] to integrate multi-scale features. This framework combines a feature pyramid network (FPN) [47] with a pyramid pooling module (PPM) [48] specifically applied to the final layer. The feature maps are then concatenated and fed into a convolutional operator to generate the semantic features with a 256-dimensional representation. The outputs are then convolved and normalized using a Softmax function [49], yielding the semantic probabilities. Lastly, the semantic segments are obtained using the Argmax operator.

With respect to the training process, instead of relying solely on input labels to supervise the network through the cross-entropy loss [50], tie-points are additionally computed to supervise the consistency in segments between input image pairs, and the loss function is constructed as:

$$\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{corre} \tag{1}$$

where $\mathcal{L}_{seg}$ and $\mathcal{L}_{corre}$ refer to the segmentation and corresponding loss, respectively. And $\mathcal{L}_{corre}$ is also functioned based on the cross-entropy loss. Accordingly, the network can be trained in a semi-self-supervised manner, which makes it robust against misalignment in the manually annotated labels.

As described in our previous work [45], a dataset comprising approximately 500 Zhurong rover images with 2048×2048 pixels was manually annotated, which was subsequently augmented into ~25,000 images with 512×512 pixels through a combination of 2D image augmentation and 3D space augmentation techniques. Specifically, 2D augmentation involves a series of image transformations, including translation, rotation, and homography transform, whereas 3D augmentation is based on roaming within a 3D mesh model, given a camera defined by interior orientation (IO) and EO parameters.

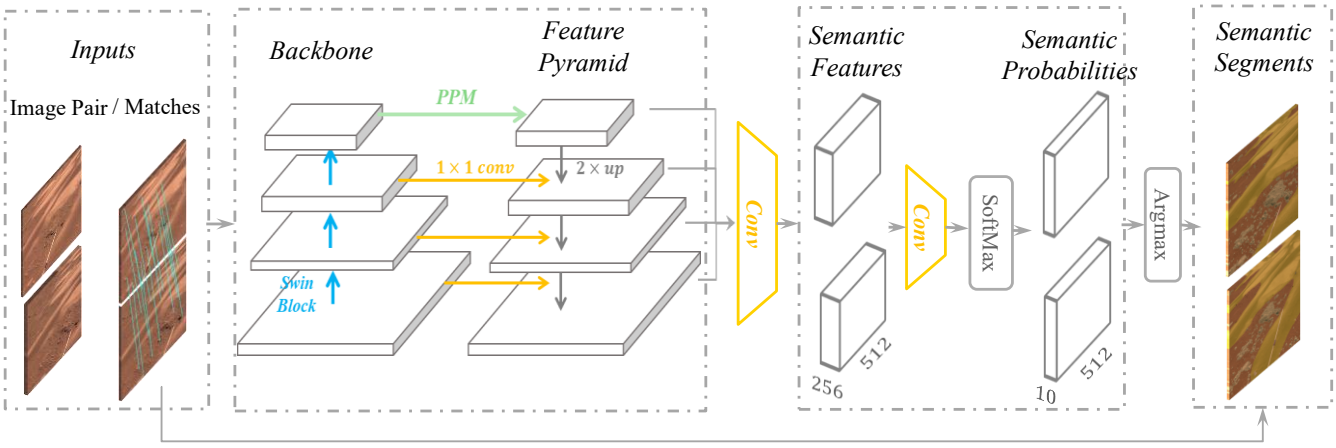


Fig. 2. Siamese Swin-transformer-based segmentation network.

C. Semantic-Aware Feature Matching between Cross-Station Images

Rover images are typically attached with an auxiliary file providing the nominal position and pointing data recorded by the onboard IMU, which is not sufficiently accurate for direct photogrammetric use. Consequently, bundle adjustment is necessary to align the cross-station images and adjust the nominal EO parameters. Whereas obtaining a sufficient number of tie-points is still a non-trivial task. The reasons for this are two-fold. Firstly, the Martian surface is predominantly covered by soil, rocks, or sand, exhibiting minimal variations in grayscale. The descriptors of the feature points are thereby alike, resulting in ambiguity in the subsequent matching process. Secondly, changes in the resolution and perspective cause the same landforms to look divergent on cross-station images, which may limit the effectiveness of locally-based descriptors for feature matching.

As shown in Fig. 3, semantic cues are integrated to enrich grayscale information throughout the matching process. Rather than relying exclusively on semantic segments for image masking, multi-level semantic cues derived from the segmentation network are leveraged. These multi-level cues not only enhance the distinctiveness of semantic information but also remain robust to potential errors in segmentation, ensuring reliable matching and reconstruction results even in the presence of imperfect segmentation. The key innovations include:

(1) The semantic features extracted from the transformer-based segmentation network are exploited to initialize the descriptor of the keypoints, incorporating both local and global features across multiple scales, which are then embedded with the scaled positional encoder to form the descriptor;

(2) The semantic segments are incorporated into the attentional aggregation module to help selectively focus on

the salient keypoints that are likely to be observed in both images, thus providing consistent information;

(3) As the matching network may yield mismatches, semantic constraints are used to ensure the reliability of the output tie-points. The distance of the semantic probabilities for each keypoint is calculated, and matches with large distances are removed.

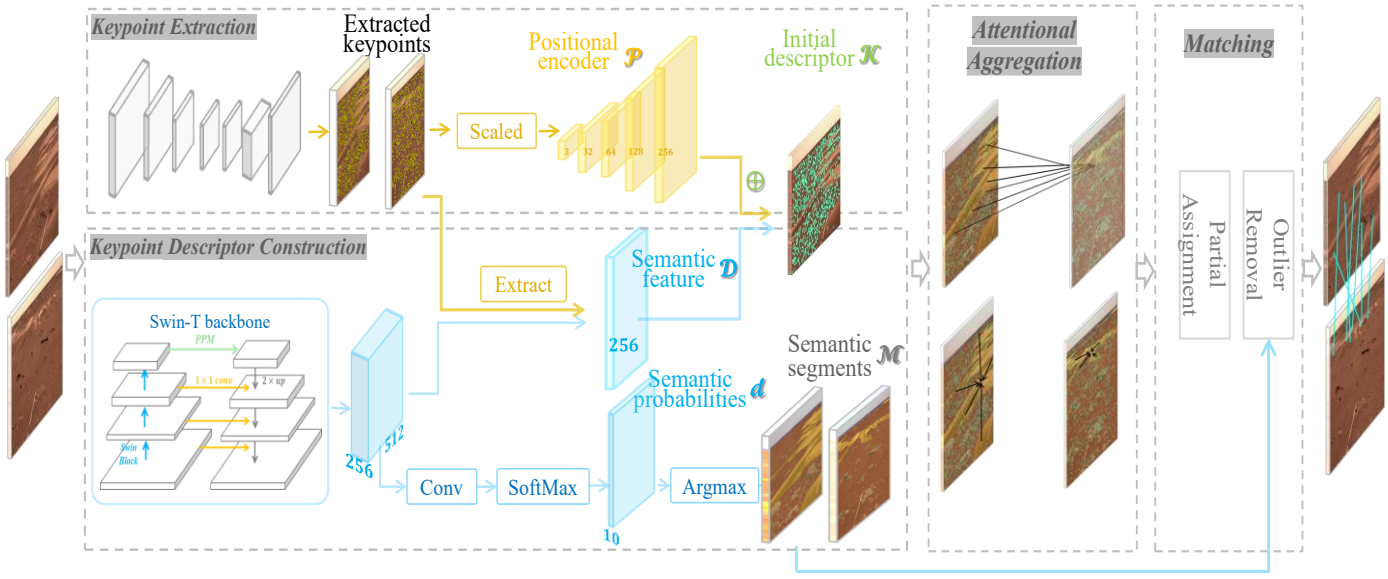


Fig. 3. Semantic-aware SuperGlue for feature matching of cross-station rover images.

The semantic-aware SuperGlue algorithm follows the state-of-the-art matching framework established by Sarlin et al., [22]. Initially, the images first go through a VGG-style backbone and decoded to retrieve the 2-D positions of the feature points [36]. The position of the i^{th} point is then extended to the 256-D positional encoder $\boldsymbol{\mathcal{P}}_i$, with descriptors $\boldsymbol{\mathcal{D}}_i$ extracted from the semantic features calculated from the segmentation network. The use of separate backbones is due to the semi-global merit offered by the segmentation network. Specifically, the original SuperGlue algorithm directly utilizes SuperPoint along with its descriptor, with both the detector and descriptor sharing the same VGG-style backbone. Although this approach is convenient and lightweight, it exhibits limited expressive capability because the convolution operations are localized, even with the pooling operator. In contrast, the FPN using Swin-T as the backbone incorporates a hierarchical structure designed to extract multi-scale features. Besides, the use of self-attention mechanisms captures long-range dependencies within the data and results in more dynamic and globally-aware feature representations. The distinctiveness of the descriptor $\boldsymbol{\mathcal{K}}_i$ is further improved by incorporating the positional encoder:

$$\boldsymbol{\mathcal{K}}_i = \boldsymbol{\mathcal{D}}_i + \mathrm{MLP}\ (\Pi(\boldsymbol{\mathcal{P}}_i)) \tag{2}$$

where MLP refers to multiple layer perceptron that increases the dimension of the positional encoder to 256 dimensions to be contacted with the $\boldsymbol{\mathcal{D}}_i$. The 2D image coordinates are augmented with the scale information by $\Pi(\cdot)$. Given the IOs and approximate pointing information, the resolution of each line and row can be estimated, enabling the use of the scaled coordinates by multiplying the image coordinates with the approximate resolution.

Regarding the attentional graph neural network, the original SuperGlue algorithm considers all keypoints in a brute-force manner. However, this seems ambiguous the descriptor as the keypoints detected in the image pair may differ significantly, thus distracting the attention mechanism. Consequently, the matches derived from the original algorithm may be affected by variations in keypoint extraction. Using the results of the semantic segmentation, the keypoints are associated with semantic labels, allowing the attention mechanism to focus on predominant landforms. For example, sand dunes or craters may be regarded as landmarks that provide robust contextual cues to enrich the initial descriptor. The aggregation can thus be described as:

$$\boldsymbol{\mathcal{D}}_i' = \boldsymbol{\mathcal{D}}_i + \mathrm{MLP}([\boldsymbol{\mathcal{D}}_i || attn(\boldsymbol{\mathcal{D}}_i, \mathcal{M}_{seman})]) \tag{3}$$

where $[\cdot \,||\, \cdot]$ represents to the concatenation operation, and attn$(\cdot)$ is the attention mechanism based on the encoded keypoint descriptors $\boldsymbol{\mathcal{D}}_i$ and the semantic mask $\mathcal{M}_{seman}$ sized $k_1 \times k_2$,, where k_1 and k_2 are the numbers of retrieved keypoints in each image. The cells corresponding to those within the distinct semantic class or salient keypoints with high scores are also set as one. The mask is multiplied by the attention weights calculated from the query and key derived from the descriptors. The attention weight then pertains to the Softmax over the semantic-masked query similarities.

The descriptors are subsequently passed through the matching module, which calculates scores for each keypoint with all other keypoints in the image pair to form a score matrix. The one-to-one matches are derived using the differentiable Sinkhorn algorithm [43]. As the SuperGlue framework does not include an outlier removal module, the tentative matches may include incorrect matches. While some studies directly enforce

semantic consistency between the keypoints, it is challenging to ensure pixel-wise consistency between the semantic segmentation results. Hence, the outputs of the vectors from the decode head after the Softmax operation are used for comparison. These vectors, which have the same dimensions as the number of semantic classes, describe the probability of classifying the pixel into its respective class. The Euclidean distance is then calculated, and the vectors within the threshold are preserved.

In terms of implementation, the two datasets used for segmentation are also used for training. The training is conducted using two NVIDIA GeForce RTX 3090 GPUs, with one handling semantic segmentation and the other running the matching algorithm. AdamW is used as the optimizer for 100 epochs with a learning rate of 0.0003.

D. Bundle Adjustment of Cross-station Rover Images

With a sufficient number of cross-station tie-points, bundle adjustment can be performed to align both the cross- and inner-station images. The feature track is subsequently generated based on the pairwise matches retrieved above. As indicated in Equation (4), two types of constraints are considered in the bundle adjustment process. As indicated in Equation (4), two types of constraints are considered in the bundle adjustment process.

$$\begin{cases} \mathcal{R}_{inner} = w_{inner} \sum_{k_{inner}} \sum_{i} \left\| \Pi(K_i, X_k) - \boldsymbol{x}_i^k \right\| \\ \mathcal{R}_{cross} = w_{cross} \sum_{k_{cross}} \sum_{i} \left\| \Pi(K_i, X_k) - \boldsymbol{x}_i^k \right\| \\ \mathcal{R}_{GCP} = w_{GCP} \sum_{k_{GCP}} \sum_{g} \left\| \Pi(K_g, \boldsymbol{X}_k) - \boldsymbol{p}_k \right\| \end{cases} \tag{4}$$

where $\boldsymbol{x}_i^k$ denotes the 2D position of the k^{th} match track on the i^{th} image. The unknown 3D coordinate X_k is projected onto the original image through $\Pi_i(K_i, X_k)$, where K_i takes into account the rotation R, translation T, and IO parameters. Despite the use of the proposed semantic-aware SuperGlue algorithm, the number of cross-station tie-points remains limited owing to the restricted overlapping region. To address this, weight parameters w_{inner} and w_{cross} are introduced to amplify the influence of cross-station tie-points, thereby compensating for the imbalance in match quantities [21]. Furthermore, ground control points (GCPs) are digitized for the salient points referenced in satellite images (e.g., HiRISE or CTX) to register the images to absolute geographic coordinates, through adjusting the projection $\boldsymbol{p}_g^k$ of the k^{th} GCP $\boldsymbol{G}_k$ on the g^{th} image. The EO parameters and sparse point clouds can be calculated by minimizing the L2 norm of the sum of these residuals.

E. Semantic-Aware Dense Matching of Rover Images

Given the consistent exterior orientation parameters, the burdensome 2D pixel-wise image correspondence problem can be simplified into a one-dimensional task through epipolar rectification [51]. This process utilizes the interior and exterior orientation parameters to compute the epipolar geometry between the image pairs, ensuring the corresponding points lie on the same horizontal scanline in these epipolar image pairs.

However, the computational demands of the algorithm remain expensive. And the industry-proven semi-global matching (SGM) algorithm uses dynamic programming (DP) from multiple directions to approximate the results [26]. The problem for each direction can be formulated as an energy minimization task, as:

$$E = w_{data} E_{data} + w_{smooth} E_{smooth} \tag{5}$$

where the overall energy E is the sum of the data cost E_{data} established on the similarity measurement and the smoothness cost E_{smooth}, which assesses the disparity continuity between neighboring pixels. Two weight parameters w_{data} and w_{smooth} are introduced to balance the contribution of these two terms. Since both terms are likely to be perturbed by the textureless surface, Fig. 4 illustrates the proposed approach to incorporate semantic cues to achieve improved disparity images.

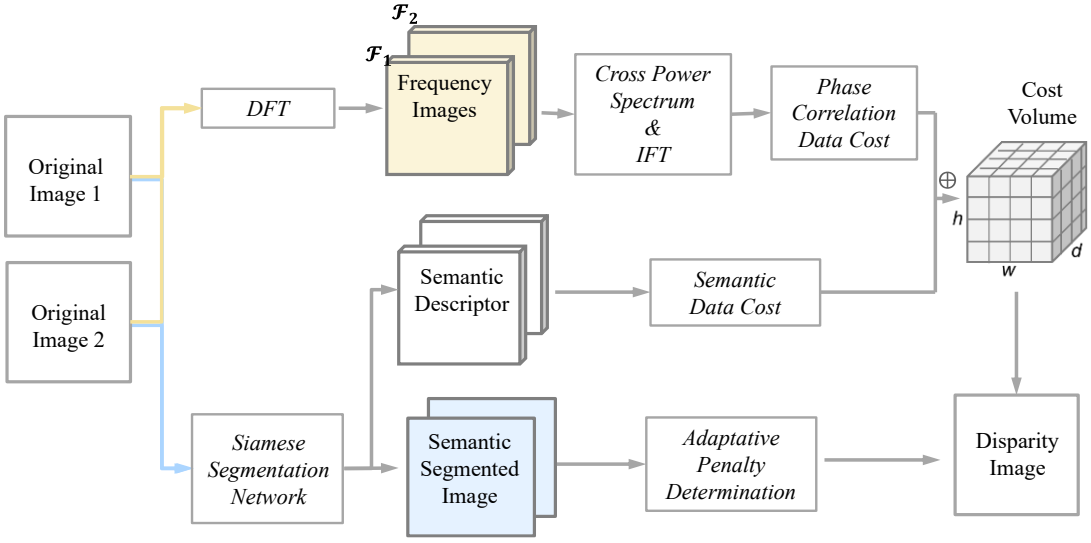


Fig. 4. Overview of the semantic-aware dense matching algorithm.

The calculation of E_{data} is typically based on a small patch with a specific center pixel and window size, using methods such as AD-Census [52] or normalized correlation coefficients in the spatial domain [31]. In contrast, the proposed algorithm transforms the patches into the frequency domain to exploit phase correlation [53] for similarity measurement. This approach is beneficial for two reasons. Firstly, the superior capability of phase correlation for image matching in textureless regions has been substantiated by numerous studies [27], [40], [54], particularly relevant to the Martian surface. Second, phase correlation directly provides the translation movement between two patches along with the similarity measurements. As only translation transformation exists in the epipolar image pair, the merit of the phase correlation is thus highlighted. To compute the phase correlation between patch $p_1(x, y)$ on the image and patch $p_2(x - a, y)$, the discrete Fourier transform (DFT) is first applied to obtain $\mathcal{F}_1(u, v)$ and $\mathcal{F}_2(u, v)$ for the following normalized cross-power spectrum $Q(u, v)$ calculation.

$$Q(u,v) = \frac{\mathcal{F}_1(u,v)\mathcal{F}_2^*(u,v)}{|\mathcal{F}_1(u,v)\mathcal{F}_2^*(u,v)|} = e^{i(au)} \tag{6}$$

The complex conjugate operator is denoted by $*$. The integral shift a can be revealed by transforming the $Q(u, v)$ to a Dirac delta function through inverse Fourier transform (IFT) $\mathcal{FT}^{-1}(\cdot)$. The phase-correlation-based data cost considers both the visual appearance and the positional translation and is expressed as:

$$Cost_{data}^{PC} = \begin{cases} \left(1 - \mathcal{FT}^{-1}\left(Q(u,v)\right)\right) * \lambda_1, & if\ a \le 1 \\ \lambda_2, & else \end{cases} \tag{7}$$

where λ_1 and λ_2 are enlarging factors to amplify the difference of the patches with translational movement less or greater than one pixel. In addition, a semantic data cost is established, akin to the semantic consistency referenced during outlier removal in feature matching. This approach utilizes an n dimensional descriptor and computes the Hamming distance, which is later concatenated to the phase-correlation-based cost to construct the overall cost volume.

The smooth cost $Cost_{smooth}$ is usually constructed as Equation (8), which is separated into the small and large disparity change parts, as:

$$\begin{aligned} Cost_{smooth} = \sum_{(i,j)\in I} \sum_{(i',j')\in \mathcal{N}_{(i,j)}} & \left(P_1^{\mathcal{E}}(i,j) \cdot \boldsymbol{C}^{\boldsymbol{r}}\left[\left(d_{(i',j')} - d_{(i,j)}\right) = 1\right]\right) \\ & + \left(P_2^{\mathcal{E}}(i,j) \cdot \boldsymbol{C}^{\boldsymbol{r}}\left[\left(d_{(i',j')} - d_{(i,j)}\right) > 1\right]\right) \end{aligned} \tag{8}$$

where the operator $\boldsymbol{C}^{\boldsymbol{r}}[\cdot]$ is zero if the specified condition $[\cdot]$ is met, and one otherwise. For each pixel (i, j) in image I and pixel (i', j') in its neighborhood $\mathcal{N}_{(i,j)}$, if the difference between the disparities $d_{(i,j)}$ and $d_{(i',j')}$ is within one pixel, a penalty P_1 is imposed. For differences larger than one pixel, a larger penalty P_2 is applied. As manually tuning these parameters is time-consuming and may not be suitable for all regions in the images, adaptation strategies are used to preserve discontinuities and enforce smoothness in the textureless regions. The semantic edges $\boldsymbol{\mathcal{E}}$ are hence incorporated to extend previous texture-aware parameter adaptation strategies based on Canny edges [29], [30], [31], as:

$$P_1^{\mathcal{E}}(i,j) = \begin{cases} P_{11}(i,j), & if\ \mathcal{E} \in Edge\ Region \\ P_{11}(i,j) + P_{12}, & else \end{cases} \tag{9}$$

where P_{12} is a fixed penalty, $P_{11}(i, j)$ is dynamically determined based on the Softmax probability values $Prob(i, j, C)$ from the semantic segmentation network:

$$P_1(i,j) = \sum_C Prob(i,j,C) \cdot P^C \tag{10}$$

Where $Prob(i, j, C)$ is the probability of pixel (i, j) belonging to class C, and P^C is the predefined penalty value for class C. To reduce computational complexity, we define semantic edge regions based on semantic segments. Specifically, we extract edge pixels from the segments and consider a τ -pixel neighborhood around these edges as the edge region:

$$Edge\ Region = \{(i,j) | distance\big((i,j), \mathcal{E}\big) \le \tau\} \tag{11}$$

The same strategy is also applied to P_2. Our implementation uses a hierarchical approach based on a resolution pyramid, where the disparity calculated at a coarser resolution level is used as a reference for the subsequent level, and a predefined search range further constrains the computation.

F. 3D Surface Reconstruction from the Matching Results

With the dense matching results, space intersection is performed using the collinearity equation [55], which establishes the mathematical relationship between the image coordinates and corresponding 3D object points. By solving this equation, the 3D coordinates of the object points can be accurately determined, thereby generating dense point clouds. These point clouds are then subjected to a series of processing steps to refine their quality and accuracy. Firstly, the point clouds are merged to combine the information from multiple image pairs. Next, a filtering process is applied to remove noise and outliers, which may arise from various sources such as image noise or matching errors. The point cloud is triangulated to form a 3D mesh model [56], which is then texture-mapped with corresponding images based on the retrieved camera EOs, resulting in a textured mesh model.

IV. Experimental Evaluation

A. Dataset Description

In this paper, data acquired from China's first Martian rover, Zhurong, is used to evaluate the performance of the proposed semantic-aware image matching algorithm. The rover is equipped with a stereo pair of navigation a stereo pair of navigation and terrain cameras (NaTeCam) [16], designed to capture the surrounding environment and ensure safe traversal. As listed in Table I, these cameras feature a focal length of 13.169 mm, a 27 mm baseline, and a high-resolution sensor with 2048 × 2048 pixels and size of 11.264 mm. Typically, the pointing direction of the cameras is tilted at an angle of 14°– 19° from the horizontal plane, allowing them to capture scenarios at a considerable distance from the rover. As the

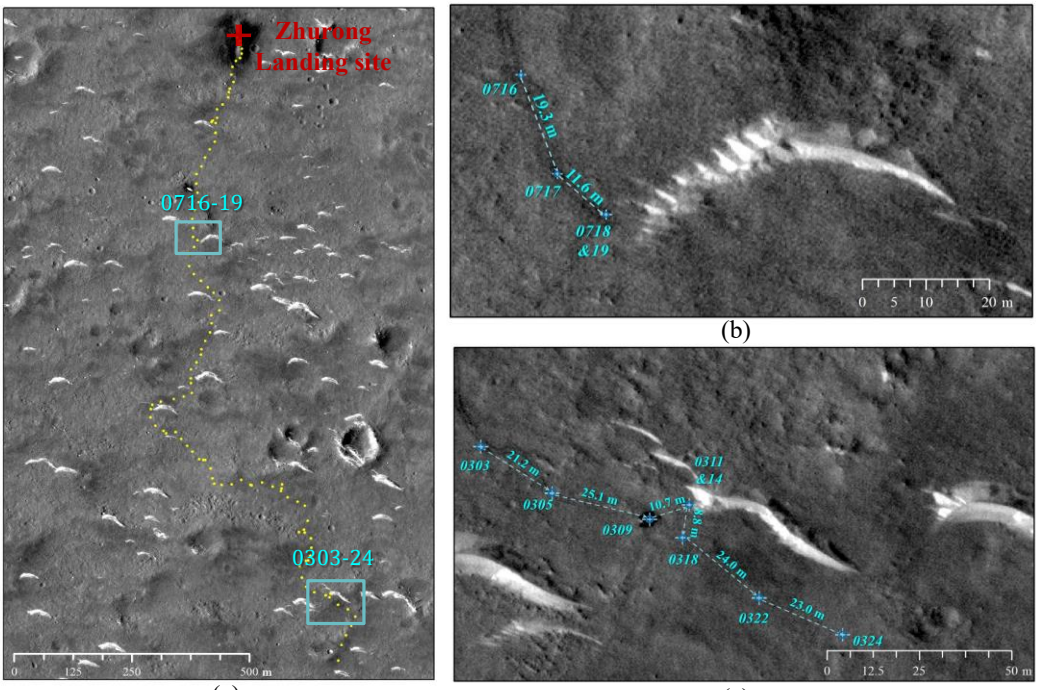
resolution varies with the angle and viewing distance, approximate resolutions are presented in the table. Additionally, there are a few stations with a horizontal tilt of nearly 29° , enabling the cameras to observe near-range scenarios with sub-centimeter resolution.

Two regions from the Zhurong traverse are exploited as the test areas, and their locations are marked in the HiRISE images in Fig. 5 (a). The detailed distribution of the stations in the two datasets is shown in Fig. 5 (b) and (c), and the corresponding statistical parameters are summarized in Table I. The 0716-19 dataset, consisting of 44 images acquired between July 16th and 19th, 2021, covers a distance of 40 m across four stations midway along the traverse. This region features a repetitive sand dune pattern, which renders image matching challenging. The 0303-24 dataset includes 78 images captured between March 3rd and 24 th, 2022, spanning a distance of over 100 m and encompassing eight rover stations at the end of the Zhurong traverse. The rover station selection employed a three-step process: (1) orbital-scale analysis using HiRISE image (0.25 m/pixel) to verify terrain homogeneity and characterizing dominant landforms (sand dunes/ rocks) [57], [58]; (2) science-driven prioritization through review of in-situ studies [59]; and (3) challenging image pair selecting based on quantitative assessment of the resolution, viewpoint and rotation variations.

TABLE I

DESCRIPTION OF THE TWO TEST DATASETS

Dataset	Date	Station Count	Image Count	Moving Distance (m)	Camera Parameters			
					Focal Length (mm)	Sensor Size (mm)	Baseline (cm)	GSD (cm)
0716-19	16th~19th July, 2021	4	44	30.9	13.169	11.264	27	0.5 (~2.5m) 1.0 (~ 5m) 4.0 (~10m) 10.0 (~15m) 20.0 (~25m)
0303-24	03rd~24th March, 2022	8	78	112.8				



(a) (b) (c)

Fig. 5. Distribution of the two test areas listed in Table I, labeled in a month-day format. (a) Illustration of the two test areas overlaid on the HiRISE image (ESP_073225_2055). (b) and (c) show the detailed station distributions for each dataset.

Fig. 6 presents six representative semantic segmentation results, yielded from the segmentation neural network. It is apparent that these test areas exhibit a diverse range of landforms, including rover, crater, rover tracks, soil, sand, rocks, far side, and other mechanical components. The diversity of landforms in these datasets makes them ideal for testing the algorithm, thereby validating its versatility and efficacy. The segmentation network accurately retrieves large and small rocks, and labels various types of sand dunes correctly, with only a few exceptions in the most distant regions. Even for the crater class, which is insufficiently labeled, the favorable results are still yielded. Furthermore, the correct segmentations suggest

the effectiveness of the semantic features and probabilities, thereby establishing a solid foundation for the subsequent semantic-aware feature and dense matching algorithms. Statistically, the mean intersection over union (mIOU) is 88.25%, demonstrating the accuracy of the segmentation [45].

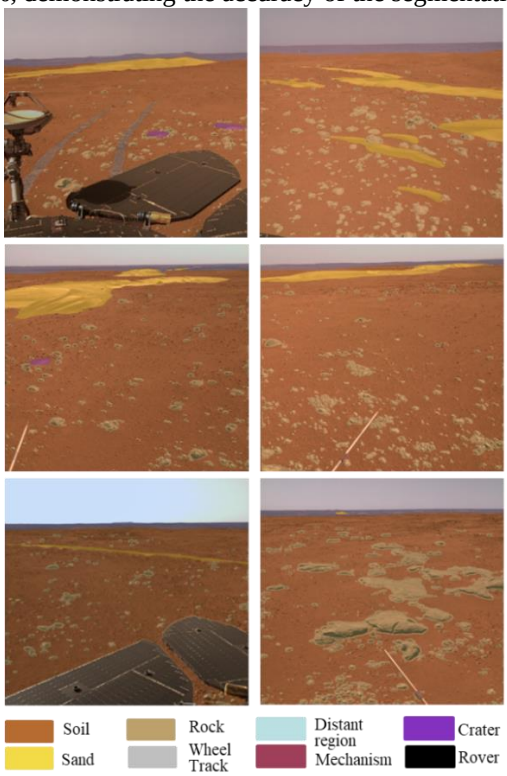


Fig. 6. Representative semantic segmentation results overlaid on the original images.

B. Evaluation of Semantic-Aware Feature Matching from Cross-station Rover Images

To comprehensively evaluate the performance of the proposed matching strategy, five representative matching pairs are selected from the two datasets, with three from the 0303-24 dataset and two from the 0716-19 dataset, respectively. The detailed information and the matching results are summarized in Table II. The viewing distance for these pairs is approximately 10 meters, yet the scenes undergo significant changes due to the rover's movement. Specifically, when the viewing vector is nearly parallel to the barren terrain, the image resolution becomes non-uniform, ranging from several millimeters to over 5 m scales. Even a small forward step can lead to significant scale changes. Accordingly, the resolution information provided in the table corresponds to the resolution marked by the yellow crosses in Fig. 7. This issue is further exacerbated by the large perspective variation, attributable to the camera's large field of view.

The proposed algorithm is benchmarked against representative state-of-the-art methods covering three key categories: (1) semantic-aware approaches (COTR [25] and Patch2Pix [24]), (2) low-texture-targeted method (LoFTR [23]), and (3) benchmark approach (SuperGlue [22]). While Patch2Pix leverages high-level semantic information at the patch-level, COTR provides a unique global semantic understanding through transformer architectures LoFTR [23] specializes in low-texture and repeated-pattern scenarios, a critical challenge for image matching of rover images. SuperGlue [22] serves as the de facto benchmark, providing fundamental performance baselines.

A comprehensive exploration of various parameter settings, including the number of keypoints, minimum distance between keypoints, and match threshold, is conducted, and the optimal configuration is selected for comparison, as illustrated in Fig. 8. Notably, the matches derived from these algorithms contain many outliers. Therefore, a manual check is conducted, and both the original matches and the inliers are reported in Table II. The original matches are marked by (·).

TABLE II
STATISTICS OF THE SEMANTIC-AWARE FEATURE MATCHING EXPERIMENTS

	Characteristic	Viewing Distance (m)	Viewing Angle Difference (°)	Resolution (m / pixel)		Match Count				
				img1	img2	Patch2Pix [24]	COTR [25]	LoFTR [23]	SuperGlue [22]	Ours
1	Pure resolution changes	10.82	5.40	0.006	0.02	40 (53)	78 (100)	107 (107)	52 (63)	67
2	Repeated pattern with a small overlap	11.06	30.11	0.01	0.05	0 (42)	0 (16)	6 (49)	5 (7)	41
3	Repeated pattern with viewing direction change	11.06	37.95	0.006	0.05	0 (26)	0 (12)	5 (39)	16 (21)	64
4	Soil region with little sand	10.82	28.81	0.007	0.05	5 (24)	0 (29)	10 (52)	18 (26)	39
5	Barren soil region	9.95	0.0	0.005	0.04	60 (77)	10 (38)	159 (159)	58 (69)	77

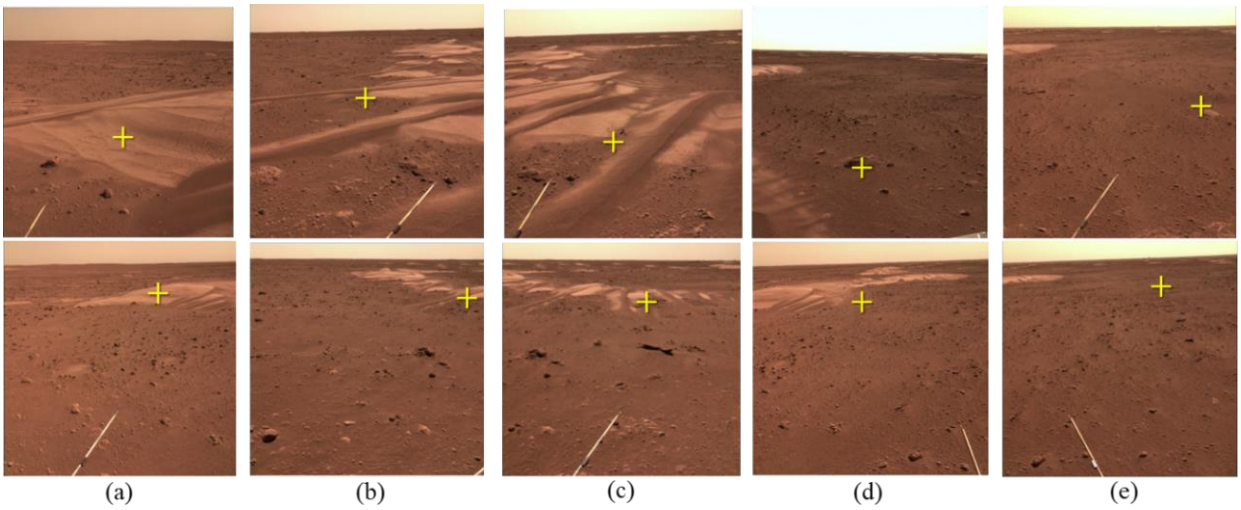


Fig. 7. Experiment image pairs for the semantic-aware SuperGlue experiment.

The first pair shown in Fig. 8 presents the region dominated by a sand dune, with the images taken from neighboring stations more than 10 meters apart, and the images exhibit significant differences at close range. The consistent distant view enables Patch2Pix, COTR, LoFTR, and the SuperGlue algorithm to retrieve a substantial number of accurate matches in the background. And SuperGlue retrieves a few correct matches in the lower-left corner of the rock. However, it also yields several wrong matches, especially in the sand dune region where large geometric distortion occurs. As the resolution increases dramatically with longer viewing distances, these unbalanced matches hinder bundle adjustment by introducing more precise matches. In contrast, the proposed algorithm successfully retrieves matches evenly across the image, even if the foreground varies due to scale changes. This success may be attributable to the scaled positional encoder, which clarifies the 3D relationships between the keypoints, and the robust descriptor. The third pair combines the challenges of the first and second pairs, with both large distortion and repetitive textureless patterns, rendering it even more difficult for humans to identify the correct matches. Results similar to those in the previous cases are observed, demonstrating the superiority of the proposed matching algorithm

The second pair features a region also dominated by repetitive sand dunes, with a similar viewing direction and an overlapping area that does not suffer from significant distorted transformation. While Patch2Pix and COTR fail to retrieve correct matches, both LoFTR and SuperGlue successfully identify approximately five matches in the far-range region, though they are still confounded by the rocks. The primary failure mechanism of Patch2Pix and COTR likely stems from the inherent semantic sparsity of Martian surface features, and significant semantic descriptor variance induced by rover mobility. In contrast, the proposed algorithm generates abundant matches that are evenly distributed across the overlapping sand dune region. The difference between the results highlights the expressiveness of the descriptor, which can effectively describe the keypoints in textureless terrain through the combination of a multi-scale feature pyramid and attentional aggregation of the distinct keypoints.

The last two pairs concern the common regions dominated by the soil, which is textureless and mostly filled with small rocks. The four benchmark algorithms detect many matches but merely in the obvious overlapped sand dune part, resulting in matches confined to a small area. Whereas the proposed algorithm detects abundant correct matches distributed across other regions. The fifth pair extremely lacks texture, and the successful matching highlights the effectiveness of the semantic cues. Besides the ability to retrieve correct matches, it is observed that incorrect matches are also prone to occur in this type of barren terrain, but they are effectively eliminated by comparing the semantic descriptors.

Quantitative analysis of the retrieved matches is performed based on bundle adjustment, assuming that if the matches exhibit high precision and reasonable distribution, the images from multiple stations can be consistently linked. Notably, the ContextCapture software cannot perform cross-station bundle adjustment owing to the lack of cross-station tie-points. And the comparison is hence conducted between the bundle adjustment before and after the introduction of the tie-points generated by our semantic-aware feature matching algorithm. Specifically, bundle adjustment is performed to resolve inconsistencies among the inner-station images based on the given EO parameters in the attached label files. These results serve as the baseline for comparing integrated bundle adjustment outcomes, and the residuals in the image space are used to evaluate the quality of the bundle adjustment. As presented in Table III, two types of tie-points are assessed: the in-use tie-points (i.e., the cross-station tie-points generated by the proposed algorithm), and checkpoints that are manually digitized and evenly distributed throughout the scene for in-depth evaluation. For each dataset, the residuals before the integrated bundle adjustment are large, exceeding 40 pixels for the 0303-24 dataset due to the stations spanning over 100 m, even when the 95th percentile evaluation is performed to eliminate abnormal observations. After the bundle adjustment, both the mean and root mean square errors (RMSEs) of the residuals drop to slightly over one pixel, suggesting a high accuracy and precision of the achieved matches.

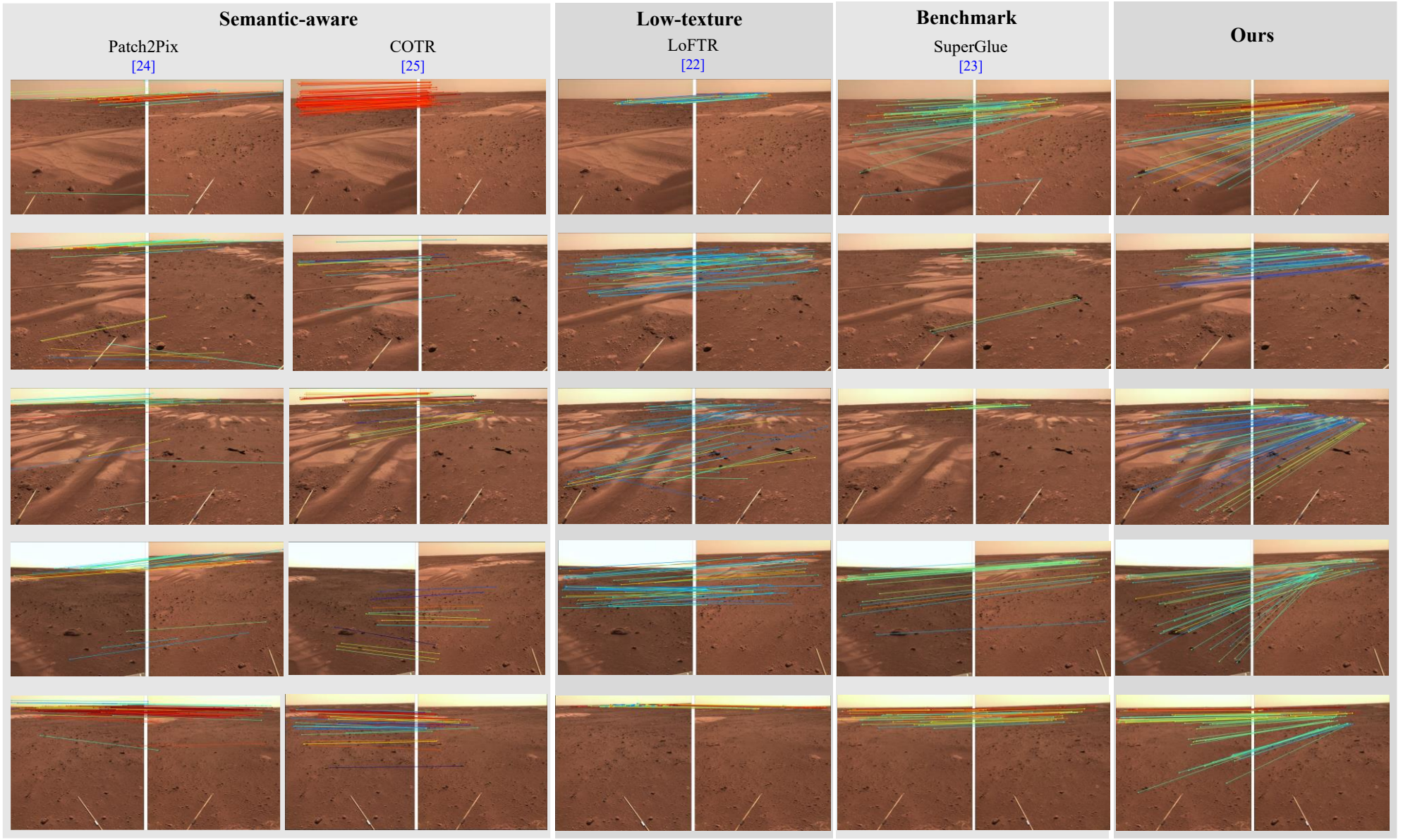


Fig. 8. Comparison of feature matching algorithms. The five columns are the results of Patch2Pix, COTR, LoFTR, SuperGlue and our algorithm.

TABLE III
COMPARISON OF IMAGE RESIDUALS OF TIE-POINTS

Dataset	Type of Cross-station Tie-points	Number of Cross-station Tie-points	Average Linked Image Count	Mean Residuals (95%) (Pixels)		RMSE of Residuals (95%) (Pixels)	
				Before	After	Before	After
0716-19							
	In-use	1500	4.6	8.03 (3.22)	1.43 (1.23)	16.79 (4.54)	1.62 (1.34)
	Check	122	5.3	19.52 (14.50)	1.78 (1.52)	25.86 (18.37)	1.87 (1.55)
0303-24							
	In-use	4552	7.75	42.04 (36.63)	1.75 (1.36)	57.09 (50.23)	1.82 (1.39)
	Check	350	8.5	54.19 (47.42)	1.58 (1.52)	64.64 (56.31)	1.62 (1.54)

C. Evaluation of Semantic-Aware Dense Matching of Rover Images

To illustrate the effectiveness and the merits of the proposed semantic-aware dense matching algorithm, two representative regions from each dataset are selected. Each region is captured by two stereo pairs, totaling four images, as shown in Fig. 9. The results are compared with the disparity results obtained using the conventional AD-Census method as well as the semantic-aware AD-Census method to demonstrate the advantages of the semantic-aware approach and the necessity of phase correlation. In the absence of ground truth data, two quantitative indicators are proposed to evaluate the dense matching algorithms: completeness and accuracy. First, the percentage of valid pixels in the disparity maps is calculated to measure the completeness of the disparity images. To highlight the differences among the results, the incompleteness is calculated. Second, we manually digitize a boundary edge of the sand dune where topographic discontinuity occurs on the epipolar image and compare it with the corresponding edge in the disparity map. The average distance between these edges quantifies alignment accuracy, validating semantic edge preservation. Furthermore, the 3D mesh model generated from these disparity images are also visualized for comparison. The 3D mesh model used here is a subset derived from merging neighboring point clouds, thereby clearly highlighting the advantages of the proposed algorithm.

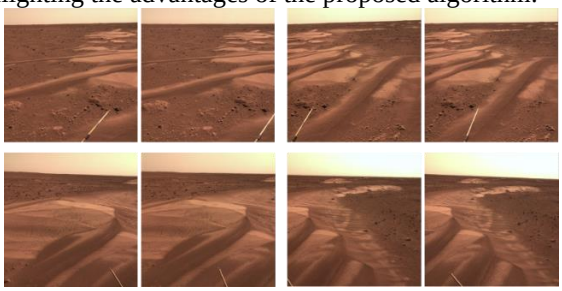


Fig. 9. Overview of images used for the semantic-aware dense matching algorithm. The first and second rows correspond to the 0716-19 and 0303-24 datasets, respectively.

The results of the 0716-19 dataset are presented in Fig. 10, the line for accuracy analysis is marked by black arrow. The first row shows the disparity maps, while the second and third rows present the 3D mesh models generated from the disparity images exploiting the EO parameters through space intersection. As shown in the first column, the disparity image obtained using AD-Census suffers from no-data and speckle issues at the boundary regions and in areas where occlusion occurs, resulting in 2.2% incompleteness and a divergence of 8.46 pixels from the ground truth at the boundary edge of the sand dune. Even when the disparity image appears smooth, noise is evident in the 3D mesh. Through the introduction of semantic cues, the textureless regions are enriched with additional information, mitigating the noise problem, as shown in the second column, both in the middle of the sand dunes and in the ridge region. As a result, the incompleteness is reduced to 1.2%, and the accuracy is improved to a divergence of 3.73 pixels. Especially in the zoomed-in view area in the third row, the two small sand dunes in the background are not as clearly reconstructed as those computed using the semantic information. In the distant region, the red peak of the sand dune, over 20 m from the viewpoint, suffers from significant blurriness and resolution changes. Nevertheless, the segmentation network can effectively segment the sand dunes, thereby enhancing the clarity of the left ridge of the peak, making it more visually similar to the original image. The ridges in the enlarged view still exhibit distortions, however, these artifacts are reduced through the phase-correlation-based semantic-aware dense matching, which decreases incompleteness to 0.8% and improves accuracy to a divergence of 2.56 pixels. This improvement highlights the superiority of the frequency domain in describing textureless regions. Moreover, the subtle details (i.e., wrinkles of the sand dunes and rock boundaries) appear clearer in the phase-correlation-based mesh model.

The abovementioned issues become more apparent in the 0303-24 dataset, which is dominated by a large sand dune that not only lacks texture but also exhibits severe occlusions. Additionally, the dark ridge is inherently prone to be confused with the background soil region, exacerbating the difficulty of preserving the discontinuities. As illustrated in Fig. 11, the disparity image is thus noisy and fragmented leading to a wavy ridge in the reconstructed mesh model, and the unexpected

protrusions in the middle of the sand dune. By enforcing semantic edges, the boundary between the sand dune and soil region is clarified enabling the continuous reconstruction of the ridge of the sand dune and preservation of the discontinuities in most areas. However, the semantic-aware approach implemented in the spatial domain still cannot eliminate all the defects (i.e., the region marked by the black bounding box), as also observed by the above 0716-19 dataset. The results generated by the proposed approach show that the disparity images are smoother compared with the other images, and the mesh model more accurately aligns with the content captured in the original image. These visual effects are supported by quantitative results. With semantic cues, incompleteness and accuracy improve to 1.8% and 3.61 pixels, respectively. Phase correlation further enhances these metrics, reducing incompleteness to 0.8% and improving accuracy to 2.29 pixels.

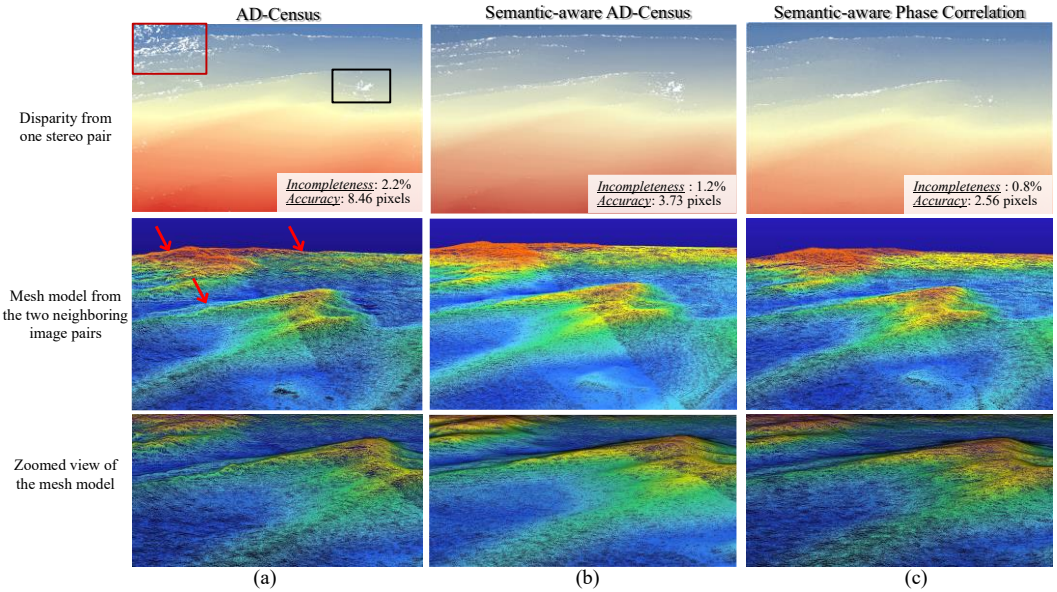


Fig. 10. Comparison of dense matching results. (a)-(c) are the results calculated from AD-Census alone, semantic-aware AD-Census, and our proposed semantic-aware phase-correlation, respectively. The first row shows the disparity images from one stereo pair, the second row presents the 3D meshes, and the third row displays the zoomed views are displayed in the third row. The striped pattern in the mesh model is a result of varying point density, attributable to certain regions being captured by two stereo images and the others observed from four different angles.

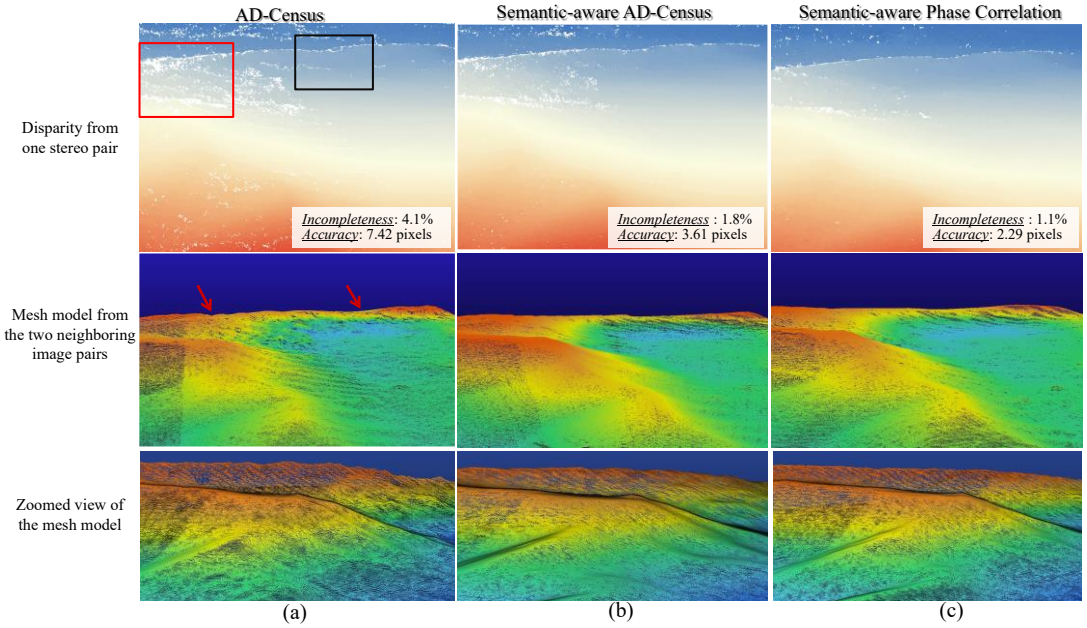


Fig. 11. Comparison of dense matching results. (a)-(c) are the results calculated from AD-Census alone, semantic-aware AD-Census, and our proposed semantic-aware phase-correlation, respectively. The first row shows the disparity images, the second row presents the corresponding 3D meshes, and the zoomed views are displayed in the third row.

D. Evaluation of the Multi-station 3D Surface Reconstruction Results

Based on the proposed semantic-aware cross-station feature matching and inner-station dense matching algorithm, large-scale 3D reconstruction is performed using all images in each dataset, and the generated DEMs are shown in Fig. 12. Owing to the GCP constraints in the bundle adjustment, the generated DEMs are well aligned with the reference HiRISE ortho-image. For other rover data, an automated localization algorithm [17] can be used to co-register the rover-derived products with satellite images for evaluation. Each DEM covers a large area, including not only extinct sand dunes but also extensive bare soil regions. The absence of apparent gaps in either DEM further demonstrates the accuracy of the integrated bundle adjustment.

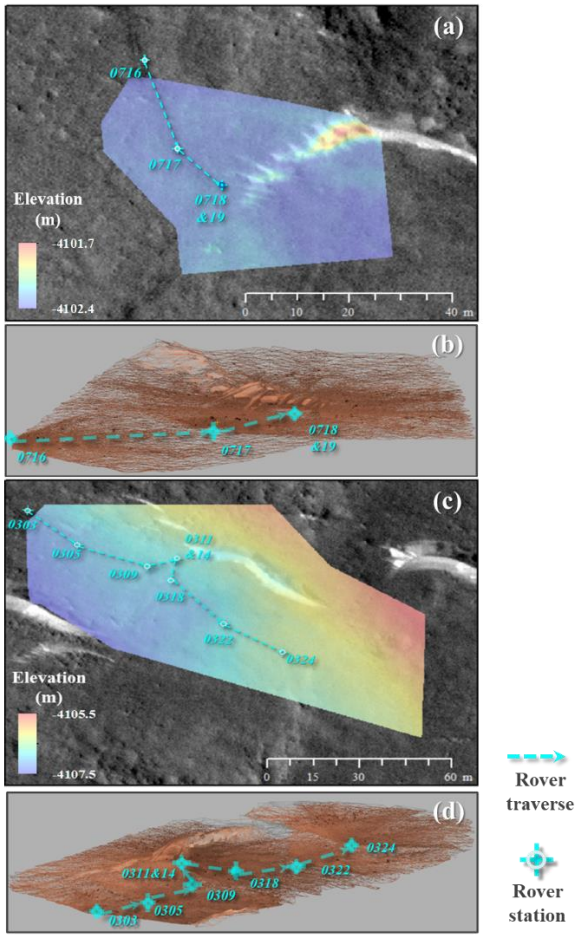


Fig. 12. The multi-station 3D reconstruction results generated using the proposed approach. (a, c) DEMs and (b, d) wireframe meshes for the (a, b) 0716-19 (4-station) and (c, d) 0303-24 (8-station) datasets, demonstrating multi-station linkage capability.

To ensure 3D geometric accuracy, a further evaluation is conducted using the HiRISE image (0.25 m/pixel) as the GT. Specifically, distances are measured from the HiRISE image and compared with those calculated from our results. Given the significant scale variation, five distinct points are carefully digitized from the HiRISE image for this evaluation. One of these points P_0, is designated as the origin, and the distances from this origin to the remaining four points are calculated for comparative analysis. The detailed distance measurements are presented in Table IV, while the spatial distribution of these points is illustrated in Fig. 13. For the 0716-19 dataset, the distances range from 3 to 50 m, while the points in the 0303-24 dataset are sampled across a range of 8 to 90 m. All measurement differences are within 0.5 meters, confirming the accuracy of the retrieved EO parameters and effectiveness of the large-scale bundle adjustment with cross-station tie-points.

To ensure 3D geometric accuracy, a further evaluation is conducted using the HiRISE image (0.25 m/pixel) as the GT. Specifically, distances are measured from the HiRISE image and compared with those calculated from our results. Given the significant scale variation, five distinct points are carefully digitized from the HiRISE image for this evaluation. One of these points, P0, is designated as the origin, and the distances from this origin to the remaining four points are calculated for comparative analysis. The detailed distance measurements are presented in Table IV, while the spatial distribution of these points is illustrated in Fig. 13. For the 0716-19 dataset, the distances range from 3 to 50 m, while the points in the 0303-24 dataset are sampled across a range of 8 to 90 m. All measurement differences are within 0.5 meters, confirming the accuracy of the retrieved EO parameters and effectiveness of the large-scale bundle adjustment with cross-station tie-points.

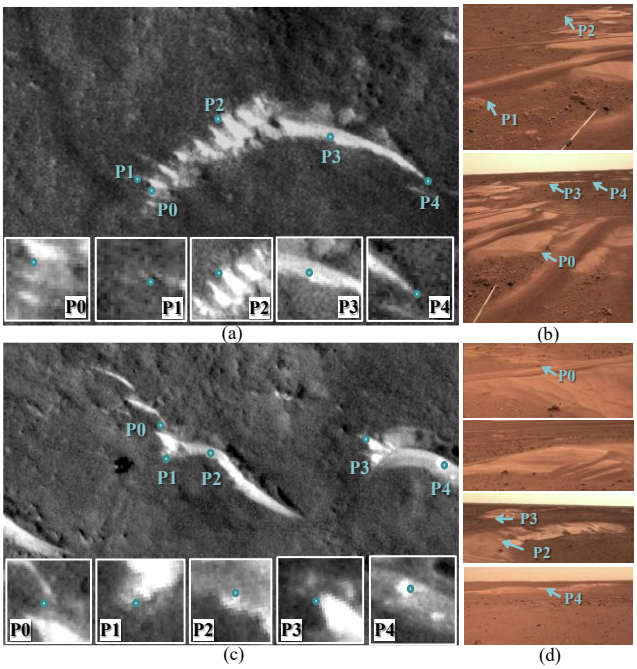


Fig. 13. The distribution of selected points for absolute 3D position verification. (a) and (b) mark points on satellite and rover images for the 0716-19 dataset, and (c) and (d) mark points for the 0303-24 dataset.

TABLE IV
ABSOLUTE 3D SCALE EVALUATION WITH REFERENCE TO THE HIRISE IMAGE

0716-19									
	P_1		P_2		P_3		P_4		Mean Difference (m)
	GT	Ours	GT	Ours	GT	Ours	GT	Ours	
Distance (m)	3.30	3.01	17.58	17.08	33.44	33.76	48.44	47.90	0.41
0303-24									
	P_1		P_2		P_3		P_4		Mean Difference (m)
	GT	Ours	GT	Ours	GT	Ours	GT	Ours	
Distance (m)	8.00	8.04	16.76	17.03	64.96	65.61	89.35	90.15	0.44

GT: Ground truth measured from HiRISE ortho-rectified images.

A comparison of the 3D textured mesh model is also conducted using the state-of-the-art photogrammetric software, ContextCapture, to demonstrate the superiority of the proposed algorithm in achieving optimal 3D reconstruction.

Fig. 14 presents regions from the 0716-19 dataset. The first column shows subsets from the original image, while the second and third columns display the textured 3D mesh models generated by ContextCapture and the proposed algorithm, respectively. To better illustrate the differences between the mesh models, the viewpoints are not strictly aligned with the image subsets, which are provided solely to clarify the ground-truth situation. The first two rows present a near-range scenario where the small rocks reconstructed by ContextCapture are all retrieved by the proposed algorithm. However, significant divergence is observed in the sand dune and soil regions. Similar to the mesh model generated by the spatial domain dense matching algorithm, the 3D mesh model produced by ContextCapture also suffers from a wrinkled effect in textureless areas with twisted ridges, as pointed out by the blue arrows. In contrast, our model is more consistent with the original image. Furthermore, a far-distance scenario is also tested in the third row, where the end of the sand dune is more than 30 m away from the nearest image viewpoints. Even though ContextCapture fails to calculate the correct 3D mesh, the proposed algorithm can generate a mesh model that extends further and maintains a reasonable shape, aligned with the original image.

Three representative regions are also selected from the 0303-24 dataset, as visualized in Fig.15. In the close-range scene depicted in the first row, noticeable holes in the middle of the small sand dune result in incorrect geometry. Our algorithm can distinguish between the sand dunes and the soil region, thus reconstructing the proper 3D mesh. In the 30-m scenario presented in the second row, while ContextCapture also reconstructs the end of the sand dune, many apparent artifacts are evident. Notably, the region indicated by the blue arrow, where the ridge is expected to descend based on the image, is not correctly retrieved. The last row presents an extreme case, where the region is approximately 40 m away from the viewpoint. Although the distant region is entirely missed by the software, our approach successfully reconstructs the area within 40 m.

Two quantitative indicators are used to compare the models generated by ContextCapture and our method: incompleteness and farthest effective distance reconstructed. These metrics measure the completeness and coverage of the reconstructed models, respectively. Notably, since the incompleteness issue primarily occurs in the sand dune areas, the incompleteness ratio is specifically calculated for these regions. As listed in Table V, the incompleteness drops significantly from ~10% to ~3%, which is consistent with the visualization results. This improvement demonstrates that our method successfully retrieves the backside of the sand dunes, which were missing in the ContextCapture models. Additionally, the farthest effective distance is extended from 20.9 meters to 34.6 meters, further highlighting the superior performance of our method.

TABLE V
QUANTITATIVE ANALYSIS OF THE MULTI-STATION3D MESH MODELS GENERATED FROM CONTEXT CAPTURE AND OUR METHOD.

	Incompleteness (%)		Farthest Effective Distance (m)	
	Context Capture	Ours	Context Capture	Ours
0716-19	11.1 %	3.2 %	20.9 m	34.6 m
0302-24	9.5 %	2.4 %	25.8 m	33.8 m

V. CONCLUSIONS AND DISCUSSION

In order to generate a large-scale optimal 3D mesh model of the Martian surface, this paper proposes introducing semantic

information into the conventional photogrammetric pipeline, specifically in the aspect of image matching. First, the transformer-based semantic segmentation is performed to obtain multiple levels of semantic features. Instead of directly utilizing the semantic segments, the proposed semantic-aware SuperGlue algorithm leverages deep features to initialize keypoint descriptors. These descriptors are first embedded with scaled positional information and then enriched through attentional aggregation of keypoints from distinct semantic classes. The resulting abundant and precise tie points facilitate cross-station bundle adjustment, which is essential for dense 3D reconstruction. By transforming images into the frequency domain and integrating semantic features, disparity images can be generated for textureless images, overcoming the over smoothness issues inherent in traditional dense matching algorithms. Finally, point clouds are derived from the disparity images and interpolated into a 3D mesh model, which is then textured accordingly.

Experiments are performed on two typical Martian datasets composed of various types of landforms. Five representative match groups are selected and compared with off-the-shelf solutions to demonstrate the superiority of the proposed semantic-aware SuperGlue method. Following bundle adjustment incorporating sufficient cross-station tie-points, the re-projection errors are significantly reduced to approximately one pixel, resulting in an absolute 3D distance difference of one meter at distances of up to 90 meters from the viewpoint, as validated by satellite imagery. Evaluations of dense matching and 3D reconstruction are conducted to compare the proposed with conventional algorithms and an off-the-shelf commercial solution, thereby demonstrating the effectiveness of incorporating semantic information in generating more accurate and realistic mesh models. Despite these advancements, the current algorithm still encounters challenges when connecting images separated by more than 15 m in the absence of distinct landforms, as substantial differences hinder the detection of corresponding points by both human and computer vision.

The proposed methods enable the integration of cross-station images for optimized large-scale 3D mapping and modeling of the Martian surface. These approaches can yield high-quality 3D Martian models, characterized by optimal accuracy, completeness, and coverage, which can support future Martian exploration missions and scientific studies.

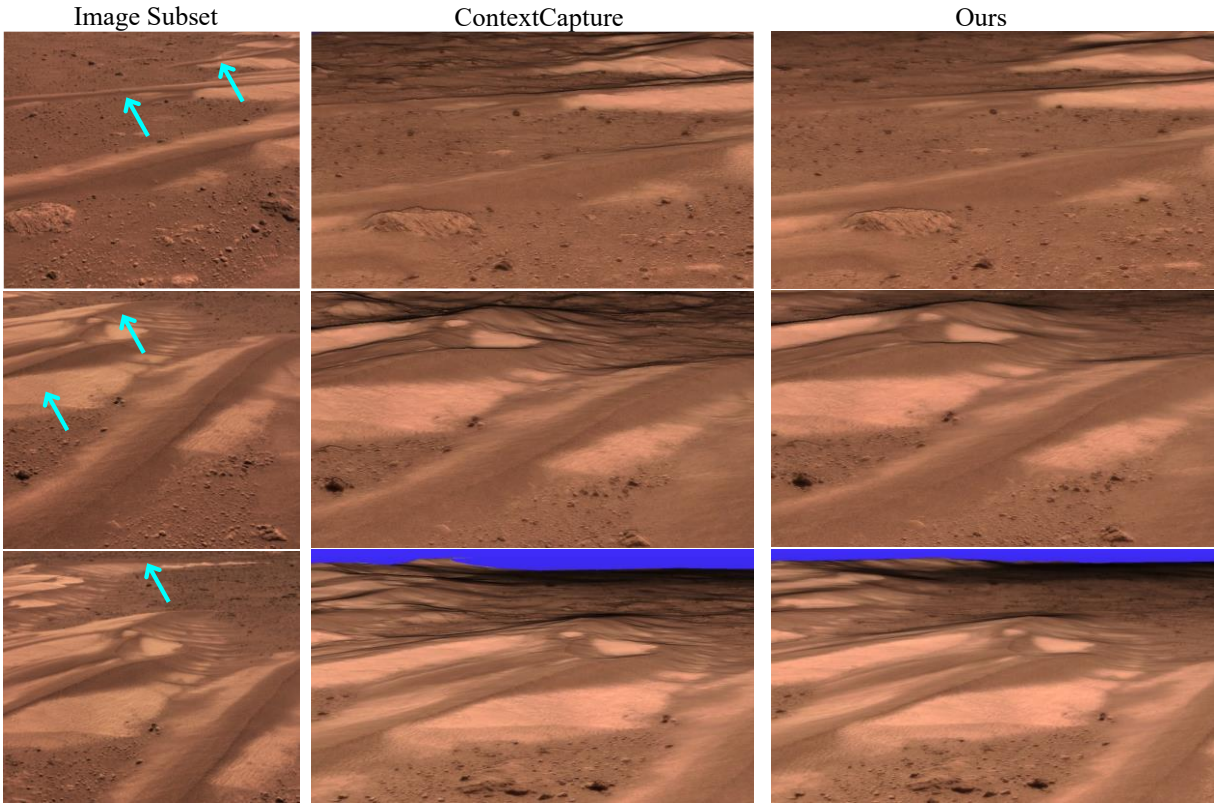


Fig. 14. Comparison of the textured 3D mesh model for the 0716-19 dataset.

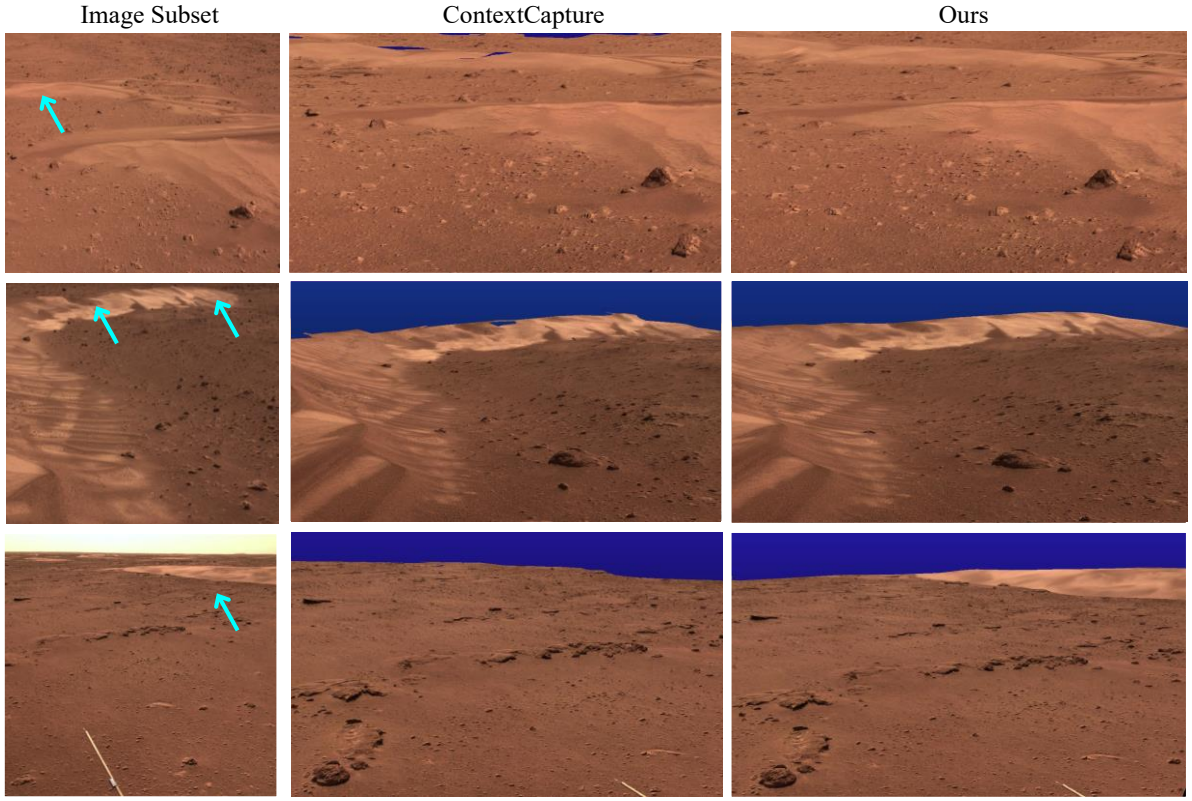


Fig. 15. Comparison of the textured 3D mesh model for the 0303-24 dataset.

REFERENCES

[1] G. Paar, T. Ortner, C. Tate et al., "Three-Dimensional Data Preparation and Immersive Mission-Spanning Visualization and Analysis of Mars 2020 Mastcam-Z Stereo Image Sequences," Earth and Space Science, vol. 10, no. 3, pp. e2022EA002532, 2023.

[2] B. Wu, J. Dong, Y. Wang et al., "Landing Site Selection and Characterization of Tianwen-1 (Zhurong Rover) on Mars," Journal of Geophysical Research: Planets, vol. 127, no. 4, pp. e2021JE007137, 2022.

[3] Y. Liu, X. Wu, Y.-Y. S. Zhao et al., "Zhurong reveals recent aqueous activities in Utopia Planitia, Mars," Science Advances, vol. 8, no. 19, pp. eabn8555, 2022.

[4] J. Pla-García, S. C. Rafkin, G. Martinez et al., "Meteorological predictions for Mars 2020 Perseverance Rover landing site at Jezero crater," Space science reviews, vol. 216, no. 8, pp. 148, 2020.

[5] Z. Li, B. Wu, W. C. Liu et al., "Photogrammetric Processing of Tianwen-1 HiRIC Imagery for Precision Topographic Mapping on Mars," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-16, 2022.

[6] K. Gwinner, R. Jaumann, E. Hauber et al., "The High Resolution Stereo Camera (HRSC) of Mars Express and its approach to science analysis and mapping for Mars and its satellites," Planetary and Space Science, vol. 126, pp. 93-138, 2016.

[7] A. S. McEwen, E. M. Eliason, J. W. Bergstrom et al., "Mars Reconnaissance Orbiter's High Resolution Imaging Science Experiment (HiRISE)," Journal of Geophysical Research: Planets, vol. 112, no. E5, 2007.

[8] W. C. Liu, and B. Wu, "Atmosphere-aware photoclinometry for pixel-wise 3D topographic mapping of Mars," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 204, pp. 237-256, 2023.

[9] R. L. Kirk, E. Howington-Kraus, B. Redding et al., "High-resolution topomapping of candidate MER landing sites with Mars Orbiter Camera narrow-angle images," Journal of Geophysical Research: Planets, vol. 108, no. E12, 2003/12/01, 2003.

[10] Z. Chen, B. Wu, and W. C. Liu, "Mars3DNet: CNN-Based High-Resolution 3D Reconstruction of the Martian Surface from Single Images," Remote Sensing, 13, 2021.

[11] Y. Tao, S. H. G. Walter, J.-P. Muller et al., "A High-Resolution Digital Terrain Model Mosaic of the Mars 2020 Perseverance Rover Landing Site at Jezero Crater," Earth and Space Science, vol. 10, no. 10, pp. e2023EA003045, 2023.

[12] M. U. Müller, N. Ekhtiari, R. M. Almeida et al., "Super-resolution of multispectral satellite images using convolutional neural networks," ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume V-1-2020, pp. 33-40, 2020.

[13] B. Wu, Y. Li, W. C. Liu et al., "Centimeter-resolution topographic modeling and fine-scale analysis of craters and rocks at the Chang'E-4 landing site," Earth and Planetary Science Letters, vol. 553, 2021.

[14] J. Bell, J. Maki, G. Mehall et al., "The Mars 2020 perseverance rover mast camera zoom (Mastcam-Z)

multispectral, stereoscopic imaging investigation," Space science reviews, vol. 217, pp. 1-40, 2021.
[15] R. Liu, Y. Xu, and Q. Yang, "High-Resolution and Spatial-Continuous 3-D Model Reconstruction of Martian Surface by Integrating Multisensor Data of Zhurong Rover," IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1-16, 2024.
[16] Y. Li, Z. Xiao, C. Ma et al., "Extraction and Analysis of Three-Dimensional Morphological Features of Centimeter-Scale Rocks in Zhurong Landing Region," Journal of Geophysical Research: Planets, vol. 128, no. 7, pp. e2022JE007656, 2023.
[17] Y. Tao, J.-P. Muller, and W. Poole, "Automated localisation of Mars rovers using co-registered HiRISE-CTX-HRSC orthorectified images and wide baseline Navcam orthorectified mosaics," Icarus, vol. 280, pp. 139-157, 2016.
[18] J. G. Zhong, J. G. Yan, M. Li et al., "A deep learning-based local feature extraction method for improved image matching and surface reconstruction from Yutu-2 PCAM images on the Moon," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 206, pp. 16-29, 2023.
[19] Q. Zhu, Z. Wang, H. Hu et al., "Leveraging photogrammetric mesh models for aerial-ground feature point matching toward integrated 3D reconstruction," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 166, pp. 26-40, 2020.
[20] B. Wu, L. Xie, H. Hu et al., "Integration of aerial oblique imagery and terrestrial imagery for optimized 3D modeling in urban areas," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 139, pp. 119-132, 2018.
[21] Z. Li, B. Wu, Y. Li et al., "Fusion of aerial, MMS and backpack images and point clouds for optimized 3D mapping in urban areas," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 202, pp. 463-478, 2023.
[22] P. E. Sarlin, D. DeTone, T. Malisiewicz et al., "SuperGlue: Learning Feature Matching with Graph Neural Networks," 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4937-4946, 2020.
[23] J. M. Sun, Z. H. Shen, Y. A. Wang et al., "LoFTR: Detector-Free Local Feature Matching with Transformers," 2021 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), pp. 8918-8927, 2021.
[24] Q. Zhou, T. Sattler, and L. Leal-Taixe, "Patch2pix: Epipolar-guided pixel-level correspondences." pp. 4669-4678.
[25] W. Jiang, E. Trulls, J. Hosang *et al.*, "COTR: Correspondence transformer for matching across images." pp. 6207-6217.
[26] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," IEEE Transactions on pattern analysis and machine intelligence, vol. 30, no. 2, pp. 328-341, 2007.
[27] Z. Ye, Y. Xu, H. Chen et al., "Area-Based Dense Image Matching with Subpixel Accuracy for Remote Sensing Applications: Practical Analysis and Comparative Study," Remote Sensing, 12, 2020.
[28] L. Ye, and B. Wu, "Integrated Image Matching and Segmentation for 3D Surface Reconstruction in Urban Areas," Photogrammetric Engineering & Remote Sensing, vol. 84, no. 3, pp. 135-148, 2018.
[29] M. Rothermel, K. Wenzel, D. Fritsch et al., "SURE: Photogrammetric surface reconstruction from imagery."
[30] Y. Yue, T. Fang, W. Li et al., "Hierarchical Edge-Preserving Dense Matching by Exploiting Reliably Matched Line Segments," Remote Sensing, 15, 2023.
[31] H. Hu, C. T. Chen, B. Wu et al., "Texture-Aware Dense Image Matching Using Ternary Census Transform," Xxiii Isprs Congress, Commission Iii, vol. 3, no. 3, pp. 59-66, 2016.
[32] P. Z. Tian, M. B. Yao, X. M. Xiao et al., "3-D Semantic Terrain Reconstruction of Monocular Close-Up Images of Martian Terrains," IEEE Transactions on Geoscience and Remote Sensing, vol. 62, 2024.
[33] Y. Liu, Q. Huang, S. Hui et al., "Semantic-aware Representation Learning for Homography Estimation," arXiv preprint arXiv:2407.13284, 2024.
[34] H.-H. Vu, P. Labatut, J.-P. Pons et al., "High accuracy and visibility-consistent dense multiview stereo," IEEE transactions on pattern analysis and machine intelligence, vol. 34, no. 5, pp. 889-901, 2011.
[35] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.
[36] J. M. Morel, and G. S. Yu, "ASIFT: A New Framework for Fully Affine Invariant Image Comparison," Siam Journal on Imaging Sciences, vol. 2, no. 2, pp. 438-469, 2009.
[37] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-Supervised Interest Point Detection and Description," Proceedings 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 337-349, 2018.
[38] Y. Zhang, X. Zhao, and D. Qian, "Searching from area to point: A hierarchical framework for semantic-geometric combined feature matching," arXiv preprint arXiv:2305.00194, 2023.
[39] Z. Ye, Y. Xu, L. Hoegner et al., "Precise Disparity Estimation For Narrow Baseline Stereo Based On Multiscale Superpixels And Phase Correlation," International Archives of Photogrammetry and Remote Sensing and Spatial Information Science., vol. XLII-2/W13, pp. 147-153, 2019.
[40] X. Wan, J. G. Liu, S. Li et al., "Phase Correlation Decomposition: The Impact of Illumination Variation for Robust Subpixel Remotely Sensed Image Matching," IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 9, pp. 6710-6725, 2019.
[41] R. K. Wei, H. D. Pei, D. J. Wu et al., "A Semantically Aware Multi-View 3D Reconstruction Method for Urban Applications," Applied Sciences-Basel, vol. 14, no. 5, 2024.
[42] D. Zhu, J. Li, X. Wang et al., "Semantic Edge Based Disparity Estimation Using Adaptive Dynamic Programming for Binocular Sensors," Sensors, 18, 2018.
[43] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," Advances in neural information processing systems, vol. 26, 2013.

[44] Z. Liu, Y. T. Lin, Y. Cao et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," 2021 IEEE /Cvf International Conference on Computer Vision (Iccv 2021), pp. 9992-10002, 2021.
[45] Z. Li, B. Wu, Z. Chen et al., "Transformer-Based Method for Semantic Segmentation and Reconstruction of the Martian Surface," Geospatial Week 2023, Vol. 48-1, pp. 1643-1649, 2023.
[46] T. Xiao, Y. Liu, B. Zhou et al., "Unified perceptual parsing for scene understanding." pp. 418-434.
[47] T.-Y. Lin, P. Dollár, R. Girshick et al., "Feature pyramid networks for object detection." pp. 2117-2125.
[48] H. Zhao, J. Shi, X. Qi et al., "Pyramid scene parsing network." pp. 2881-2890.
[49] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," Neurocomputing: Algorithms, architectures and applications, pp. 227-236: Springer, 1990.
[50] C. M. Bishop, Neural networks for pattern recognition: Oxford university press, 1995.
[51] A. Fusiello, E. Trucco, and A. Verri, "A compact algorithm for rectification of stereo pairs," Machine Vision and Applications, vol. 12, no. 1, pp. 16-22, 2000.
[52] S. Birchfield, and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 4, pp. 401-406, 1998.
[53] B. S. Reddy, and B. N. Chatterji, "An FFT-based technique for translation, rotation, and scale-invariant image registration," IEEE Transactions on Image Processing, vol. 5, no. 8, pp. 1266-1271, 1996.
[54] X. H. Tong, Z. Ye, Y. S. Xu et al., "A Novel Subpixel Phase Correlation Method Using Singular Value Decomposition and Unified Random Sample Consensus," IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 8, pp. 4143-4156, 2015.
[55] B. Wu, "Photogrammetry: 3-D From Imagery," International Encyclopedia of Geography, pp. 1-13, 2017.
[56] D. T. Lee, and B. J. Schachter, "Two algorithms for constructing a Delaunay triangulation," International Journal of Computer & Information Sciences, vol. 9, no. 3, pp. 219-242, 1980.
[57] J. Liu, C. Li, R. Zhang *et al.*, "Geomorphic contexts and science focus of the Zhurong landing site on Mars," *Nature Astronomy*, vol. 6, no. 1, pp. 65-71, 2022/01/01, 2022.
[58] Y.-Y. S. Zhao, J. Yu, G. Wei *et al.*, "In situ analysis of surface composition and meteorology at the Zhurong landing site on Mars," *National Science Review*, vol. 10, no. 6, pp. nwad056, 2023.
[59] S. Gou, Z. Yue, K. Di *et al.*, "Transverse aeolian ridges in the landing area of the Tianwen-1 Zhurong rover on Utopia Planitia, Mars," *Earth and Planetary Science Letters*, vol. 595, pp. 117764, 2022.

Zhaojin Li received the B.S. degree in remote sensing from the Wuhan University of China. She is currently pursuing a Ph.D. degree with a major in photogrammetry and remote sensing with the Hong Kong Polytechnic University. Her research interests include planetary mapping, photogrammetry and computer vision

Bo Wu is a professor and director of the Planetary Remote Sensing Laboratory of the Hong Kong Polytechnic University. His research interests are mainly in planetary remote sensing and mapping, photogrammetry and robotic vision. He chairs the ISPRS Working Group "Planetary Remote Sensing and Mapping" and serves as Associate Editor or Editorial Board Member of several international journals.