BRIDGING 3D RECONSTRUCTION AND SEMANTIC SEGMENTATION FOR PLANETARY SURFACE CHARACTERIZATION

ZHAOJIN LI

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Land Surveying and Geo-Informatics

Bridging 3D Reconstruction and Semantic Segmentation for Planetary Surface Characterization

Zhaojin LI

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

March 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

LI Zhaojin (Name of student)

Abstract

The study of planetary surfaces holds great significance, as it ensures the safety of various in-situ space exploration missions and uncovers the evolutionary history of planets. Over the past several decades, a wealth of data has been collected, revealing the surfaces of celestial bodies such as the Moon, Mars, Mercury, and some asteroids. Among the various types of products, topographic maps and landform datasets derived from optical images have garnered significant attention. These products combine geometric and semantic information, providing a more complete description of surface characteristics.

For decades, three-dimensional (3D) reconstruction of planetary surfaces has been a focus of extensive research, utilizing techniques such as laser altimetry, photogrammetry, and photoclinometry to produce numerous topographic products across various planets. It is only in the past decade that semantic segmentation of planetary surfaces has gained significant attention, propelled by the rise of deep learning. This advancement has facilitated more robust results and decreased the dependence on human labor. Despite thorough research in both 3D reconstruction and semantic segmentation, challenges persist due to the poorly textured nature of planetary surfaces. While both 3D reconstruction and semantic segmentation originate from the same data source, they explore the underlying information in distinct ways.

Hence, this research aims to develop innovative approaches that bridge 3D reconstruction and semantic segmentation to achieve enhanced performance in planetary surface characterization, described through both geometric and semantic perspectives. Starting with high-resolution 3D reconstruction through the fusion of laser altimetry, photogrammetry, and photoclinometry, the resulting topographic models are employed to improve semantic segmentation via enhanced training data and geometric supervision. Conversely, the derived semantic cues are fed back to refine 3D reconstruction, to tackle complex features and dense matching scenarios.

In the first approach, a rigorous and pixel-wise 3D reconstruction is performed by integrating laser altimetry data, grayscale images, and radiance data. During the photogrammetric processing, we propose an exterior-orientation-parameter-guided feature matching algorithm and an object-based dense matching strategy to address the challenges of feature correspondence caused by the poorly textured nature of planetary surfaces. The resulting photogrammetric digital elevation model (DEM) is further refined using a photoclinometry process, enabling the generation of a topographic product with pixel-wise resolution and enhanced geometric detail. A comprehensive evaluation based on various satellite images verifies the generalizability and effectiveness of the proposed algorithm. The overall geometric difference is within 10% relative to publicly available DEM references, and qualitative assessments indicate the retrieval of pixel-wise details.

Building upon the first approach, semantic segmentation is enhanced by incorporating 3D information. A semi-automatic dataset construction method is proposed, leveraging both 3D mesh models and recovered parameters of the cameras from the 3D reconstruction stage. This approach simultaneously generates textured RGB images, semantically labeled images, depth images, and XYZ images, provided there are enough manually labeled images. To augment the manually labeled segments for training dataset construction, a depth-enhanced transformer-based network is designed to generate additional semantic segments. Furthermore, a Siamese transformer-based network is proposed to extract transform-invariant multi-level semantic features, introducing tie-points as constraints to supervise semantic class consistency. Approximately 400 images are manually annotated and then augmented to around 25,000 images to construct the training dataset. The segmentation network achieved an overall accuracy of 88%, validating the effectiveness of the proposed method. Additionally, the consistency between overlapping images reaches 97.1%, compared with 91.2% for the original

Swin Transformer. This demonstrates the necessity of the Siamese architecture and tie-point supervision.

In the third approach, the retrieved semantic cues are further utilized for enhanced 3D reconstruction. During the feature matching phase, these semantic cues are integrated to enhance the construction of feature descriptors, contextual aggregation, and outlier removal, resulting in robust cross-station matches. These matches facilitate accurate bundle adjustment, effectively linking more challenging images. In the dense matching stage, a frequency-domain similarity measurement is proposed and combined with semantic cues to enhance matching reliability and preserve surface discontinuities. Finally, the disparity maps generated from the matching results are used to derive 3D point clouds, which are then meshed to create 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 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.

In summary, data from laser altimetry, visual camera, radiance information, and the frequency domain are integrated to advance current 3D reconstruction and semantic segmentation techniques for planetary surfaces, optimally revealing both geometric and contextual information. This research holds the potential to significantly enhance the characterization of planetary surfaces, improve the surface operations of exploration missions, and deepen our understanding of the relevant geomorphological and geological implications.

Publications Arising from the Thesis

Journal Papers:

[1] **Z. Li**, B. Wu, 2025. Semantic-aware Image Matching for Large-scale 3D Reconstruction of the Martian Surface from Rover Images. ***IEEE Transactions on Geoscience and Remote Sensing***, 63, 1-17.

[2] **Z. Li**, B. Wu, Y. Li, Z. Chen, 2023. Fusion of Aerial, MMS and Backpack Images and Point Clouds for Optimized 3D Mapping in Urban Areas, ***ISPRS Journal of Photogrammetry and Remote Sensing***, 202, 463-478.

[3] **Z. Li**, B. Wu, W. C. Liu, L**.** Chen, H. Li, J. Dong, W. Rao, D. Wang, Q. Meng, J. Dong, 2022. Photogrammetric Processing of Tianwen-1 HiRIC Imagery for Precision Topographic Mapping on Mars. ***IEEE Transactions on Geoscience and Remote Sensing***, 60, 1-16.

[4] **Z. Li**, B. Wu, W. C. Liu, Z. Chen, 2022. Integrated Photogrammetric and Photoclinometric Processing of Multiple HRSC Images for Pixel-wise 3D Mapping on Mars, ***IEEE Transactions on Geoscience and Remote Sensing***, 60,1-13.

[5] Y. Ma, **Z. Li***, B. Wu, R. Duan, 2025. DepthFormer: Depth-Enhanced Transformer Network for Semantic Segmentation of the Martian Surface From Rover Images, ***Earth and Space Science***, 12, 6.

[6] B. Wu, J. Dong, Y. Wang, W. Rao, Z. Sun, S. Krasilnikov, **Z. Li**, Z. Tan, Z. Chen, C. Wang, M. Ivanov, J. Zhu, W. C. Liu, L. Chen, H. Li, 2024. A Probable Ancient Nearshore Zone in Southern Utopia on Mars Unveiled from Observations at the Zhurong Landing Area, ***Scientific Reports***, 14, 24389.

[7] R. Duan, L. Chen, **Z. Li**, Z. Chen, B. Wu, 2024. A Scene Graph Encoding and Matching Network for UAV Visual Localization, ***IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing***, 17, 9890-9902.

[8] S. Chen, B. Wu, H. Li, **Z. Li**, Y. Liu, 2024. Asteroid-NeRF: A deep-learning method for 3D surface reconstruction of asteroids, ***Astronomy & Astrophysics***, 687, A278.

[9] Z. Chen, B. Wu, S. Krasilnikov, W. Xun, Y. Ma, S. Liu, **Z. Li**, 2024. A Global Database of Pitted Cones on Mars for Research on Martian Volcanism, ***Scientific Data***, 11, 942.

[10] W. C. Liu, B. Wu, **Z. Li**, J. Dong, W. Rao, 2022. Pre and Post-Landing Atmospheric Optical Depths at the Zhurong Landing Site on Mars Retrieved Using a Single-Image-Based Approach, ***Icarus***, 387, 115223.

[11] B. Wu, J. Dong, Y. Wang, W. Rao, Z. Sun, **Z. Li**, Z. Tan, Z. Chen, C. Wang, W. C. Liu, L. Chen, J. Zhu, H. Li, 2022. Landing Site Selection and Characterisation of Tianwen-1 (Zhurong Rover) on Mars. ***Journal of Geophysical Research: Planets***, 127(4), e2021JE007137.

[12] Z. Chen, B. Wu, Y. Wang, S. Liu, **Z. Li**, C. Yang, J. Dong, W. Rao, 2022. Rock Abundance and Erosion Rate at the Zhurong Landing Site in Southern Utopia Planitia on Mars, ***Earth and Space Science***, 9, e2022EA002252.

[13] Y. Lu, K. S. Edgett, B. Wu, Y. Wang, **Z. Li**, G. G. Michael, H. Yizhaq, Q. Jin, Y. Wu, 2022. Aeolian disruption and reworking of TARs at the Zhurong rover field site, southern Utopia Planitia, Mars, ***Earth and Planetary Science Letters***, 595,117785.

[14] B. Wu, J. Dong, Y. Wang, **Z. Li**, Z. Chen, W.C. Liu, J. Zhu, L. Chen, Y. Li, W. Rao, 2021. Characterization of the Candidate Landing Region for Tianwen-1 – China's First Mission to Mars, ***Earth and Space Science***, 8, e2021EA001670.

[15] C. Ding, Z. Xiao, B. Wu, **Z. Li**, Y. Su, B. Zhou, K. Liu, J. Cui, 2021. Rock Fragments in Shallow Lunar Regolith: Constraints by the Lunar Penetrating Radar onboard the Chang'E-4 Mission, ***Journal of Geophysical Research - Planets***, 126, e2021JE006917.

Conference Papers:

[1] **Z. Li**, B. Wu, S. Chen, 2025. 3D Gaussian Splatting for Detailed Reconstruction of Planetary Surfaces from Orbiter Images. In European Geosciences Union (EGU), 27 April – 2 May, Vienna, Austria.

[2] **Z. Li**, B. Wu, Z. Chen, Y. Ma, 2023. Transformer-Based Method for Semantic Segmentation and Reconstruction of the Martian Surface, Vol. XLVIII-1/W2-2023, pp. 1643–1649, doi:10.5194/isprs-archives-XLVIII-1-W2-2023-1643-2023.

[3] **Z. Li**, B. Wu, Y. Li, 2020. Integration of Aerial, MMS, and Backpack Images for Seamless 3D Mapping in Urban Areas, International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XLIII-B2-2020, pp. 443-449, doi:10.5194/isprs-archives-XLIII-B2-2020-443-2020.

Acknowledgments

Pursuing a PhD in Hong Kong has been the most valuable and unforgettable experience of my life. It has been a long and arduous journey, and I would have found it much more difficult to complete without the help, support, and guidance I received over the past five years.

Foremost, I would like to express my sincere gratitude to my supervisor, Prof. Bo Wu. I appreciate the guidance and opportunities he provided, which enabled me to grow and develop my skills and knowledge. I am particularly grateful for the chance to participate in the Tianwen-1 mission. The entire study was motivated by the need to address the real-world problems faced by the Tianwen-1 mission, and many experiments were conducted using its data. The experience and spirit gained from participating in this mission have continued to inspire and motivate me throughout my PhD studies, and will remain with me for the journey ahead.

I am grateful to Prof. Jan-Peter Muller and Prof. Jie Shan, the external examiners of this thesis, whose thorough review and thought-provoking feedback have not only improved the rigor and clarity of this dissertation but also inspired new directions for my future work. I also wish to thank Dr. Wai Chung Liu, Dr. Yuan Li, and Dr. Yiran Wang for their valuable and patient guidance, as well as their mental support, which has been instrumental in helping me navigate the challenges of my research.

I would also like to extend my gratitude to all my groupmates for their unconditional support and help. In particular, I would like to thank Dr. Long Chen, Mr. Zeyu Chen, and Mr. Hongliang Li for the methodological discussions and camaraderie, and Mr. Jiaming Zhu and Ms. Yi Liu for their kindness and assistance in my daily life.

Finally, I am grateful to my beloved family for their constant understanding, support, and pride in me. Many thanks are also given to Dr. Xibo Sun for all of his insightful academic vision, powerful coding skills, and quality companions.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Research Background

Characterization of planetary surfaces in terms of their three-dimensional (3D) geometry and semantic information is a prerequisite for advancing the frontiers of planetary exploration and scientific investigations (Rothrock et al., 2016; Swan et al., 2021; Wu et al., 2021). While 3D information reconstructs the scene's absolute geometric structure, semantic information provides contextual meaning, enabling a more intuitive comprehension. With numerous space missions launched over the past decades, abundant satellite and rover images have been acquired, providing unprecedented insights into planetary surfaces (Daly et al., 2017; Li et al., 2011). These datasets have sparked intense interest and necessitate the development of more efficient and optimal methods for extracting the underlying 3D and semantic information.

The most representative method for the 3D reconstruction of planetary surfaces is photogrammetry, which requires multiple observations of the same place sharing a certain viewing angle (Gwinner et al., 2016; Li et al., 2011). Photogrammetry algorithms are widely used for their robustness under various scenarios, and their rigorous mathematical process, which not only guarantees the reliability of the results but also facilitates the assessment of the errors. Typically, four main phases are involved in the process: feature matching, bundle adjustment, dense image matching, and topographic production generation (Hu and Wu, 2018). Initial feature matching involves the detection of distinct keypoints and the establishment of correspondences across images, which subsequently serve as constraints in the bundle adjustment stage to correct the nominal exterior orientation (EO) parameters and generate sparse point clouds. Subsequently, dense image matching is performed to establish pixel-wise correspondences, which are then used to generate a dense point cloud via the collinearity equation and interpolated into a 2.5-D digital elevation model (DEM) or 3D mesh model. It is

apparent that correspondence retrieval or image matching is pivotal to the entire photogrammetric process, in respect of both geometric accuracy and product resolution, and thus continuous efforts have been made to generate better matching results. Comprehensive investigations have been undertaken to tackle the matching problem in the context of Earth's environments, where large coverage, resolution, and viewpoint variations, as well as occlusion, pose significant challenges (Li et al., 2023c). The planetary surface, characterized by a poorly textured nature, further exacerbates the matching issues, making it even harder to identify the correspondence across images. Complementary information must therefore be incorporated into the gray-scale information to enhance the distinctiveness of the pixel-level description.

To overcome the limitation of stereo observation and improve the resolution of the reconstruction, photoclinometry has emerged as a promising approach. Unlike photogrammetry, which relies directly on gray-scale information from images, photoclinometry endeavors to uncover the underlying mechanism of light propagation and its interaction with the surfaces (Horn, 1990). At its core, photoclinometry aims to invert the process by which sunlight is reflected by landforms and captured by the camera, thereby uncovering the intricate relationships between illumination, topography, and image formation (Wu et al., 2018). Hence, this method is theoretically single-image, allowing it to perform calculations on a pixel basis and retrieve pixel-wise subtle details. Moreover, when multiple images with varying illumination conditions are available, they can provide complementary information, enabling a more comprehensive reconstruction of areas that may be obscured in a single image (Alexandrov and Beyer, 2018). Early photoclinometry studies focused on near-range object 3D reconstruction, and this approach was later extended to large-scale remote sensing applications over the past decade. However, when it comes to recovering absolute elevation on planetary surfaces, the strength of photoclinometry’s local-focusing capability becomes a limitation for large-scale elevation recovery. Accordingly, the general trend is combining the merits of

photogrammetry and photoclinometry to generate accurate DEMs that are rich in detail (Liu and Wu, 2020). Photoclinometry has already been successfully applied to Lunar (Alexandrov and Beyer, 2018; Wu et al., 2018) and asteroid surfaces (Gaskell et al., 2008), supporting multiple exploration missions. Recent research efforts (Liu and Wu, 2023) have concentrated on developing methods to account for atmospheric influences in photoclinometry, to extend the method to the Martian surface. This involves sophisticated modeling of how light interacts with both the atmosphere and the surface.

Alongside the success of 3D information retrieval, semantic information has gained increasing attention. Semantic information enables the direct utilization of the image by providing context and meaning, as users or robots can interact with images in a way that aligns with vision perception and understanding. Moreover, automatic large-scale segmentation reveals patterns across extensive areas that might not be immediately apparent, facilitating scientific understanding and interpretation. Whereas semantic segmentation has made significant strides in Earth-based scenarios, its application to planetary surfaces remains an ongoing area of research. There are two primary reasons for this challenge. The success of segmentation tasks, which are predominantly machine learning-based, depends heavily on the quality and quantity of training datasets. For Earth applications, there is a wealth of publicly available datasets that have been meticulously annotated, allowing machine learning models to be trained effectively. In contrast, datasets for planetary surfaces are limited. This scarcity is due to the relatively few missions that have captured high-resolution images of other planets, as well as the substantial effort required for human experts to annotate these images accurately. Second, planetary surfaces often exhibit continuous and homogeneous characteristics, making it challenging to distinguish specific semantic classes. Unlike Earth's diverse landscapes, which include easily identifiable landforms with clear boundaries, planetary surfaces may consist of vast expanses of similar terrain. This homogeneity complicates the task for both

human annotators and machine learning algorithms, as subtle differences must be detected to accurately segment and classify the surface landforms.

Previous semantic retrieval tasks have often been approached as object detection problems, focusing on identifying specific types of landforms such as craters, rocks, sand dunes, or volcanoes (Chen et al., 2024c; Robbins et al., 2014). The development of refined detection algorithms and the release of global-scale catalogs have significantly benefited the research community by reducing the manual effort required for data analysis. Recently, NASA has taken steps to advance beyond object detection by exploring the segmentation of planetary images into multiple semantic classes. This effort leverages advancements in deep learning to facilitate the analysis of potential landing sites and guide rover traverses (Swan et al., 2021). A significant challenge is concurrently posed by the requirement for substantial training data. Even when satisfactory semantic segmentation results are achieved, there is an ongoing debate about the value of using these results solely for visualization purposes. Therefore, a growing trend has emerged of incorporating semantic information into vision tasks, aiming to achieve optimal results by extending the original algorithm to a semantic-aware one (Cong et al., 2024; Kundu et al., 2014).

Intuitively, both 3D reconstruction and semantic segmentation of a planetary surface are confronted with the same challenge. The scarcity of planetary surface data necessitates the integration of multi-modal data to augment the existing information. Extensive efforts are continuing to improve the 3D reconstruction and semantic segmentation results individually. The imperfect but acceptable result, obtained with significant time and labor, serves as data support for downstream applications, rather than improving the vision tasks *per se*. In contrast, 3D reconstruction and semantic segmentation naturally complement each other. 3D reconstruction provides detailed spatial information that can enhance semantic segmentation by adding depth and context to 2D images. This can help distinguish features that appear

homogeneous in 2D but are distinct in 3D. Conversely, semantic segmentation can provide meaningful labels and features that guide the reconstruction and refine 3D models, revealing underlying structures and patterns that might not be apparent from geometry alone. Therefore, a more elegant integration of multi-modal data could lead to better outcomes for both 3D reconstruction and semantic segmentation, yet systematic research in this area remains limited.

1.2 Research Objectives

The aim of this research is to propose a framework to integrate multi-modal data to generate optimal 3D model and semantic segmentation results, for further engineering or scientific applications. The objectives of this research are as follows:

(1) To develop a 3D reconstruction method integrating multi-modal data (laser altimetry, photogrammetry, and photoclinometry), ensuring absolute geometric accuracy while also capturing the subtle details present in the images;

(2) To design a semantic segmentation network that not only classifies images into semantic categories but also generates a series of multi-level semantic features and exhibits transformation-invariant properties;

(3) To extend the 3D reconstruction method with semantic information, by developing a semantic-aware image matching method that connects challenging image pairs and generates better dense matches, enabling subsequent optimal dense 3D reconstruction;

(4) To validate the developed approaches systematically using actual images of representative planetary scenes.

1.3 Contributions and Innovations of the Research

The primary innovation of this research lies in the exploitation of multi-modal data. By leveraging these data, the approach effectively addresses the challenges posed by the poorly textured nature of the planetary surface, thereby facilitating accurate 3D reconstruction and semantic segmentation of the actual scene. The contributions of this research are as follows:

(1) A generic pipeline is proposed for integrating laser altimetry and photogrammetry, resulting in a large-scale topographic DEM with high geometric accuracy. By generating a camera model that accounts for the rough terrain provided by laser altimetry data, the learning-based feature matching process can be guided more effectively. Using epipolar images rectified within the space established by the laser altimetry DEM, the dense matching process is further enhanced based on landforms to compensate for textureless images during disparity retrieval;

(2) This research investigates the integration of photogrammetry and photoclinometry, accounting for atmospheric effects. The algorithm first initializes atmospheric parameters and albedo using photogrammetric DEMs, and then jointly optimizes the pixel-wise DEM, albedo, and atmospheric parameters. By implementing the algorithm within the TensorFlow framework, the Adam optimizer is employed to facilitate fast and accurate convergence. Pixel-wise DEM can be achieved with favorable accuracy, together with pixel-wise albedo, optical depth, and atmospheric scatter;

(3) By leveraging the 3D reconstruction results, the semantic segmentation is enhanced in two aspects: dataset construction and neural network training. A semi-automatic dataset construction approach is proposed to augment the dataset and enrich it with more semantic segments, complementary depth, and XYZ coordinates. Furthermore, a Siamese transformer-based network is proposed and trained to extract multi-level semantic features by incorporating tie-point supervision, which is calculated with the

assistance of 3D information. This approach leads to improved semantic segmentation results in terms of both consistency and accuracy;

(4) Attempts to incorporate the semantic information into the 3D reconstruction pipeline are made, in both feature and dense image matching stages. The semantic features derived from the semantic segmentation network are used to initialize the descriptors, and the semantic segments are leveraged to guide the aggregation of the global information for local descriptor enhancement. By integrating the underlying semantic information into gray-scale and frequency domain information, robust descriptors can be achieved and the matching can be guided in a more concentrated manner, thereby achieving more precise correspondence.

The contributions of this study enable the generation of accurate planetary topographic products with pixel-wise spatial resolution and coverage, along with semantic contextual information. These products can facilitate more detailed and precise analysis of planetary surfaces for scientific research.

1.4 Thesis Organization

This thesis consists of six chapters. Following the introduction chapter, the remainder of the thesis is organized as follows.

Chapter 2 provides a comprehensive review of previous related works, covering both 3D reconstruction and semantic segmentation. It reviews the origins and development of these algorithms individually, as well as the most recent advancements in their integration.

Chapter 3 elaborates on the methods developed in this research for multi-modal data integration in 3D reconstruction, including approaches for integrating laser altimetry with photogrammetry, as well as photogrammetry with photoclinometry. Experiments are conducted

using various images to assess the geometric accuracy and detail performance, demonstrating the algorithm's generalizability and superiority.

Chapter 4 proposes an improved semantic segmentation method based on the 3D information obtained in the previous chapter. Leveraging the Zhurong dataset, comprehensive experiments are conducted, verifying the effectiveness and feasibility of the proposed semi-automatic dataset construction, depth-enhanced transformer, and Siamese transformer for semantic segmentation.

Chapter 5 incorporates the aforementioned semantic information to optimally reconstruct 3D scenes, including semantic-aware feature matching and dense matching. Thorough experiments conducted on the Zhurong dataset validate that the proposed feature matching algorithm achieves favorable accuracy, and the dense mesh is optimized using semantic information.

Chapter 6 provides a comprehensive summary of the concluding remarks, along with an in-depth discussion of the findings and potential directions for future research.

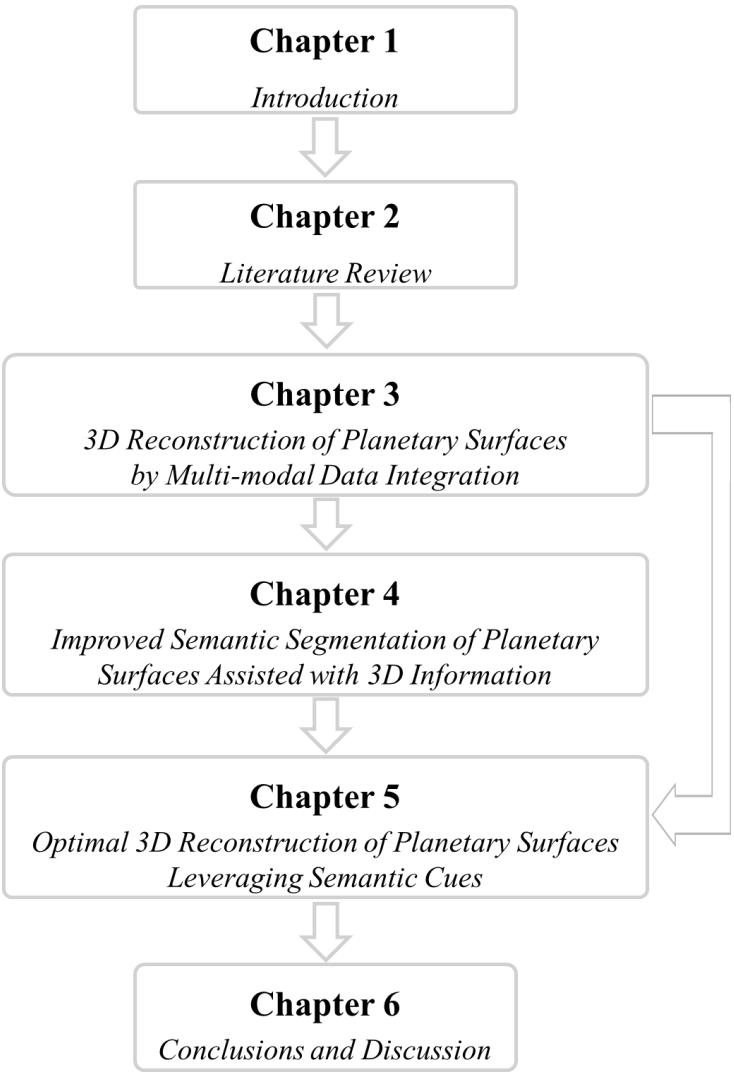


Figure 1.1 Schematic of the structure of this thesis.

Chapter 2 Literature Review

Both 3D reconstruction and semantic segmentation are broad research areas that have been extensively studied individually. While 3D reconstruction has a long history spanning several decades, the development of semantic segmentation has been significantly accelerated by recent advancements in machine learning. The improvements in both areas have opened up new possibilities for integrating these two fields. The following section reviews the historical and current developments in both areas, as well as their integration. This overview aims to provide a comprehensive understanding of how 3D reconstruction and semantic segmentation have evolved individually and how their integration has advanced the field.

2.1 3D Reconstruction of the Planetary Surfaces

The 3D reconstruction of planetary surfaces has a long history dating back to the 1970s, when the Apollo missions first left Earth and entered lunar orbit. These missions provided some of the earliest data used to create 3D models of the lunar surface, laying the groundwork for future advancements in planetary surface mapping and analysis. Currently, four main techniques are involved: laser altimetry, photogrammetry, photoclinometry, and learning-based algorithms. Below is a review of each technique.

2.1.1 Laser Altimetry

Light detection and ranging (LiDAR) technology has been an essential approach for topographic mapping of the Earth, lunar, Martian, and other planetary surfaces (Zhou et al., 2017) since the Apollo missions (Kaula et al., 1973; Sjogren and Wollenhaupt, 1973). LiDAR uses a laser beam to measure the distance between the sensor and the target surface. The laser altimeter, consisting of a laser transmitter and a receiver, emits a laser beam toward the target

surface (Wehr and Lohr, 1999). The reflected signal is then detected by the receiver, and the time-of-flight of the laser pulse is measured. By using the speed of light, the distance between the sensor and the target surface is calculated. It is an active remote sensing technique that is not influenced by sunlight, shadow, or surface characteristics, making it robust and effective for mapping a wide range of planetary surfaces.

The first laser altimetry system, the Mars Orbiter Laser Altimeter (MOLA) (Smith et al., 2001; Zuber et al., 1992), was launched in 1992 as part of the Mars Observer mission (Jonathan, 1986), with the goal of mapping the entire Martian surface (Smith et al., 1998). Although this spacecraft experienced a communication loss, global mapping was resumed four years later in 1997 with the Mars Global Surveyor (MGS) mission (Albee, 2002), which continued to utilize the MOLA system. In 1994, the Clementine LiDAR (Ledebuhr et al., 1995; Smith et al., 1997) carried the Lunar Orbiter Laser Altimeter (LOLA) to map the lunar surface, which is also the first laser altimetry system using multiple laser beams. Attempts were made to map the Mercury surface using laser altimetry with the Mercury Laser Altimeter (Zuber et al., 2012) onboard the MESSENGER spacecraft (Gold et al., 2001) in 2006. Laser altimetry was also utilized to map the shape of the Bennu and the Ryugu asteroids as part of the Origins Spectral Interpretation Resource Identification Security Regolith Explorer (OSIRIS-REx) mission (Daly et al., 2017), and the Hayabusa2 mission (Mizuno et al., 2017), respectively. In addition to these extraterrestrial laser systems, Earth-orbiting laser altimetry systems, such as ICESat-1 (Schutz et al., 2005), and ICESat-2 (Neumann et al., 2019), have been widely used for applications beyond topographic mapping, including water-level (Scherer et al., 2022) and forest analysis (Mulverhill et al., 2022).

There are also numerous other laser altimetry systems onboard various satellites orbiting different planets (Anthony et al., 2012; Ping et al., 2009), each providing unique insights into the topography and composition of their respective celestial bodies. The widespread use of

laser altimetry demonstrates its superiority over other methods. It offers a unique combination of high accuracy, high resolution, and large-scale coverage, making it an ideal tool for retrieving topographic data over vast areas, such as entire continents or planetary surfaces. The advancements in laser altimetry in the past decades can be summarized into the following aspects. To enhance the resolution offered by laser data, the pulse repetition rate is increased, allowing for a higher number of pulses to be emitted and counted within a given time period. Furthermore, earlier laser systems were primarily based on linear detection, which required high-energy laser consumption. However, the limited repetition frequency of these systems resulted in low data density. Recent breakthroughs in quantum information technology (Hadfield, 2009) have enabled the integration of photon detection with laser radar technology, resulting in reduced energy consumption and payload weight (Degnan et al., 2007). This technique also allows for achieving multi-beam detection, and a 51-beam laser system has already been implemented.

With respect to the laser data processing and DEM generation, cross-track adjustment and outlier removal are of significance to preserve the resolution offered by the original laser points (Barker et al., 2021). The first step is to adjust each track accordingly to eliminate all the possible inconsistencies, and the second is to remove the outliers (Xie et al., 2022). These outliers may be attributable to various factors, such as multipath effects, atmospheric interference, and multiple reflections. As an active technique, laser altimetry still has some other limitations. One major drawback is its high energy consumption, which is greater than that of passive methods, and this limits its resolution when mapping planetary surfaces. Additionally, laser altimetry relies heavily on the nominal orbit trajectory, but accurately determining this trajectory can be complex, particularly for planetary surfaces. Furthermore, the lack of visualization images is a significant weakness, which limits the technique's ability to perform 3D mapping and analysis, and hinders its use in visual-based scientific analysis.

2.1.2 Photogrammetry

Photogrammetry is also known as multi-view geometry, a name coined by Meydenbauer (1867). This technique utilizes a rigorous mathematical process to recover the entire 3D scene, enabling accurate and detailed reconstructions of complex environments (Torlegård, 1992). Its advantages lie in its rigorous theoretical derivation and robustness to arbitrary scenes, making it widely applied in various scenarios, including Earth, lunar, Martian, and asteroid environments since the last century (Chandler and Cooper, 1989; LaChapelle, 1962). To achieve these advantages, the computational pipeline of photogrammetry is relatively complex compared with the other methods reviewed in this chapter, and can be broadly divided into two main parts, namely, structure from motion (SfM) and multi-view stereo (MVS).

SfM aims to recover the camera positions and poses when capturing the images, and recover the sparse 3D scene in point cloud format (Snavely et al., 2008). The technique integrates both conventional photogrammetry and multi-view geometry theory from the computer vision field. First, the camera model should be defined and established. There are typically three main types of cameras used in various applications: frame camera, which capture a single image at a time; linear pushbroom cameras, which capture a sequence of images in a linear fashion; and fisheye cameras, which capture a wide-angle view of the scene. The frame camera and the linear pushbroom camera are the two main types of camera orbiting planetary surfaces.

A significant difference between line scanners and frame cameras lies in the line-by-line variation of EO parameters, which complicates the entire photogrammetric process. First, to describe and rectify the exterior orientation parameters, the trajectory must be fitted using polynomial functions (Li et al., 2011). However, this approach introduces additional issues, as modifications to the camera result in a series of changes in subsequent photogrammetric stages.

To overcome this limitation, the rational function model (RFM) was developed to provide a generalized camera model and simplify the subsequent process, a breakthrough that was contemporaneous with the launch of IKONOS satellite, which was the pioneering commercial satellite. (Grodecki, 2001). By directly relating 3D object space coordinates to 2D image coordinates, the RFM model provides a seamless and accurate interface between image data providers and users, fulfilling the requirements of IKONOS. Benefiting from its independence from specific sensors and platforms, as well as its coordinate system, this camera model is highly flexible and has been widely implemented in numerous photogrammetric software packages (Yang, 2000). The RFM has stood the test of time, remaining a widely used standard in satellite products. The camera models for both frame and pushbroom have remained unchanged for decades due to their outstanding capabilities.

Another key module in SfM is feature matching, which is decisive for the precision and quality of the following bundle adjustment, and in turn accounts for the overall quality of the final topographic product. As suggested by Ma et al. (2021), the development of feature matching can be broadly summarized as a transition from handcrafted features to deep learning-based features. Most early manually defined features are corner features. The earliest Moravec features (Moravec, 1977) utilize gray-scale variance to detect local extrema of gray-scale changes by scanning a rectangular window. The gray-scale variance of a pixel is calculated in eight directions, with the minimum value calculated as the corner response function, and local non-maximum suppression is employed to eliminate non-maximum values. Harris and Stephens (1988) later improved upon Moravec's corner detector, which also employs a predefined window, and calculates the first derivative to obtain the gradient in both vertical and horizontal directions. The gradient is then convolved with a Gaussian template to compute the response, and a threshold is applied to extract the features. Mathematically, the Harris feature possesses rotation and translation invariance, but lacks scale invariance. Before the advent of

scale-invariant feature transform (SIFT) (Lowe, 2004), the Harris corner detector was a popular choice for feature extraction, owing to its simplicity and robustness, and was further refined and extended in many research studies. The SIFT detector consists of four primary stages: scale-space extrema detection, keypoint localization, orientation assignment, and descriptor computation. The SIFT algorithm computes local extrema in the resolution pyramid, which allows it to possess the desirable property of scale invariance. Moreover, the feature direction is estimated by analyzing the dominant orientation of a large neighborhood, providing a robust representation of the feature. Since then, it has revolutionized the entire vision-related field with its unparalleled stability and performance, and it is still widely used in many computer vision or photogrammetric tasks. In comparison with other handcrafted features, SIFT's major drawback is its high computational complexity, which has led to the widespread adoption of ORB in many online applications (Rublee et al., 2011). The ORB algorithm utilizes the Harris response function to identify a set of FAST corners as the final feature points. The main direction of each ORB feature point is then established by calculating the centroid of the surrounding image block and the center pixel, generating a vector that enables the computation of the similarity between ORB binary descriptors.

Benefiting from the advancements in convolutional neural networks (CNNs), which enable the extraction of deep features, the paradigm established by SIFT was overturned in 2015 (Verdie et al., 2015). TILDE is a supervised learning-based keypoint detector, which uses SIFT to collect the supervised key points and is robust to changes in light, viewpoint, and weather. The TILDE architecture consists of a CNN that takes an image as input and produces a feature map that highlights the detected features. The network is trained using a large dataset of images with annotated features, and the loss function is designed to encourage the network to produce invariant features. With the development of deep learning, more learning-based detectors with more convolutional layers have been proposed. DetNet (Lenc and Vedaldi, 2016)

is a pioneering work that presents a fully generic framework for learning local covariant features; by reformulating the detection task as a regression problem, it introduces a covariance constraint that enables the automatic learning of stable anchors for local feature detection, which is robust to geometric transformations. DeTone et al. (2018) proposed a novel approach, wherein a single convolutional encoder is employed to extract deep features, which are then fed into two separate convolutional decoders to detect keypoints and compute descriptors, respectively. Following this work, the same team later proposed the SuperGlue (Sarlin et al., 2020) network for end-to-end matching thereby opening a new era since the SIFT. SuperGlue draws inspiration from the natural language processing (NLP) field and integrates an attention mechanism into the matching pipeline, complemented by a positional encoder. Additionally, it leverages optimal transport to solve the matching problem in a differentiable and end-to-end manner. Two consecutive modules are involved in this network, namely, the attentional graph neural network and the optimal matching layer. After encoding the keypoint with the union of the 2D coordinates (u,v) on the image and the 256-dimension descriptor derived from the SuperPoint, the self- and cross-attention, illustrating the connections among the features within one image and across images, are constructed. Owing to this encoder, the relative positions of the keypoints are considered, which alleviates the dependency of the descriptor, thus being suitable for images that do not share similar contextual cues. Furthermore, the well-preserved structural information can flexibly adapt to changes in scenarios. Subsequently, the matching scores of each feature pair are calculated in a brute-force manner. The one-to-one match selection problem is then formulated as an optimal partial assignment based on the Sinkhorn algorithm (Peyré and Cuturi, 2019).

Similarly, dense image matching (DIM) has transitioned from traditional methods to deep learning-based approaches. The conventional milestone was set by semi-global matching (SGM) (Hirschmuller, 2005), which combines the advantages of both local and global methods.

SGM achieves this by aggregating matching costs along multiple 1-D paths, rather than relying on a single, 2-D optimization. The matching costs are separated into two main terms, namely, data cost and smoothness cost. While data cost measures the similarity between each patch, it is also dependent on the matching process, where the deep-learning is more powerful than the human-crafted ones (Chen et al., 2015; Kendall et al., 2017). In 2018, a landmark learning-based dense image matching algorithm, MVSNet, was proposed (Yao et al., 2018; Yao et al., 2019). Its pipeline is analogous to that of conventional DIM, comprising feature extraction, cost volume construction, cost aggregation, and depth inference. MVSNet made several significant contributions to the field. First, it introduced the use of a shared CNN to extract features from multi-view images. Second, it proposed the concept of differentiable homography within the neural network, which connects deep features with 3D cost volume regularization, enabling end-to-end 3D reconstruction training. Third, it addressed the issue of uncertain input image numbers by converting arbitrary numbers of pixel feature vectors into a single matching cost vector, thereby allowing the same network to accommodate any number of input images.

Deep learning in photogrammetry faces limitations in computational resources, speed, and reliability. Firstly, computational resource demands are high; models like MVSNet, with multiple convolutional layers, require GPU acceleration and substantial memory, restricting deployment on resource-constrained devices. Training and inference can take hours or days, depending on dataset size and hardware. Secondly, processing speed is slower compared to traditional methods like SGM. The complex feature extraction and cost volume construction in deep learning models, such as MVSNet's 3D cost volume regularization, significantly increase computation time, hindering real-time applications. Lastly, reliability depends heavily on training data and scene variability. Deep learning models need large annotated datasets, and performance may degrade with insufficient data or changes in conditions like lighting or

viewpoint, affecting matching and reconstruction stability. Optimization strategies, including model pruning, quantization, and efficient cost aggregation, are needed to reduce resource demands, enhance speed, and improve reliability across diverse scenarios, making deep learning more viable for practical photogrammetric applications.

With regard to the photogrammetric reconstruction of planetary surfaces, many investigations have concentrated on refining particular aspects of the process (Hu and Wu, 2018; Li et al., 2011) or creating more accurate and detailed topographic products since the 1970s using data acquired by Apollo 15,16,17 (Light, 1972). As few ground control points (GCPs) are available on the planetary surface, the general trend is to use a large-scale but low-resolution DEM as the constraint for photogrammetric bundle adjustment (Gwinner et al., 2016; Light, 1972; Wu et al., 2011). With the launch of numerous satellites and the subsequent acquisition of vast amounts of data, NASA released the modern ISIS3 software (Anderson et al., 2004). This software is optimized for current computer systems and offers a broad range of photogrammetric functions, as well as integration with the SPICE kernel, which records ancillary data associated with the images (Acton, 1998). The Ames Stereo Pipeline (ASP) was open-sourced in 2018 (Beyer et al., 2018), enabling direct generation of DEMs and DOMs, thereby aiding both engineers and scientists in the planetary field to produce high-quality topographic products.

2.1.3 Photoclinometry

Photoclinometry, also known as shape from shading (SfS), has been used for the 3D reconstruction of the Lunar surface since the 1960s (Hapke, 1965; Hapke and van Hoen, 1963; Van Diggelen, 1951). In contrast to photogrammetry, which relies on the geometric stability of structures to recover 3D information, photoclinometry exploits the shading information inherent in 2D images to infer 3D shapes and structures. Specifically, the terrain does not emit

energy but rather reflects energy from the source (i.e., the Sun). This reflection interaction determines the amount of energy (i.e., radiance) received by the sensor (i.e., the camera), which is influenced by both the geometry of the topography and the photometric properties (e.g., albedo) (Grumpe and Wöhler, 2014).

The scene radiance as a function of the surface gradient is referred to as a reflectance map, and can be derived from a phenomenological model (Horn and Sjoberg, 1979). The Lambertian surface model, first proposed by Lambert (1760), is a seminal reflectance model that assumes a surface emits light (either as a source or by reflection) in a manner that is proportional to the cosine of the angle between the surface normal and the direction of incidence. This model is thus characterized by a cosine-based relationship between the incidence angle and the emitted light intensity. The later Lommel-Seeliger model (Chandrasekhar, 1960), a widely used radiative transfer model, assumes that the radiance observed by a sensor is the result of light scattered by all particles within the medium that lie within the sensor's field of view. This model is designed to model the reflectance of the lunar surface and is still actively used in planetary photometric studies. The Lunar-Lambert model integrates the above two models linearly by assigning empirical weights to balance their contributions (Weaver and Meador, 1977). Later, McEwen (1986) proposed the use of the phase angle to calculate the empirical weight. While some studies model this weight using a polynomial function (Lohse et al., 2006), the other construct an exponential function (Gaskell et al., 2008). Meanwhile, the Hapke model (Hapke, 1981), which accurately captures the photometric behavior of a wide range of planetary surfaces, remains a widely studied topic. However, the model's numerous and interdependent parameters hinder its widespread adoption. McEwen (1991) demonstrated that the Hapke model can be approximated by the Lunar-Lambert model under specific conditions. Subsequent studies (Soderblom et al., 2006) have also shown that the Lunar-Lambert model can be used as a simplification for the Martian surface.

Apart from the reflectance, other properties of the surface also influence the overall radiance received by the camera. Significantly, as illustrated in Kirk (1987), the accuracy of the reflectance model and the local variations in albedo are the primary sources of error in photoclinometry. Although single-image photoclinometry is usable for photoclinometry products, the problem of solving the gradients of two directions is mathematically ill-posed. Multi-image photoclinometry, which incorporates more observations to solve the albedo parameters rigorously, is thus extensively studied (Woodham, 1980). Multi-image photoclinometry is also named photometric stereo SfS (PS-SfS). Liu et al. (2018) and NASA Ames Stereo Pipeline (ASP) (Alexandrov and Beyer, 2018) proposed a novel PS-SfS approach that leverages multiple images to simultaneously solve for elevation and albedo, thereby addressing a long-standing limitation in shape from shading research and building upon previous work in the field. (Heipke et al., 2001; Wöhler, 2004). Besides the solving of albedo, the multi-view stereo also enables the calculation of the atmospheric parameters of the Martian surface (Walter et al., 2015). Another intuitive benefit is that multiple observations with different illumination conditions reveal more details of the planetary surface, particularly for regions like the lunar South Pole (Chandraker et al., 2007).

While photoclinometry excels at recovering subtle details at the pixel scale, large-scale 3D reconstruction poses a significant challenge for this algorithm. To address this limitation, the coarse-to-fine strategy is widely employed in current photoclinometry methods (Alexandrov and Beyer, 2018; Jiang et al., 2017; Wu et al., 2018), which utilize a rigorous but not highly detailed laser or photogrammetric DEM as input. By upsampling the DEMs by a factor of two at each layer, favorable details can be added to the original DEM product while preserving the absolute elevation. A more elegant integration method that has been attempted involves using the photometric difference to measure the similarity in dense image matching to overcome the drawbacks of the Census-like or MI-like features (Liu and Wu, 2020).

In summary, photoclinometry is a long-standing 3D reconstruction algorithm with a rich history dating back nearly 300 years. Extensive research has been conducted to enhance its capabilities, including the development of advanced reflectance models, solving for albedo, incorporating low-resolution DEMs, accounting for the atmosphere (Liu and Wu, 2023), and adapting training strategies. Despite its maturity, photoclinometry continues to receive growing interest in the context of Lunar South Pole exploration missions and small-scale scientific studies.

2.1.4 Learning-based Algorithms

As the above three approaches involve several stages and depend on rigorous and precise parameters to conduct 3D reconstruction, learning-based approaches have become the center of interest for their ability to learn from networks and apply them to the images directly to generate reasonable topographic products in an end-to-end manner.

Numerous works have investigated various scenarios, which can be generally categorized into two main types, namely, single-view and multi-view (Wang et al., 2024). As mentioned in Section 2.1.3, single-view 3D reconstruction is not only an interesting concept, but also of great significance for planetary surfaces, for which the multi-view high resolution images are not guaranteed. As suggested by previous paper, stereo coverage is below 1.5% at high-resolution (CTX, HiRISE) of Martian surface. Multiple images are even rarer (less than 1% of all stereo coverage). As the entire scene is not comprehensively captured, the problem is typically formulated as depth estimation. The pioneering work was conducted by Saxena et al. (2008), first segmenting the images into many small planar surfaces using superpixels, and leveraging the five nodes Markov random field (MRF) to capture both 3D location and 3D orientation. Eigen et al. (2014) later proposed a two-branch CNN architecture, where one branch focuses on global information prediction and the other refines the prediction locally. Subsequent work

has fully exploited the superiority of CNNs in feature capture, using MRFs to predict depth by minimizing the negative log-likelihood (Liu et al., 2015). With the development of CNNs, depth prediction has also been explored as an application for VGG (Eigen and Fergus, 2015) and ResNet (Laina et al., 2016) architectures. In 2018, the Megadepth dataset was released, featuring diverse scenes and high-quality depth annotations, which has facilitated the development of more accurate and robust models in various applications (Li and Snavely, 2018). Recent advancements in single-view depth prediction have focused on the introduction of attention mechanisms, which enable the model to selectively focus on relevant regions of the image and improve depth estimation accuracy (Huynh et al., 2020). A generative adversarial neural network was also used for the estimation of a dense depth map given a color image (Abdulwahab et al., 2020). Furthermore, some approaches have achieved zero-shot learning (Guizilini et al., 2023), allowing the model to generalize to new, unseen environments without requiring additional training data. In the planetary community, convolutional neural networks (CNNs) have also been introduced and implemented to generate depth images, which are subsequently normalized to an absolute scale based on a reference low-resolution DEM (Chen et al., 2021).

Neural network-based multi-view stereo, which is reviewed in the photogrammetry section, replaces conventional MVS methods, but still strictly adheres to the traditional multi-view geometry pipeline. Remarkably, recent advancements in multi-view stereo have enabled direct output of point clouds or mesh products, ushering in a new era for this field. The Neural Radiance Field (NeRF) represents a 3D scene using a 5D function that encodes both 3D position and 2D radiance emission (Mildenhall et al., 2021). The input of the NeRF is the adjusted camera and exterior orientation parameters, and it is trained by a combination of reconstruction loss and regularization terms. Much work has since been conducted on the NeRF, focusing on its architecture, including the use of hierarchical representations (Barron et al.,

2021; Xu et al., 2022), improved training (Cong et al., 2024; Jiang et al., 2023), and the incorporation of additional scene information (Deng et al., 2022; Zhang et al., 2021). The success of this approach can be attributed to several key aspects. Firstly, it eschews complex physical models and instead utilizes a simple multi-layer perceptron to learn the scene representation. Second, its rendering process is differentiable, allowing it to be driven by data. Asteroids, characterized by their small, single 3D shape with redundant views, present a fascinating challenge for the NeRF, and it has been applied to the study of Bennu, Itokawa, and Ryugu (Chen et al., 2024a; Chen et al., 2024b). Recent advancements have leveraged 3D Gaussian distributions to efficiently describe scenes, yielding faster rendering times through a technique named Gaussian Splatting (Kerbl et al., 2023). Specifically, Gaussian Splatting represents the scene as a collection of Gaussian splats, each parameterized by a mean, covariance, and weight. This representation is more expressive than the simple density-based representation employed in the NeRF, enabling more accurate and robust reconstructions. Consequently, Gaussian Splatting achieves more accurate geometry reconstructions, particularly for scenes featuring complex or thin structures.

In summary, the current learning-based 3D reconstruction methods can achieve favorable dense topographic results with absolute scale, and are still promising approaches. However, the learning-based algorithms rely on the position and pose parameters offered by the rigorous photogrammetric process, limiting direct use this kind of algorithm to reconstruct the rigorous topographic product.

2.1.5 Summary of 3D Reconstruction Approaches

3D geometry understanding is a crucial aspect of 3D scene understanding, and has been a subject of study for decades. Long-standing techniques, such as laser altimetry and photogrammetry, which are grounded in rigorous mathematical equations, remain the

mainstream methods for 3D reconstruction, particularly in planetary 3D reconstruction where high accuracy is paramount and the analysis of errors and uncertainties is essential. With the continuous development and refinement of topographic products, photoclinometry has garnered significant attention in recent years, owing to its capability to retrieve pixel-wise elevation and enhance the resolution of existing products. Moreover, the rapid advancement of deep learning techniques has significantly benefited the field of 3D reconstruction, not only enhancing the performance of individual modules within the conventional photogrammetric pipeline but also revolutionizing the field with the introduction of state-of-the-art methods, such as NeRF and Gaussian Splatting. While each of these approaches possesses some inherent drawbacks, the pursuit of rigorous and detailed topographic products remains an active area of research.

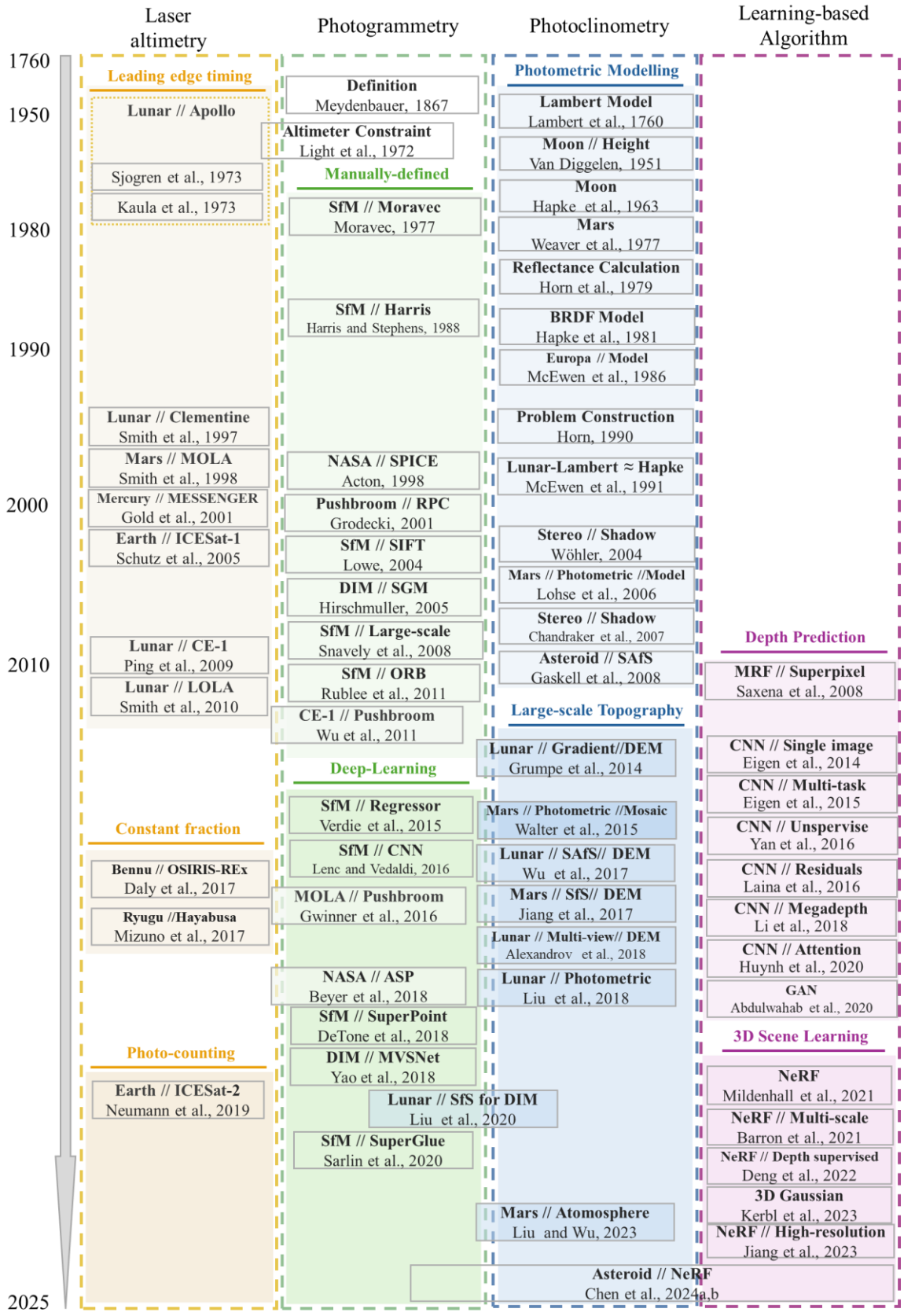


Figure 2.1 Summary of representative works related to 3D reconstruction of planetary surfaces.

2.2 Semantic Segmentation of Planetary Surfaces

Semantic segmentation has been a cornerstone of computer vision research since the early 2000s (Shotton et al., 2006; Zhao et al., 2004). Before the era of the neural network, many segmentation tasks were formulated as an energy optimization problem in a conditional random field (CRF), which is a natural fit for labeling inference tasks to model the interactions between the outputs directly (Krähenbühl and Koltun, 2011). That period was followed by an abundance of studies that combined neural networks and CRFs, aiming to leverage their strengths in feature extraction and reasoning, respectively (Chen et al., 2018; Vemulapalli et al., 2016; Zheng et al., 2015). Subsequently, the majority of researchers shifted their focus toward purely neural network-based approaches, as many parameters remained unlearnable, and relied on human understanding. Moreover, neural networks themselves can process input images and produce semantic segments in an end-to-end manner, which is more direct and convenient compared with integrated methods. Hence, this section primarily concentrates on neural network-based semantic segmentation approaches, as well as the datasets that facilitate the training of these networks.

2.2.1 Methods for Semantic Segmentation

The concept of neural networks has a rich history, dating back to the pioneering work of McCulloch and Pitts (1943). However, the modern concept of neural networks, as we understand it today, began to take form in the 1980s. This period saw the introduction of the backpropagation algorithm (LeCun, 1989), a crucial innovation that laid the foundation for the development of contemporary neural network architectures. This work was subsequently leveraged to develop LeNet, a pioneering CNN that recognized handwritten digits with five fully adaptive connections (LeCun et al., 1989). Notably, LeNet also established the overall architecture of CNN, laying the foundation for future developments in the field. This paper

built the classical architecture consisting of convolutional, pooling, and fully connected layers. Notably, the authors provide a clear explanation for the inclusion of the pooling layer, which is designed to achieve distortion and translation invariance, thereby ensuring that the recognition of digits is not affected by such transformations. Subsequent improvements in the LeNet-4 and LeNet-5 architectures were modest, involving the addition of a single fully connected layer and a Gaussian connection layer, respectively (LeCun et al., 1998). Furthermore, this approach adhered to the paradigm established by traditional pattern recognition, which decomposes classification tasks into two sequential modules: a feature extraction module and a trainable classifier module. This paradigm was subsequently adopted by subsequent generations of researchers.

However, due to limitations in computational resources and training datasets, the neural network approach failed to gain widespread acceptance during the early 21st century (Deng et al., 2009). Although sustained efforts continued to be made, it was not until the emergence of AlexNet (Krizhevsky et al., 2012) that the new era of computing all vision tasks including semantic segmentation was opened up. The contributions made by AlexNet can be summarized into three main aspects. First, GPUs were leveraged to accelerate the training process. Second, it was trained and evaluated on ImageNet, a large and standard dataset. Third, the ReLU activation function (Nair and Hinton, 2010) together with dropout regularization (Wang and Manning, 2013) was introduced to improve the convergence and the training of the neural network. Subsequently, VGG-16/19 (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), and DeepLabv3 (Chen, 2017) were proposed in succession, all of which leveraged the ImageNet dataset as a benchmark for evaluation. This facilitated the measurement and understanding of the strengths of neural networks, enabling researchers to systematically assess and compare their performance. In 2015, the U-Net was designed, primarily targeting the segmentation of medical images (Ronneberger et al., 2015). This work pioneered the

incorporation of the encoder–decoder architecture into the semantic segmentation pipeline, which is still widely used in many vision tasks.

Since then, many further efforts have been made to adapt the architecture to improve the overall performance. In 2017, a team from Google proposed the attention mechanism (Vaswani, 2017), which had shown favorable performance in the NLP field. This mechanism was later integrated into the Vision Transformer (ViT) (Dosovitskiy et al., 2020), achieving state-of-the-art performance in various vision tasks. Building upon ViT, the Swin Transformer (Liu et al., 2021) combined the advantages of CNNs and transformers, significantly advancing visual representation learning. In 2023, a notable advancement was made with the introduction of the Segment Anything Model (SAM) by Meta, which demonstrated promising results. Although SAM is also built upon the transformer architecture, it can be leveraged in any dataset zero-shot learning (Kirillov et al., 2023). As it performs mask-based segmentation, the semantic information is negligible.

In the context of semantic segmentation of planetary surfaces, early research efforts primarily framed the problem as an object detection task, with a focus on extracting specific landforms, such as craters (Robbins et al., 2014), rocks (Golombek et al., 2020), rockfalls (Bickel et al., 2020), and volcanoes. This choice of focus was not only based on computational resources but was also constrained by the datasets. In 2016, NASA published a paper introducing their semantic segmentation work, titled Soil Property and Object Classification (SPOC) (Rothrock et al., 2016). The primary architecture employed was DeepLabV3, which was applied to images captured by navigation cameras or satellite cameras. Additionally, a lightweight support vector machine was also introduced, which could be implemented on the ROS platform to facilitate terrain classification by the rover.

2.2.2 Datasets for Semantic Segmentation

As neural networks are inherently data-driven, the volume and quality of datasets play a crucial role in determining the overall performance of the approach. Concurrent with the development of the architecture of neural networks, past decades have also witnessed great advancement in datasets. The earliest dataset, MNIST, was published together with LeNet in 1994, comprising 58,527 images of digits written by 500 different writers to evaluate the performance of various digit detection algorithms (Bottou et al., 1994). Later, in 2002, the classic dataset Middlebury for the stereo matching algorithm (Scharstein and Szeliski, 2002) was published. PASCAL VOC (Everingham et al., 2006; Everingham et al., 2010), which was initially released in 2005 and underwent annual updates until 2012, consisted of 20 object classes and comprised over 10,000 images. LabelMe (Russell et al., 2008), released in 2008, not only provided a publicly available dataset but also introduced a Web-based segmentation tool that remains actively used by many researchers. Moreover, its latest version has incorporated the SAM into the tool, thereby simplifying the manual segmentation process.

In 2009, ImageNet (Deng et al., 2009) was published, based on its creators' former work the Caltech 101 dataset (Fei-Fei et al., 2006), but with more than 14,000,000 images and more than 20,000 synsets. This dataset is a landmark dataset in the field of computer vision, playing a pivotal role in the development of deep learning-based image recognition systems. The significance of ImageNet lies in its impact on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), an annual competition that evaluates the performance of image classification algorithms. The success of ImageNet has also led to the development of pre-trained models, such as VGG-16/19 and ResNet, which have become the foundation for many state-of-the-art vision tasks. As a result, ImageNet has become a benchmark for evaluating the performance of deep learning models, and its influence can be seen in many areas of computer vision research and applications. Cityscapes (Cordts et al., 2016) a dataset targeting vision tasks

in outdoor scenarios, was released in 2016, comprising over 20,000 high-resolution images that cover 50 European cities and feature 30 classes. With the increasing capability of neural networks, 3D tasks have also garnered attention, and consequently, datasets such as Scannet (Dai et al., 2017) and Megadepth (Li and Snavely, 2018) have been constructed.

With respect to planetary surface segmentation or detection, large datasets have also been constructed to facilitate engineering evaluation and scientific studies (Chen et al., 2024c; Robbins et al., 2014). AI4MARS (Swan et al., 2021), a large-scale semantic segmentation dataset, was a crowd-sourcing initiative organized by the Jet Propulsion Laboratory and Caltech, aimed at facilitating the analysis and understanding of the Martian terrain. In addition to the semantic labels, the dataset is also enriched with depth information calculated from stereo camera imagery.

2.2.3 Summary of Semantic Segmentation

Over the past two decades, significant advancements have been made in semantic segmentation techniques and datasets, with progress in one area driving innovation in the other. The advent of CNNs revolutionized the field of computer vision, leading to the proposal and validation of numerous variants aimed at improving the performance of the original models. However, it was not until the emergence of Transformers from the natural language processing (NLP) field that computer vision underwent another paradigm shift, transitioning from convolutional to attention-based architectures. In recent years, another major breakthrough has been the introduction of large models, which have addressed the transfer learning issues inherent in deep learning, achieving zero-shot segmentation with promising results.

2.3 Integration of Semantic Segmentation and 3D Reconstruction

Intuitively, semantic segmentation 3D reconstruction are not separate tasks, as discussed in the review in Sections 2.1 and 2.2 (Zheng et al., 2022). In the era of the CRF method, many

studies focusing on jointly reasoning semantic and depth information have been conducted (Hane et al., 2013; Kundu et al., 2014; Sandhu et al., 2011). The advantages of these studies lie in their ability to manually incorporate rules that leverage semantic information to guide 3D reconstruction, and vice versa. Benefiting from the multi-scale and comprehensive features extracted from CNNs, the use of multi-task algorithms, which yield multiple products at the same time (e.g., semantic segments, depth maps, and normal maps), has also increased (Eigen and Fergus, 2015). This has primarily been driven by the fact that feature extraction is a fundamental component that underlies all vision tasks, including both segmentation and 3D reconstruction (Abdulnabi et al., 2015; Hane et al., 2013; Zhao et al., 2023). Chen et al. (2019) fully exploited the content consistency between depth estimation and semantic segmentation to train a network in a self-supervised fashion. Shvets et al. (2024) extracted semantic features and MVS features separately, and then embedded one into the other's feature space before the decoding process.

Recently, with the rapid advancement of the NeRF and Gaussian splatting, researchers have endeavored to use them to learn semantic information, which can also be viewed as leveraging 3D information to facilitate semantic segmentation. The semantic-NeRF (Zhi et al., 2021) and semantic ray (Liu et al., 2023) formulate semantic segmentation as an inherently view-invariant function, mapping the relationship between 3D points in the world and their corresponding semantic labels. Elsewhere, Wang et al. (2022) proposed the Generalizable NeRF Transformer, which achieves generalized neural representation across scenes. This work has subsequently inspired the development of 3D-aware semantic representation (Cong et al., 2024), yielding multi-view depth estimation and semantic segmentation.

Apart from sharing the feature extraction backbone, the semantic segmentation and 3D reconstruction tasks are also integrated in the form of offering complementary information to one another to enrich the original RGB/gray images. Beyond architectural innovations,

researchers have also investigated methods to enhance the input data itself, particularly by incorporating depth information to improve the accuracy of semantic label inference. This trend, which emerged around 2010, initially involved conducting 3D reconstruction to generate corresponding depth maps for each image, as no direct depth information was available. With the development of the Microsoft Kinect RGBD camera in the same year (Ren et al., 2012), it became possible to simultaneously acquire RGB and depth images, thereby facilitating the widespread adoption of depth-enhanced segmentation in numerous studies (Couprie et al., 2013; Silberman et al., 2012). As for planetary surface segmentation or detection, 3D information has also been widely used in the form of DEMs, which can also be regarded as depth information. With the assistance of DEM, Wang and Wu (2019) proposed a crater detection algorithm surpassing the performance of purely image-based algorithms. Ma et al. (2024a) implemented the depth-aware attention mechanism to generate better rock detection results. In contrast, many studies have also focused on leveraging the semantic prior to improve the 3D reconstruction results. Song et al. (2024) first extracted boundaries from semantic segments, and then leveraged them to facilitate depth estimation.

Semantic segmentation and 3D reconstruction are often regarded as two separate tasks, but in reality, they are closely intertwined. Early studies have demonstrated that jointly reasoning about semantic and depth information can improve the performance of both tasks. In recent years, with the development of technologies such as NeRF and Gaussian point clouds, researchers have begun to explore the use of these technologies to learn semantic information and leverage 3D information for semantic segmentation. Meanwhile, researchers are also investigating ways to integrate semantic segmentation and 3D reconstruction tasks, such as by sharing feature extraction backbones or providing complementary information to enrich the original images. Additionally, researchers are exploring ways to utilize depth information to improve the accuracy of semantic label inference, by using RGBD cameras or DEMs to obtain

depth information. In summary, semantic segmentation and 3D reconstruction are two closely related tasks, and joint reasoning and integration can improve the performance of both tasks.

2.4 Summary

Based on the above review, a few key points can be summarized. First, a variety of 3D reconstruction techniques have been proposed and successfully exploited to recover the 3D information of planetary surfaces. While these techniques have demonstrated significant merits, their limitations and drawbacks are also not negligible. The integration of laser altimetry and photogrammetry is a mainstream approach, and the further incorporation of photoclinometry has also gained increasing attention. Although learning-based methods have not been extensively studied in the context of 3D reconstruction of planetary surfaces, their considerable capability and power will undoubtedly advance planetary mapping in the future. Second, the semantic segmentation of planetary surfaces is typically conducted using object detection algorithms. The major limitation lies in the volume and quality of the dataset. Although the AI4MARS dataset is publicly available, it only includes rover images, and the scene variability is significant due to differences in rover, lighting, atmospheric conditions, and camera configurations. Third, it can be inferred from Figure 2.2 that 3D-based semantic segmentation has received more attention than semantic-aware 3D reconstruction. The primary reason for this disparity is the fact that depth information is easier to obtain than pixel-wise semantic information. While some researchers have focused on the joint generation of semantic and depth maps, the laboriously acquired semantic maps are not being fully utilized.

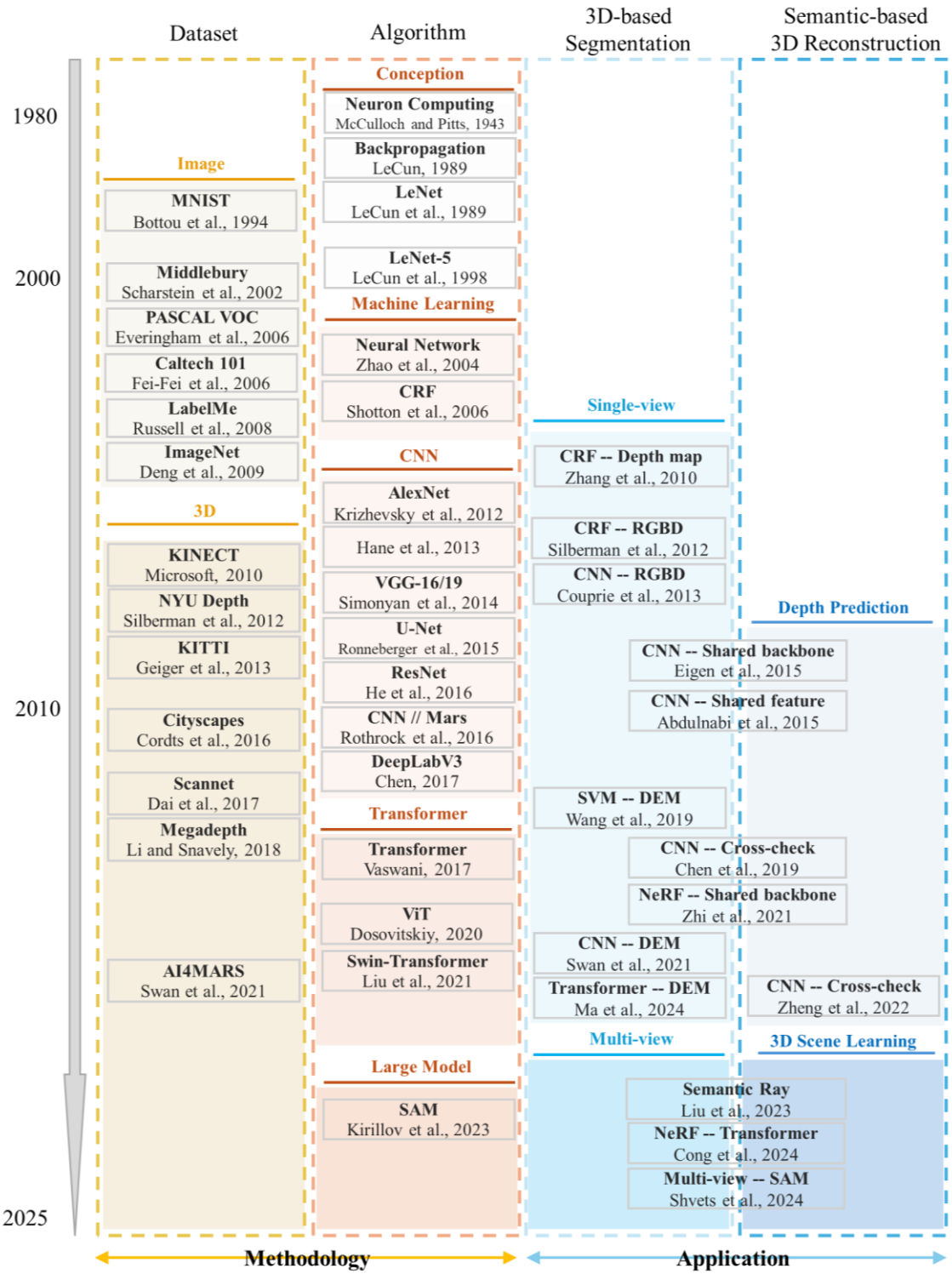


Figure 2.2 Summary of representative works related to the semantic segmentation of planetary surfaces.

Chapter 3 3D Reconstruction of Planetary Surfaces by Multi-modal Data Integration

Recovering 3D information from 2D images, thereby enriching the available information and facilitating a deeper understanding of the scene, has been a topic of intense interest for decades (Gwinner et al., 2016; McEwen et al., 2007). Extensive research has been conducted, yielding numerous DEM products with varying resolutions, ranging from hundreds of meters to tens of meters and even meters. Typically, mainstream DEM products are generated using three primary techniques: laser altimetry, photogrammetry, and photoclinometry.

Laser altimetry data are considered the most reliable data source because they are independent of both the characteristics of the planetary surface and the hardware configuration (Smith et al., 2001). Photogrammetry, which involves deriving DEMs from stereo images via rigorous mathematical equations, improves the resolution offered by laser DEMs and provides ortho-rectified images. Meanwhile, photoclinometry is typically based on the above-mentioned DEM and radiance information to calculate a pixel-wise DEM. Although the overall pipeline is quite mature, many issues still hinder effective 3D reconstruction. The integration of these three techniques is rarely discussed. Hence, a generic method is proposed herein to integrate laser altimetry, image, and radiance data to achieve rigorous and detailed 3D reconstruction even for atmosphere-covered planetary surfaces.

There follows a description of the 3D reconstruction approach developed for integrating multi-modal data for rigorous and detailed 3D reconstruction of planetary surfaces. Section 3.1 first gives an overview of the approach. Section 3.2 elaborates on the approach to integrate the laser and photogrammetric data to achieve rigorous 3D reconstruction, which is further extended to incorporate the photometric data to generate pixel-wise topographic products as described in Section 3.3. In Section 3.4, systematic experimental evaluation is conducted using

images with various configurations. The concluding remarks of this chapter are presented in Section 3.5.

3.1 Overview of the Approach

The proposed approach integrates laser altimetry, photogrammetry, and photoclinometry in a two-stage process, using multiple images along with radiance information, nominal EO parameters, laser data, and solar parameters as inputs. First, a generic photogrammetric process is performed using laser altimetry as reference data, taking into account the typical configuration of planetary images. Subsequently, photoclinometry is applied to the resulting DEMs and corresponding images, accounting for atmospheric effects, to produce a refined DEM that reveals pixel-wise details. As shown in Figure 3.1, the overall terrain is reconstructed by integrating laser altimetry and photogrammetry, and subtle details can be further refined by incorporating photoclinometry.

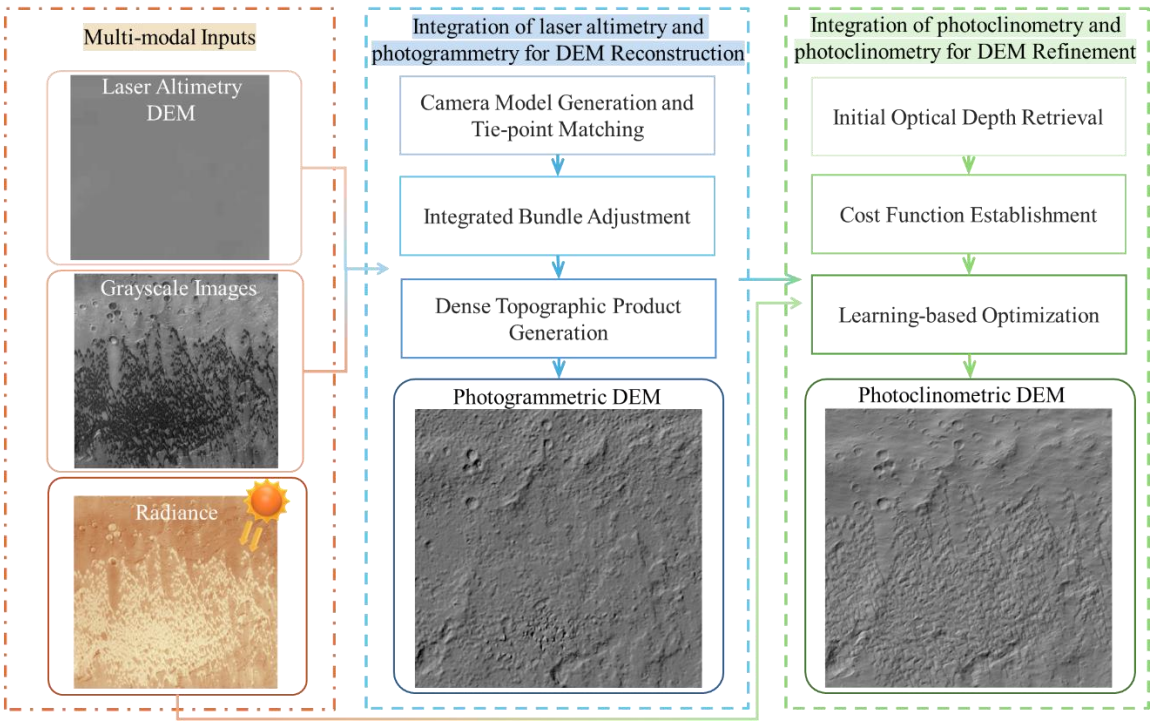


Figure 3.1 Overview of the approach.

3.2 Laser Altimetry and Photogrammetry Integration for DEM Generation

The proposed approach to integrate laser altimetry and photogrammetry for DEM generation is illustrated in Figure 3.2, where laser altimetry data is utilized throughout the entire photogrammetric process.

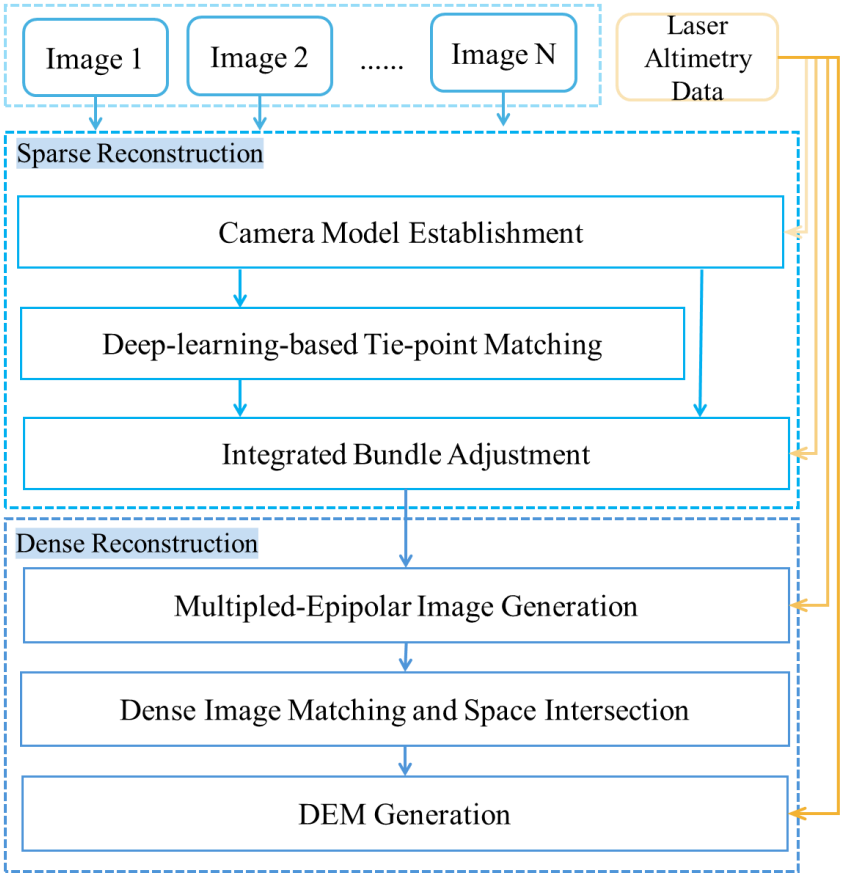


Figure 3.2 Overview of the pipeline of integration of laser altimetry and photogrammetry.

The benefits of this approach can be summarized in two key aspects. Firstly, the rough terrain provided by the laser altimetry data facilitates the construction of a precise calculation space for the camera model and epipolar calculation, thereby establishing more reasonable controls (Grodecki and Dial, 2003; Hu et al., 2019) and removing significant outliers in the DEM generation stage to ensure accuracy. Specifically, even if these two fitting problems could be solved within the virtual space, the rough terrain provided by the laser altimetry data guarantees the successful projection of the 3D points onto the images, thus enhancing the constraints. Secondly, the laser altimetry data serves as a widely recognized topographic datum,

and its integration into the bundle adjustment stage not only corrects the camera poses but also aligns the photogrammetric product with the laser product.

3.2.1 EO-guided Deep-learning based Tie-point Matching

Tie-point matching is a non-trivial work for all kinds of vision tasks, and the textureless planetary surface further complicates this issue, which emphasizes the distinctiveness of each keypoint. Traditional scale-invariant feature transform (SIFT) (Lowe, 2004) can succeed in matching large overlapped regions, but for the narrowed textureless scenario its ability is constrained. The reason may be attributed to two factors. The SIFT matching process is typically followed by a random sample consensus (RANSAC) algorithm, which assumes that the relationship between the two images can be described by a certain transformation matrix. However, the entire image captured by the pushbroom camera does not conform to this rule, leading to filtering many correct matches (Hu and Wu, 2019). Furthermore, despite the careful design of SIFT, its descriptor's expressiveness is still outperformed by those extracted from deep neural networks, which have a greater capacity to capture complex features. Insufficient tie-points may result in the consistency among images, potentially leading to the introduction of artifacts in the final DEM product.

To achieve robust and sufficient tie-point matching results, we extend the state-of-the-art deep-learning-based algorithms, Superpoint and SuperGlue (DeTone et al., 2018; Sarlin et al., 2020), for image matching. Superpoint is for feature point extraction, and SuperGlue is used together with Superpoint for feature point matching. Unlike SIFT-like detectors, Superpoint features are trained using basic shapes (e.g., triangles, cubes, checkerboards, stars) and have succeeded in extracting abundant features in narrow overlapped image regions (e.g., corners of sand dunes, crater rims). Furthermore, the matching algorithm SuperGlue combines the merits of self- and across-attention natural language processing to describe the relationships between

the features within an image and across images to be matched based on graph theory. As an end-to-end matching algorithm, no additional outlier filter module is required, thereby avoiding the aforementioned issue and ensuring a sufficient number of matches. However, the large size of typical satellite images poses a challenge for GPU-based processing, as they exceed the available memory capacity. To address this issue, nominal exterior orientation parameters are employed to reduce the computational memory requirements and enable the efficient processing of the matching algorithm. Specifically, the images are partitioned into small patches with the corresponding latitude according to the collinearity equation and the EOs.

Figure 3.3 presents a comparison of the tie-point matching results obtained using the Superpoint and SuperGlue algorithms, with the HiRIC image captured by the Tianwen-1 satellite serving as an illustrative example. The CCD images exhibit a significant challenge due to the extremely narrow overlap region between adjacent images, with a mere 100 pixels overlapping out of a total of 6144 pixels. The blue dots represent the extracted tie-points, while the red lines indicate the retrieved matches. It is apparent that the keypoints extracted by SuperPoint are significantly more numerous than those extracted by SIFT. Furthermore, for the two overlapping regions, the SuperGlue algorithm retrieves evenly distributed matches, whereas SIFT only yields a few matches, which is insufficient to facilitate a robust bundle adjustment.

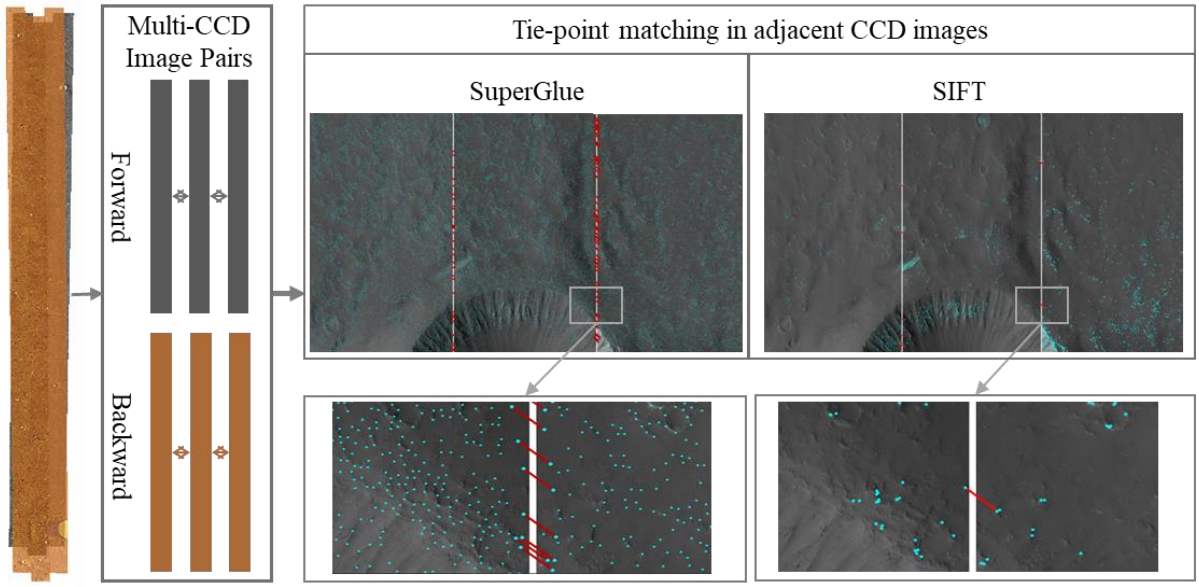


Figure 3.3 Illustration of the EO-guided SuperGlue matching algorithm.

Utilizing these tie-points, corresponding DEMs are generated, demonstrating the importance of incorporating the SuperGlue algorithm, as illustrated in Figure 3.4. Apparent artifacts are present when using SIFT matches, indicating that the inconsistencies among the CCDs are not fully eliminated.

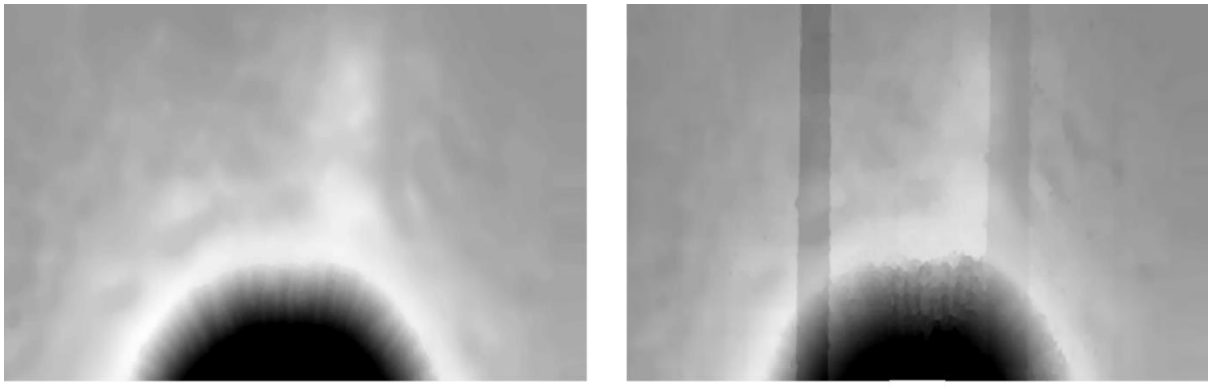


Figure 3.4 Comparison of the DEMs generated with the tie-points retrieved by (a) SuperGlue and (b) SIFT. The dark strips in (b) are caused by the insufficient and unevenly-distributed tie-points.

In Figure 3.5, the red points represent the matches extracted by SuperGlue for multi-orbit images. These feature tracks denote the correspondences between multiple images, highlighting the successful matching of features across different orbits.

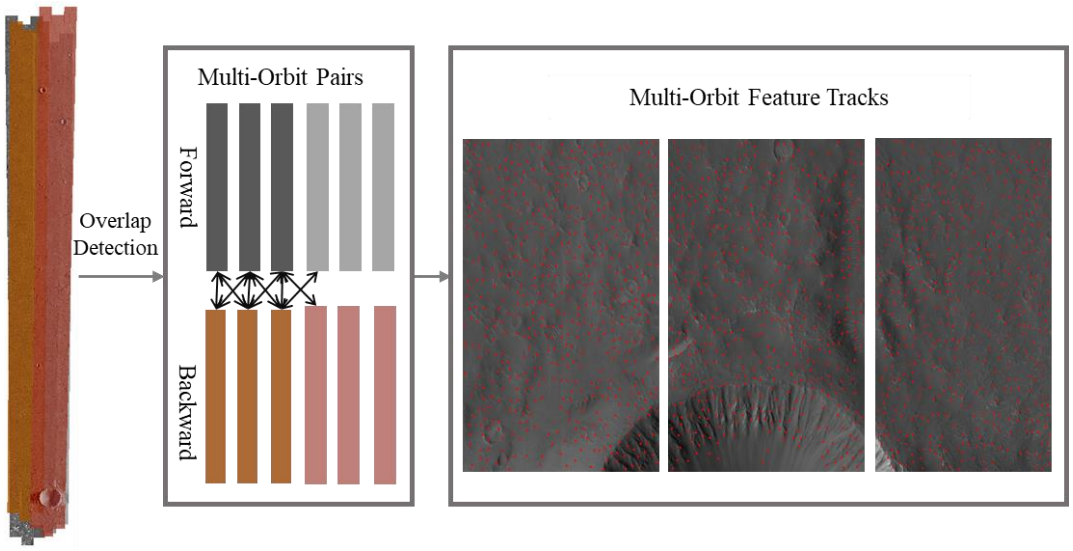


Figure 3.5 Illustration of the distribution of the feature track generated from the EO-guided SuperGlue algorithm for cross-track images.

3.2.2 Integrated Bundle Adjustment

Bundle adjustment is important in the overall photogrammetric procedure for its ability to remove inconsistencies among images and further enforce consistency with the reference data. As for the planetary surface 3D reconstruction, the bundle adjustment is even more necessary since the nominal exterior orientation parameters (EOs) are not accurate or consistent. Therefore, three aims of the integrated bundle adjustment should be achieved. First, the inconsistency among the EOs of the images captured by different CCDs of different orbits is supposed to be mitigated. While the second and third constraints ensure that the 3D coordinates of the observed points and the resolved EOs remain consistent with the reference data (laser DEM) and the observed EOs, respectively, without significant divergence.

Accordingly, three kinds of constraints are involved. The first constraint is derived from the collinearity equation, which enforces that corresponding points across different images intersect at the same 3D point in world space. This equation can be mathematically formulated as:

$$v_{proj} = \boldsymbol{x}_{ij} - \mathbb{\Pi}\left(\boldsymbol{X}_i\right) \tag{3-1}$$

where $\boldsymbol{x}_{ij}$ denotes 2D coordinates of the i^{th} keypoint track $\boldsymbol{X}_i$ projected on the j^{th} images. $\mathbb{\Pi}(\cdot)$ denotes the projection operation.

In the absence of ground control points (GCPs) on the planetary surface, the absolute datum constraint is typically established using keypoints, which are used to retrieve the corresponding height information from the laser-based DEM. This height information serves as a constraint, thereby highlighting the necessity of retrieving sufficient and evenly-distributed keypoints, which forms the foundation of the entire bundle adjustment process.

$$v_{control} = h_i - \widehat{h_i} \tag{3-2}$$

where h_i and $\widehat{h_i}$ are the heights above the reference geoid, with h_i calculated from the process and $\widehat{h_i}$ retrieved from the laser DEM, respectively. And the observation constraint is constructed, as:

$$v_{observ} = \boldsymbol{O}_i - \widehat{\boldsymbol{O}_i} \tag{3-3}$$

Hence, the overall constraint could be formulated, as:

$$v_{total} = w_{proj}v_{proj} + w_{control}v_{control} + w_{observ}v_{observ} \tag{3-4}$$

where w_{proj}, $w_{control}$, and w_{observ} are the three weight parameters to balance the contribution of these constraints. Furthermore, despite the improvement in tie-point matching achieved by the EO-guided SuperGlue, the overall number of tie-points for cross-CCD or cross-orbit images still falls short of those obtained from stereo image pairs. To address this issue, different

weights are assigned to balance the contributions of each type of image pair, thereby ensuring a more equitable comparison, which can be calculated based on pre-defined weight λ, prior precision σ, and the normalization factor N:

$$w = \sqrt{\lambda} \Big/ \left(\sqrt{N} \times \sigma\right) \tag{3-5}$$

This formulation explicitly states that the type with fewer tie points is expected to have a higher weight, which will balance the weights of the above two types of tie-points in the bundle adjustment process to simultaneously achieve inner-orbit and cross-orbit consistency.

3.2.4 Object-based Dense Matching

Sparse point clouds are simultaneously calculated with the integrated bundle adjustment. While these points are derived from the sparse feature points, most pixels in the image are not used, leading to low resolution in the point clouds. Therefore, dense correspondence retrieval is crucial to fully exploit the images and improve the resolution of the point cloud. However, correspondence retrieval is a two-dimensional search problem, and the computational volume is enormous, particularly for large satellite images. Epipolar rectification is hence conducted to align the corresponding points into the same line, and simplify the searching problem into one dimension (Wang et al., 2011). This is achieved by iteratively projecting the points onto the epipolar space defined by the laser altimetry data to retrieve the epipolar direction, followed by an affine transformation of the entire image, as illustrated in the following figure.

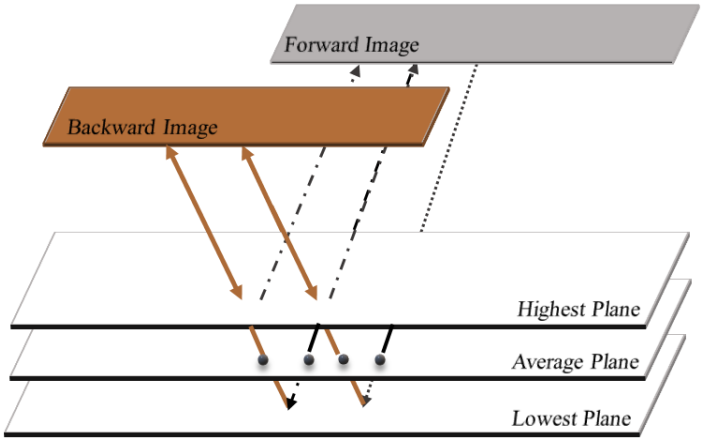


Figure 3.6 Illustration of the epipolar line retrieval algorithm.

With the rectified epipolar images, dense image matching could be performed to retrieve pixel-wise correspondence for the following point cloud calculation. However, the industry-proven semi-global matching (SGM) algorithm (Hirschmuller, 2005) has the disadvantages that (1) there is ambiguity in regions with insufficient texture; (2) it is difficult to preserve discontinuity; and (3) it is difficult to tune the penalties P1 and P2, which imposed on small and large disparity changes between adjacent pixels, respectively. As for cost calculation, AD-Census, is selected for its robustness in the textureless region and computational efficiency. One of the primary advantages of AD-Census is its ability to effectively handle ambiguous and noisy regions in the image, which can lead to incorrect matches in traditional stereo matching algorithms. By using a census transform to encode the local structure of the image, AD-Census can better capture the underlying patterns and relationships between pixels, resulting in more accurate and reliable cost calculations. Another significant merit of AD-Census is its high computational efficiency, which can be attributed to its implementation using efficient bitwise operations.

While the latter two issues primarily arise from the inherent smoothness assumption, which enforces smoothness among neighboring pixels under the supposition that disparities should be close due to the continuity of the 3D world. However, this assumption is against the presence

of discontinuities, such as occlusions and landform boundaries. Even if the penalty term can be adapted to facilitate discontinuity preservation, the optimal parameters for different regions are still distinct, and a uniform solution may not be favorable for the entire image. A texture-aware strategy is therefore proposed to adaptively adjust the penalties based on the edge and defined texture. The Canny edge detection algorithm is leveraged to detect the boundaries of the landforms, and a small penalty can thus be enforced to promote continuity. As aforementioned, the epipolar rectification is not strictly rigorous for the satellite image. Therefore, least-square matching is implemented to retrieve the disparity in the vertical direction and achieve sub-pixel precision in the horizontal direction based on the above rough disparity.

The above process, the integration of SGM and LSM, can produce satisfactory disparity maps in most usual circumstances (i.e., a clear atmosphere). However, for images influenced by haze or aerosol problems, the disparity estimation remains a problem, even for large landform objects (e.g., craters and cones) recognizable in the images. This is a severe problem for the 3D reconstruction of the planetary surface. First, the omitting of a large landform leads to a misunderstanding of the region, which may cause incorrect measurements of the potential landing sites. Although photoclinometry can add details to the photogrammetric DEM, the deviation of the modification is limited to ensure the reliability of the photoclinometric DEM. Consequently, it cannot recover large-scale landforms with significant elevation differences. An object-based matching method is therefore developed to compensate for atmospheric defects. The workflow of this method is shown in Figure 3.7. First, landform objects are regarded as centrosymmetric convex shapes and extracted using algorithms such as the Hough circle detection algorithm (Ni et al., 2016; Smereka and Dulęba, 2008) or Douglas–Peucker algorithm (Jung et al., 2019; Liu et al., 2019). The external rectangle is the bounding box, and the farthest distance between the bounding box and the center is taken as the size of the object.

From the known epipolar geometry, the centers of corresponding objects should be on the same epipolar line within a certain horizontal disparity and thus be matched efficiently. Then, to make full use of the gray-value information generated by shadows, features such as those obtained using the features from an accelerated segment test (FAST) (Rublee et al., 2011) are extracted inside each matched bounding box. Here, we define FAST-like features as the pixels whose intensity at N consecutive points on a circle of radius R_f is larger than the intensity I_P plus a threshold T or smaller than $I_P - T$. Rather than assigning certain values to R_f, I_P, and T, the method adjusts these parameters automatically according to the size of the objects. The normalized correlation coefficient is then calculated to match the existing features.

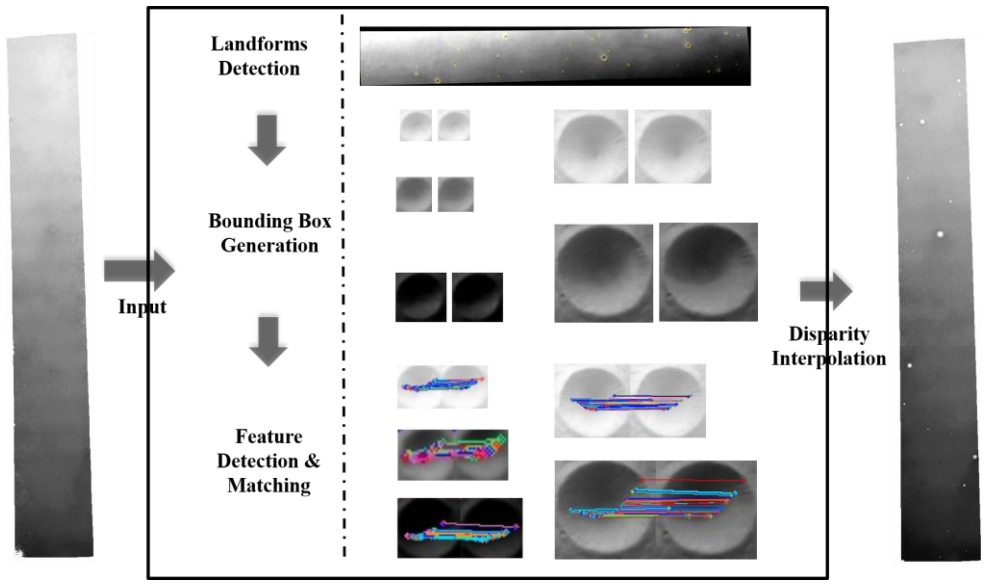


Figure 3.7 Workflow of the proposed pair-wise object-based matching.

3.2.5 Point Clouds Generation and DEM Interpolation

Knowing the disparities between the epipolar images, their pixel-wise correspondence is obtained and passed to the original images using $P'_{epi2lam}$. Space intersection is then conducted to calculate the 3D coordinates of the matching points based on the refined EOs after bundle adjustment and block adjustment. Multi-baseline triangulation is adopted to leverage the

redundant disparity information and retrieve the 3D coordinates of the matched points as a linear problem $\boldsymbol{AX} = B$.

$$A = \begin{bmatrix} x\boldsymbol{P}_1^{3T} - \boldsymbol{P}_1^{1T} \\ y\boldsymbol{P}_1^{3T} - \boldsymbol{P}_1^{2T} \\ x\boldsymbol{P}_2^{3T} - \boldsymbol{P}_2^{1T} \\ \cdots \cdots \\ y\boldsymbol{P}_n^{3T} - \boldsymbol{P}_n^{1T} \end{bmatrix}, B = \begin{bmatrix} x\boldsymbol{P}_1^{(3,4)} - \boldsymbol{P}_1^{(1,4)} \\ y\boldsymbol{P}_1^{(3,4)} - \boldsymbol{P}_1^{(2,4)} \\ x\boldsymbol{P}_2^{(3,4)} - \boldsymbol{P}_2^{(1,4)} \\ \cdots \cdots \\ y\boldsymbol{P}_n^{(3,4)} - \boldsymbol{P}_n^{(2,4)} \end{bmatrix} \tag{3-6}$$

where $\boldsymbol{P}$ is the projective matrix defined by both interior and exterior orientation parameters, and $\boldsymbol{X}$ denotes the unknown point coordinates. Finally, DEMs can be generated by interpolating the 3D point clouds. The corresponding orthoimages can be rectified using the derived DEMs and optimized EO parameters.

3.3 Photogrammetry and Photoclinometry Integration for Refined 3D Reconstruction

Aiming at revealing pixel-wise topographic details, photoclinometry or shape-from-shading methods have been developed for the high-resolution 3D mapping of planetary surfaces in recent years (Alexandrov and Beyer, 2018; Grumpe and Wöhler, 2014; Liu et al., 2018). By simulating the surface scattering and imaging process, the underlying information of each pixel's intensity, such as the albedo and topographic gradient, can be obtained and then used to refine the existing 3D shape (Pentland, 1984). In addition to its use in pixel-wise reconstruction, the photoclinometric method can retrieve 3D information from a single image and is thus applicable to more scenarios than photogrammetric methods. Nevertheless, photoclinometric methods suffer from common problems, such as an accumulation of errors, a reliance on albedo information, and a lack of theoretical-based accuracy (Wu et al., 2018). Additionally, the atmosphere above the Martian surface introduces extra difficulties in the application of photoclinometric methods on Mars, as the recovered gradients may be contaminated by the atmosphere (Liu and Wu, 2020). In light of this, the proposed

photoclinometric method is performed on the basis of the photogrammetric DEM, to preserve the geometric accuracy offered by photogrammetry and further enhanced with the photoclinometric details.

3.3.1 Overview of the Photoclinometric Approach

Figure 3.8 illustrates the workflow of the photoclinometric approach for refining the photogrammetric DEM generated in the previous step, taking into account the atmospheric effect using along-track stereo images. We adopt the iterative hierarchy strategy to achieve the pixel-wise reconstruction layer by layer. Initially, the images are down-sampled to a resolution consistent with the resolution of the photogrammetric DEM. The surface albedo is initialized by the weighted average according to the input DEM and the images. Atmospheric parameters such as optical depths are estimated from multiple images based on a stereo method (Hoekzema et al., 2010; Hoekzema et al., 2004). Then, with the emission angles, sun incidence angle, and sun azimuth associated with each image, four types of cost functions are established to calculate and optimize five parameters for each pixel: horizontal gradient, vertical gradient, albedo, optical depth, and aerosol scattering parameter. These parameters could be up-sampled and serve as the initial value of the next layer until the pixel resolution is reached. Once the optimization of the last layer is achieved, the elevation of each pixel can be calculated according to the horizontal gradients and the vertical gradients (Frankot and Chellappa, 1988).

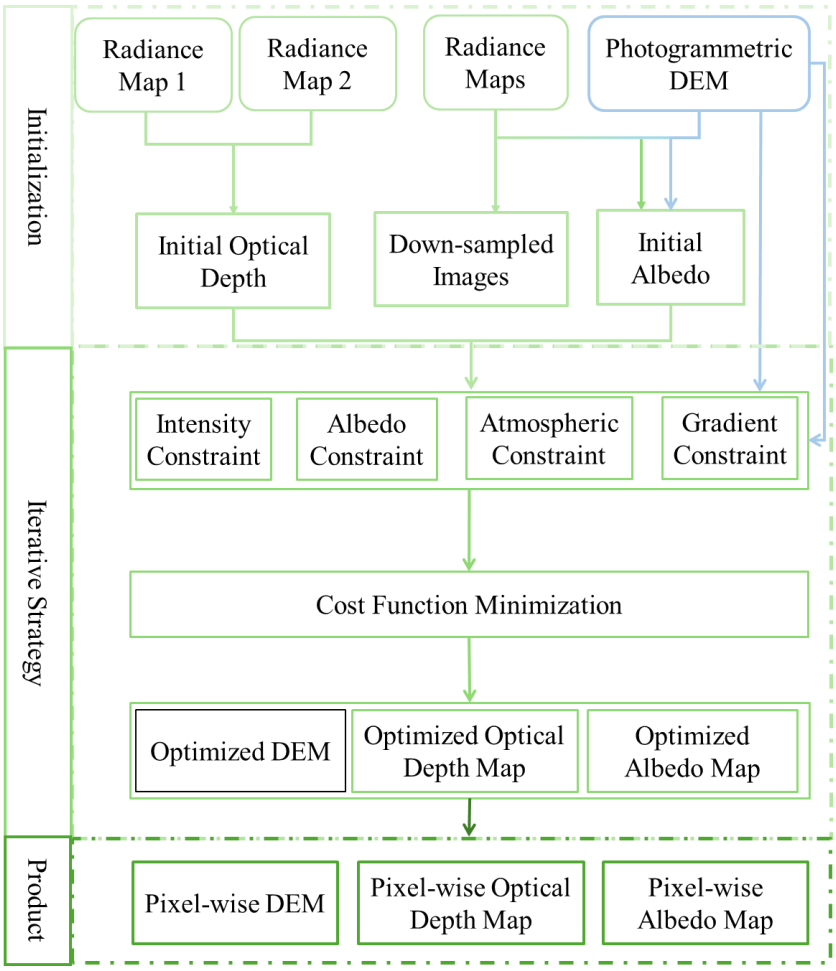


Figure 3.8 The overall workflow for the proposed photoclinometric approach.

3.3.2 Photometric Modeling of the Planetary Surface

The core idea of photoclinometry is to simulate the entire process of the sunlight being reflected by the terrain surface and captured by the camera. A mathematical model to describe this process is thus required, which is typically simplified as:

$$I = AR \tag{3-7}$$

where I is the intensity , or radiance factor, of the observed image derived from the observed digital number, scaling, and offset factor. A is the terrain-dependent albedo and R is the reflectance of light from a surface. Notably, this reflectance model simulates the surface reflectance that is illuminated by direct sunlight and then propagated directly to the sensor.

Various photometric models can be used to describe R, and the lunar-lambert model which has robust performance on the lunar and Martian surface, is selected in this study.

$$R = (1 - \varepsilon)\mu_{sun} + \varepsilon \frac{2\mu_{sun}}{\mu_{sun} + \mu_{camera}} \tag{3-8}$$

where R is a product of μ_{sun} (the cosine of the incidence angle), μ_{camera} (the cosine of the emission angle), and ε (a function of the phase angle). The incidence angle is defined as the angle between the direction of the incident sunlight and the local normal to the surface at the point of illumination. The emission angle is defined as the angle between the line of sight from the sensor to the target and the local normal to the surface at the point of observation. Defining all operations in the local Cartesian coordinate system, the sunlight vector v_{illu} and camera vector v_{cam} could be defined as:

$$v_{illu} = \begin{bmatrix} \cos(azi) & -\sin(azi) & 0 \\ \sin(azi) & \cos(azi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ \sin(zen) \\ \cos(zen) \end{bmatrix} \tag{3-9}$$

where $\cos(azi)$ and $\sin(azi)$ refer to the cosine and sine operator of the azimuth, and $\cos(zen)$ and $\sin(zen)$ refer to the cosine and sine operator of the elevation. And the local normal is defined by the horizontal gradient p and the vertical gradient q, as:

$$\begin{cases} p = \dfrac{dZ}{dx} = \left(Z_{i,j+1} - Z_{i,j}\right)/s \\ q = \dfrac{dZ}{dy} = \left(Z_{i,j} - Z_{i+1,j}\right)/s \end{cases} \tag{3-10}$$

And s denotes the resolution of the DEM. While the phase angle is defined as the angle between the above incident sunlight and the line of sight from the sensor to the target, typically measured in the plane containing the sun, the target, and the sensor. The local normal is thus required to be defined.

3.3.3 Initial Optical Depth Retrieval

Unlike lunar images, it is quite ubiquitous for Martian images to be affected by the atmosphere. Therefore, gauging the optical depths to correct the influences of the atmosphere is of vital importance for the photoclinometric processing of Martian images (Hoekzema et al., 2010; Hoekzema et al., 2004). With the inherent along-track stereo trait, multiple images are taken for the same place within a short time, which favors the use of a stereo-method for the estimation of optical depths. And the model described in Equation (3-7) could be extended, as:

$$I = Be^{-\frac{\tau}{\mu_{camera}}} + \lambda \tag{3-11}$$

where B is the original intensity without atmospheric intervention, τ is the optical depth, μ_{camera} is the cosine of the emission angle, and λ represents the aerosol scattering parameter. Since aerosol distribution is usually very small, the contrast in λ can be ignored, leading to:

$$C_I^i \approx e^{-\frac{\tau}{\mu}} C_B^i \tag{3-12}$$

If the contrast of B of the i^{th} image C_B^i can be approximately equal to the j^{th} image C_B^j, then the relationship between the contrast of I of the i^{th} image C_I^i and the j^{th} image C_I^j can be modeled as:

$$\ln C_I^i - \ln C_I^j \approx \tau \frac{\mu^i - \mu^j}{\mu^j \mu^i} \tag{3-13}$$

Thus, τ can be calculated through:

$$\tau \approx \frac{\mu^j \mu^i}{\mu^i - \mu^j} \ln \frac{C_I^i}{C_I^j} \tag{3-14}$$

It is worth noting that the stereo-based method yields one optical depth for each input region, which can provide an initial value for the following photoclinometric process.

3.3.4 Photoclinometric Functions Establishment

Inherently, photoclinometry is an optimization problem, which leverages multiple observations to solve for the unknown parameters. The optimization functions thereafter play a decisive role in achieving favorable results. Typically, four types of constraints are involved, namely: intensity constraint, atmospheric constraint, gradient constraint, and albedo constraint.

(1) Intensity constraint

The intensity constraint measures the similarity between the observed image $\widehat{I^i}$ captured by the camera, and the image I^i calculated by the aforementioned mathematical problem. Ideally, if the mathematical model rigorously depicts the entire light transmission process, these two intensities should be identical, and the mean absolute error (MAE) is thus a widely used metric for calculating the difference.

$$v_{Intensity} = f_{SSIM}\left(I^i, \widehat{I^i}\right) + f_{MAE}\left(I^i, \widehat{I^i}\right) \tag{3-15}$$

Instead of using only the absolute error to quantify the difference, this study applies the structural similarity index measure (SSIM) (Wang et al., 2004) to the above constraint. The SSIM is a commonly used index of structure information having three main parts: the illumination $l\ (x, y)$, contrast $c\ (x, y)$, and structure $s\ (x, y)$. These parts are modeled as:

$$\begin{aligned} l(x,y) &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \\ c(x,y) &= \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \\ s(x,y) &= \frac{2\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \end{aligned} \tag{3-16}$$

where μ_x and μ_y denote the mean intensity, σ_x and σ_y denote the standard deviation,σ_{xy} is the covariance, $C_1 = (K_1L)^2$, $C_2 = (K_2L)^2$, $C_3 = C_2/2$, $K_1 \ll 1$, $K_2 \ll 1$, and L is 255 for 8-bit grayscale images. The SSIM index is thus calculated as:

$$f_{SSIM}(x,y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_x\sigma_y + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{3-17}$$

There are two benefits of using the SSIM: (1) The results of photoclinometry rely heavily on the intensity of each pixel, but it is complicated to recover the actual surface intensity of each image as a result of the atmosphere. Therefore, for images with strong atmospheric effects, such as those acquired during dust storms, the use of the pixel-based mean absolute error may further exacerbate this condition by introducing the noise into the DEM; (2) The orthoimages may not guarantee conformity well for the inherent occlusion problem and for small inaccuracies in the photogrammetric DEM. Conflicts thus arise among multiple images. Therefore, we suggest combining the mean squared error (MSE) loss and the structural similarity (SSIM) loss to maintain geometric accuracy while minimizing noise introduction.

(2) Atmosphere constraint

Second, atmospheric modeling is necessary when adopting photoclinometry for the accurate reconstruction of the Martian surface. An atmospheric constraint is introduced for the estimation of the optical depth at each pixel. The optical depth should be proportional to the elevation (Hoekzema et al., 2010), and the SSIM loss is thus generated to enforce structural consistency between the optical depth O and height H as:

$$v_{Intensity} = f_{SSIM}(O,H) \tag{3-18}$$

The aerosol scattering parameter λ is initially assigned a small value around zero and then optimized with other parameters. It is assumed that if the gradient and albedo are well constrained, the remaining component will be left to λ, thus eliminating the effect on the slope. A stricter imaging model is formulated with the pixel-wise optical depth and aerosol distribution.

(3) Gradient constraint

As previously mentioned, the unconstrained photoclinometric process may produce results that greatly deviate from the ground truth (Grumpe and Wöhler, 2014). It is thus necessary to control the reconstructed gradients ($p = \frac{dZ}{dx}$, $q = \frac{dZ}{dy}$) at all points in the overall cost function, and the weight w is used to balance the importance of the two gradients:

$$v_{Gradient} = w\|p - \hat{p}\|^2 + (1-w)\|q - \hat{q}\|^2 \tag{3-19}$$

Additionally, integrability is enforced for faster convergence and lower error (Frankot and Chellappa, 1988):

$$v_{Integral} = \left\|p_q - q_p\right\|^2 \tag{3-20}$$

where $p_q = \frac{\partial}{\partial y}p$ and $q_p = \frac{\partial}{\partial x}q$ are the second partial derivatives of the surface. This formula simply describes that the elevation of a certain point is unrelated to the path of integration.

(4) Albedo constraint

The albedo is initialized by the original photogrammetric DEM and the images according to Equation (3-11). Therefore, unlike the albedo measured by the Thermal Emission Spectrometer onboard the *Mars Global Surveyor*, which reveals the inherent geological properties directly, the albedo here is terrain-dependent and may contain a certain amount of shadow information. An albedo constraint based on the total variation (Rudin et al., 1992) is established to model the albedo more accurately. The first term limits the deviation of the optimized albedo A_n from the initial value $\widehat{A_n}$. The second term assumes that adjacent pixels in continuous landforms share similar albedo.

$$v_{Albedo} = \frac{1}{n}\sum_{n}(A_n - \widehat{A_n})^2 + \sum_{i,j}\sqrt{|A_{i+1,j} - A_{i,j}|^2 + |A_{i,j+1} - A_{i,j}|^2} \tag{3-21}$$

Besides the cost functions, the weights for each constraint determine the visual texture and the geometrical accuracy as well and need to be adjusted using values obtained through experimental analysis.

3.3.5 Optimization of the Photoclinometric Functions

The image pyramid is employed to facilitate the optimization process, with each layer featuring increasing resolution. Instead of enriching the observation constraint, the down-sampled pyramid allows the photoclinometry algorithm to deviate significantly from the input photogrammetric DEM. Another reason is that inevitable noises exist in the generated photogrammetric DEM due to the dense matching algorithm and the exterior orientation parameters; the down-sampled DEM not only preserves the large-scale landforms derived from the DEM but also mitigates the noise effect. Specifically, the input images are first down-sampled to fit the resolution of the input photogrammetric DEM, and pyramids of the image and DEM are established. The cost function is minimized for each pyramid layer until the maximum number of iterations is reached or the loss threshold is satisfied.

Apart from the cost functions, the optimization algorithm is also important for the calculation of the results. Conventional optimization methods, such as Gradient Descent (Wu et al., 2018) and Newton's method and its variants (Alexandrov and Beyer, 2018), are prone to getting stuck in local optima and exhibit significant performance gaps compared to recently developed optimizers. The Adam optimizer (Kingma, 2014) has become the default configuration for many deep learning tasks. Its main feature is that it takes into account both first-order and second-order momentum, ensuring stable and fast convergence of the iterative process. Assuming the parameters to be updated in the optimization function f are θ, the updated gradient values are g:

$$g_t = \nabla_\theta f_t(\theta_{t-1}) \tag{3-22}$$

wherein t is the current gradient update step. For the obtained gradient g_t, its first-order momentum m_t is:

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \tag{3-23}$$

where β_1 is the exponential decay rate, controlling the weights of different momentum batches, with a default value of 0.9. The formula for the calculation of the second-order momentum v_t is defined as:

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \tag{3-24}$$

The value β_2 is the exponential decay rate, which defaults to 0.999. Equations (3-23) and (3-24) are not applicable in the initial condition $t = 0$ where the first-order and second-order moments are zero or the accumulated quantities are too small to be stable. To stabilize the gradient update in the initial stage, the actual update formulas for the first-order and second-order moments are:

$$\widehat{m_t} = m_t/(1 - \beta_1^t) \tag{3-25}$$

$$\widehat{v_t} = v_t/(1 - \beta_2^t) \tag{3-26}$$

Accordingly, the update of the θ_t could be established as:

$$\theta_t = \theta_{t-1} - \alpha \cdot \widehat{m_t}/(\sqrt{\widehat{v_t}} + \varepsilon) \tag{3-27}$$

The addition of a small value ε is used to prevent division by zero. From Equation (3-27), it can be seen that the first-order momentum controls the direction of gradient updates, while the second-order gradient takes into account the magnitude of past gradient updates, i.e., the larger the amplitude of past gradient oscillations, the smaller the updated gradient, which helps to stabilize the gradient update process and accelerate convergence.

As for the calculation efficiency, even the above cost functions are similar to those used in bundle adjustment, each pixel of the images in photoclinometry is subject to four types of constraint, leading to an exponential growth in computation, which is comparable to optimizing approximately 10,000 parameters in bundle adjustment. Therefore, the conventional method of using the Ceres library (Agarwal et al., 2012) supplemented by OpenMP for multi-threaded processing on a central processing unit is not a feasible choice. Instead, the TensorFlow software library is used here to take advantage of graphics processing units and compute the minimization problem in parallel. TensorFlow's GPU acceleration capabilities significantly outperform OpenMP-based CPU parallelization, achieving substantial speedups in deep learning computations due to the massively parallel processing capabilities of modern GPUs. By leveraging the thousands of cores on a GPU, TensorFlow can perform complex matrix operations and neural network computations much faster than OpenMP, which is limited to the number of CPU cores available. This results in TensorFlow with GPU acceleration being 10-100 times faster than OpenMP on CPU for many deep learning workloads.

Additionally, by leveraging the TensorFlow framework, the parameters could be automatically fine-tuned the hyperparameters (e.g., the learning rate, maximum iterations, and overall loss) and complex system parameters (e.g., the Lambert weight ε and gradient weight w) are auto-adjusted using the Neural Network Intelligence (NNI) toolkit (Gridin, 2022), which is a popular open-source toolkit for automated machine learning (AutoML) and hyperparameter tuning. Its strength lies in its ability to efficiently and effectively fine-tune model parameters using a variety of optimization algorithms, such as grid search, random search, and Bayesian optimization. By leveraging NNI, users can automate the tedious process of hyperparameter tuning, accelerating the development and deployment of high-performance machine learning models.

3.4 Experimental Evaluation

To evaluate the performance of the proposed algorithm, various images are leveraged for comprehensive analysis, namely, CTX, HiRIC, and HRSC. CTX has only one CCD and employs a cross-track stereo configuration, enabling the production of sub-ten meter imaging of the planetary surface, resulting in an approximate 20-meter topographic product. In contrast, HiRIC is also a cross-track stereo camera, featuring three CCDs mounted in one focal plane to achieve a large field of view while acquiring sub-meter resolution images. Meanwhile, HRSC is the only along-track stereo camera orbiting the Moon and Mars, consisting of nine parallel CCDs. Even its resolution is only 12.5 m/pixel, the global stereo coverage makes it an indispensable data source for 3D mapping of the planetary surface. A comprehensive evaluation of the proposed algorithm can be achieved by conducting experiments on a variety of cameras, which will help to assess its effectiveness and limitations. The detailed parameters are listed in Table 3.1.

Table 3.1 The parameters of the cameras used for the experimental evaluation.

	Resolution (m / pixel)	CCD amount	CCD pixels	Stereo Configuration	Organization
CTX	6	1	5064	Cross-track	NASA
HiRIC	0.7	3	6199	Cross-track	CNSA
HRSC	10 - 20	9	5184	Along-track	DLR

3.4.1 Experimental Evaluation of the CTX Dataset

a. Dataset description

The Mars Reconnaissance Orbiter's (MRO) Context Camera (CTX) is a high-resolution imaging instrument designed to capture detailed images of the Martian surface. Having orbited Mars for nearly two decades, it has acquired numerous images, enabling the creation of a global mosaic covering 99.5% of the surface from 88°S to 88°N. With a resolution of 6 m/pixel and a

single CCD in the focal plane, the CTX facilitates efficient 3D reconstruction, effectively revealing Martian landforms at an appropriate scale. The detailed parameters are listed in Table 3.2.

Table 3.2 Parameters of the CTX.

Parameters	**CTX**
Focal length	350 mm
Active pixels per CCD line	5064
Pixel size	7.00 μm
Spatial resolution	6.0 m/pixel @ 300 km
Swath per orbit	30 km @ 290 km
Spectral filters	one panchromatic
Panchromatic	500-800 nm

To evaluate the performance of the proposed algorithm, four high-resolution stereo image pairs covering the entire McLaughlin crater region were selected based on their optimal overlap and spatial resolution. These image datasets were used to test the algorithm's ability to integrate laser altimetry and photogrammetry, and to assess its accuracy and robustness in generating high-quality DEMs of the Martian surface. The detailed information on these stereo pairs is listed in Table 3.3, with each stereo pair sharing a minimum 15 degree intersection angle.

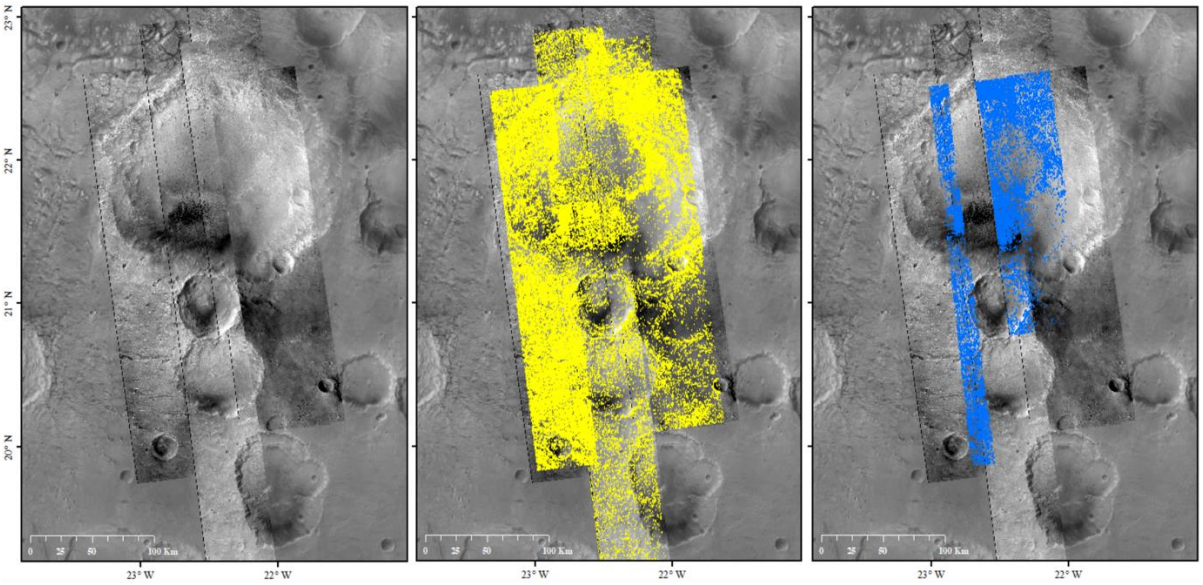
Table 3.3 Details about the experimental images.

	Image ID	Captured Time	Emission Angle	Length (pixels)
pair 1	F18_042780_2014_XI_21N023W	2015-09-11	16.01°	28672
	B17_016183_2014_XN_21N023W	2010-01-08	0.35°	26624
pair 2	F02_036569_2004_XN_20N022W	2014-05-16	23.23°	52609
	F03_036859_2005_XN_20N022W	2014-06-07	2.28°	52224
pair 3	F01_036015_2020_XN_22N022W	2014-04-02	15.72°	20480
	P14_006597_2020_XI_22N022W	2007-12-23	0.1°	32768
pair 4	J06_047158_2025_XN_22N022W	2016-08-18	2.03°	43008

J07_047448_2024_XN_22N022W	2016-09-09	26.68°	45056

b. Experimental evaluation in the McLaughlin crater region

The McLaughlin region is of central interest in the planetary community, as it may have been an ancient lake, potentially preserving hints of the existence of water on Mars. The region's unique geological features, including its impact crater and sedimentary deposits, suggest that it may have played a key role in the planet's hydrological history. The rigorous 3D maps are thus important to support all kinds of scientific analysis, and further landing site selection evaluation. The distribution of these four stereo pairs and the generated DEM are visualized in Figure 3.9 (a). The generation of this DEM involves multiple stages, including RPC generation, tie-point matching, block adjustment, epipolar rectification, dense image matching, and DEM interpolation. These reasonable and favorable results reflect the effectiveness of all the above algorithms. As aforementioned, the matched tie-points serve as the foundation for the entire bundle adjustment process, to retrieve the correct and consistent EOs. As presented in Figure 3.9 (b), for all four inner-track stereo pairs, the tie-points are evenly matched throughout the entire region, ensuring not only the consistency between the stereo pairs but also the sampling of a sufficient number of ground control points from the laser altimetry DEM. In terms of the cross-track tie-points shown in Figure 3.9 (c), dense tie-points are matched between pair 1 and pair 2, and between pair 2 and pair 3. However, the overlapping region covered by pair 3 and pair 4 suffers from severe textureless issues, resulting in sparse tie-points. However no apparent gaps are observed from the generated DEM, this may be attributed to the effectiveness of the balancing weight and the precise inner-track bundle adjustment.



(a) (b) (c)

Figure 3.9 (a) The experiment CTX images overlaid on the CTX global mosaic, (b) and (c) are the distributions of the tie-points for inner-track and cross-track stereo pair, respectively.

The DEM product is visualized in a 3D manner and presented in Figure 3.10, which includes one orthographic view and two specific perspectives as shown in Figure 3.10 (a). The complete product is composed of the four stereo tracks mentioned above. Notably, there are no visible gaps among these three views, demonstrating that a substantial number of tie-points were detected. The subsequent bundle adjustment effectively utilizes these tie-points to resolve any inconsistencies among the EOs of these tracks. Owing to the capability of the dense matching, abundant details are retrieved, which is more apparent in Figure 3.11.

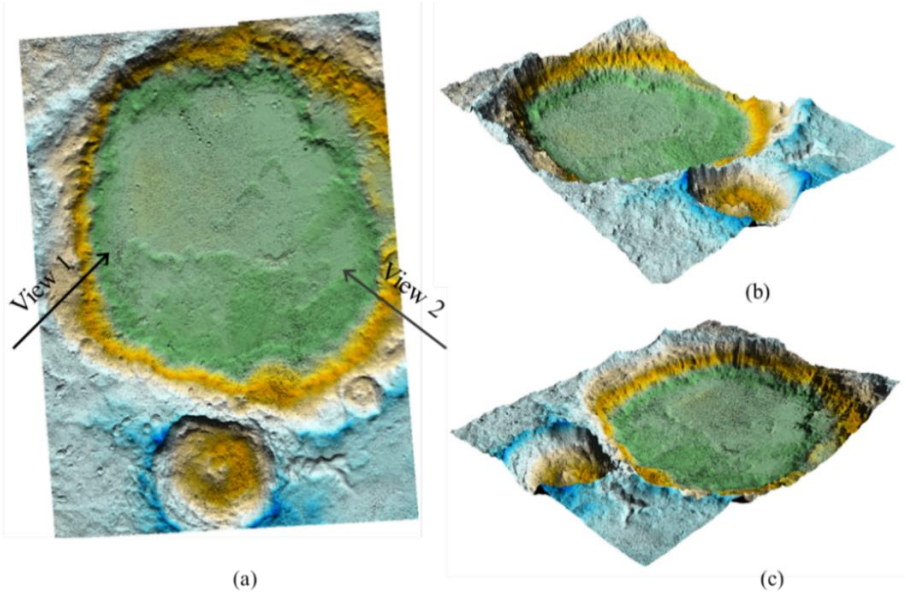


Figure 3.10 3D views of the generated DEM. (a) the orth view, (b) and (c) are the views corresponding to the arrows marked in (a).

In Figure 3.11, a qualitative comparison is made between the generated DEMs and those from MOLA (463 m/pixel) and HRSC (150 m/pixel). The first row displays the hillshade of the DEM, while the second row provides a zoomed view of the lower left corner. The figure indicates that the CTX DEM generated by the proposed algorithm effectively captures the landforms depicted in the images, significantly enhancing quality compared to the two widely used products. The details, such as the ridges of the craters and the smaller craters within the larger one, are all well-reconstructed with accurate shapes.

Quantitative analysis is also performed, with two representative profiles drawn vertically and horizontally. These profiles are compared with the two benchmark datasets mentioned earlier. The profiles of the three DEMs are well-aligned, and the CTX DEM generated by our algorithm is closer to the laser DEM. As it includes more details, the profiles are more oscillatory. The statistics are shown in

Table **3.4**, and the mean differences are around 50 meters. According to the profiles, the major differences occur at the boundaries of the craters and in regions with large landforms. The two benchmark DEMs overlook these details due to their limited resolution.

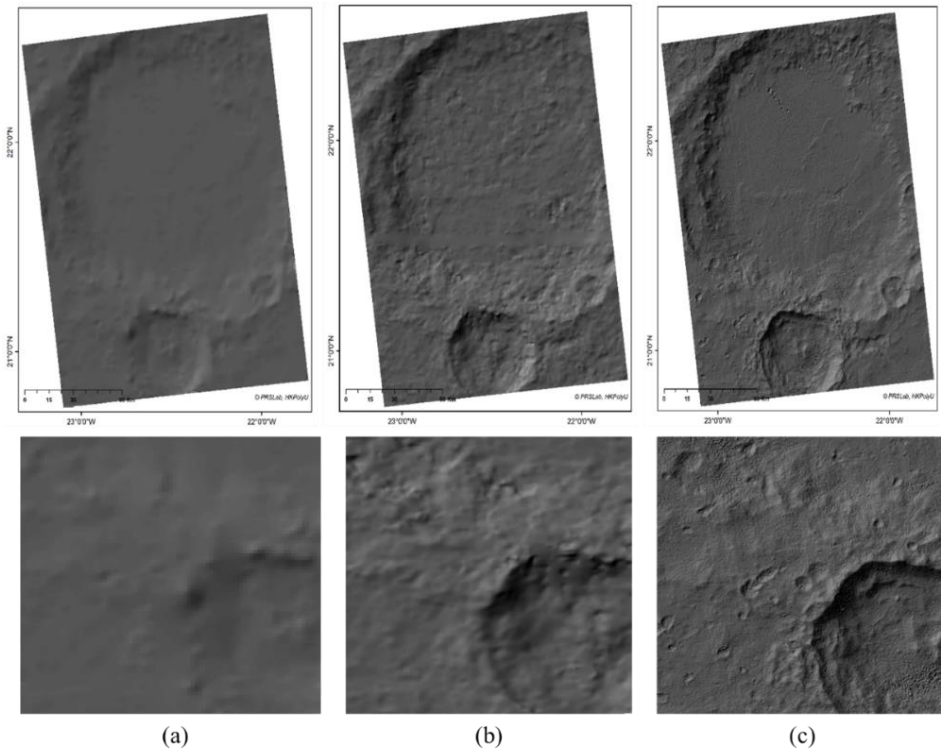


Figure 3.11 The Comparison of the Hillshade of the DEM products. (a) MOLA DEM (463 m/ pixel), (b) HRSC DEM(150 m/pixel), and (c) Our DEM (20 m/pixel).

Table 3.4 Quantitative analysis of the two profiles.

	Difference (m)		
	Mean	Maximum	MSE
	HRSC / MOLA	HRSC / MOLA	HRSC / MOLA
Profile 1	50.08 / 48.6	351.1 / 328.0	76.21 / 73.52
Profile 2	44.74 / 40.8	186.7 / 267.0	57.59 / 60.54

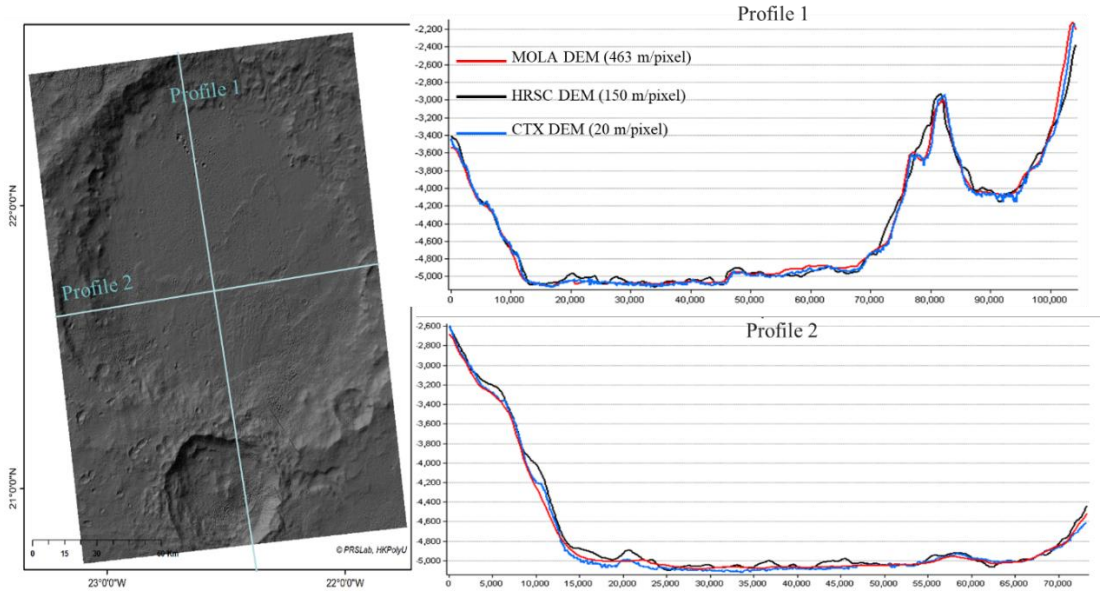


Figure 3.12 Profiles of the DEM products.

3.4.2 Experimental Evaluation of the HiRIC Dataset

a. Dataset description

China's first Martian exploration mission, the Tianwen-1 probe carrying the Zhurong rover, was launched in July 2020 and entered the Martian orbit in February 2021, successfully landing Zhurong in southern Utopia Planitia on May 15, 2021 (Liu et al., 2022). From mid-February to early May 2021, the Tianwen-1 probe performed detailed investigations of the potential landing site in southern Utopia Planitia using its onboard instruments. Among these investigations, images obtained by the High-Resolution Imaging Camera (HiRIC) were used for high-precision and high-resolution 3D topographic mapping to facilitate the selection of the best landing site for the Zhurong rover (Gwinner et al., 2016; McEwen et al., 2007; Wu et al., 2022) and provide data support for various scientific studies (Wan et al., 2020).

The HiRIC onboard the Tianwen-1 orbiter is a pushbroom camera that can achieve high-resolution imaging (0.5 m/pixel at an altitude of 265 km) with a focal length of 4,640 mm (Meng et al., 2021). The HiRIC contains three charge-coupled devices (CCDs), named CCD1, CCD2, and CCD3, on the same image plane to achieve a swath of up to 9 km. There are several

pixel offsets between any two adjacent CCDs perpendicular to the flying direction. Moreover, CCD2 in the middle has an offset of 0.47° (equal to more than 4,000 pixels) with the other two CCDs along the flying direction, which cannot form a unified image with the other two CCDs for photogrammetric processing. With respect to stereo imaging, the HiRIC adopts a side-slewing stereo strategy by revisiting the region at a 22° convergent angle. However, the overlapping region between the stereo pair is uncertain because it includes six CCD images, each 6,144 pixels wide and more than 240,000 pixels long. In addition, the overlapping region between adjacent CCDs is as narrow as approximately 140 pixels, which makes it difficult to preserve the internal consistency of multiple CCD images within one orbit. The detailed parameters are listed below.

Table 3.5 Parameters of the HiRIC.

Parameters	**HiRIC**
Focal length	4,640 mm
Active pixels per CCD line	6144
Pixel size	8.75 μm
Spatial resolution	0.5 m/pixel @ 265 km
Swath per orbit	9 km @ 265 km
Spectral filters	one panchromatic, four color
Panchromatic	450–900 nm
Blue (BL)	450–520 nm
Green (GR)	520–600 nm
Red (RE)	630–690 nm
Near-infrared (IR)	760–900 nm

During the initial flight, the designed slewing stereo convergent angle for the HiRIC is 22°, with a two-day-long revisit time. With an approximately 300-km orbit height, the HiRIC image resolution can reach 0.7 m/pixel, which allows high-resolution mapping to reveal the detailed topography. Three stereo pairs of HiRIC images of the Zhurong landing region are selected for

evaluation, as shown in Figure 3.13. The forward- and backward-looking images of the first orbit (marked by the red box) covering the Zhurong landing site (red cross in Figure 3.13 (a)) were acquired on March 24 and 26, 2021, with more than 18,000 × 250,000 pixels for each image. The second and third stereo pairs have a length of up to 260,000 pixels, covering the eastern region of the first stereo pair. Detailed information for the three stereo pairs of HiRIC images is listed in Table 3.6. Owing to the clear weather during these days, the HiRIC images are of satisfactory quality (e.g., favorable signal-to-noise ratio, consistent illumination). Therefore, a DEM with a resolution of 3.5 m/pixel (five times the image resolution) is generated using the photogrammetric method, as shown in Figure 3.13 (e).

To evaluate the performance of the proposed photogrammetric method, the geometric accuracy and reconstruction details of the HiRIC DEM are compared with reference DEMs generated from the MOLA data and HiRISE images. The MOLA DEM provides the most accurate Martian topography reference data (Gwinner et al., 2016). Several HiRISE images with resolutions of 0.25 m/pixel are available within the Zhurong landing region, and DEMs with a resolution of 1 m/pixel are generated by photogrammetric processing (McEwen et al., 2010; McEwen et al., 2007).

Table 3.6 Information on the HiRIC images used for the experimental analysis.

Orbit No.	Acquisition Date	Image Length (pixels)	Center Latitude and Longitude
0306	06/03/2021	241,000	24.744°, 110.169°
0308	08/03/2021	264,000	24.745°, 110.157°
0316	16/03/2021	253,000	24.715°, 110.033°
0318	18/03/2021	280,000	24.782°, 110.039°
0324	24/03/2021	254,000	24.784°, 109.891°
0326	26/03/2021	280,000	24.730°, 109.879°

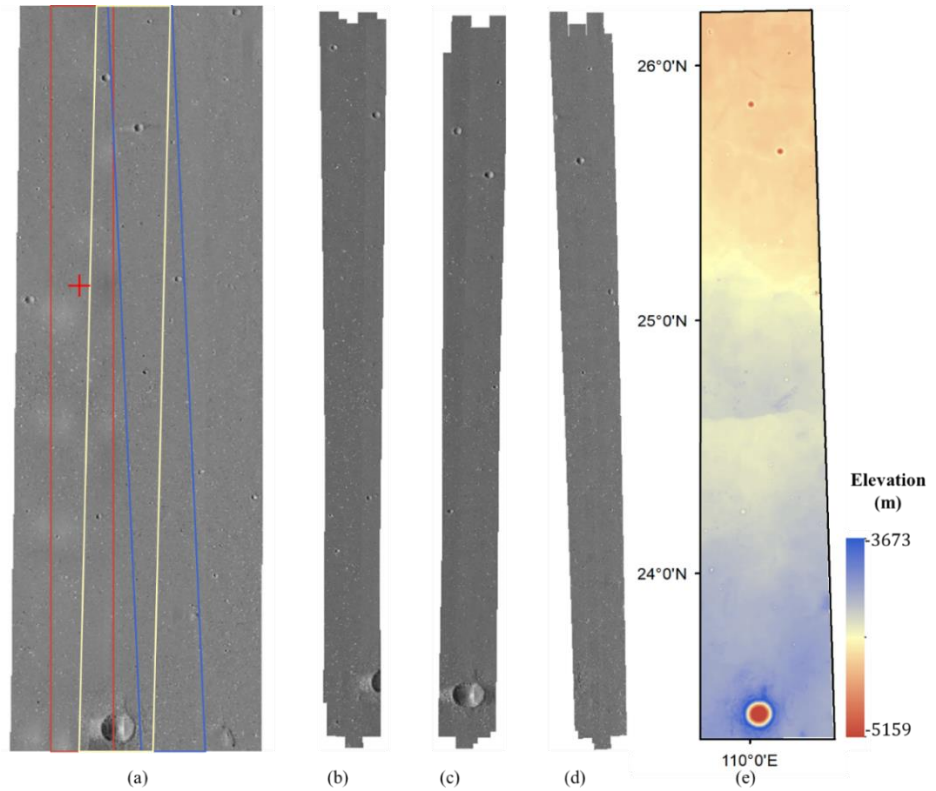


Figure 3.13 The HiRIC images used for the experimental analysis. (a) The HiRIC image coverage overlaid on a HiRIC image mosaic covering the Zhurong landing region. The Zhurong landing site is marked by the red cross; (b)–(d) The separate pairs of HiRIC images with orbit numbers 0324-0326, 0316-0318, and 0306-0308, corresponding to the red, yellow, and blue boxes marked in (a). (e) The DEM (3.5 m/pixel) generated from the HiRIC images.

b. Evaluation of the Block Adjustment of Multi-Orbit Images

Figure 3.14 shows the results of the block adjustment of the three stereo orbits covering the Zhurong landing region, and the quantitative analysis is presented in Table 3.7. Beginning with the adjusted RPCs derived from the single stereo orbit adjustment, the block adjustment eliminates inconsistencies among different orbits of images. In the experiment, orbit 0324-26 in the westernmost is treated as the reference to sequentially adjust the other two orbits. As shown in Figure 3.14 (b), the back-projection error in the image space among the orbits dramatically decreases from dozens of pixels in maximum to sub-pixel, demonstrating the

effectiveness of the block adjustment. This improvement is visualized in the zoomed view in Figure 3.14 (c), where long error vectors are shortened to small points after the adjustment. As no ground control point measured from MOLA is leveraged in the block adjustment to register the HiRIC orbit strictly, the process is still a free network block adjustment.

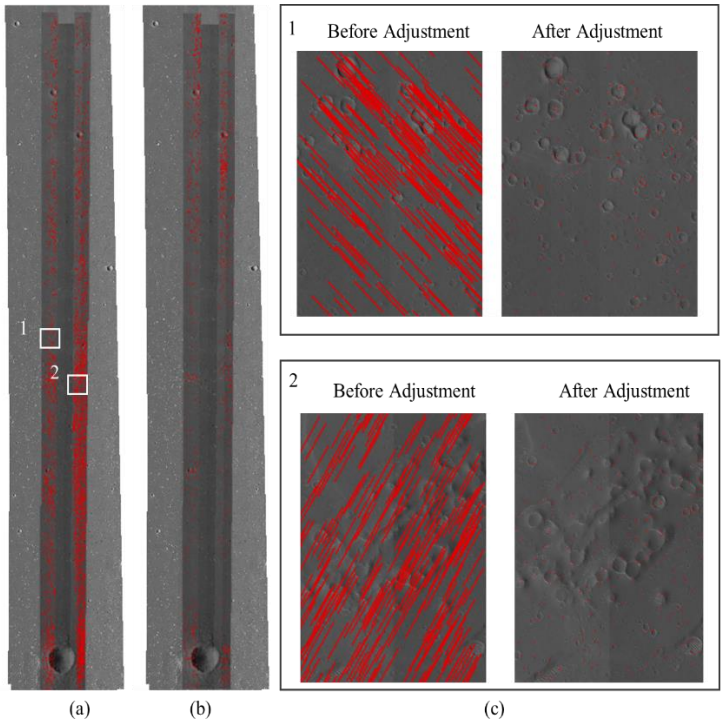


Figure 3.14 (a)–(b) Error vectors illustrating image residuals before and after block adjustment (exaggerated 40 times); (c) zoomed view of the boxes marked in (a).

Table 3.7 Statistics of image residuals of the block adjustment of multi-orbit HiRIC images.

		Image Residuals (pixels)		
		Mean	Maximum	RMSE
Orbit 0306-08				
Before BA	*X*	1.72	15.89	1.04
	Y	2.38	17.85	1.37
	2D	3.02	19.84	1.86
After BA	*X*	0.15	0.55	0.12
	Y	0.18	0.51	0.14

	2D	0.24	0.61	0.21
Orbit 0324-26				
Before BA	*X*	3.02	24.92	1.42
	Y	4.06	27.59	1.89
	2D	6.12	27.90	3.30
After BA	*X*	0.22	0.79	0.14
	Y	0.28	0.83	0.18
	2D	0.43	0.85	0.44

c. Evaluation of the Generated DEM

Figure 3.15 (b) presents the generated HiRIC DEM, revealing an overall smooth slope down to the north and providing a suitable place for rover landing (Wu et al., 2021). The 3D information of distinct landforms (i.e., the large craters, a series of dunes, and ridges) is properly reconstructed. Verified by the corresponding MOLA DEM in Figure 3.15 (a) and the difference maps in Figure 3.15 (c), the global trend of the HiRIC DEM agrees well with the reference MOLA DEM. Apart from the overall difference map, the four profiles marked in Figure 3.15 (a) are also analyzed quantitatively in Figure 3.15 (d). Three statistical indexes (i.e., the maximum, mean, and RMSE) of the profiles and the DEMs are also calculated (Table 3.8). Both the mean deviation and the RMSE are below 10 m, indicating that the imaging geometry is correct, and the overall workflow is feasible. However, orbit 0316-18 suffers from a relatively large deviation from the MOLA DEM due to the large craters at the bottom of the region. Despite the omitted laser point in MOLA on the right side of the crater, the large resolution difference leads to a deviation inside the crater where elevation changes dramatically. However, the overall deviation is still within a reasonable range (Gwinner et al., 2009). Additionally, the more than 100-fold resolution difference results in some detailed textures (i.e.,

small craters, cones, troughs) being recovered by the HiRIC DEM but missed by MOLA. The subtle misalignment between the two DEMs may exacerbate this difference.

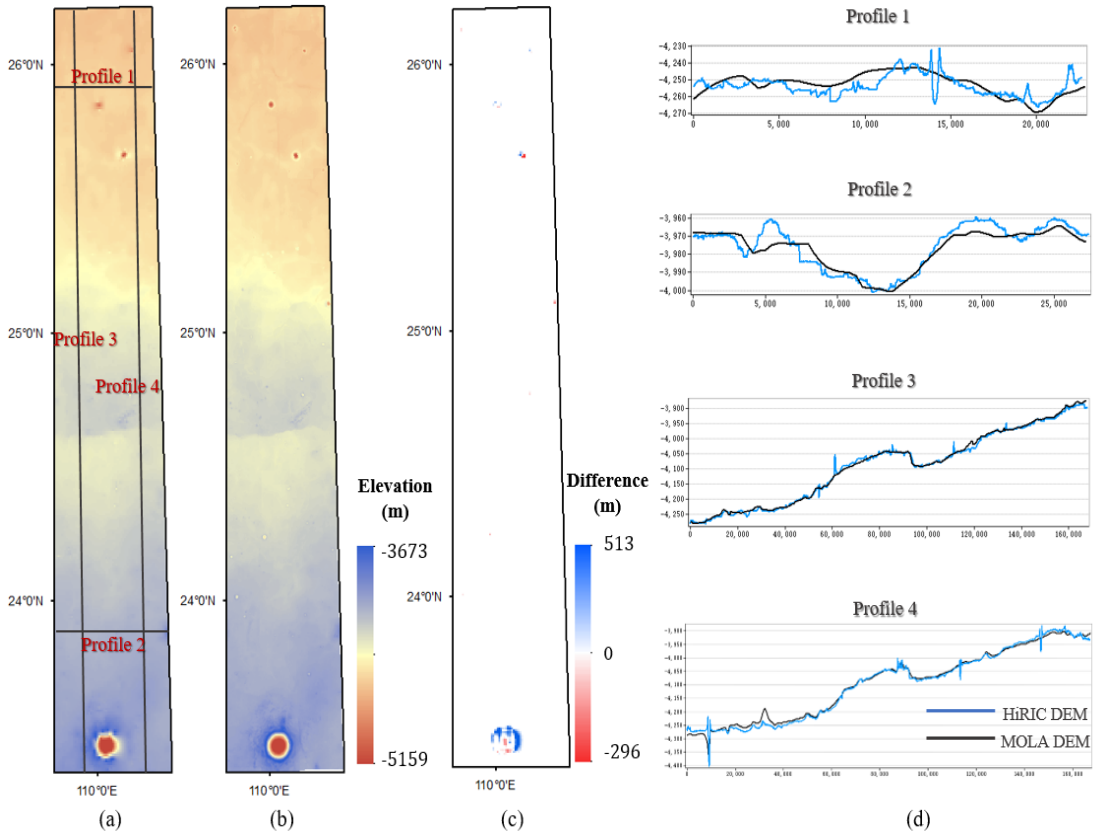


Figure 3.15 The accuracy analysis of the produced HiRIC DEM. (a) The corresponding MOLA DEM (463 m/pixel); (b) the produced HiRIC DEM (3.5 m/pixel); (c) the difference map between the produced HiRIC DEM and MOLA DEM; and (d) the profiles of the lines marked in (a).

Table 3.8 Statistics of the accuracy analysis of the HiRIC DEM

	Absolute Elevation Differences (m)		
	Mean	Maximum	RMSE
Profile 1	5.46	21.48	4.23
Profile 2	5.36	23.36	4.43
Profile 3	8.01	55.93	6.35
Profile 4	4.65	9.69	5.17
Difference Map	8.69	513.21	19.77

To examine the details of the generated HiRIC DEM, a local region surrounding the Zhurong landing site is selected and compared with the 1 m/pixel HiRISE DEM. The region shown in Figure 3.16 is a subset of the 0324-26 orbit, where the Zhurong landing location is indicated by the red cross. As Figure 3.16 (a) shows, this region features a series of troughs, craters, and cones, despite the overall flat trend. These landforms may provide evidence for the eruption of magma or muddy debris (Mills et al., 2021), offering an appropriate landing site in terms of both landing safety and scientific significance. This subset region contains more than 18,000 × 40,000 pixels across the three CCD images. Within this relatively small region, the overlapping region for adjacent CCDs from stereo pairs is approximately a rectangle more than 1,000 pixels wide. As illustrated in Figure 3.16 (b) and (c), the 3D view of the HiRISE DEM and the HiRIC DEM is visually similar; however, the HiRIC DEM is inevitably smoothed due to the original image resolution. Facilitated by sub-pixel LSM, all the distinctive landforms captured by the HiRIC images are well-reconstructed with sufficient detail (i.e., the sand dunes inside the troughs) and precise elevation.

To quantitatively study the geometrical accuracy, two profiles are drawn vertically and horizontally through the entire region (Figure 3.17) for comparison of the HiRIC and HiRISE DEMs, and results are shown in Figure 3.17 (c). Fig. 10 (b) also shows a direct difference map of the two DEMs. Figure 3.17 (b) and (c) show a good level of consistency between the HiRIC DEM and the HiRISE DEM, with only 2.08 m mean error and 1.86 m RMSE, as listed in Table 3.9. Because the 1 m/pixel HiRISE DEM is considered the ground truth of the region, the difference between the two DEMs is the most likely source of error, rather than a simple deviation. The maximum error occurs in the border region with the large cone and crater, owing to the slightly deviated elevation and misalignment.

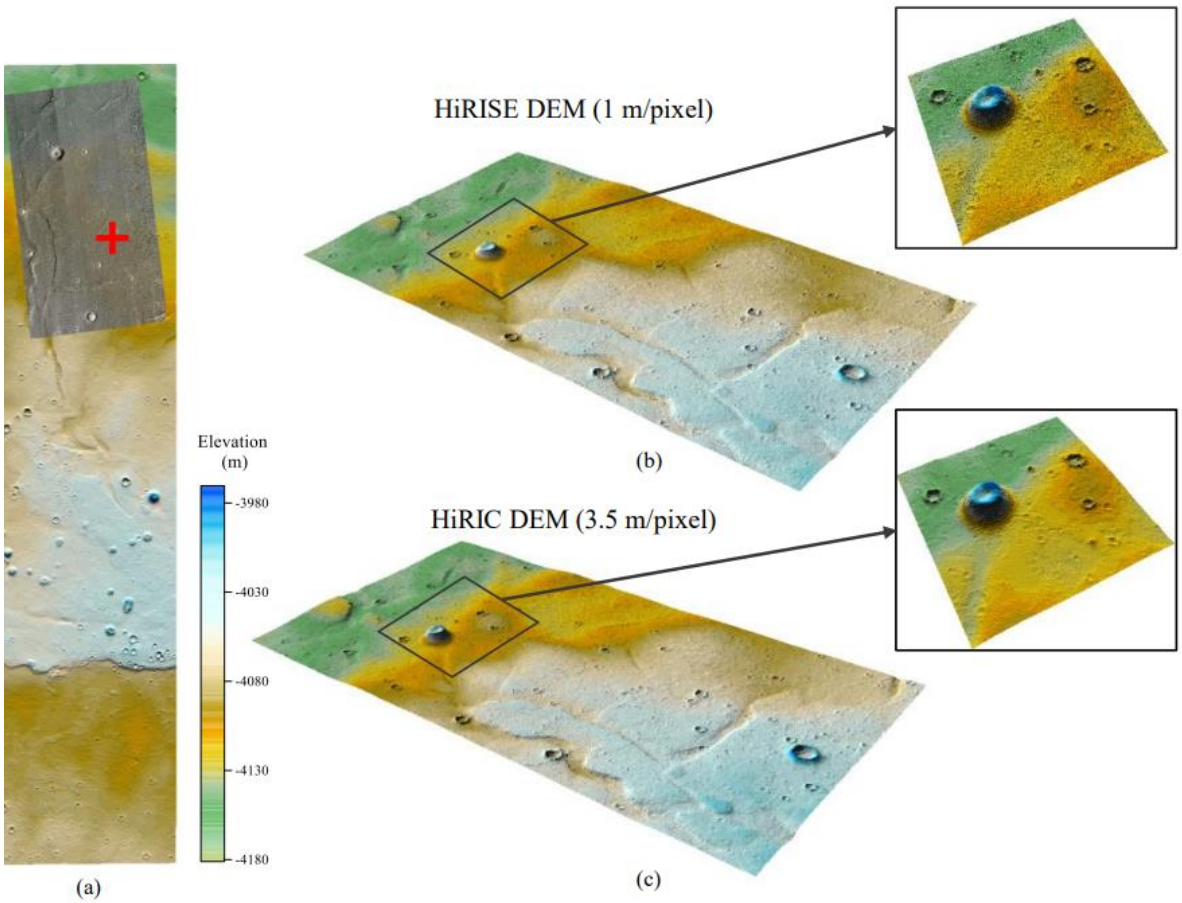
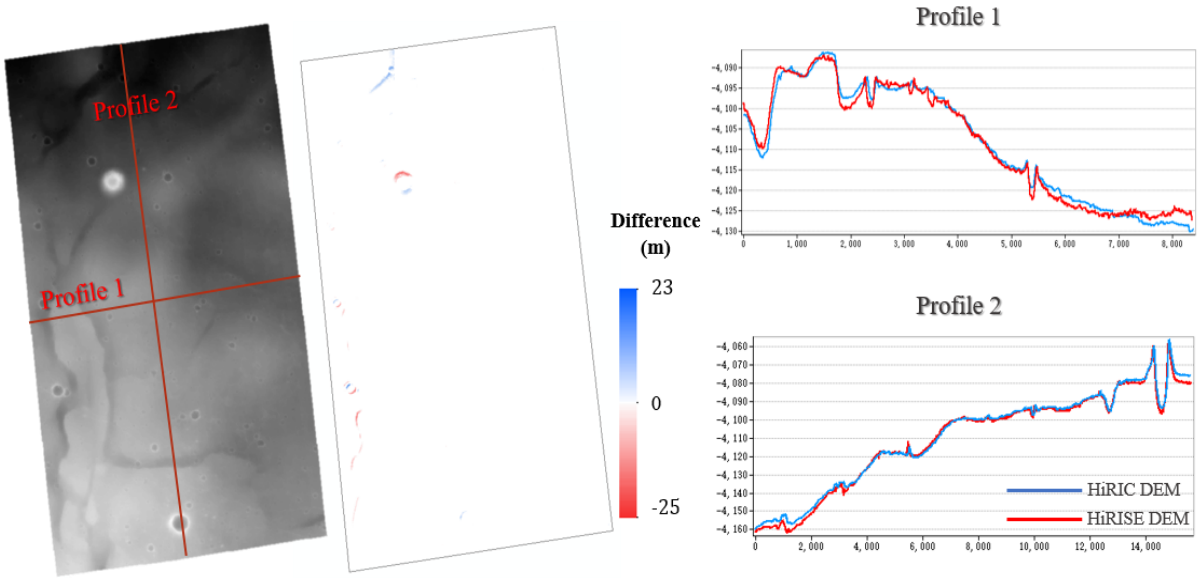


Figure 3.16 3D views of the HiRIC and HiRISE DEMs at Zhurong landing region. (a) The HiRIC DEM covering the Zhurong landing site with the HiRISE images overlaid; The Zhurong landing site is marked by the red cross; (b) 3D view of the HiRISE DEM (1 m/pixel) used for comparison; and (c) 3D view of the HiRIC DEM (3.5 m/pixel) cropped to the same area of the HiRISE DEM.

Table 3.9 Statistics of the comparison between the HiRIC DEM and the HiRISE DEM.

	Mean (m)	**Maximum (m)**	**RMSE (m)**
Difference map	2.08	25.25	1.86
Profile 1	1.48	9.03	1.01
Profile 2	1.62	7.48	1.37



(a) (b) (c)

Figure 3.17 Comparison between the HiRIC and the HiRISE DEMs. (a) Profiles shown on the HiRISE DEM for reference; (b) the difference map between the HiRISE and HiRIC DEMs; and (c) a comparison of the profiles marked in (a).

3.4.3 Experimental Evaluation of the HRSC Dataset

The High Resolution Stereo Camera (HRSC) camera on board *Mars Express* was designed by the DLR and has been orbiting Mars since January 2004 (Neukum and Jaumann, 2004). Aiming at imaging the Martian surface and unmasking the geological evolution of Mars, five panchromatic charge-coupled device (CCD) lines (S1, P1, ND, P1, and S2; with stereo angles of ±18.9° for S1 and S2 and ±12.8° for P1 and P2) and four spectral CCD lines (near-infrared, green, red, and blue) were mounted on the HRSC camera for simultaneous photo-snapping to avoid changes in the illumination and atmospheric conditions. A resolution of 10–20 m/pixel can be reached for most orbits. Given that each CCD line has 5184 active pixels and that there are more than 40,000 scanlines in one orbit strip, one image covers a region with an area larger than 60 × 400 km^2. These merits have led to a wealth of scientific achievements (Neukum and

Jaumann, 2004). After almost 20 years' operation in orbit, the HRSC has recorded observations for approximately 95% of the Martian surface. The three-dimensional (3D) mapping of the Martian surface from HRSC images may offer global mapping products, e.g., digital elevation models (DEMs) with a medium-resolution (50–100 m/pixel) (Oberst et al., 2008), superior to other global DEM data for Mars (e.g., the Mars Orbiter Laser Altimeter (MOLA) DEM with a spatial resolution of 463 m/pixel) (Smith et al., 2001). The detailed parameters are listed below.

Table 3.10 Parameters of the CTX.

Parameters	HRSC
Focal length	175 mm
Active pixels per CCD line	5184
Pixel size	7.00 μm
Spatial resolution	10 m/pixel @ 250 km
Swath per orbit	52.5 km @ 250 km
Spectral filters	five panchromatic, four color
Panchromatic	675±90 nm
Blue (BL)	440±40 nm
Green (GR)	540±45 nm
Red (RE)	750±25 nm
Near-infrared (IR)	955±40 nm

Two representative HRSC datasets are used in a systematic experimental analysis to evaluate the performance of the proposed approach. The orbit h5145 passes over Arabia Terra, which contains abundant remarkable landforms (e.g., valleys, basins, craters, and cones) and this dataset has been studied by many researchers (Gwinner et al., 2009). The acquisition date of the orbit h5145 dataset is between the northern hemisphere Martian spring and summer, a period in which there is favorable weather without visible dust storms. To verify the importance and practicability of the proposed approach, another more challenging dataset (orbit hd674) is also used. Orbit hd674 covers the landing region of the Chinese Mars rover (Zhurong) (Wu et

al., 2021) in the Utopia Planitia, which features less distinctive landforms, making image matching difficult. The orbit hd674 imagery was taken between the northern hemisphere autumn and winter, a period that is more likely to have prevailing aerosols and dust (Montabone et al., 2015). Furthermore, the incidence angle of orbit hd674 is larger than that of orbit h5145, which may reduce the visibility from the former orbit. Among the five panchromatic CCD lines, S1, S2, and ND with emission angles of 18.9°, 18.9°, and 0° respectively are selected in consideration of the computational resources. To further evaluate the accuracy, DEMs with a resolution of 20 m/pixel derived from higher-resolution Context Camera (CTX) images (6 m/pixel) and HRSC DEMs (50 m/pixel) (Gwinner et al., 2016) produced by the German Aerospace Center (DLR) are used as references for comparison. The detailed information of these two sets of HRSC images is listed in Table 3.11. The solar azimuth is not available (Hoekzema et al., 2004), so we infer the direction from the shadow of the landforms. The HRSC DEMs produced by the DLR are available at (https://maps.planet.fu-berlin.de/#map=3/2074498.35/0).

Table 3.11 Information about HSRC images - orbit h5145 and orbit hd674.

Orbit No.	Acquired Time	Center Latitude	Center Longitude	Sun incidence angle	Sun azimuth	Resolution (m/pixel)
h5145	2008-01-03	20°	-15°	65.18°	270°	12.5
hd674	2014-10-09	24°	110°	72.59°	270°	12.5

a. Evaluation of the photogrammetric approach

To quantitatively evaluate the performance of the multi-image bundle adjustment (BA), triple-matched tie points are chosen as checkpoints. The 3D positions of the checkpoints obtained from the space intersection of the S1 and S2 channels are back-projected to the ND

channel to compare with the ground truth. The image residuals of the tie points before and after BA are visualized in Figure 3.18 and summarized in Table 3.12, respectively. Following table shows that the discrepancies among multiple images are reduced from about seven pixels before BA to less than a pixel after BA, illustrated by the shorter residual vectors in Figure 3.18 (b) and (d). It is worth noting that, as hd674 is hazed and lacks apparent textures, the number of triple-matched tie points is much less than those on h5145.

Table 3.12 Statistics of multi-image bundle adjustment (BA).

		Absolute Error (Pixels)	
		Mean	**Max**
Orbit h5145			
Before BA	Residual *X*	4.60	10.10
	Residual *Y*	4.80	10.15
	Residual vector	7.32	10.19
After BA	Residual *X*	0.51	1.24
	Residual *Y*	0.54	1.22
	Residual vector	0.83	1.24
Orbit hd674			
Before BA	Residual *X*	4.35	11.63
	Residual *Y*	3.42	12.68
	Residual vector	6.09	12.76
After BA	Residual *X*	0.58	1.33
	Residual *Y*	0.61	1.37
	Residual vector	0.94	1.40

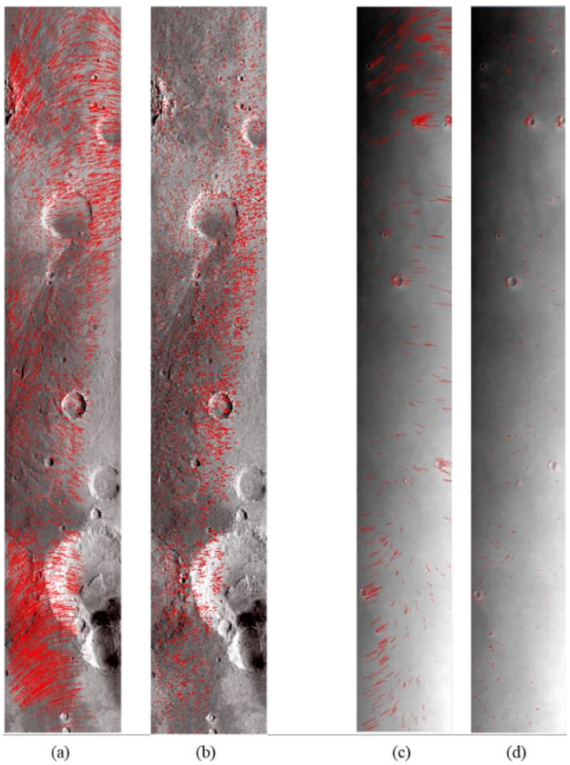


Figure 3.18 Residual vectors of the tie-points before and after the multi-image bundle adjustment (BA) shown on the images. (a) Orbit h5145: before BA, (b) Orbit h5145: after BA, (c) orbit hd674: before BA, and (d) orbit hd674: after BA. The vectors are all exaggerated 50 times for better visualization.

Figure 3.19 shows the photogrammetric results for orbit h5145. The MOLA DEM (Figure 3.19 (d)) and DLR HRSC DEM (Figure 3.19 (e)) are used as references for the quantitative evaluation of the geometric results. A difference DEM is generated by subtracting our DEM (Figure 3.19 (f)) from the MOLA DEM, as shown in Figure 3.19 (g). The mean differences between the DLR HRSC DEM and our DEM with respect to the MOLA DEM are 25.56 m and 25.90 m, respectively, which are considered typical of HRSC DEMs (Gwinner et al., 2016). Two representative profiles are selected and analyzed as shown in Figure 3.19 (h) and (i), respectively. The statistical results are summarized in Table 3.13. The profile comparisons re-emphasize that our results are close to DLR's but may deviate a little from the MOLA DEM. The explanations for these differences could be that: (1) the resolution and imaging difference

between MOLA and photogrammetric DEMs result in detail revealed in one DEM being missed by the other; (2) there might be errors in the MOLA DEM generated by the laser strip; and (3) the positioning of the produced DEM has a small misregistration with the MOLA data. It is noted that the DLR DEMs have been refined by photoclinometry and have a resolution of 50 m/pixel, and they thus possess more details than the photogrammetric DEM. However, as the resolution of the original image can reach 12.5 m/pixel, we make full use of this original information and refine the resolution of the produced DEM to 12.5 m/pixel.

Table 3.13 Statistics of Profile Comparison with MOLA DEM for h5145.

Profile	**Mean Absolute Difference (m)**		**Maximum Absolute Difference (m)**	
	Ours	DLR	Ours	DLR
1	55.14	48.07	472.48	495.14
2	50.17	60.88	553.69	724.64

The results for orbit hd674 are shown in Figure 3.20 and Table 3.14. As the DLR has not published a DEM covering the orbit hd674 region, we compare our results with the MOLA DEM directly. It is seen that noise is reduced effectively by enlarging the penalty in SGM. The large craters in Table 3.14 (a) and (b) are recovered through object-based matching. The consistent profiles verify the feasibility and accuracy of the proposed algorithm. Most craters visible at this scale are well retrieved using our approach, resulting in a favorable mean intersection error of 23.51 m. However, as the object-based feature matching is restricted by the object extraction algorithm, some irregular landforms are not recovered, especially at the bottom of the orbit hd674 region, which may also contribute to the intersection errors.

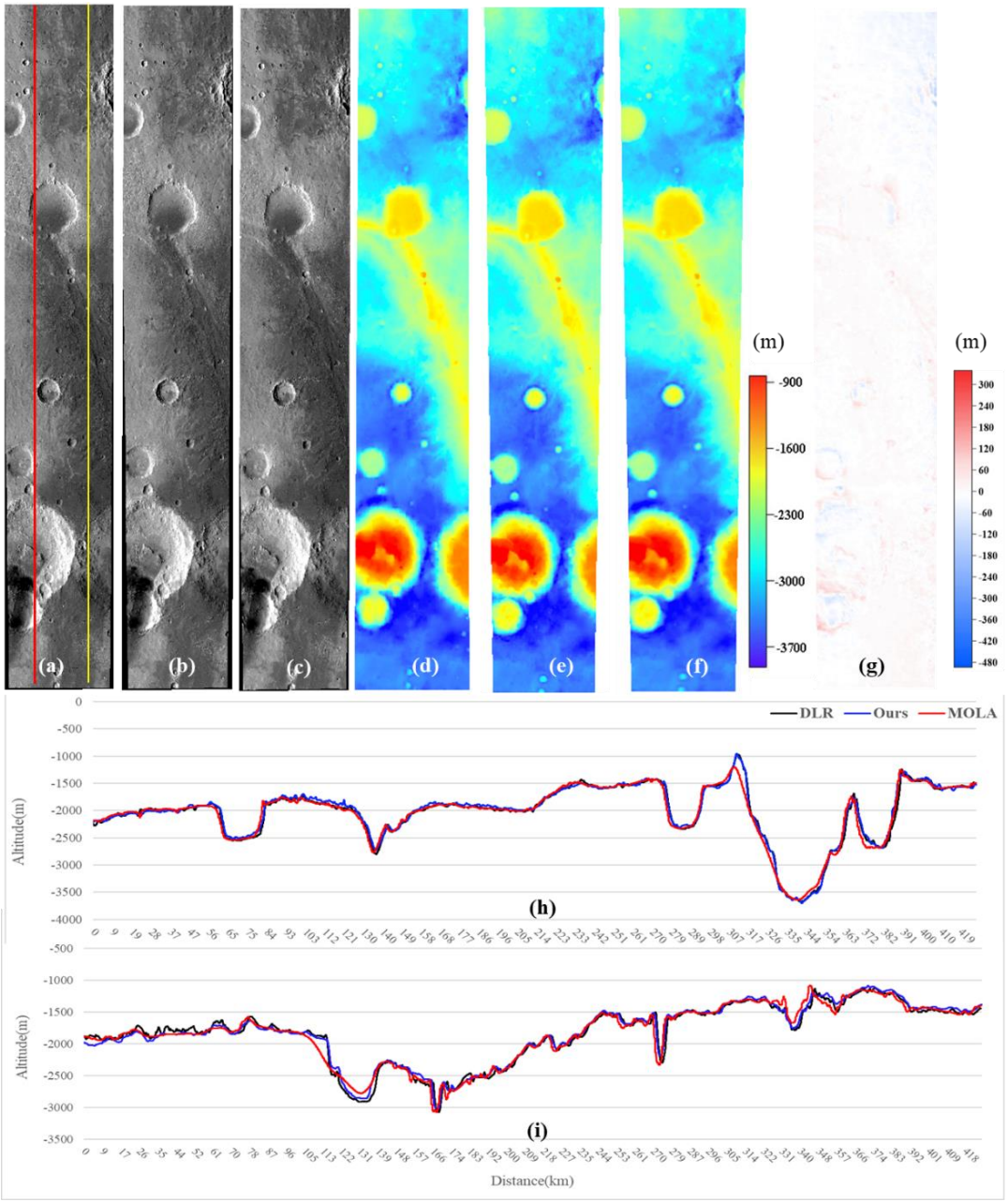


Figure 3.19 Photogrammetric results of orbit h5145: (a) nadir image (12.5 m/pix, 5176 pix×32368 pix), (b) S1 image (12.5 m/pix, 5176 pix×32280 pix), (c) S2 Image (12.5 m/pix, 5176 pix×32808 pix), (d) a MOLA DEM (463 m/pixel) covering the region, (e) DEM produced by the DLR on the basis of multiple orbits (50 m/pixel), (f) DEM derived using the proposed photogrammetric approach (50 m/pixel), (g) difference DEM between our DEM and the MOLA DEM, (h) and (i) are the profiles of the yellow and red line marked in (a).

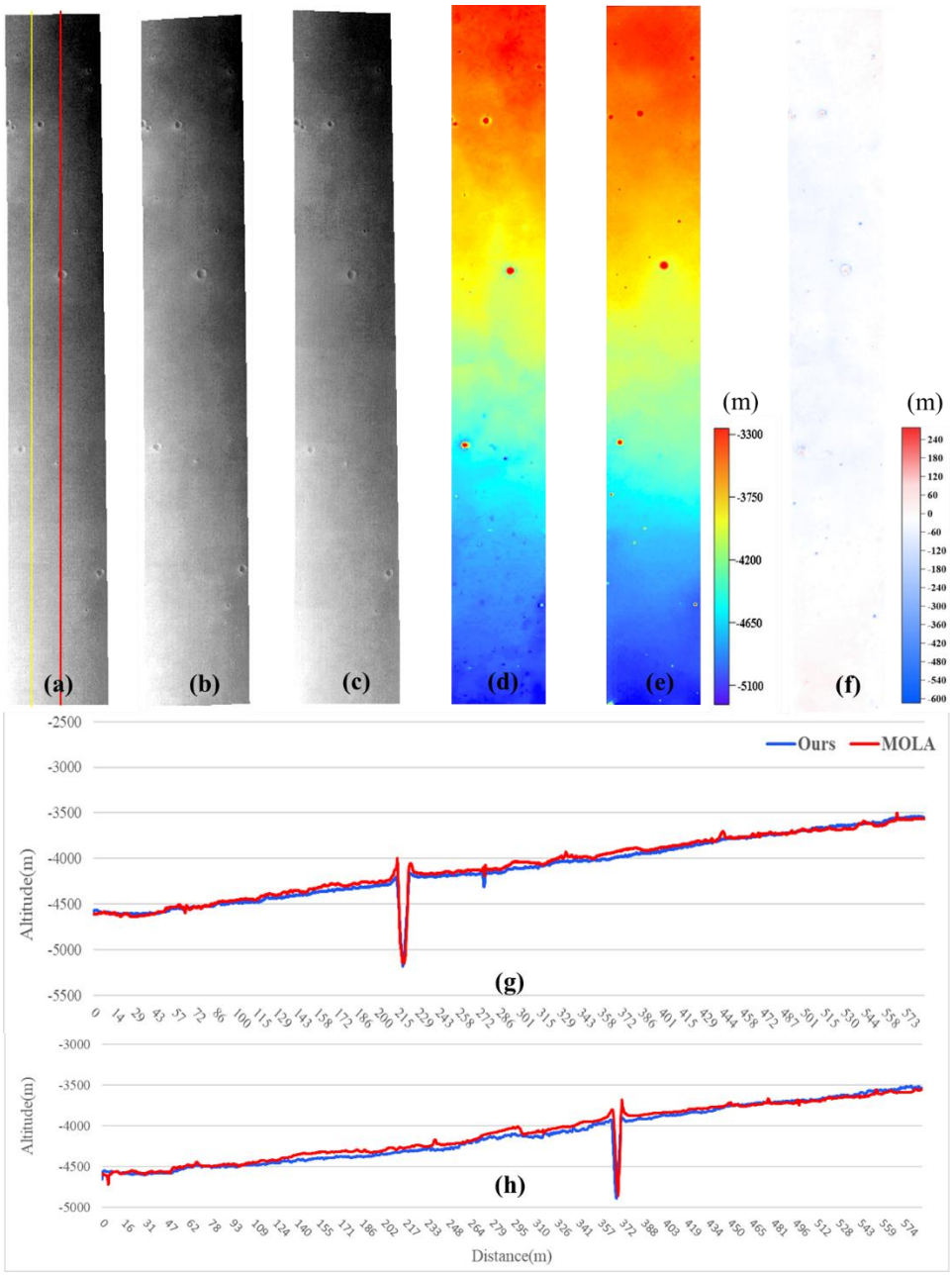


Figure 3.20 Photogrammetric results of orbit hd674: (a) nadir image (12.5 m/pix, 5176 pix×39424 pix), (b) S1 image (12.5 m/pix, 5176 pix×38704 pix), (c) S2 image (12.5 m/pix, 5176 pix×40184 pix), (d) a MOLA DEM (463 m/pixel) covering the region, (e) DEM derived using the proposed photogrammetric approach (100 m/pixel), (f) difference DEM between our DEM and the MOLA DEM, (g) and (h) are the profiles of the yellow and red line marked in (a).

Table 3.14 Statistics of Profile Comparison with MOLA DEM for hd674.

Profile	Absolute Difference (m)	
	Mean	**Maximum**
1	42.75	528.77
2	33.48	340.52

Figure 3.21 is an enlarged view of a local region to show the performance of the object-based matching. While Figure 3.21 (b) presents DEM directly yielded from SGM and LSM, Figure 3.21 (c) presents the DEM with the improvement made by the object-based matching algorithm. Even the wrinkles of the craters add some texture to the region, but from the perspective of image processing, it remains severely textureless. Moreover, since the SGM is a global inference algorithm, the neighboring textureless region may also perturb the disparity retrieval of this region. Apparently, the original photogrammetric DEM is unsuitable for both scientific studies and engineering evaluation. However, with object-based matching, a reasonable shape of the crater is retrieved, including the small crater in the top-left corner. Furthermore, as suggested by the profile in Figure 3.20, which crosses the large crater with a depth of over 500 meters, the geometric accuracy is also ensured by this algorithm. This rigorous DEM subsequently provides favorable input for photoclinometry processing, leading to the reconstruction of a pixel-wise DEM, as compared to the original image in Figure 3.21 (a).

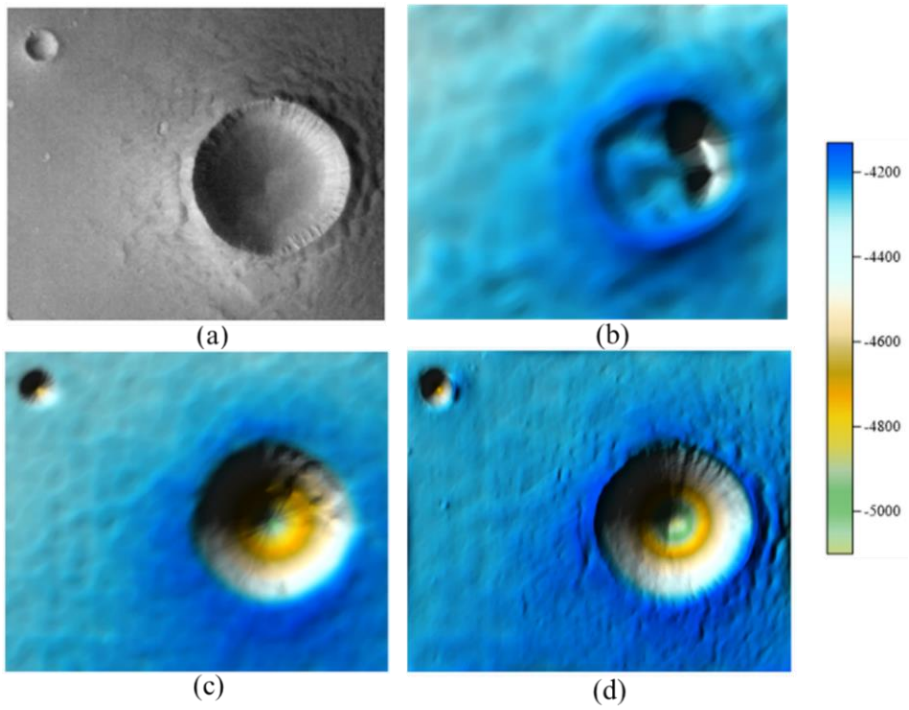


Figure 3.21 The performance of the object-based matching: (a) the nadir image of orbit hd674 (12.5 m/pixel), (b) the DEM without object-based matching or photoclinometry process (100 m/pixel), (c) the DEM after object-based matching but without photoclinometry process (100 m/pixel), and (d) the DEM after object-based matching and photoclinometric refinement (12.5 m/pixel).

b. Evaluation of the photoclinometric approach

To evaluate the performance of the proposed photoclinometric approach, two local regions of approximately 2000×2000 pixels are selected from each orbit for both qualitative comparison and quantitative analysis. The corresponding terrain-dependent albedo and the optical depth are also shown. A tile for orbit h5145 is evaluated in Figure 3.22 in comparison with the 20 m/pixel CTX DEM generated through the photogrammetric processing of CTX images. With the input of the low-resolution photogrammetric DEM, the apparent noise is eliminated and some missing details (e.g., small textures and craters) are reconstructed completely, whereas even the CTX DEM fails to reconstruct several of the missing details.

Within the tile, five craters generated by photoclinometry are chosen for the investigation of error. Figure 3.22 and Table 3.15 show that the deviations between our results and the CTX DEM are almost within 10% of the maximum depth, which demonstrates that our method not only unfolds the pixel-wise texture visually but also guarantees pixel-wise geometric accuracy.

The second tile in Figure 3.23 covers an almost textureless region of orbit hd674 with severe dust leading to a low signal-to-noise ratio. The photogrammetric result is unsatisfactory, with apparent noise generated by inaccurate matching results. The photogrammetric DEM is therefore down-sampled twice before use in photoclinometry. After adopting the photoclinometry process, the small craters and cones appear to have reasonable shape, and the profile comparisons of the landforms align well with the reference CTX DEM (Figure 3.23 and Table 3.16), while the height of the right side of the bottom cone is erroneously reduced because of the shadow and severe noise. It is noted that even if the optical depth and aerosol distribution are excluded in the cost function (e.g., using $I = AR$ instead of Equation (3-11)), the results of the hd5145 dataset are not greatly inferior to the present results. But these two parameters play a decisive role in the hd674 region, and ignoring them may enforce the atmospheric influences into the terrain.

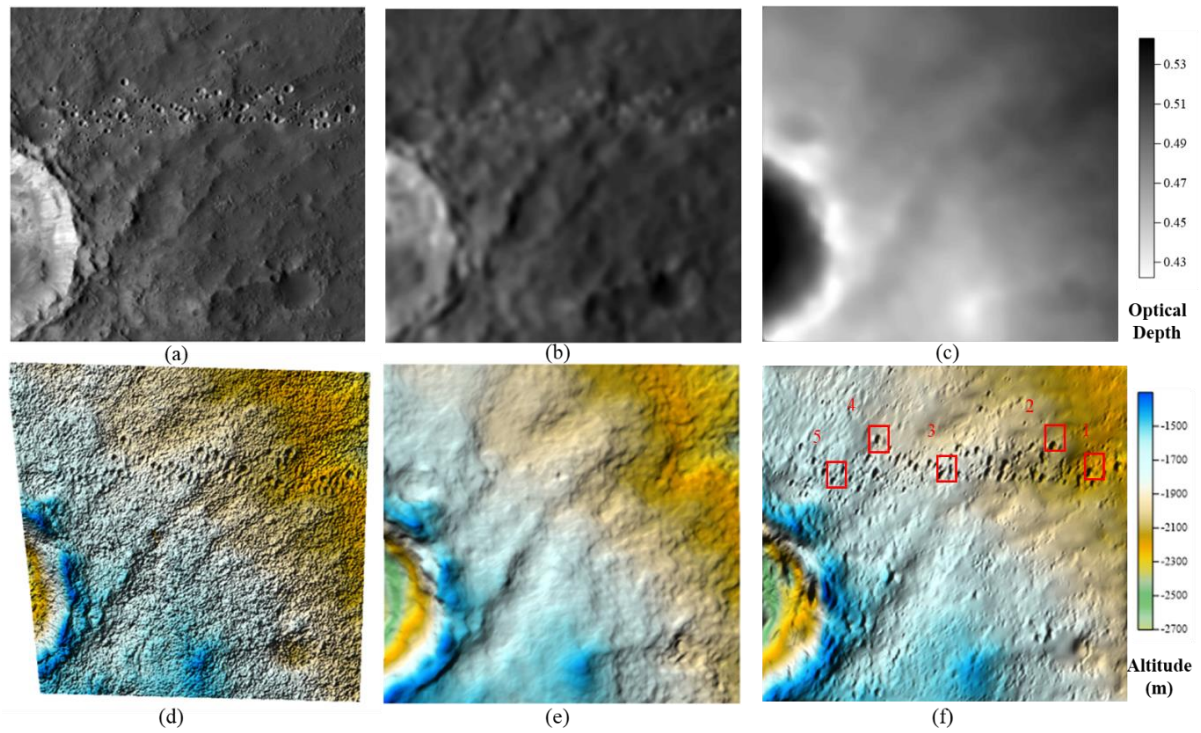


Figure 3.22 Examples of photoclinometric results for orbit h5145: (a) nadir image (12.5 m/pixel), (b) optimized albedo (12.5 m/pixel), (c) optimized optical depth (12.5 m/pixel), (d) CTX DEM (20 m/pixel) for reference, (e) generated photogrammetric DEM (50 m/pixel), and (f) refined DEM (12.5 m/pixel) by photoclinometry.

Table 3.15 Statistics of profile comparison for the Craters in Figure 3.22.

Craters	Depth: Ours/CTX (m)	Depth Difference (m)	Relative Difference	Absolute Difference of Elevations (m)		
				Mean	Min	Max
1	152/163	-11	6.7%	6.90	0.54	35.59
2	148/156	-8	5.1%	19.48	0.00	10.52
3	172/180	-8	4.4%	4.44	0.65	11.02
4	130/134	-4	2.9%	12.39	0.00	37.03
5	92/100	-8	12.5%	7.40	0.42	22.02

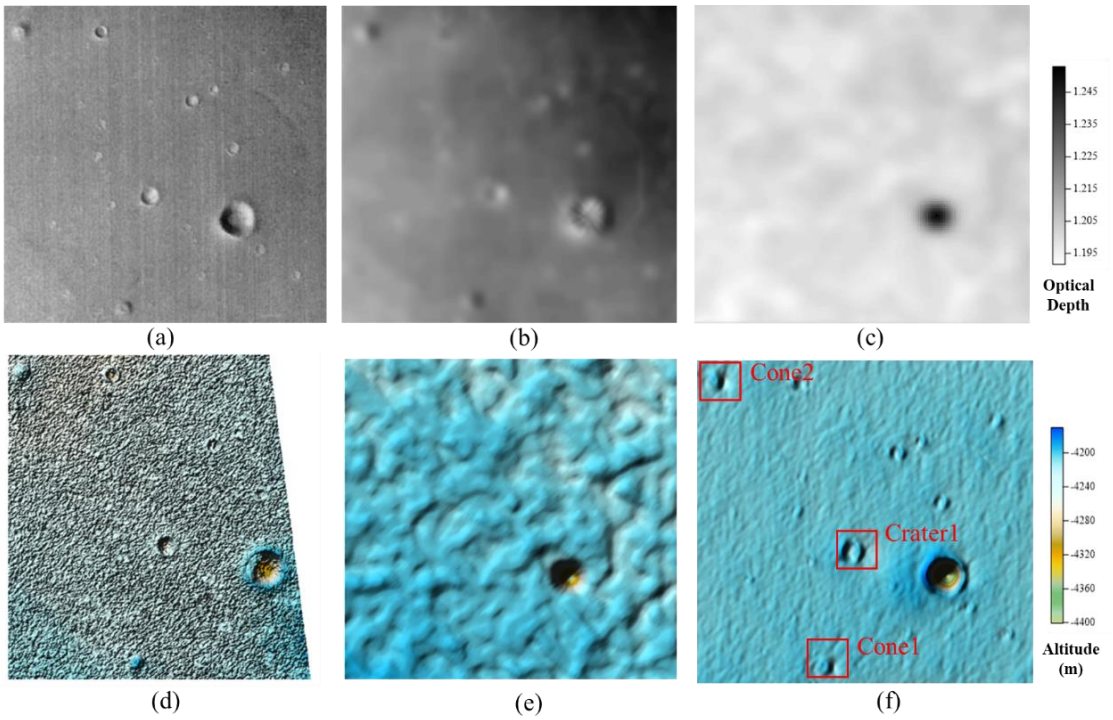


Figure 3.23 Profile comparison for selected landforms: (a) photoclinometric results with landform numbers marked and (b)–(d) profiles for cone 1, cone 2, and crater 1, (e) generated photogrammetric DEM (50 m/pixel), and (f) refined DEM (12.5 m/pixel) by photoclinometry.

Table 3.16 Statistics of profile comparison for the Landforms in Figure 3.23 (f).

Land-forms	Depth: Ours/CTX (m)	Depth Difference (m)	Relative Difference	Absolute Difference of Elevations (m)		
				Mean	Min	Max
1	65/67	-2	3.0%	9.16	0.02	30.82
2	79/80	-1	1.25%	6.91	0.10	20.65
3	74.5/71	3.5	4.93%	9.00	0.07	28.62

3.5 Summary

Aiming to generate rigorous DEMs with abundant details for various kinds of cameras, a generic framework for integration of multi-modal data is proposed, which is divided into two main stages.

The first stage involves the integration of laser altimetry and photogrammetry for large-scale rigorous DEM generation. An EO-guided learning-based feature matching approach is proposed to achieve evenly distributed tie-points, particularly in textureless and narrow overlap regions. Based on these tie-points, bundle adjustment can be performed to integrate laser data and retrieve the actual EOs of the cameras. Epipolar rectification is extended to generate multiple epipolar images, and fuse images captured by multiple CCD cameras into a single image, enabling the direct application of texture-aware dense image matching to obtain disparity images. The second stage involves incorporating photoclinometry into the pipeline to refine the subtle details of the photogrammetric DEM. Atmospheric effects, comprising both optical depth and scattering, are carefully considered in the photoclinometry pipeline. Furthermore, a learning-based optimizer and framework are introduced to achieve better optimization, yielding pixel-wise estimates of albedo, scattering, optical depth, and gradients.

Experiments were conducted on three typical cameras, namely CTX, HiRIC, and HRSC, for both quantitative and qualitative analysis, covering representative regions. The results revealed that subpixel accuracy was achieved for image residuals among multiple-camera or multiple-orbit images. Comparison with reference data (HiRISE and MOLA DEMs) revealed a mean deviation of approximately 20 m in terms of geometric accuracy. Furthermore, the accuracy of the photoclinometric depth exceeded 90%. The final DEMs exhibited subtle textures without apparent noise and conformed well to the original images. The presented method offers a generic and rigorous solution for utilizing various types of data for planetary topographic mapping.

Chapter 4 Improved Semantic Segmentation of Planetary Surfaces Assisted with 3D Information

Semantic segmentation is the process of object extraction with respect to semantic class (e.g., craters, sand dunes, rocks, volcanoes) and it is of great significance in planetary exploration missions. Although many mature semantic segmentation networks are publicly available, they have predominantly been trained on conventional Earth datasets. The latest large learning model, namely the segment anything model (SAM) (Kirillov et al., 2023), can segment planetary images successfully, but misses semantic information.

Typically, most semantic segmentation algorithms are applied to 2D images, but the available information remains limited, particularly when dealing with planetary surfaces where inherent features are scarce and homogeneous. When 3D reconstruction results have already been obtained, an intuitive way approach is to enrich the 2D images with 3D information. In light of this, an improved semantic segmentation method using 3D information is developed in this research. The 3D information used here comprises not only the reconstructed model but also the refined EO parameters. By taking advantage of 3D information, the proposed method can therefore be trained in a semi-supervised manner and obtain improved results.

This chapter presents the improved semantic segmentation algorithm and is organized as follows. Section 4.1 briefly introduces the overview of the proposed approach, where 3D models and the corresponding EOs are used throughout the segmentation pipeline. Section 4.2 shows how this 3D information is used to facilitate the construction of the training dataset. The detailed design of the depth-enhanced and Siamese semantic segmentation networks for semantic label improvement is proposed in Section 4.3 and Section 4.4, to further enhance the

training dataset and achieve consistent segmentation. Experimental evaluations are presented in Section 4.5. Concluding remarks are summarized in Section 4.6.

4.1 Overview of the Approach

As illustrated in Figure 4.1, the proposed approach consists of two primary phases: training dataset construction and neural network training. In the first phase, a comprehensive training dataset is constructed, specifically designed for semantic segmentation of planetary surfaces. This involves generating 3D original-textured and semantic-masked mesh models through a rigorous photogrammetric process, which is facilitated by a set of manually labeled images. By leveraging the OpenGL pipeline (Burns and Osfield, 2004), more realistic images can be generated based on these meshes, together with the semantic label images, XYZ images, and depth images. These images serve as the foundation for the subsequent phase. In the second phase, the generated depth information is incorporated with the RGB image used to train the transformer-based segmentation network (Liu et al., 2021), to retrieve more labeled images with precise labels.

Subsequently, a Siamese transformer-based neural network is trained on the constructed dataset in a pairwise manner, where each pair comprises two images with varying degrees of overlap. Initially, tie-points between the input images are calculated using a feature matching algorithm and filtered based on EO parameters. Contrastive learning then proceeds in a self-supervised manner, leveraging the tie-points as supervision through the Siamese segmentation neural network. The tie-points are updated concurrently with the semantic segmentation. Finally, the output semantic segments are filtered and cross-checked against the tie-points, yielding both semantic segments and tie-points as the final products.

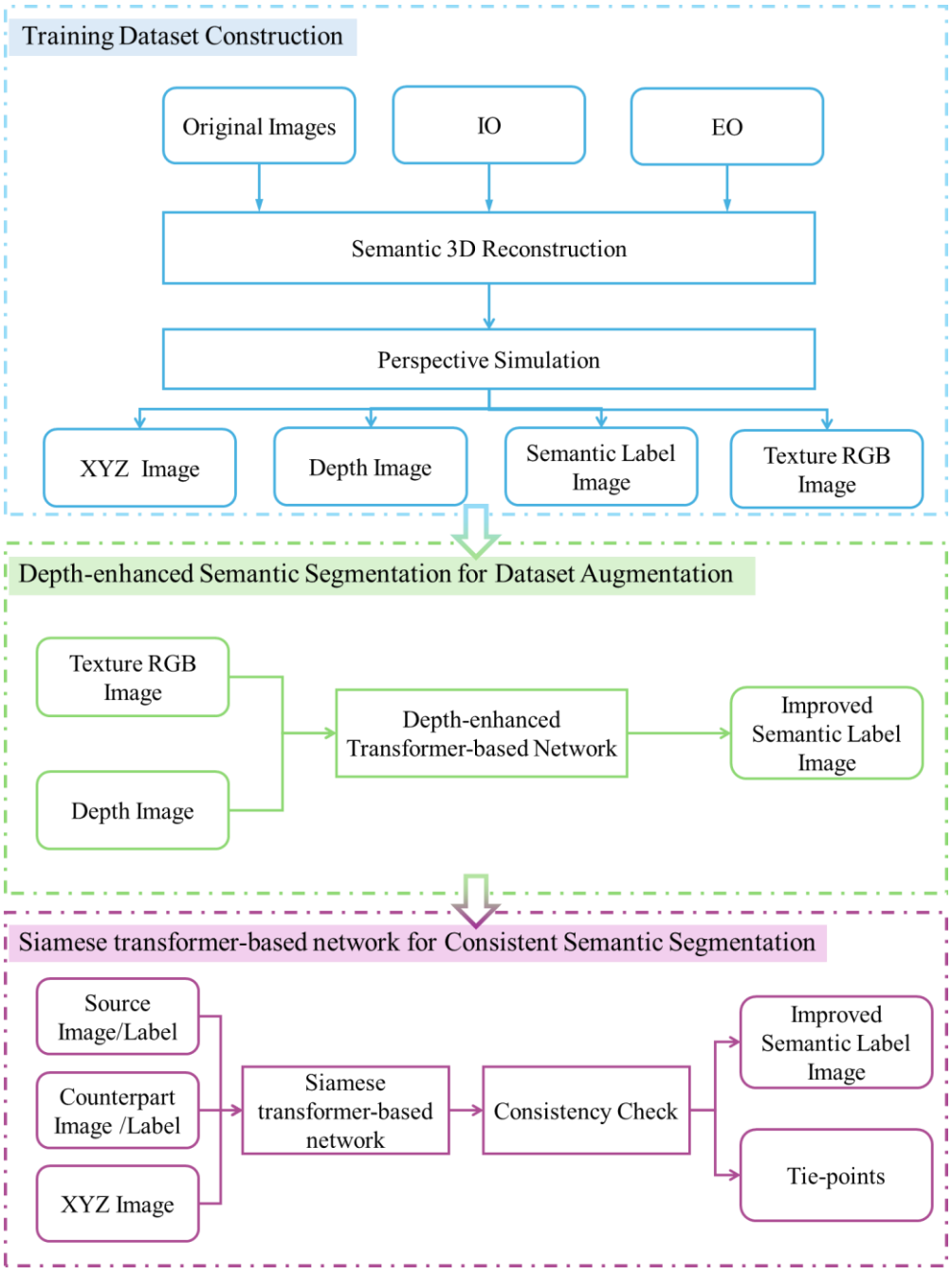


Figure 4.1 Overview of the proposed workflow.

4.2 Semi-automatic Dataset Construction

As deep learning is inherently a data-driven method, training with a segmentation dataset constructed with planetary images is indispensable. However, this is still challenging for two reasons. One is the fact that constructing a pixel-wise semantic segmentation dataset requires

substantial human labor, and the number of planetary images is severely limited (Ma et al., 2024b). In 2019, the ESA pioneered the public LabelMars project to organize a large labeled dataset based on thousands of images from Spirit, Opportunity, and Curiosity. The dataset was criticized by NASA for its overly specialized categorization and the resulting small volume. The AI4MARS dataset (Swan et al., 2021) was thus proposed based on similar data, but with more intuitive labels, namely, sand, bedrock, soil, and big rock. The associated depth data were also provided, which enhanced the versatility of the dataset. Although it has been claimed that the dataset comprises ~35,000 images, only ~18,000 images are available, and the detailed distribution of each class is unavailable. The involvement of tremendous human labor in generating such a dataset makes it hard to further augment with more images from the latest rovers. Simulation strategies have hence been considered, and Ma et al. (2024b) used the OAISYS simulator to add some rocks to the designed surface to generate a large dataset. However, the images vary considerably from the real scene on Mars. A simulation method based on a real 3D scene is thus desirable.

4.2.1 Semantic 3D Model Generation

Instead of directly augmenting the 2D images through perspective transformations (i.e., translation, rotation, scale transform), a more realistic approach to boosting the volume of the dataset is proposed. The core idea of semi-automatic semantic dataset construction is to fully exploit the results of 3D reconstruction of the images, which relies on the premise that the traverse of the rover is typically continuous or at least several images share some overlapping regions. Following the *ad hoc* SfM pipeline (Agarwal et al., 2009), bundle adjustment can be performed based on the tie-points among images, and a 3D textured mesh model can be formed from the dense point clouds calculated with the MVS algorithm (Vu et al., 2011). A semantic-masked 3D model can also be obtained with some manually labeled images.

The pipeline for the generation of the semantic 3D model is illustrated in Figure 4.2. Based on the 3D reconstruction pipeline illustrated in Chapter 3, the textured mesh model is generated with the refined EOs of the images. Some images are also expected to be manually labeled. The basic requirement is that at least one observation is labeled for each region. The 3D model can thus be annotated with semantic information according to the collinearity equation.

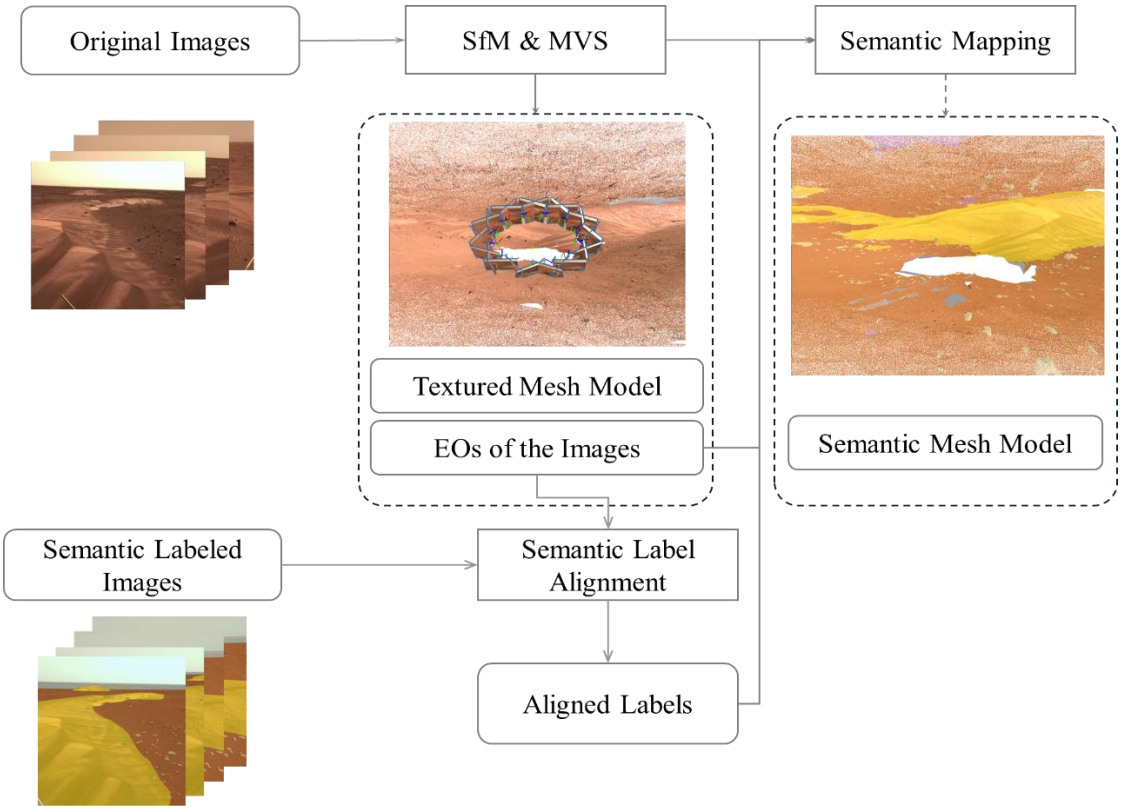


Figure 4.2 Illustration of semantic mesh model generation.

It is worth noting that the landforms on planetary surfaces are complex and confusing. Even if the images captured by stereo images are nearly identical, the labels generated by humans are not guaranteed to be the same. This issue is even more severe for images captured from different viewpoints, showing severe distortion and scale variations. To address this, overlap detection is first performed to ensure that each overlapping region is only labeled in one image. Then, the semantic segments are passed to the other unlabeled images based on

photogrammetry. For each pair of overlapping images, epipolar rectification is first conducted and the disparity images are calculated to retrieve the pixel-wise correspondence. The algorithm then iterates through the entire disparity image to verify consistency across images. When a misaligned pixel is detected, a four-direction depth-first-search is initiated to assign a distinct label to the corresponding normal pixel. For instance, if a small rock is overlooked by one image but labeled by another, the rock should be segmented accordingly.

4.2.2 Image Rendering using OpenSceneGraph (OSG) based on Real 3D Information

In the typical photogrammetric process, a 3D point $\boldsymbol{X}$ in the real world can be projected to a 2D point $\boldsymbol{x}$ on the focal plane according to the collinearity equation based on the protocol of BlockExchange (Zhu et al., 2020):

$$\boldsymbol{x} = \boldsymbol{f}\boldsymbol{D}\left(\prod\left(\boldsymbol{R}(\boldsymbol{X} - \boldsymbol{C})\right)\right) + \boldsymbol{x_0} \tag{4-1}$$

where $\boldsymbol{f}$ denotes the focal length, and $\boldsymbol{x_0}$ refers to the principal points (x_0, y_0). The 3D points are first subjected to the translation transformation $\boldsymbol{C}$ and rotation transformation $\boldsymbol{R}$. $\prod(\cdot)$ is a function calculating the normalized 2D points, and $\boldsymbol{D}(\cdot)$ deals with the distortion defined by $(k_1, k_2, k_3, p_1, p_2)$. In summary, the process involves two key steps. First, the 3D model is transformed to align with the desired view using the view matrix. Second, the transformed 3D model is projected onto a 2D image plane using the model matrix, resulting in a 2D representation of the original 3D scene.

The process of rendering an image from a 3D model involves a similar procedure. Specifically, the 3D model undergoes a transformation to align with the desired viewpoint, utilizing the view matrix to establish the camera's position and orientation. Subsequently, the transformed 3D model is projected onto a 2D image plane via the model matrix, yielding a 2D image of the original 3D scene. During the photogrammetric process, the field of view (FOV)

is often overlooked, as it is implicitly determined by the focal length and sensor size. However, when generating simulation images, it is essential to explicitly define the FOV, along with the near and far viewing depths, to ensure accurate and realistic rendering. The definition of these parameters is shown in Figure 4.3.

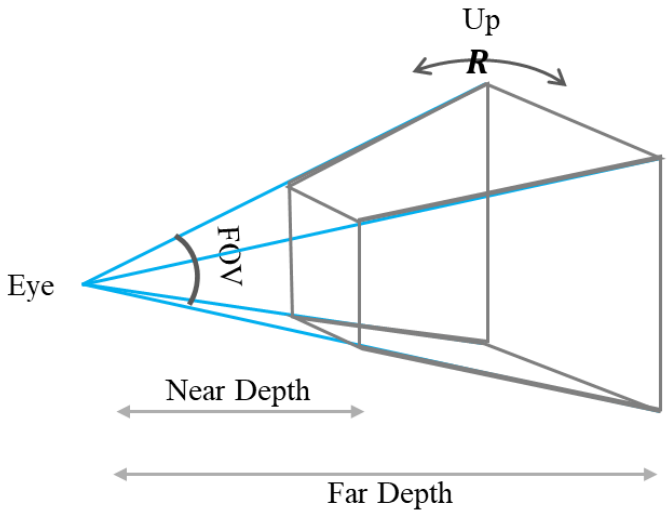


Figure 4.3 Definition of camera parameters.

When using OSG for 3D scene rendering, eye, center, and up are three key parameters that collectively define the camera's position and direction. Specifically, the eye parameter represents the camera's position, which is the location of the observer's eye. It is typically defined as a 3D vector that indicates the camera's position in 3D space. The center parameter represents the camera's focal point, which is the target location being observed by the camera. It is also a 3D vector, and indicates the target location in 3D space. The up parameter represents the camera's up direction, which is the camera's up axis. It is typically defined as a 3D vector that indicates the camera's up direction in 3D space. The corresponding definitions of these three factors and the photogrammetric parameters are listed below:

$$\begin{cases} Eye = C \\ Center = C + z_m R^T e_z \\ Up = -R^T e_y \end{cases} \tag{4-2}$$

where z_m is the depth of the middle of the scene, and e_z and e_y are the unit vectors parallel to the y-axis and z-axis, respectively. With the given width w and height h of the image, the FOV and the aspect ratio ρ of the image can be calculated, as:

$$\begin{cases} FOV = 2 \arctan(h/2f) \\ \rho = w/h \end{cases} \tag{4-3}$$

In summary, eye, center, up, and interior camera configurations are the key parameters that define a specific camera and its position and direction in OSG, which collectively affect the rendering result of the scene. By adjusting these parameters, different camera views and observation effects can be achieved.

4.2.3 Semi-automatic Training Dataset Construction

The pipeline for semi-automatic training dataset construction is illustrated in Figure 4.4. By imposing the same camera on the original and the semantic-masked mesh model, the aligned RGB and the semantic images can be acquired simultaneously. The semantic-masked images are then transformed into label images according to the color of the semantic mask. Furthermore, in addition to the labeled image output, this method can also generate a depth image that measures the absolute distance of each pixel in the image to the camera center, as well as an XYZ image that records the absolute 3D coordinates of each pixel.

The semi-automatic pipeline for training dataset construction offers several advantages in the context of deep learning. First, it significantly reduces the manual annotation effort required to create high-quality training data, which is a major bottleneck in many computer vision

applications. By automating the process of generating labeled images, depth images, and XYZ images, this pipeline enables researchers to focus on more complex tasks such as model development and hyperparameter tuning. Additionally, the semi-automatic pipeline allows for the creation of large-scale datasets with consistent annotation quality, which is essential for training robust and accurate deep learning models.

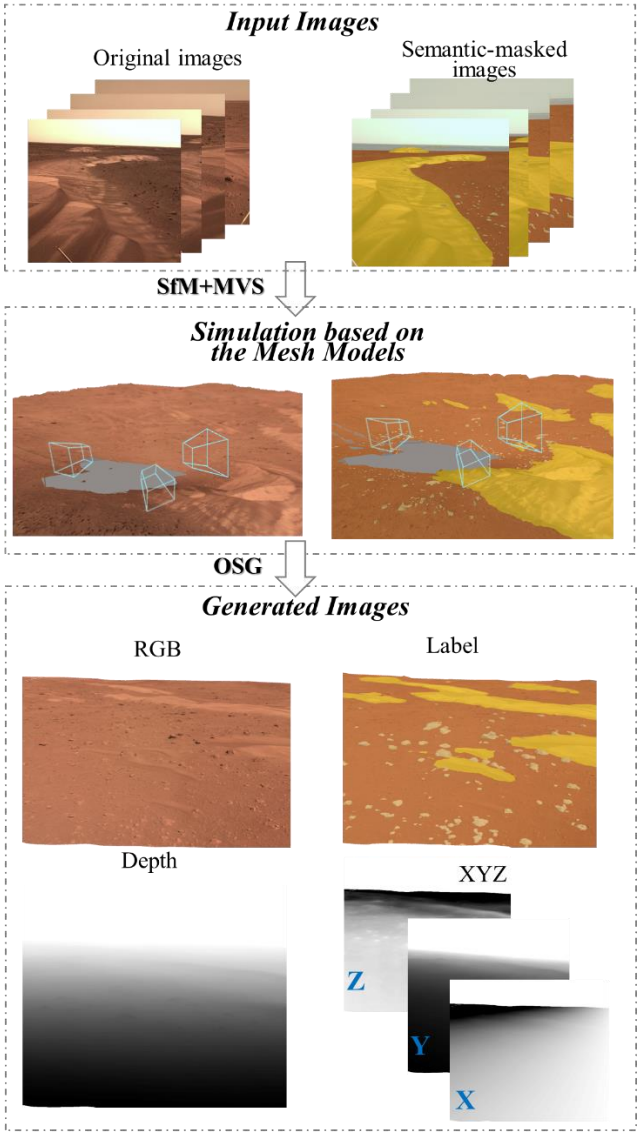


Figure 4.4 The pipeline for semi-automatic dataset construction.

Although the 3D augmentation is more realistic and has more advantages, the 3D reconstruction of planetary surface is still non-trivial, and conventional 2D image augmentation techniques, such as transformation matrices (e.g., scaling, translation, rotation, affine, and

homography) and image processing methods (e.g., cropping, resizing, and blurring), are still essential. Our dataset comprises three primary categories of images. The first category consists of images captured by cameras, which, although lacking accurate 3D information and consistent semantic labels, represent the most realistic and complex scenarios. The second category comprises images augmented from actual images using 2D approaches, which, while missing 3D information, retain transformation information between the original and augmented images. In contrast, the third category consists of images generated through the aforementioned 3D augmentation process, which possess comprehensive 3D information and consistent semantic information. It is assumed that a diverse and comprehensive set of images can be acquired for the construction of the training dataset through these three methods, thereby ensuring a robust and representative dataset for training purposes.

4.3 Depth-enhanced Transformer-based Semantic Segmentation

As semi-automatic dataset construction still relies on manual labels, the initial number of labels significantly impacts the overall labels input into the neural network. Given that the generated dataset comprises not only RGB textured images and corresponding semantic images but also attached depth images, an intuitive approach is to further incorporate the depth information to enrich the RGB information, thereby extracting more semantic labels for subsequent training.

4.3.1 Architecture of Swin-transformer

Recently, transformers and their variants have been widely adopted for semantic segmentation tasks. We introduce a novel variant of the transformer tailored for depth-enhanced semantic segmentation, as depicted in Figure 4.5. The backbone of our network is based on the state-of-the-art Swin-Transformer architecture (Liu et al., 2021), which

synergistically combines the strengths of convolutional and transformer-based networks, thereby enabling both local and global contextual understanding. Considering the poorly textured nature of planetary surfaces, the utilization of a strong backbone enables the extraction of more comprehensive and discriminative deep features, thus enhancing the understanding of the image. As illustrated in Figure 4.5, the input of Swin-transformer is the three-channel RGB image, which is segmented into patches of size of four pixels. Since ViT (Dosovitskiy, 2020) first leveraged the transformer structure for vision tasks, patch partitioning has become a common strategy to reduce the number of tokens input into the transformer, driven by both memory and efficiency considerations. These patches are later embedded with positional information to capture neighboring cues and then undergo linear embedding to increase their dimensionality to 96 dimensions (for Swin-Tiny (Swin-T)). The core Swin transformer block is then applied, and its output has the same dimension as the input. To capture multi-dimensional features, which are essential for various computer vision tasks, the patch merging mechanism, akin to the pooling operator in CNNs, is proposed to expand the receptive field. After three repetitions of this stage, the resolution is downsampled 16 times leading to a relatively global understanding.

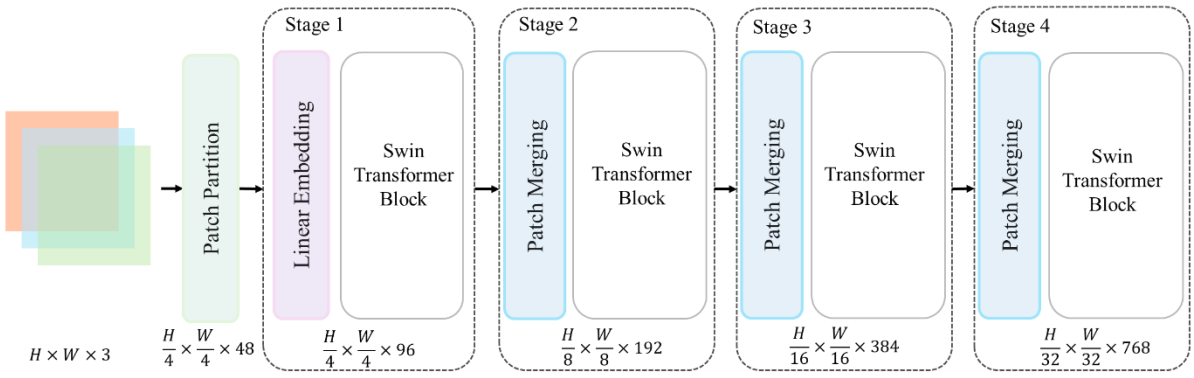


Figure 4.5 Overview of the Swin-transformer.

The core Swin transformer block, visualized in Figure 4.6, comprises three key components underlying its superior performance. Each Swin block contains the consecutive modules illustrated in Figure 4.6 (a), which are similar to the original transformer. These four small modules consist of layer normalization, multi-head attention, layer normalization, and multi-layer perceptron. While the first module's multi-head attention is conducted in the normally partitioned window, the following modules are partitioned by the shifted window, as illustrated in Figure 4.6 (b). Conventional transformer architectures, including those tailored for image classification, rely on global self-attention mechanisms that exhaustively compute relationships between all tokens. However, this approach is computationally expensive, scaling quadratically with the number of tokens, which makes it impractical for vision tasks. In light of this, the Swin-Transformer only performs self-attention within a local window (of a predefined size) comprising several patches. To address the issue of disconnection between local windows, a shifted window is introduced to reform the windows for local attention calculation, thereby enhancing global reasoning ability.

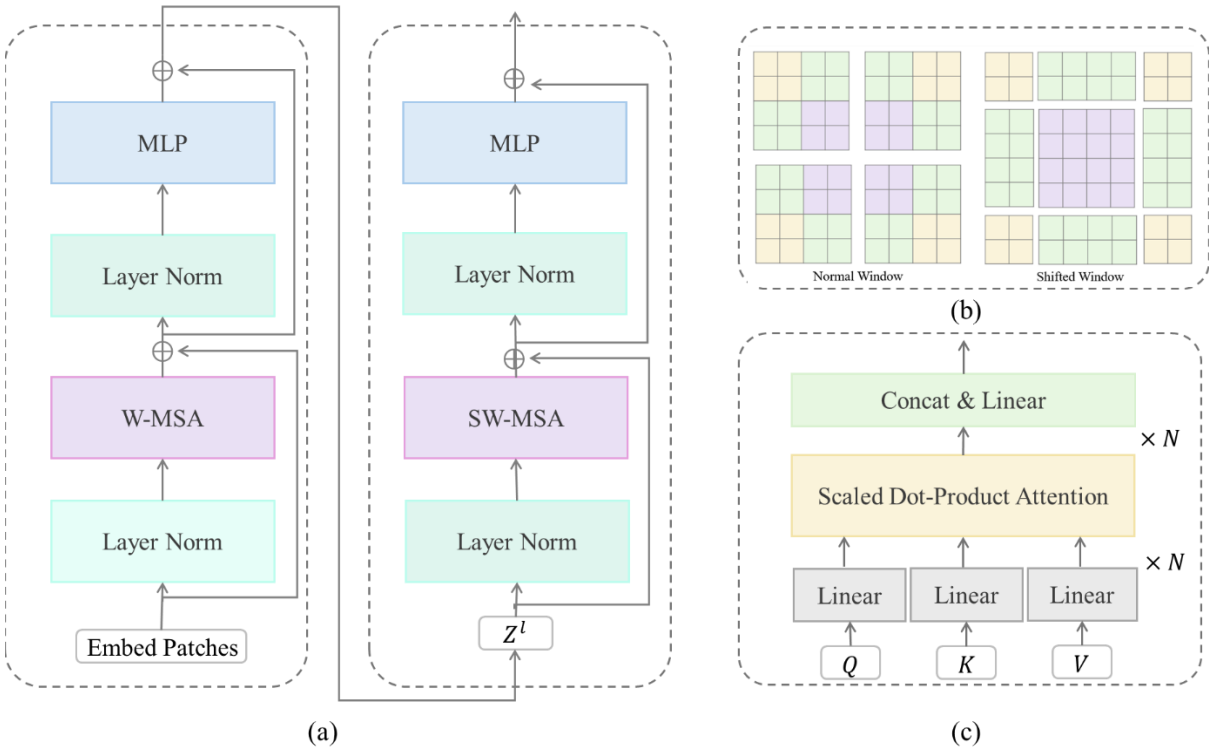


Figure 4.6 Important modules for Swin-transformer. (a) Two consecutive Swin blocks, (b) shifted window strategy, and (c) the self-attention operation.

Inside each calculation window, the self-attention is calculated as presented in Figure 4.6 (c). Multi-head attention is employed to enhance the expressibility of the network for various downstream tasks, with the query, key, and value first linearly transformed to lower dimensions, as follows:

$$MultiHead(Q,K,V) = Concat(head_1, head_2, \dots, head_h)W^O \tag{4-4}$$

where $head_n$ denotes the outputs from each head, and W^O are the projection parameters to map the results to the original full dimension. The scaled-dot product can be calculated as:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{4-5}$$

where Q, K, V are the query, key, and value. The similarity measurement is conducted by calculating the product of the query and the key, which is normalized by the square root of the dimension of the output of each head d_k, and then applying the SoftMax operation to calculate the weight for each token (Vaswani, 2017). These weights are later assigned to the values of all tokens to compute the final output, taking into account all the other tokens. The scaled-dot product attention mechanism has proven to be a powerful tool in transformer architectures, allowing for efficient and effective computation of attention weights. By scaling the dot product of the query and key vectors by the square root of the dimensionality of the vectors, this mechanism enables the model to capture abstractive relationships between input elements while avoiding the vanishing gradient problem. This leads to improved performance and stability in a wide range of natural language processing and computer vision tasks. The outputs of the multi-head attention are finally concatenated and linearly transformed to match the input dimension.

4.3.2 Architecture of the Depth-Enhanced Transformer

The overall architecture of the proposed depth-enhanced transformer network model is illustrated in Figure 4.7. Given a textured RGB image and its corresponding depth image as inputs, the model outputs a segmented image with semantic labels. Two main parts are involved, namely, encoding and decoding. Among the variants of the Swin-transformer, the Swin-T is leveraged as the backbone, which provides a robust and efficient feature extraction mechanism. To effectively integrate multi-scale features, we adopt the UperNet framework, a state-of-the-art approach that combines the strengths of feature pyramid networks (FPNs) (Lin et al., 2017) and pyramid pooling modules (PPMs) (Zhao et al., 2017). Specifically, we apply the PPM to the final layer of the FPN, allowing for the aggregation of features at different scales and spatial resolutions. The resulting feature maps are then concatenated and fed into a convolutional

operator, which generates semantic features. These features are subsequently convolved and normalized using a Softmax function, yielding a probability distribution over the semantic classes. Finally, the semantic segments are obtained by applying the Argmax operator, which selects the class with the highest probability for each pixel, resulting in a precise and accurate segmentation mask.

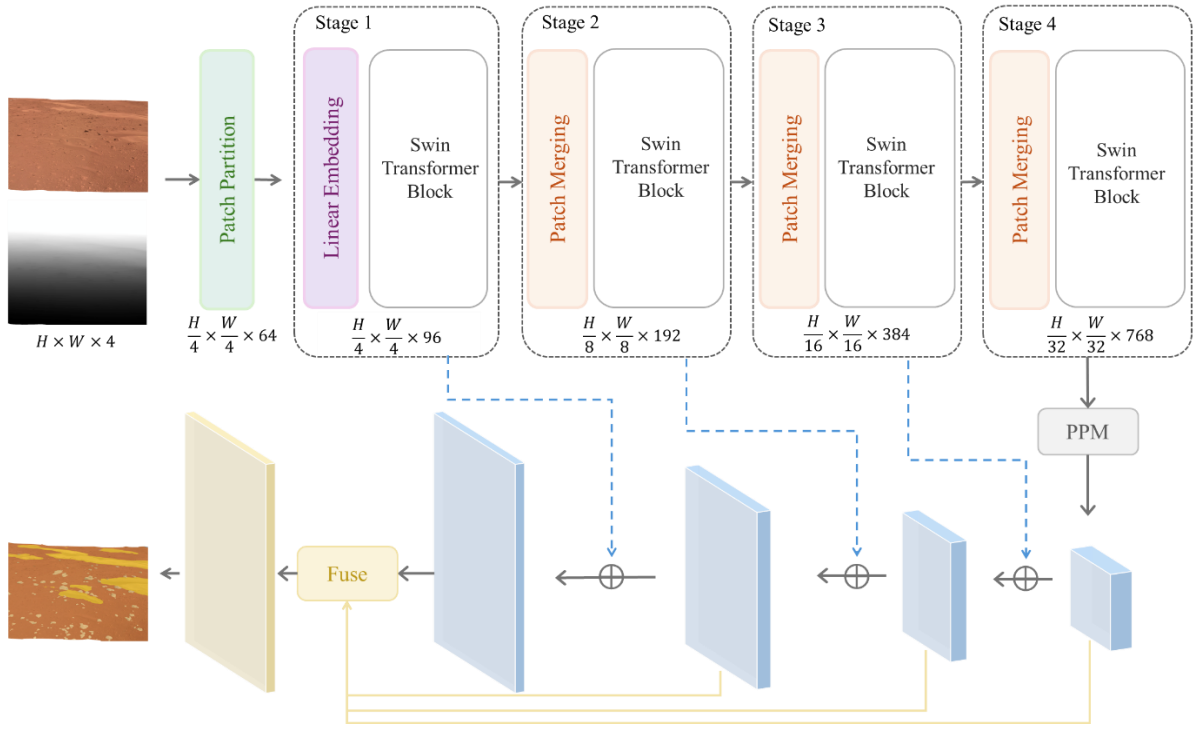


Figure 4.7 The overall architecture of the DepthFormer.

In the encoding process, the $M \times M \times 3$ sized RGB images are first concatenated with the corresponding $M \times M \times 1$ depth image, and then processed by the partitioning operator to form a $\frac{M}{4} \times \frac{M}{4} \times 64$ tensor. This step not only reduces the number of tokens to improve computational efficiency, but also helps to preserve the spatial information. Specifically, flattening an image into a 1D sequence would result in loss of the spatial relationships between pixels, which is crucial for image understanding. With the patch partition, each patch retains its spatial relationships with neighboring patches, allowing the transformer to capture local and

global contextual information. These tensors are subsequently enriched to higher dimensions via linear embedding, thereby enhancing their representational capacity, and then processed by the core Swin block, which leverages self-attention mechanisms to capture contextual relationships and dependencies. Subsequent Swin blocks are applied after each patch merging operator to fuse neighboring patches and extract both local and global features.

To capture multi-level features for a robust and comprehensive understanding of the scene, the feature pyramid network is adopted, which is a hierarchical structure that combines features from different scales and levels to produce a robust and accurate representation of the input image. The feature map of each layer is computed by summing the convolutional output of the original input and the feature map from the subsequent layer, and the overall feature map is obtained by aggregating all feature maps from all resolution layers.

The 2D convolutional layer is then applied to the feature map to generate an $\mathcal{N}$ (equal to the number of semantic classes) dimensional vector. Finally, the SoftMax operation normalizes the vector to a probability distribution, where each element is a value between 0 and 1, and the elements sum to 1, as:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{4-6}$$

SoftMax applies the exponential function $e^{(\cdot)}$ to each element of the vector, resulting in a probability distribution where the highest output value corresponds to the highest input value, and all negative values are transformed into positive values within the range of zero to infinity.

4.3.3 Loss Function and Training Strategy

To quantify the loss during training, the multi-class cross-entropy is leveraged. In a multi-class classification scenario, the cross-entropy loss measures the dissimilarity between the true

one-hot encoded labels and the predicted probability distributions. It encourages the model to assign higher probabilities to the correct classes and penalizes incorrect predictions. This loss function is particularly effective for training models in tasks where the goal is to classify inputs into one of several categories.

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(p_i) \tag{4-7}$$

where y_i is the true probability distribution (one-hot encoded, so only y_i is 1 and the rest are 0), and p_i is the predicted probability for class i.

$$\mathcal{L} = -\sum_{b=1}^{B}\sum_{i=1}^{N} w_i y_i \log(p_i) \tag{4-8}$$

where B is the given batch size. To balance the contribution of all classes due to their significance and sample imbalances, the weighted cross-entropy loss is employed as a useful extension. By assigning appropriate weights w_i to each class, the model is guided to focus more on the underrepresented classes, thereby improving performance in those classes. This approach is particularly effective in planetary surface segmentation scenarios, where the soil region dominates a large area, and the quantity of landforms varies significantly due to geological and data-related reasons. By emphasizing the minority classes, the model can better capture the abstractive details of diverse landforms, leading to more accurate and comprehensive segmentation results. Consequently, this method enhances the model's ability to generalize across different terrains, making it a robust solution for planetary exploration and analysis. And the weight of each class is proportional to the inverse of its frequency in the dataset.

4.4 Siamese Transformer-based Semantic Segmentation

While the depth-enhanced transformer directly leverages 3D depth information to augment the expressiveness of 2D grayscale images, retrieving accurate depth information can be challenging, and this approach may be susceptible to certain limitations. For instance, regions with low texture may be prone to large geometric errors, which can potentially confuse the network. Furthermore, supervised learning places high demands on training dataset construction in the aspects of both volume and consistency. Therefore, a more elegant approach to extracting high-level deep features is also necessary, and should satisfy the following requirements.

(1) Directly extract high-level semantic features from RGB images while preserving the transformation-invariant characteristic, enabling their application in diverse scenarios;

(2) Harness the complementary benefits of 3D information to augment semantic segmentation performance, while avoiding over-reliance on the 3D model;

(3) Design a self-supervised training process, to alleviate the reliance on the training dataset.

4.4.1 Architecture of the Neural Network

Given the existence of overlapping regions in the available planetary images, it is natural to consider leveraging this constraint to supervise the segmentation. A straightforward approach to address this challenge is to feed the original and transformed images to a neural network utilizing a Siamese architecture. By leveraging the Siamese structure, two primary benefits can be achieved. First, the two branches of the network can cross-validate each other, reducing the reliance on the training dataset and ensuring consistency in the output. Secondly, even if a single branch of the network is capable of segmenting images into reasonable semantic classes, the Siamese structure enables the exploration of transform-invariant properties, which cannot be achieved by a single branch network alone. For each branch, the UperNet framework

is still employed with the Swin-T architecture as the backbone, whose advantages have been extensively discussed in the preceding section.

Typically, the contrastive loss can then be established by measuring the similarities between these segmentation results. However, this strategy is not applicable to any two images sharing overlapping observations, for which the pixel-wise transformation is hard to obtain. Specifically, the relationship between images taken in the actual 3D space is too complex to be described only with a transformation matrix. Moreover, wrapping an image during training consumes a large amount of memory. Tie-points are hence introduced to find the corresponding points between the images. Instead of conducting matching on the original images directly, the images are transformed into semantic-masked images to guide the distribution of tie-points to specific classes and filter out the inevitable false matches. To further include more images for constraints in the training stage, input image pairs are designed. For each image, an overlapping image in the dataset is randomly chosen to form the counterpart image, as shown in Figure 4.8.

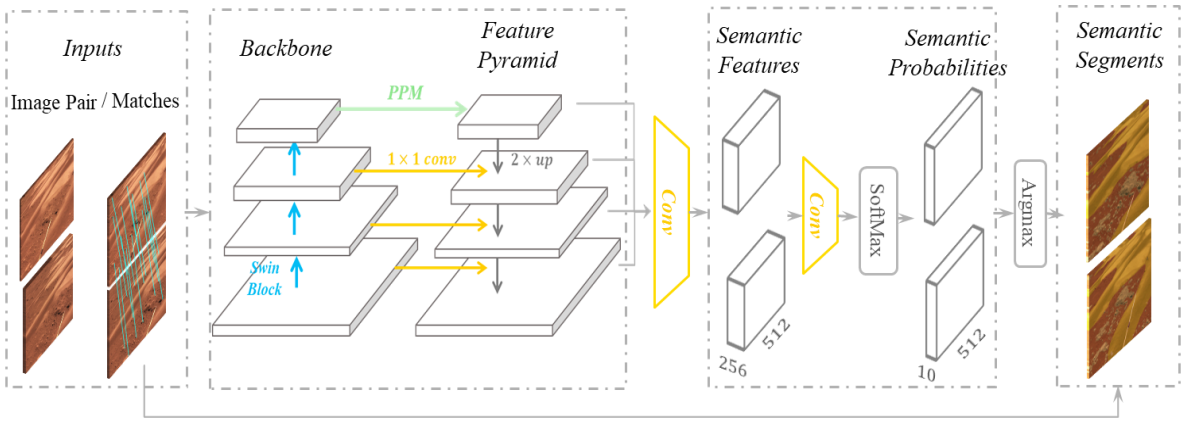


Figure 4.8 Overall architecture of the Siamese transformer for semantic segmentation.

4.4.2 Tie-point Matching Strategy

To achieve match-based contrastive learning, the number and accuracy of tie-points are vital. Although randomly selecting points can accomplish the task, the concept of keypoints is

introduced to focus on distinctive points. The state-of-the-art GPU-based keypoint extraction algorithm, SuperPoint, is leveraged to efficiently calculate the keypoints for each image.

For the matching part, two situations are considered. For the simulated images generated from the semi-automatic dataset construction pipeline, the XYZ coordinates are inherently attached, allowing for the direct retrieval of strictly corresponding points. With respect to the actual images without 3D information, nominal EOs are first used to check whether these two images overlap. Upon confirmation of overlap, SuperGlue matching is utilized to establish tie-points between the images. To mitigate the inherent outliers yielded by SuperGlue matching, a semantic verification step is implemented. This involves retaining tie-points that share the same semantic class, while discarding those that do not. By adopting this approach, the potential for incorrect matches to introduce confusion during training is minimized, as the semantic consistency of the tie-points is maintained, thereby ensuring the reliability of the network's learning process. Specifically, even if two keypoints are not corresponding points, their semantic classes should be identical according to the manual labels. This can also serve as a useful constraint.

Moreover, to further augment the constraint, multiple overlapping image counterparts of a source image can be simultaneously fed into the Siamese network, thereby providing a richer set of correspondences and enhancing the robustness of the learning process. The overall tie-point retrieval pipeline is thus organized as shown in Figure 4.9. Given the source image, the overlapping images can be generated or retrieved from the dataset, and the tie-points can be calculated accordingly. These tie-points are then combined into a single image to serve as supervision.

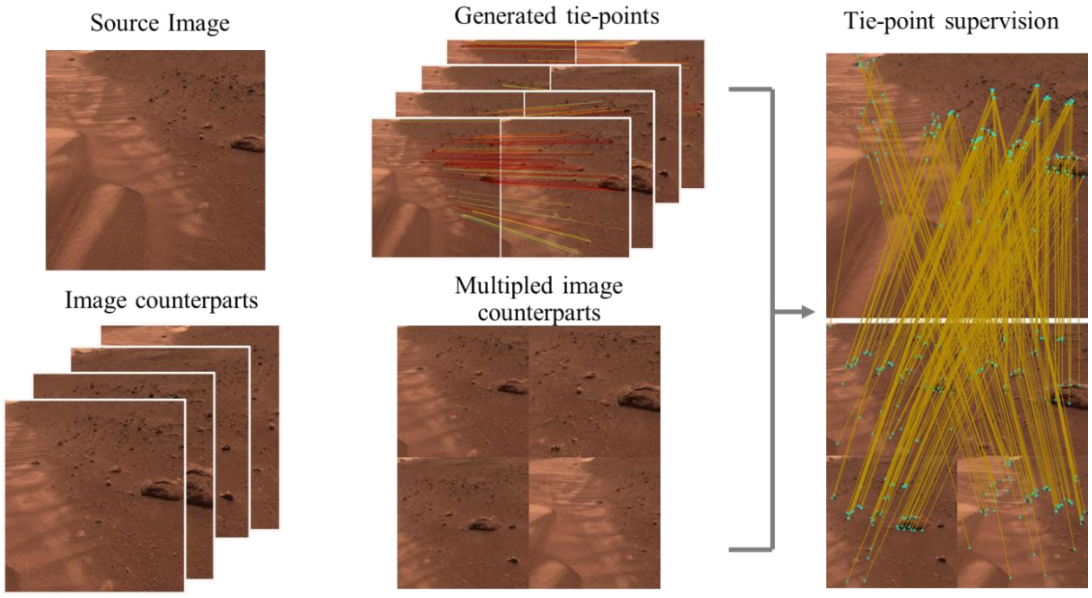


Figure 4.9 Illustration of the semantic-guide tie-point matching strategy.

4.4.3 Iterative Training Strategy

The loss function $\mathcal{L}_{all}$ is thus composed of two parts. While the first part $\mathcal{L}_{Label}$ examines the cross-entropy loss between the images and the supervised labels, the second part $\mathcal{L}_{corre}$ punishes inconsistent segmentation between the input images. After several warm-up epochs, the weight of the first part is expected to decrease and the images are made to supervise each other to mitigate the incomplete labeling issue caused by the complexity of the landforms.

$$\mathcal{L}_{all} = \mathcal{L}_{Label} + \mathcal{L}_{corre} \tag{4-9}$$

Note that during the training process, some unlabeled landforms may be retrieved, requiring the calculation of tie-points based on the union of the predicted and labeled masks. Consequently, the output of the entire training process is twofold: not only is a segmentation network obtained, but also a reliable tie-point dataset is generated, which may be valuable for future applications. Considering the volume of the SuperGlue network and the Swin Transformer, two GPUs are utilized for the overall training process.

4.5 Experimental Evaluation

4.5.1 Dataset Description

In this study, the data acquired from China's first Martian rover, Zhurong, are used for evaluating the performance of the proposed semantic-aware image matching algorithm. The rover is equipped with a stereo navigation camera to capture the surrounding environment and ensure safe traversal. As listed in Table 4.1, this camera features a 13.169 mm focal length, a 27 cm baseline, and a high-resolution sensor comprising 2048×2048 pixels and an 11.264 mm sensor size. Typically, the camera's pointing direction is tilted $14 - 19°$ from the horizontal plane, enabling the capture of scenarios at a considerable distance from the rover. As the resolution varies with the angle and the viewing distance, the approximate resolution is listed in the table. However, there are also a few stations with a nearly 29° horizontal tilt, which allow for careful observation of near-range scenarios with sub-centimeter resolution.

Table 4.1 Parameters of the camera (PCAM) onboard Zhurong rover.

Parameters	PCAM
Focal length	13.172 mm
Pixel size	5.5 μm
Sensor size	2048×2048 (pixels)
Baseline	27 cm
Pointing direction	19° / 29°

4.5.2 Dataset Construction Results

The 3D mesh models are first generated through the SfM and MVS pipelines (Bentley, 2019). However, holes and over-interpolation may occur due to the inevitable occlusion problem resulting from the perspective of the rover. Examples of these situations are visualized in Figure 4.10. Virtual cameras are thus defined on the basis of the original cameras to avoid these defects. Empirically, the distance between the positions of the virtual and the original camera along the direction away from the rover should be within 8 meters. The rotation and IO parameters can be more flexible if a favorable position is selected. Although the algorithm is limited by the quality of the 3D mesh, the number of training images is enriched 20 times to ~20,000 in a 3D manner. It is worth noting that the number of virtual cameras for one 3D model is not constant. The semantic labels are also used to balance the number of samples in each class, and the amount is hence adapted automatically. Considering the memory of the GPU, the images are then cropped into patches of 512×512 pixels through three levels of scale pyramids.

To alleviate the labor-intensive human labeling process, the SAM algorithm (Kirillov et al., 2023) is employed within the LabelMe software (Ma et al., 2024b). By leveraging the SAM, homogeneous regions can be automatically extracted from a single point within the region, significantly reducing the time required to annotate polygon boundaries. Notably, the SAM excels in segmenting small rocks with distinct boundaries, but it struggles to extract large landforms, such as sand dunes and craters, which often lack clear boundaries. Consequently, manual polygon drawing remains necessary, and the semi-automatic dataset construction algorithm continues to play a crucial role in the labeling process.

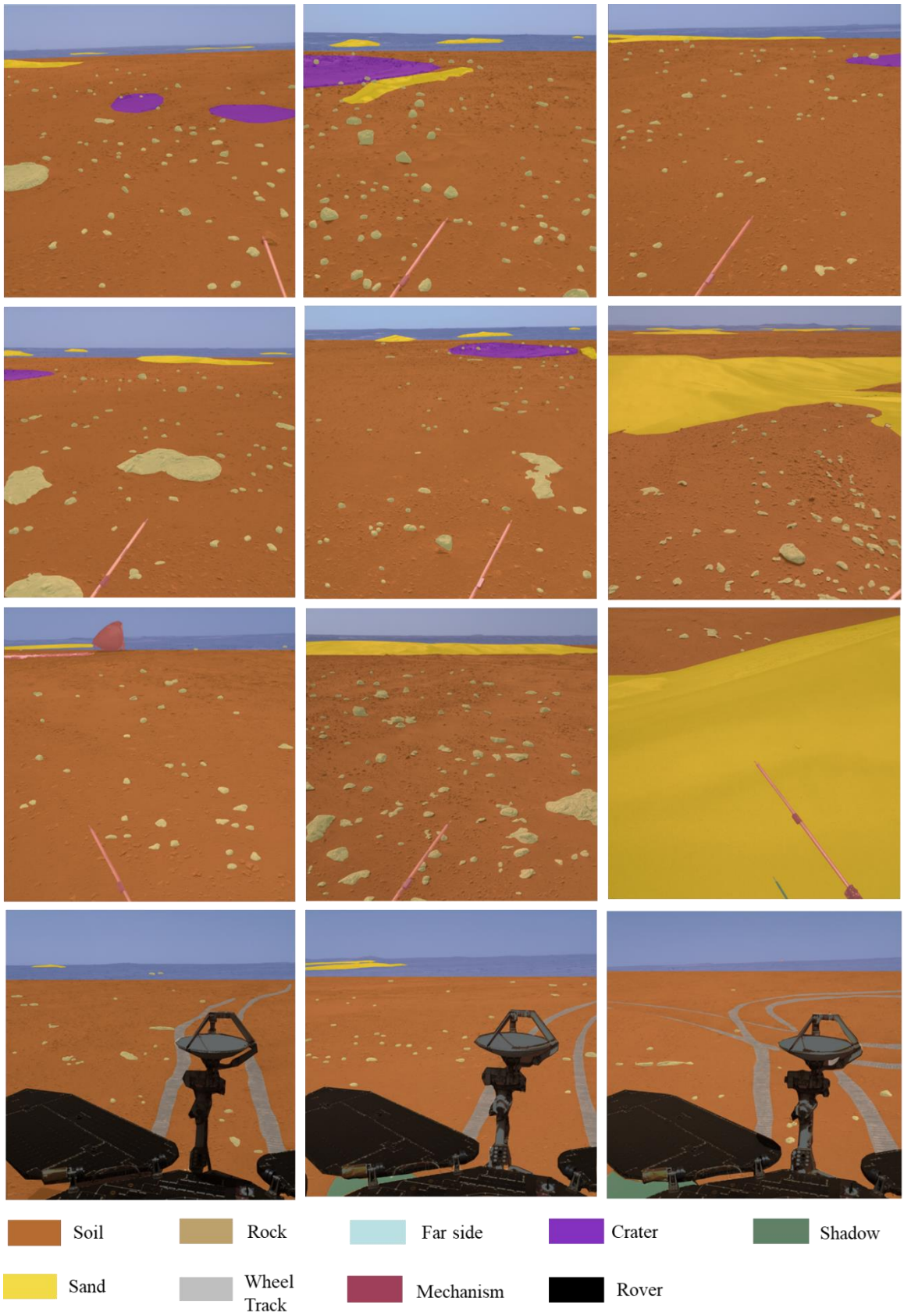


Figure 4.10 Representative labeled images.

4.5.3 Experimental Evaluation of the Depth-enhanced Transformer-based Semantic Segmentation

To evaluate the performance of the depth-enhanced transformer, we train and compare various state-of-the-art segmentation neural networks, including OCRNet, fully convolutional networks (FCNs), DeeplabV3, SegFormer, Mask2Former, Swin Transformer, and SGACNet, to serve as baselines for our proposed algorithm. The ground truth is manually labeled.

Two classical CNNs are used. FCNs (Long et al., 2015) are a type of deep learning architecture specifically designed for tasks involving image segmentation. Unlike traditional CNNs that typically end with fully connected layers, FCNs replace these layers with convolutional layers that produce spatially dense outputs. This allows FCNs to take input images of arbitrary size and generate output maps of the same size, where each pixel is classified into a specific category, and to perform pixel-wise classification. By leveraging techniques like upsampling and skip connections, FCNs can capture both high-level semantic information and fine-grained details, resulting in more accurate and detailed segmentation maps. While FCNs laid the groundwork for semantic segmentation by introducing a fully convolutional approach, DeepLabV3 significantly enhances this foundation with advanced techniques like Atrous Spatial Pyramid Pooling (Chen, 2017). These innovations enable DeepLabV3 to capture multi-scale contextual information more effectively, resulting in more precise and detailed segmentation maps.

Object-contextual representations (OCRNet) is a pioneering transformer-based model that demonstrates the feasibility of the transformer-based encoder-decoder architecture (Yuan et al., 2020). By leveraging self-attention, OCRNet effectively models the relationship between objects and their background, capturing long-range dependencies to enhance accuracy. SegFormer (Xie et al., 2021) integrates the transformer with multi-layer perceptron (MLP) decoders to extract multi-scale features while avoiding complex decoders, resulting in a

lightweight overall network. And it is also positional-encoding-free architecture. Meanwhile, MaskFormer (Cheng et al., 2022) introduces a mask to capture the shape and position of objects, and utilizes multi-head attention to aggregate features in the image. Furthermore, the original Swin-transformer (Liu et al., 2021) is used as a comparison to illustrate the strengths of the proposed depth-enhanced transformer. To verify the necessity of the transformer-based approach, we also train and test SGACNet (Zhang et al., 2023), which takes cross-modal data as input, where the RGB and depth images share a common encoder. The statistical performances of these segmentation networks are compared in Table 4.2.

Table 4.2 Statistical comparison with classical semantic segmentation neural networks.

Methods	Semantic Class			
	Soil	Rock	Sand	Others
OCRNet (Yuan et al., 2020)	6.62	47.45	11.26	67.25
Deeplabv3 (Chen et al., 2017)	74.81	99.00	69.20	98.82
FCN (Long et al., 2015)	74.22	98.98	69.82	98.80
SegFormer (Xie et al., 2021)	96.31	42.96	93.26	98.55
Mask2Former (Cheng et al., 2022)	96.17	57.21	92.46	98.59

Swin Transformer (Liu et al., 2021)	92.67	54.09	91.47	95.14
SGACNet (Zhang et al., 2023)	95.70	34.83	87.81	96.76
DepthFormer	93.82	60.25	92.66	94.78

The results for three representative images are shown in Figure 4.11. The CNN, FCN, DeeplabV3, SegFormer, and Mask2Former present stable performance, segmenting the subtle small rocks yet not yielding evidently false labels. However, OCRNet not only misclassifies many sand pixels into the sand dune class but also misses many small rocks. Notably, the original Swin-transformer fails to accurately segment the lower right portion of the large rock in the second row, and commits a significant error by misclassifying a substantial area of sand as soil. This phenomenon can be attributed to several factors. First, the fixed window size of the Swin Transformer may not be sufficient to capture the overall features of the large landforms. Moreover, although the Swin Transformer attempts to obtain global features by fusing window features, the fusion process may lead to loss or confusion of feature information when the features of large objects are split across different windows. In contrast, the depth-enhanced model, DepthFormer, is designed to address these limitations by incorporating depth information into the feature fusion process. Meanwhile, even with depth information, SGACNet cannot segment the image into a favorable number of semantic segments, missing many small rocks. Conversely, our depth-enhanced transformer not only accurately segments the overall sand region but also retrieves a favorable amount of small rocks, including even the tiny rocks in the upper right part of the second row. This is due to the transformer's ability to effectively incorporate depth information into the feature fusion process, allowing it to capture

subtle details and nuances in the image. Furthermore, this high accuracy enables the use of more semantic labels for subsequent dataset construction, which in turn facilitates the training of other tasks by providing more detailed and informative annotations.

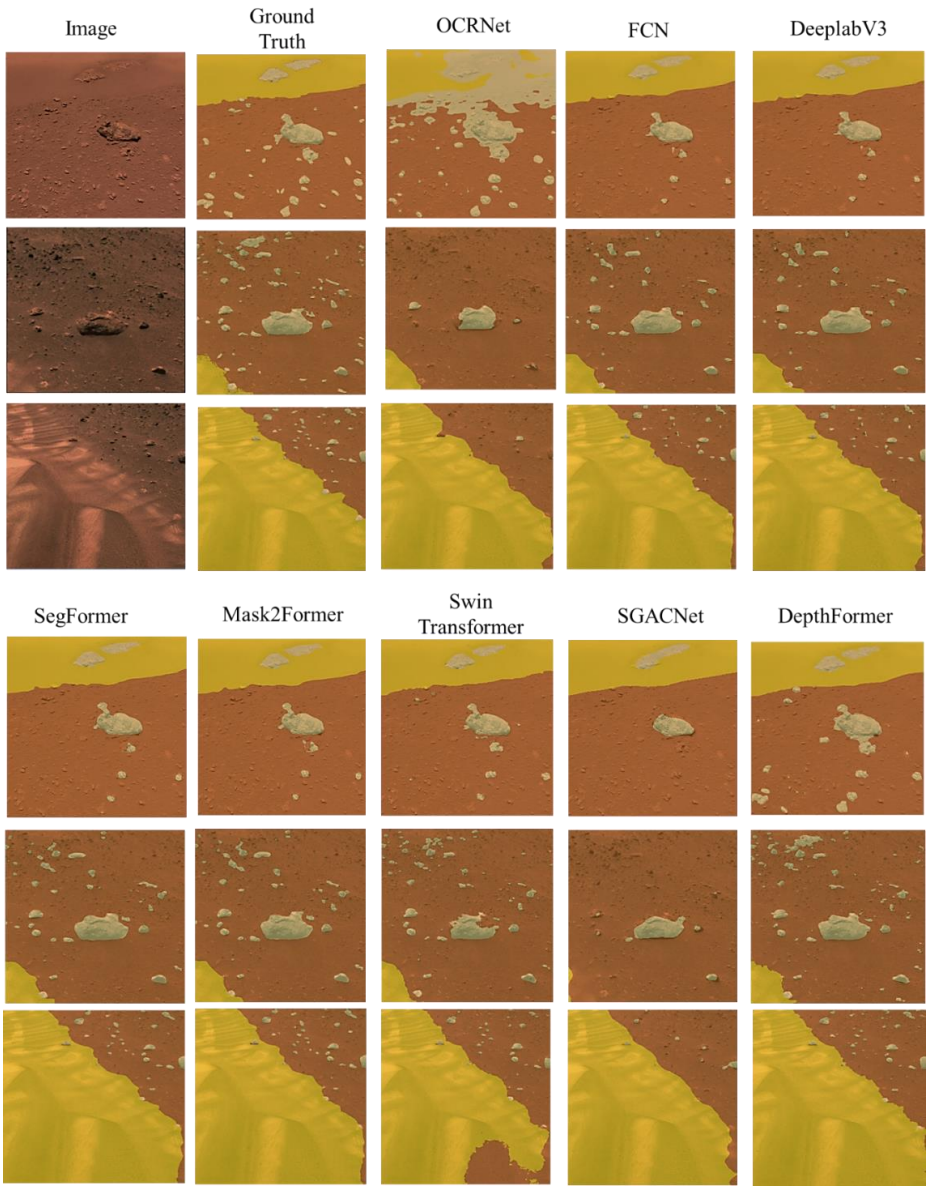


Figure 4.11 Qualitative comparison of the segmentation results from the FCN, Deeplabv3, Swin Transformer, and depth-enhanced transformer.

4.5.4 Experimental Evaluation of the Siamese Transformer-based Semantic Segmentation

The overall performance of the semantic segmentation achieved by the Siamese transformer is visualized in Figure 4.12, which presents results encompassing a wide range of semantic classes, including soil, rock, sand, rover, crater, wheel track, shadow, mechanism, and far side. Visually, the results demonstrate that many rocks are accurately segmented, indicating that the transformer has successfully extracted favorable features. Moreover, even craters with limited supervised labels are correctly segmented. In the sand dune region, large dunes on the near side are well-segmented with reasonable boundaries, while those on the far side are mostly labeled correctly, with only a few pixels missed by the network. Notably, for the last images, there is a small shadow region, which is also correctly attached with the corresponding labels.

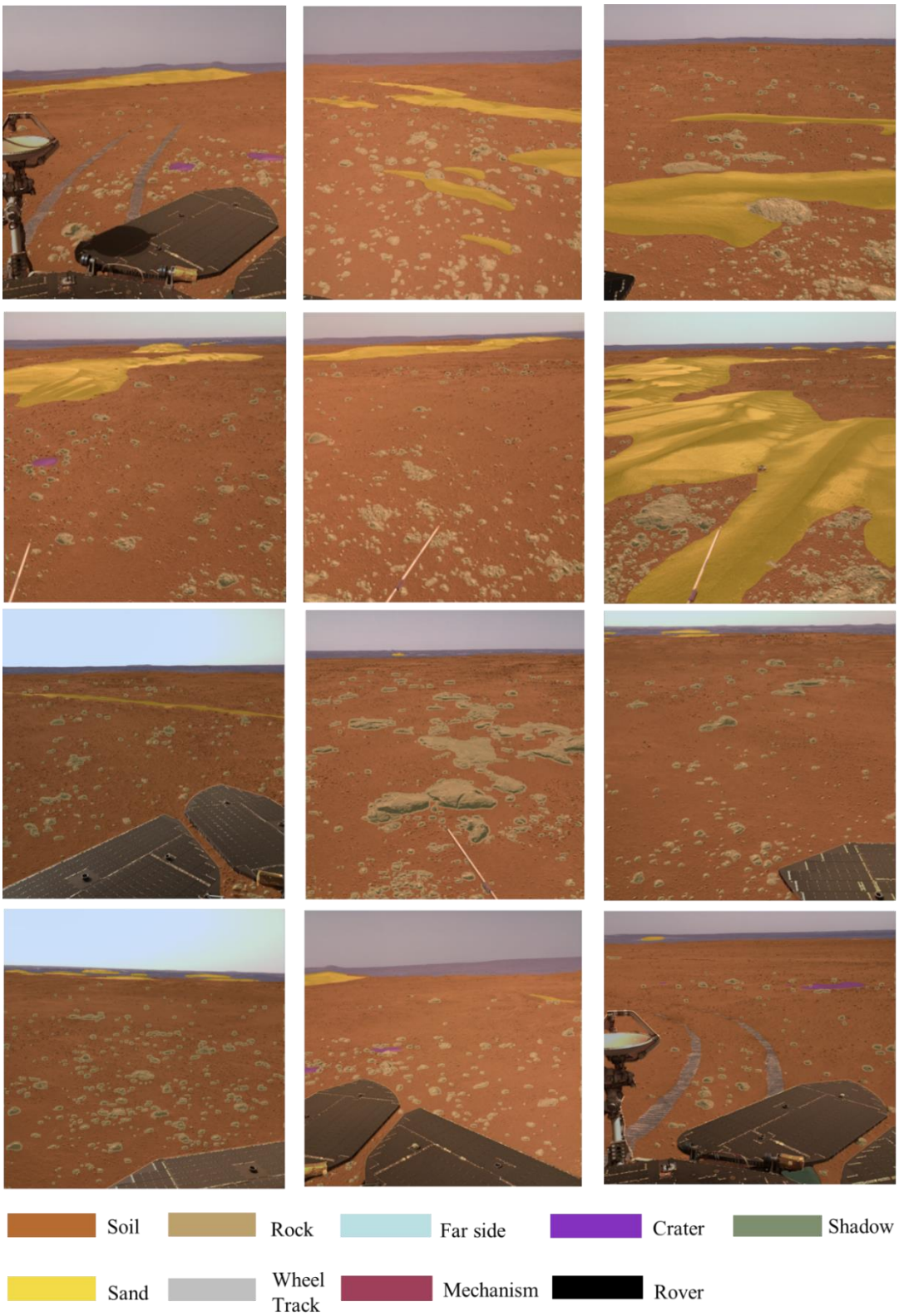


Figure 4.12 Representative semantic segmentation results generated from the Siamese transformer-based neural network.

The improvement in segmentation results demonstrates the effectiveness of the self-supervised framework. Although the network has already learned the features, using incomplete semantic labels can lead to confusion. However, through mutual supervision between images, the network can enhance its understanding of the category, thereby enabling the extraction of more semantic classes without human intervention. Consequently, more rocks and sand dunes on the far side can be accurately labeled, rather than being misclassified as soil. This is particularly important for geological analysis, as accurate identification of rock and sand dune formations is crucial for understanding the geological history and processes of a region.

Furthermore, the evaluation of the transformed images is performed to test the transform invariance of our approach. Three experiments involving all the translation, rotation, scale, and real-world transformations are exhibited in Figure 4.13. Two highlighted regions are marked by the white and blue ellipses, respectively. Despite the incomplete or incorrect retrieval problem, the semantic labels in the ellipses calculated by the Swin-T are not strictly aligned across the experiments. In contrast, the Siamese Swin-T tends to maintain a consistent pattern even when there are more segmented rocks, which not only improves the accuracy of the segmentation but also ensures the transformation invariance of the extracted features. This is particularly important for downstream photogrammetric applications, which rely heavily on the consistency among overlapping regions and the invariant description of corresponding points across different images. The transformation invariance of the extracted features is crucial for ensuring that the features are robust to changes in viewpoint, lighting, and other environmental factors, which is essential for accurate photogrammetric reconstruction.

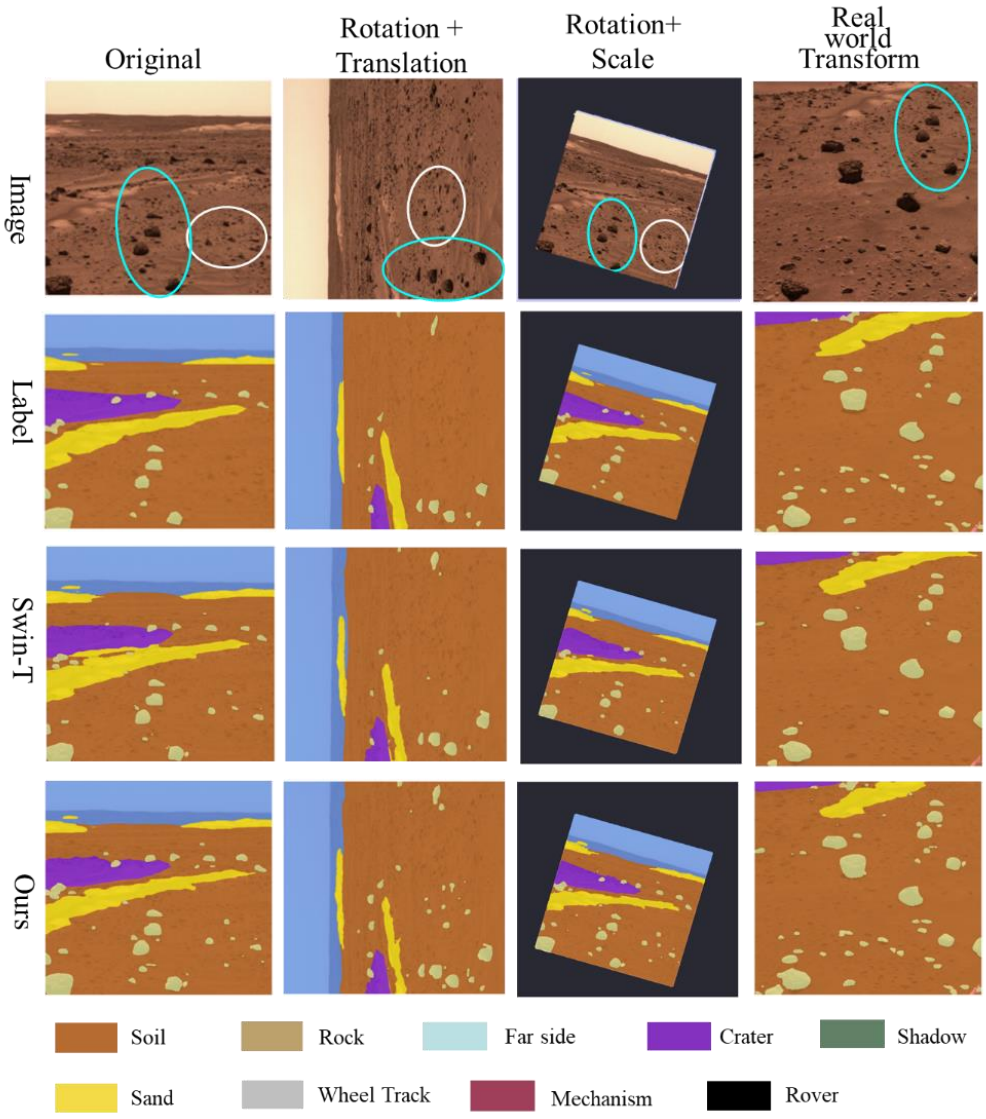


Figure 4.13 The illustration of the consistency of the semantic segmentation results.

Quantitative analysis is also conducted in the aspects of both accuracy and consistency, as listed in Table 4.3. For accuracy evaluation, the mean intersection over union (mIOU), a commonly used evaluation index is used. In image segmentation, the intersection is typically used to calculate the overlapping region between two segmentation results, while "union" refers to all existing pixels in both the ground truth and prediction. Using the same training dataset and test dataset, the mIoU of the original Swin-T is 86.08%, which is further improved to 88.25% by the proposed Siamese transformer. Consistency is evaluated through the

consistency ratio and KL divergence. Tie-points are extracted using the SuperGlue algorithm, and their semantic labels are compared to determine consistency. The consistency ratio is calculated as the percentage of tie-points with matching semantic labels, and the mean and standard deviation are reported. To measure the similarity between the extracted features, the KL divergence is calculated using Equation (4-10):

$$D_{KL}(P||Q) = \sum_{i} P(i) log \frac{P(i)}{Q(i)} \tag{4-10}$$

where P and Q are the probability distribution of the features of the image pair, and $P(i)$ and $Q(i)$ represent the probability values of two probability distributions in the i^{th} dimension. $log \frac{P(i)}{Q(i)}$ is the relative entropy. As demonstrated in Table 4.3, training with overlapped supervision results in an increased consistency ratio and a decreased KL divergence, highlighting the effectiveness of the proposed Siamese transformer. With high accuracy and favorable consistency, these semantic segments and features can be effectively used to enhance the photogrammetric process.

Table 4.3 Statistics of the proposed Siamese Swin-transformer for semantic segmentation.

	Accuracy		Consistency (%)		KL Divergence	
	mIoU (%)	MPA(%)	Mean	Std	Mean	Std
Swin-T	86.08	94.33	91.2	9.4	0.067	0.013
Ours	88.25	95.78	97.1	8.6	0.043	0.012

4.6 Summary

Regarding the extraction of semantic information from planetary surfaces, this chapter focuses on addressing two primary challenges: dataset construction and improved semantic segmentation. A semi-automatic dataset construction algorithm is proposed, leveraging 3D

models and camera 3D information to generate more labeled images from a few initial manual labels, together with the depth information. Building on this, a depth-enhanced transformer-based semantic segmentation approach is extended to retrieve additional semantic segments that complement human-labeled masks, and enrich the segmentation dataset. To cater to subsequent photogrammetric applications, which require consistency among overlapping regions, a Siamese transformer-based semantic segmentation framework is designed, incorporating tie-points to supervise consistency.

Experimental evaluation is conducted using images captured by the Zhurong rover. Ten semantic classes are used to label approximately 400 images, which are later augmented using the proposed method to construct a large dataset. The depth-enhanced method is trained on this dataset, and its superiority is demonstrated when compared with other methods. The evaluation of the Siamese transformer segmentation shows an overall mIOU of 88% and a consistency ratio of 97.1%.

The significance of the method developed in this chapter lies in its ability to retrieve enhanced semantic contextual information, thereby facilitating image understanding, while also providing a reliable foundation for subsequent optimal 3D reconstruction processes.

Chapter 5 Optimal 3D Reconstruction of Planetary Surfaces Leveraging Semantic Cues

In the previous chapter, efforts are made to achieve semantic segmentation of an image, revealing the contextual information to facilitate the analysis of the image. Intuitively, this underlying information could also compensate for the textureless defects faced by the 3D reconstruction mentioned in Chapter 3.

Utilizing semantic information to facilitate 3D reconstruction is not a novel concept, as pioneering works have explored simultaneously optimizing these two products by sharing the same backbone (Eigen et al., 2015). Recent studies have discussed the effectiveness of introducing semantic information into specific algorithms, such as feature matching or dense matching. As interesting and favorable results have been obtained, the application of semantic cues to connect planetary images and achieve high-quality, large-scale 3D reconstruction of the planetary surface is worth exploring. With the segmentation neural network trained in the previous chapter, the semantic cues can be extended from 1D segments to multi-level semantic cues derived from different stages in the network, thereby catering to different requirements in each stage of 3D reconstruction.

In light of this, this chapter presents the proposed optimal 3D reconstruction leveraging multi-level semantic cues, and is organized as follows. Section 5.1 briefly introduces an overview of the proposed approach, where 3D models and the corresponding EOs are used throughout the segmentation pipeline. Section 5.2 briefly reviews the semantic segmentation proposed in Chapter 4. The proposed semantic-aware sparse and dense image matching algorithms are elaborated in Section 5.3 and Section 5.4, respectively. Experimental evaluations are presented in Section 5.5. Concluding remarks are summarized in Section 5.6.

5.1 Overview of the Approach

The photogrammetric process for generating 3D models from rover images can be decomposed into four consecutive stages: sparse image matching, bundle adjustment, dense image matching, and point cloud generation. Specifically, bundle adjustment ensures the consistency of images based on the tie-points extracted through sparse image matching. In contrast, dense image matching aims to match images on a pixel-by-pixel basis, producing disparity images for subsequent point cloud generation. However, the presence of the textureless surface and significant resolution and viewpoint variation hinders both sparse and dense image matching.

In this study, we surmount the aforementioned challenges by harnessing semantic information. The overview of the proposed semantic-aware image matching algorithm is illustrated in Figure 5.1. Beginning with the stereo images captured from different stations, the Siamese transformer-based neural network is employed to extract the multi-level semantic cues from the images. Regarding tie-point matching, keypoints are detected using SuperPoint (DeTone et al., 2018), and initially described by the semantic feature yielding 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 (Cuturi, 2013), and filtered by measuring the distance of the semantic probabilities. Subsequently, the matches are aggregated into tie-point tracks, followed by integrated bundle adjustment that connects cross-station images by inferring a consistent exterior orientation (EO). Based on the refined EOs, dense image 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 are extracted from the segments, which are utilized to adapt the parameters and preserve the discontinuities between the landforms. From the disparity images, the point clouds are calculated through space intersection, and thereafter interpolated to produce 2.5-D DEM and 3-D model, which are then textured with corresponding imagery.

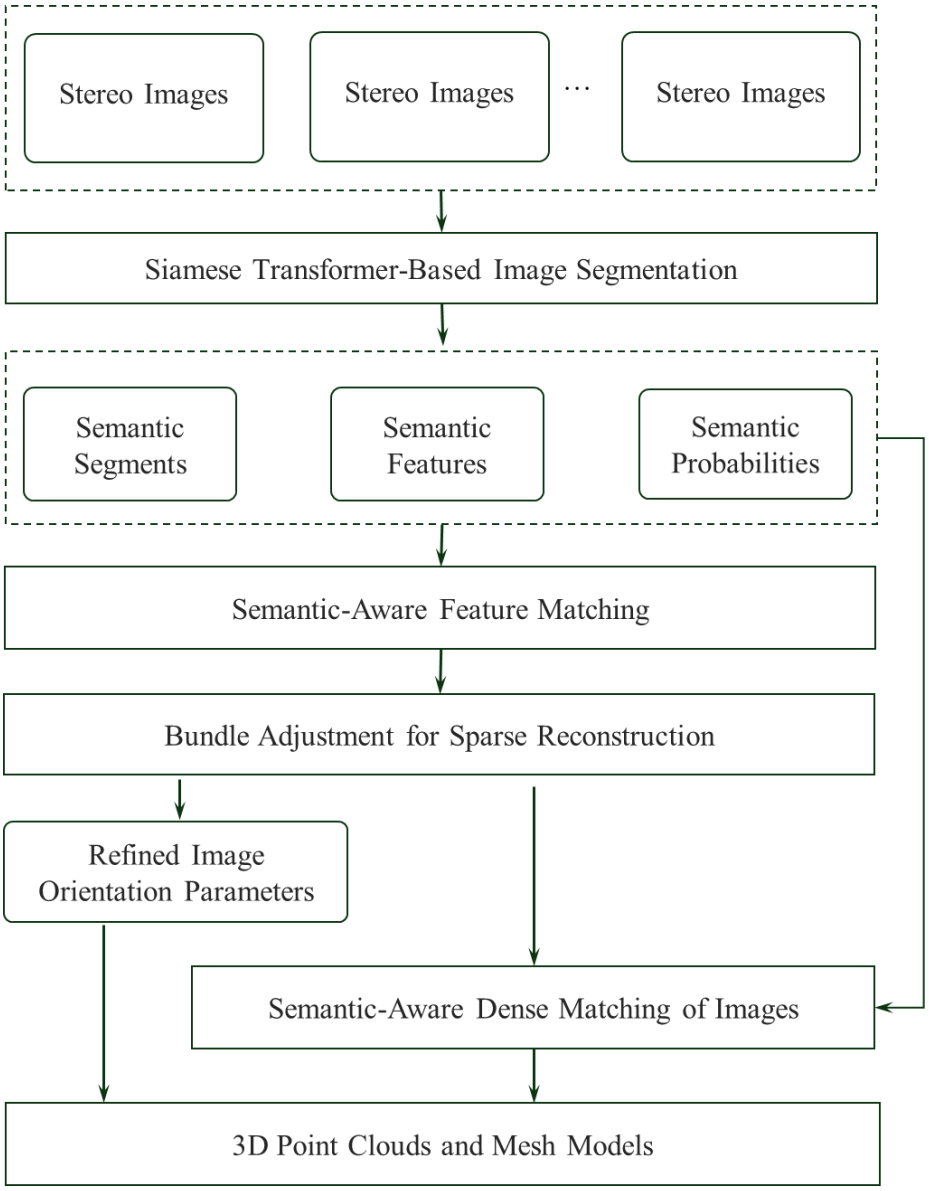


Figure 5.1 Overview of the proposed approach.

5.2 Transformer-Based Semantic Segmentation of Images

To automatically segment a 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. As illustrated in Chapter 4, the cutting-edge Swin-Transformer framework (Liu et al., 2021), 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 (Li et al., 2023b), the network is designed in a Siamese architecture, taking two overlapping images as the input, as illustrated in Figure 5.2.

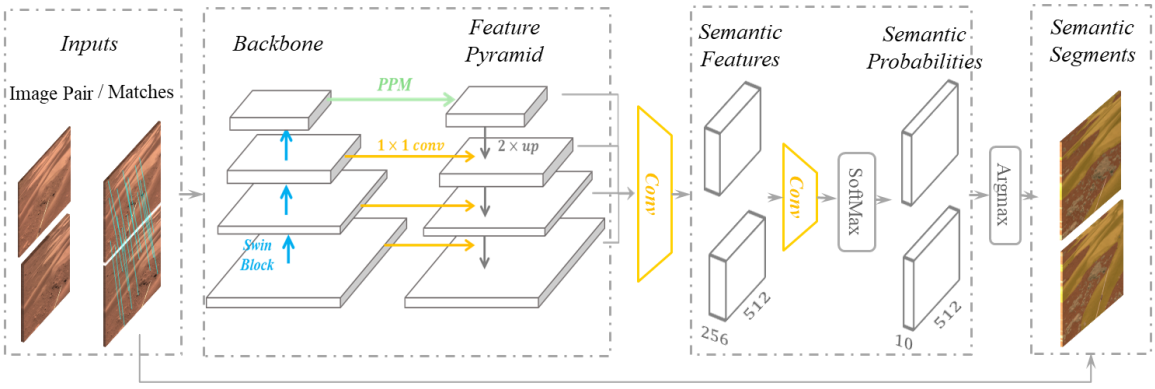


Figure 5.2 Siamese Swin-transformer-based segmentation network.

For each branch in the network, we employ Swin-T (Liu et al., 2021) as the backbone and utilize the UperNet framework (Xiao et al., 2018) to integrate multi-scale features. This framework combines a feature pyramid network (FPN) (Lin et al., 2017) with a pyramid pooling module (PPM) (Zhao et al., 2017) 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 (Bridle, 1990), yielding the semantic probabilities. Lastly, the semantic segments are obtained using the Argmax operator.

5.3 Sparse 3D Reconstruction Using Semantic Cues

5.3.1 Semantic-Aware Feature Matching

Planetary 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 planetary surface is predominantly covered by regolith, 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 images captured by different camera, which may limit the effectiveness of locally-based descriptors for feature matching.

As shown in Figure 5.3, semantic information is hence introduced to enrich the grayscale information to construct a more comprehensive descriptor, which has been exploited throughout the matching process including:

(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 initial descriptor;

(2) Rather than relying exclusively on semantic segments for image masking, they are incorporated into the attentional aggregation module to constrain message passing, selectively focusing on the salient keypoints that are likely to be observed in both images, thus providing consistent information;

(3) As the matching network may yield wrong matches, the 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.

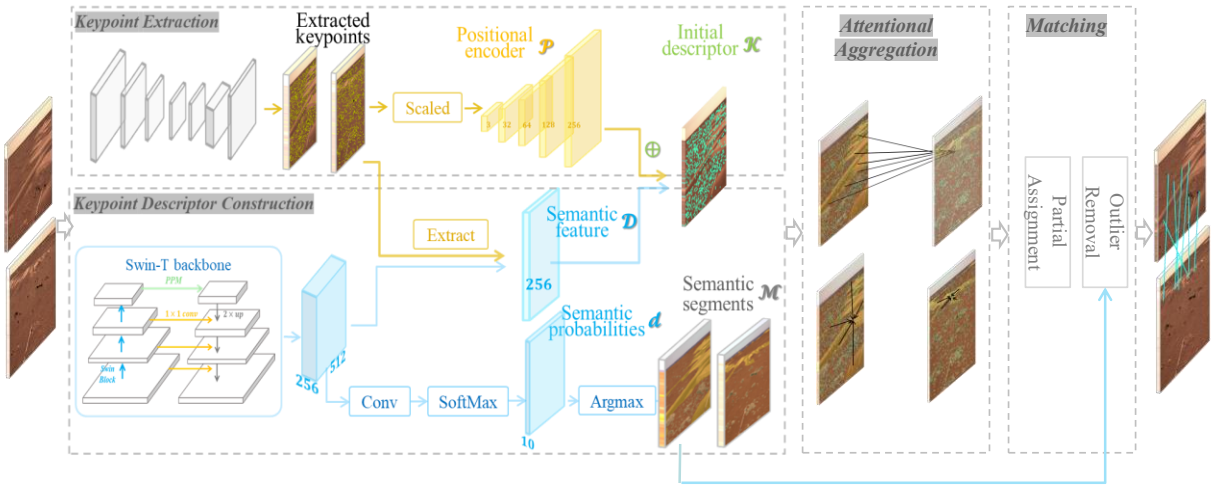


Figure 5.3 Semantic-aware SuperGlue for feature matching of cross-station rover images.

The semantic-aware SuperGlue algorithm follows the state-of-the-art matching framework established by Sarlin et al. (2020). Initially, the images first go through a VGG-style backbone and decoded to retrieve the 2-D positions of the feature points (DeTone et al., 2018). These positions are then extended to the 256-D positional encoder $\boldsymbol{\mathcal{P}}$, with descriptors $\boldsymbol{\mathcal{D}}$ 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:

$$\boldsymbol{\mathcal{K}}_i = \boldsymbol{\mathcal{D}}_i + \text{MLP}\ (\Pi(\boldsymbol{\mathcal{P}}_i)) \tag{5-1}$$

where MLP refers to multiple layer perceptron that increases the dimension of the positional encoder to 256 dimensions to be contacted with the $\boldsymbol{\mathcal{D}}_i$. $\Pi(\cdot)$ augments the 2D image coordinates with the scale information. 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 ambiguate 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:

$$\boldsymbol{\mathcal{D}}_i' = \boldsymbol{\mathcal{D}}_i + \text{MLP}([\boldsymbol{\mathcal{D}}_i || attn(\boldsymbol{\mathcal{D}}_i, \mathcal{M}_{seman})]) \tag{5-2}$$

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 (Cuturi, 2013). 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.

5.3.2 Bundle Adjustment for Sparse Reconstruction

With a sufficient number of tie-points, bundle adjustment can be performed to align both the cross- and inner-stereo images. The feature track is subsequently generated based on the pairwise matches retrieved above. As indicated in Equation (5-3), two types of constraints are considered in the bundle adjustment process.

$$\begin{cases} \mathcal{R}_{inner} = w_{inner} \sum_{k_{inner}} \sum_{i} \left\| \Pi_i(K_i, X_k) - \boldsymbol{x}_i^k \right\| \\ \mathcal{R}_{cross} = w_{cross} \sum_{k_{cross}} \sum_{i} \left\| \Pi_i(K_i, X_k) - \boldsymbol{x}_i^k \right\| \\ \mathcal{R}_{GCP} = w_{GCP} \sum_{k_{GCP}} \sum_{g} \left\| \Pi(K_g, \boldsymbol{X}_k) - p_k \right\| \end{cases} \tag{5-3}$$

where $\boldsymbol{x}_i^k$ denotes the 2D position of the k^{th} match track on the i^{th} image. $\Pi_i(K_i, X_k)$ projects the unknown 3D coordinate X_k onto the original image, which 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- stereo 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 (Li et al., 2023c). 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. The EO parameters and sparse point clouds can be calculated by minimizing the L2norm of the sum of these residuals.

5.4 Dense 3D Reconstruction Using Semantic Information

5.4.1 Semantic-Aware Dense Matching

Given the consistent EOs, the burdensome 2D pixel-wise image correspondence problem can be simplified into a one-dimensional task after the epipolar rectification (Fusiello et al., 2000). 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 (Hirschmuller, 2005). The problem for each direction can be formulated as an energy minimization task, as:

$$E = w_{data}E_{data} + w_{smooth}E_{smooth} \tag{5-4}$$

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. 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, Figure 5.4 illustrates the proposed approach to incorporate semantic cues to achieve improved disparity images.

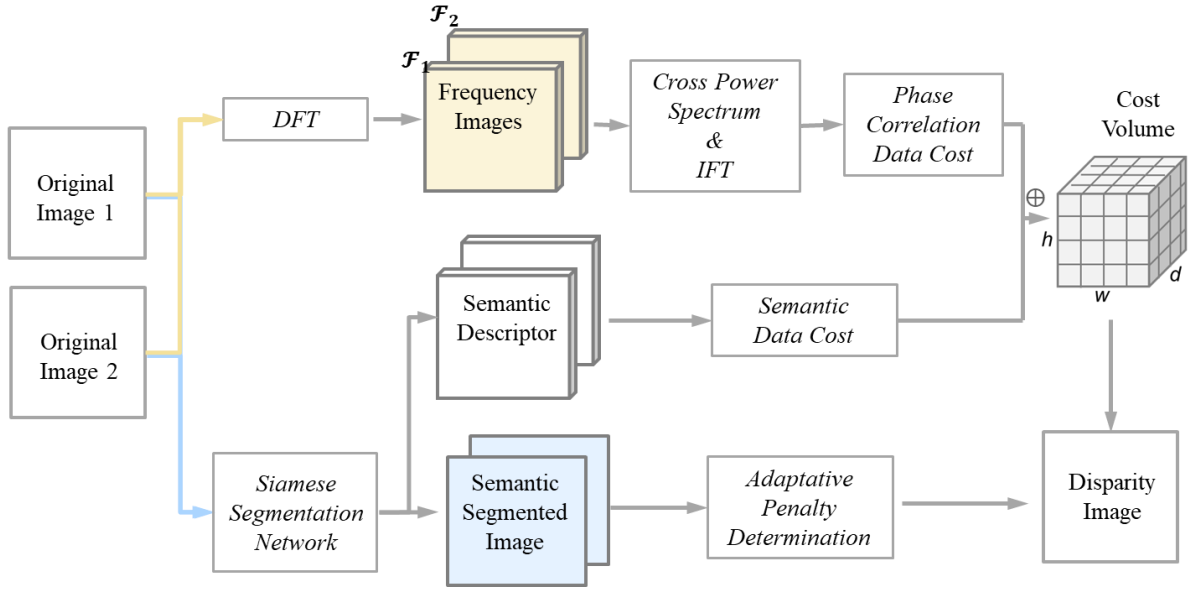


Figure 5.4 Overview of the semantic-aware dense matching algorithm.

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 (Birchfield and Tomasi, 1998) or normalized correlation coefficients in the spatial domain (Hu et al., 2016). In contrast, the proposed algorithm transforms the patches into the frequency domain to exploit phase correlation (Reddy and Chatterji, 1996) 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 (Tong et al., 2015; Wan et al., 2019; Ye et al., 2019), particularly relevant to the planetary 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.

$$Q(u, v) = \frac{\mathcal{F}_1(u, v)\mathcal{F}_2^*(u, v)}{|\mathcal{F}_1(u, v)\mathcal{F}_2^*(u, v)|} = e^{i(au)} \tag{5-5}$$

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:

$$Cost_{data}^{PC} = \begin{cases} \left(1 - \mathcal{FT}^{-1}(Q(u, v))\right) * \lambda_1, & if\ a \leq 1 \\ \lambda_2, & else \end{cases} \tag{5-6}$$

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 (5-7), which is separated into the small and large disparity change parts, as:

$$Cost_{smooth} = \sum_{(i,j)\in I} \sum_{(i',j')\in \mathcal{N}_{(i,j)}} \left(P_1(\mathcal{E}) \cdot \boldsymbol{C^r}\left[\left(d_{(i',j')} - d_{(i,j)}\right) = 1\right]\right) + \left(P_2(\mathcal{E}) \cdot \boldsymbol{C^r}\left[\left(d_{(i',j')} - d_{(i,j)}\right) > 1\right]\right) \quad (5\text{-}7)$$

where the operator $\boldsymbol{C^r}[\cdot]$ 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 (Hu et al., 2016; Rothermel et al., 2012; Yue et al., 2023). Whereas if a semantic edge is detected, a small penalty is applied, and a larger penalty is expected within the same semantic class region. The adaptation could be formulated as a linear mapping function between the user's predefined ranges.

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.

5.4.2 3D Surface Reconstruction from the Matched Results

With the dense matching results, space intersection is performed using the collinearity equation, 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, which is then texture-mapped with corresponding images based on the retrieved camera EOs, resulting in a textured mesh model.

5.5 Experimental Evaluation

5.5.1 Dataset Description

In this section, 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 and terrain cameras (NaTeCam) (Li et al., 2023a), designed to capture the surrounding environment and ensure safe traversal. As listed in Table 5.1, these cameras feature a focal length of 13.169 mm, a 27 mm baseline, and a high-resolution sensor with 2048×2048 pixels and a 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 Figure 5.5 (a). The detailed distribution of the stations in the two datasets is shown in Figure 5.5 (b) and (c), and the corresponding statistical parameters are summarized in Table 5.1. The 0716-19 dataset, consisting of 44 images acquired between

July 16 and 19, 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 3 and 24, 2022, spanning a distance of over 100 m and encompassing eight rover stations at the end of the Zhurong traverse.

Table 5.1 Description of the two test datasets.

Dataset	Date	Station Count	Image Count	Moving Distance (m)	Camera Parameters			
					Focal Length (mm)	Sensor Size (mm)	Base line (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				

Figure 5.6 presents five 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. 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 (Li et al., 2023b).

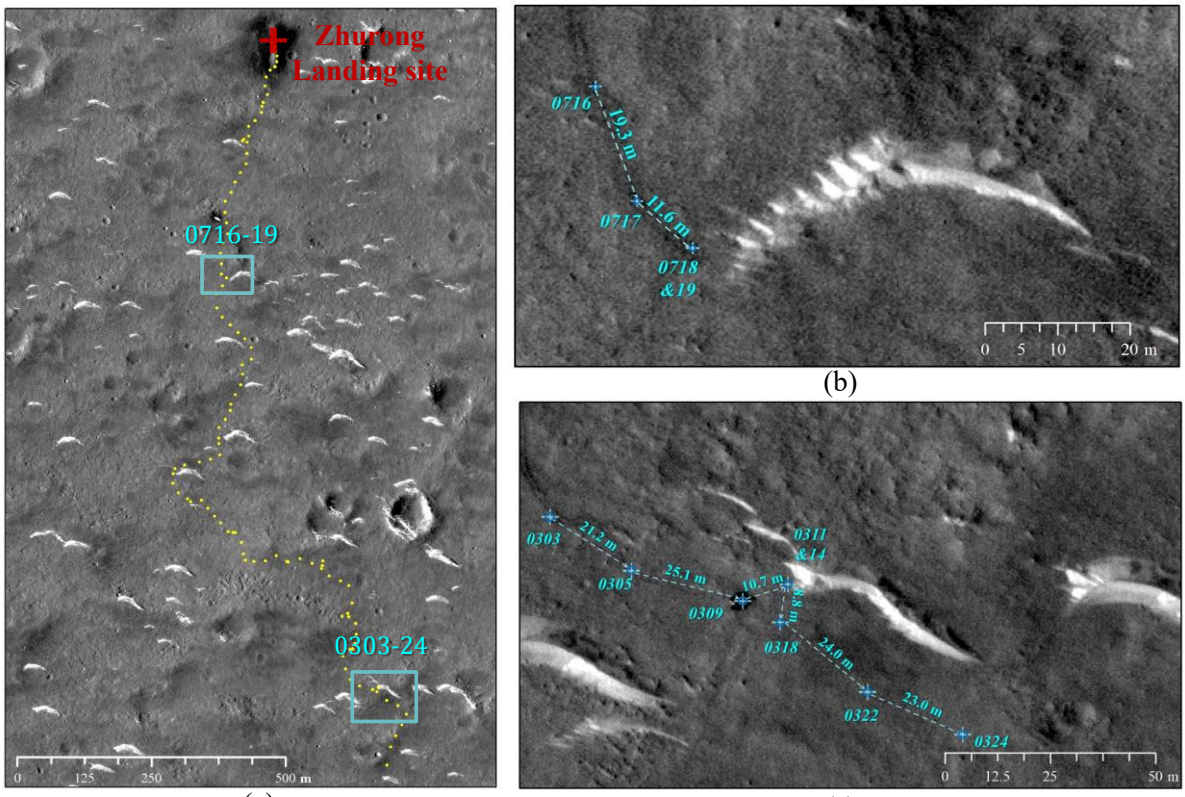


Figure 5.5 Distribution of the test areas. (a) Illustration of the two test areas overlaid on the HiRISE image (ESP_073225_2055). (b) and (c) show the detailed station distributions for each dataset.

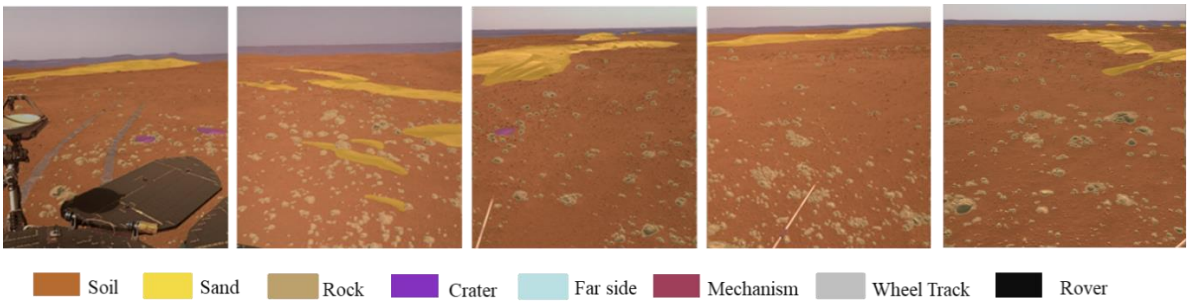


Figure 5.6 Overview of the semantic-aware dense matching algorithm.

5.5.2 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 5.2. 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 Figure 5.7. This issue is further exacerbated by the large perspective variation, attributable to the camera's large field of view.

Table 5.2 Statistics of the semantic-aware feature matching experiments.

	Characteristic	Viewing Distance (m)	Viewing Angle Difference (°)	Resolution (m / pixel)		Match Count	
				img1	img2	Original (Sarlin et al., 2020)	Ours
1	Pure resolution changes	10.82	5.40	0.006	0.02	52 (63)	67
2	Repeated pattern with a small overlap	11.06	30.11	0.01	0.05	5 (7)	41
3	Repeated pattern with	11.06	37.95	0.006	0.05	16 (21)	64

	viewing direction change						
4	Soil region with little sand	10.82	28.81	0.007	0.05	18 (26)	39
5	Barren soil region	9.95	0.0	0.005	0.04	58 (69)	77

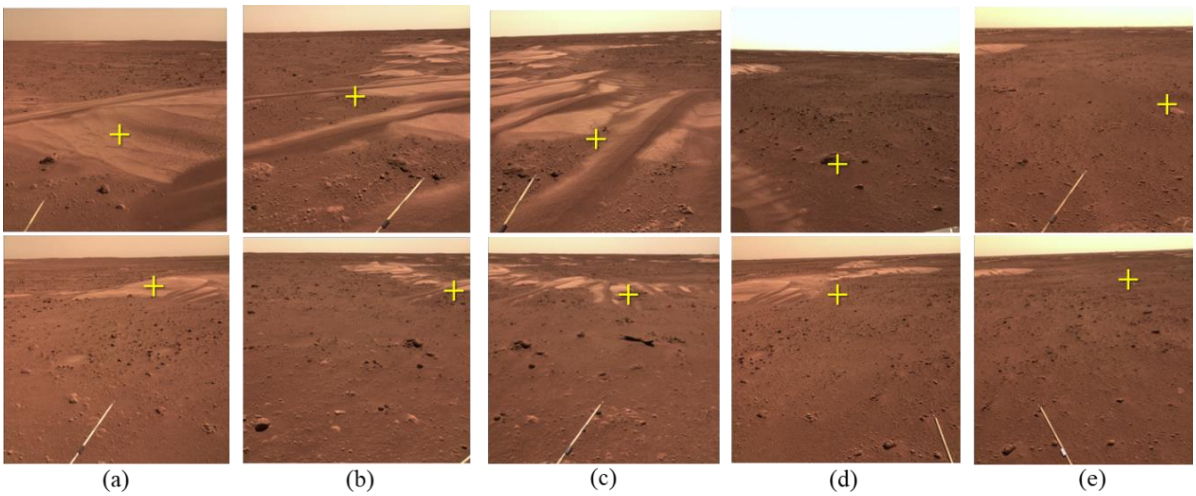


Figure 5.7 Experiment image pairs for the semantic-aware SuperGlue experiment.

The proposed algorithm is compared with the original SuperGlue algorithm (Sarlin et al., 2020), as the traditional SIFT fails to generate meaningful matches in the considered scenarios. As the model trained on the outdoor scenes performs better in this context than that trained on indoor scenes, it is selected to calculate the matches. Various parameter settings are explored, and the optimal results are selected for comparison, as presented in Figure 5.8.

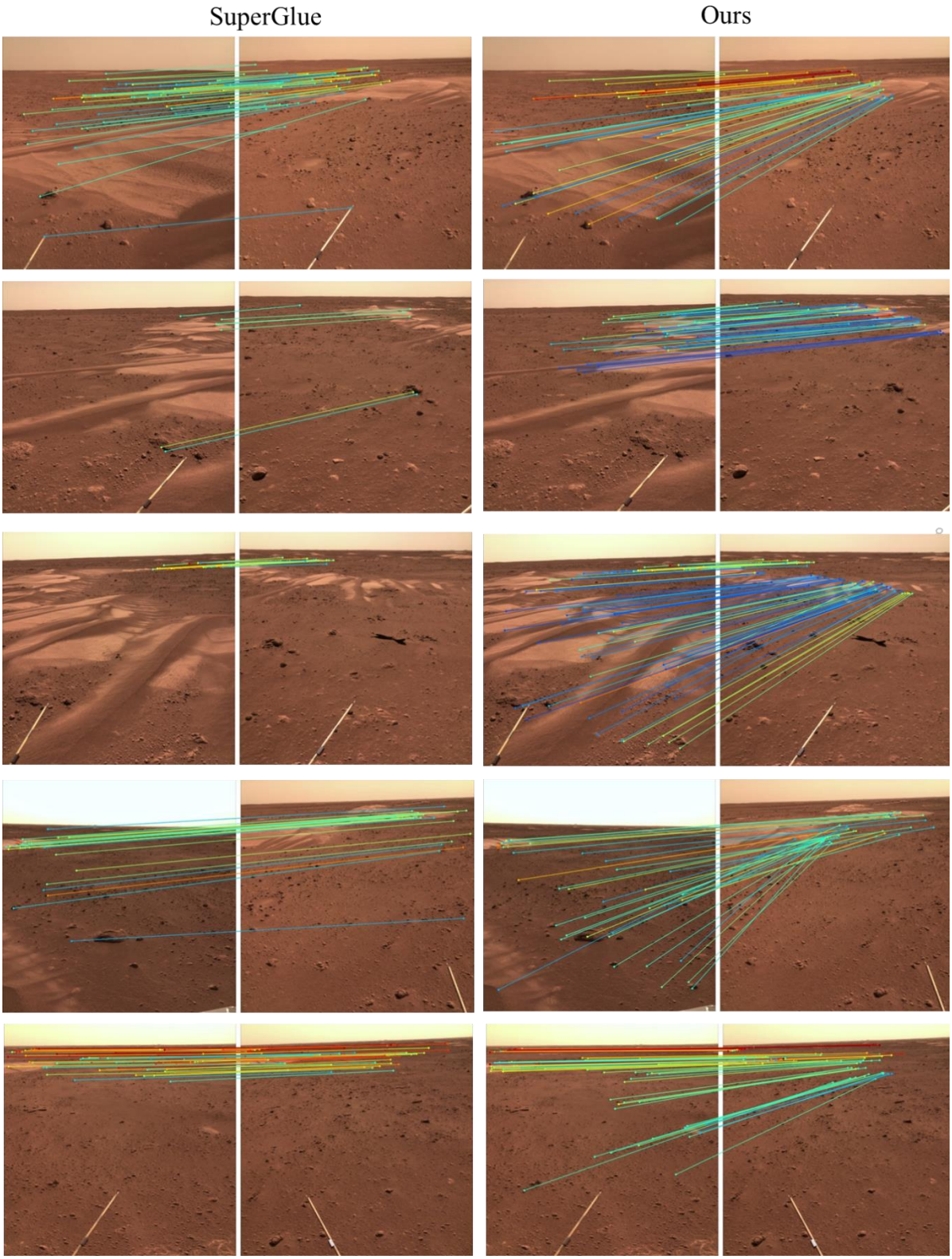


Figure 5.8 Results of the feature matching. (a) and (b) are the results derived from the original SuperGlue and the semantic-aware SuperGlue, respectively.

The first pair features a region dominated by repetitive sand dunes, with a similar viewing direction and an overlapping area that does not suffer from significant distorted transformation. The original SuperGlue algorithm successfully identifies four correct matches in the distant background, but is confounded by a large rock in the foreground, mistakenly matching features that are not the same. 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 second pair presents the region is also 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 allows the original SuperGlue algorithm to retrieve a substantial number of accurate matches, even at 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 last two pairs concern the common regions dominated by the soil, which is textureless and mostly filled with small rocks. Unlike the SuperGlue algorithm merely matches the

obvious overlapped sand dune part correctly, 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 5.3, 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 95^{th} 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. The observed residuals exceeding one pixel are primarily attributed to two factors: (1) limited texture in planetary surface regions, which

hampers sub-pixel matching accuracy, and (2) the relatively short baseline (27 cm) of the stereo system, resulting in rapidly declining depth precision at distances beyond 30 meters. As demonstrated by geometric analysis, the small disparity at long ranges amplifies the impact of pixel-level matching errors on 3D reconstruction accuracy.

Table 5.3 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)

5.5.3 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 Figure 5.9. 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.

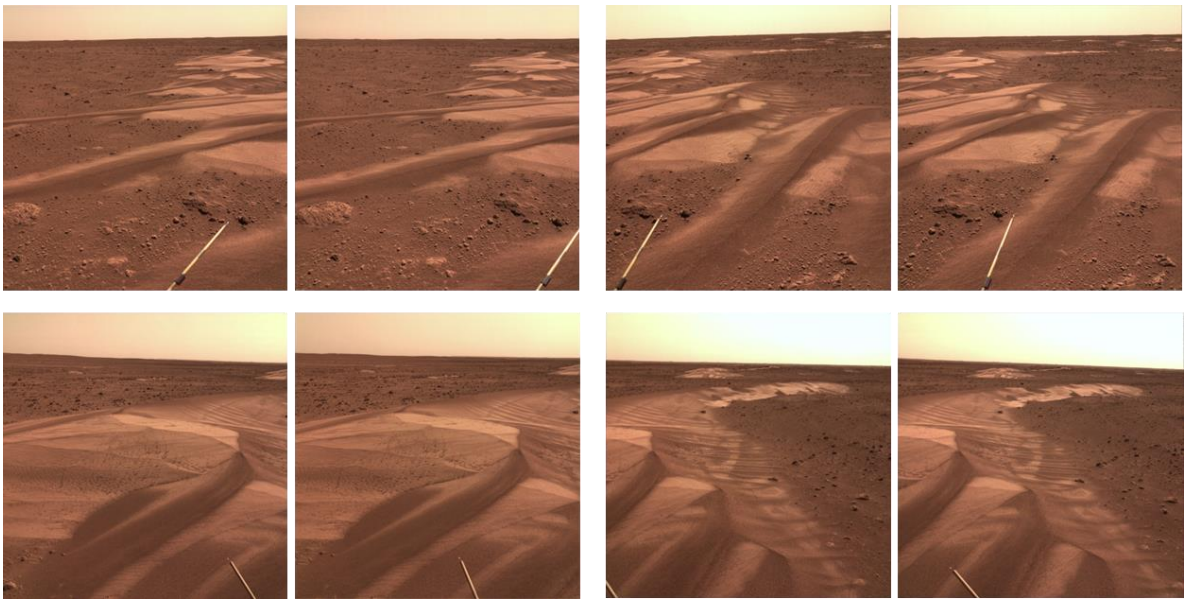


Figure 5.9 Overview of images used for the semantic-aware dense matching algorithm. The first and second rows correspond to the 0716-19 and 0303-24 datasets, respectively.

The results of the 0716-19 dataset are presented in Figure 5.10. 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. Notably, in the absence of ground truth (GT) data for quantitatively analyzing the accuracy of disparity images, the 3D mesh model, which transforms the 2D results into the 3D space, provides an intuitive way to visualize the differences among the algorithms and comparisons with the original

images. The 3D mesh model used here is a subset derived from merging neighboring point clouds, thereby clearly highlighting the advantages of the proposed algorithm.

As shown in the first column in Figure 5.10, the disparity image obtained using AD-Census suffers from no-data and speckle issues at the boundary regions and in areas where occlusion occurs. 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. 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 zoomed view still exhibit distortions, but these artifacts are reduced through the phase-correlation-based semantic-aware dense matching. 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.

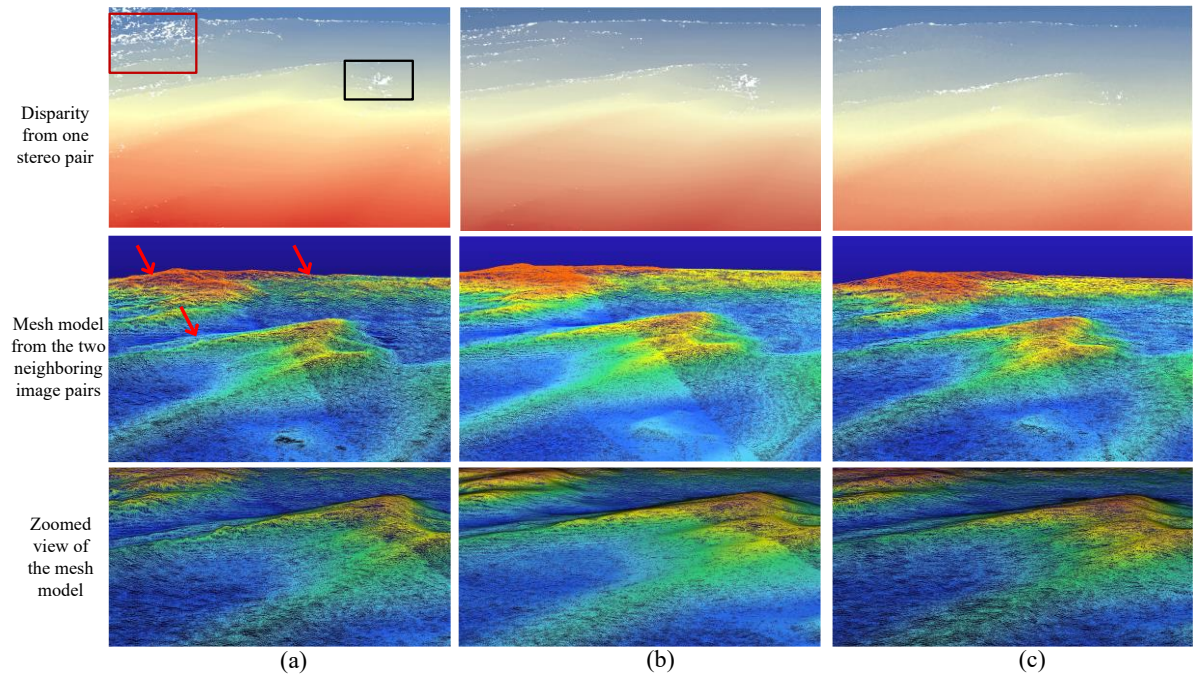


Figure 5.10 Comparison of dense matching results. (a)-(c) are the results calculated from AD-Census alone, semantic-aware AD-Census, and proposed semantic-aware phase-correlation, respectively. The first row shows the disparity images from one stereo pair, the second row presents the 3D meshes, and the 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.

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 Figure 5.11, 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.

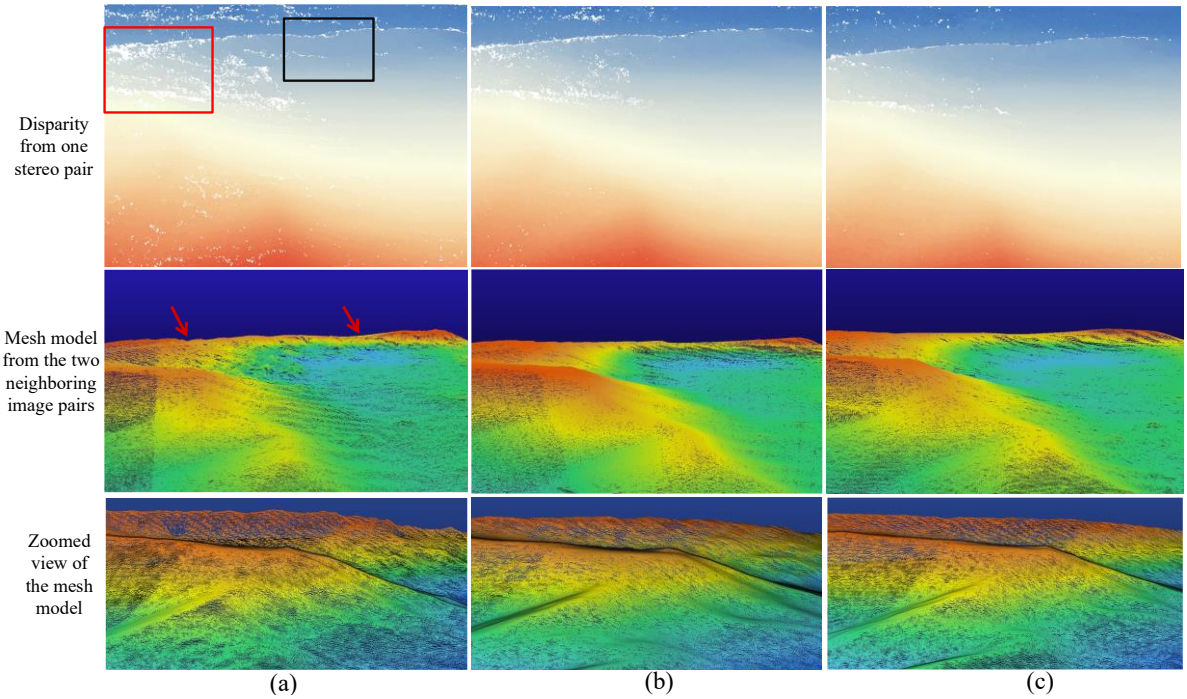
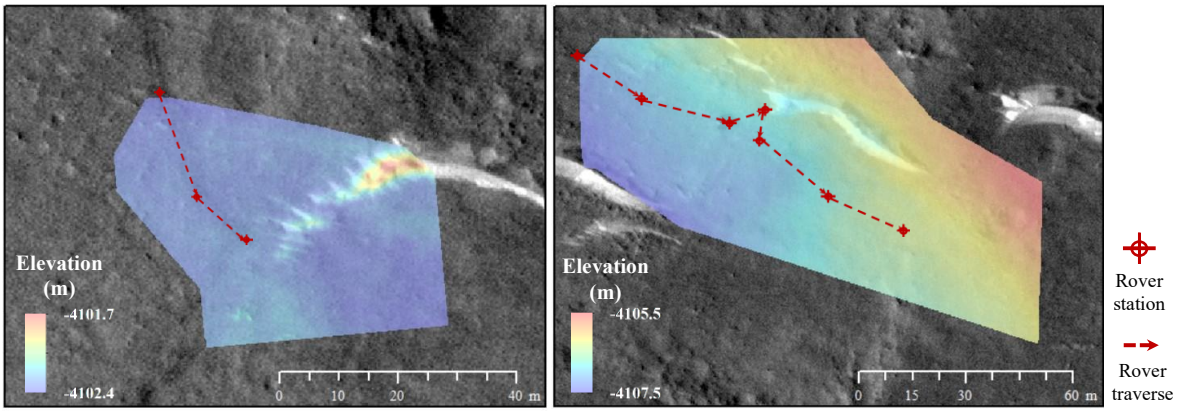


Figure 5.11 The DEMs generated using the proposed approach. (a) and (b) are the results of the 0716-19 and 0303-24 datasets, respectively.

5.5.4 Evaluation of the 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 Figure 5.12. 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.



(a) (b)

Figure 5.12 The DEMs generated using the proposed approach. (a) and (b) are the results of the 0716-19 and 0303-24 datasets, respectively.

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 5.4, while the spatial distribution of these points is illustrated in Figure 5.13. 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.

Table 5.4 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)	3.3	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

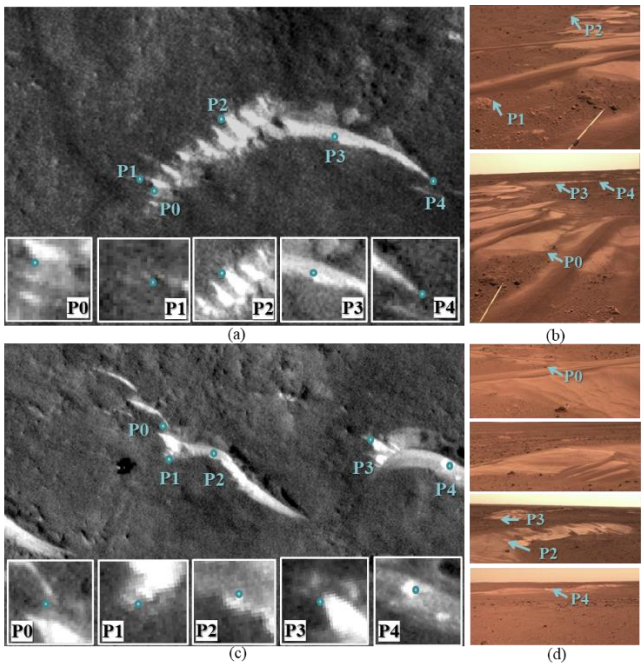


Figure 5.13 The distribution of selected points for absolute 3D position verification. (a) and (b) mark points on satellite and rover images for the 0716-19 dataset, and (c) and (d) mark points for the 0303-24 dataset.

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. Figure 5.14 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.

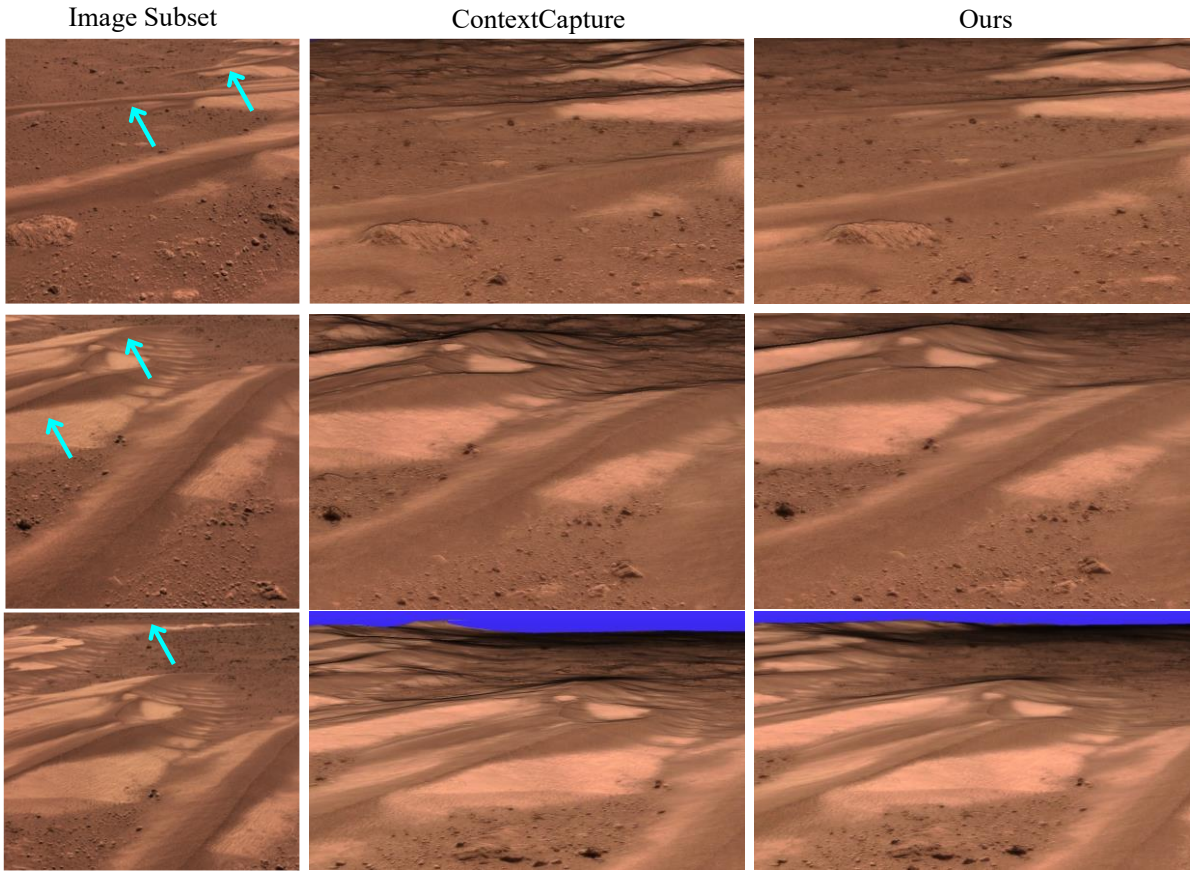


Figure 5.14 Comparison of the textured 3D mesh model for the 0716-19 dataset.

Three representative regions are also selected from the 0303-24 dataset, as visualized in Figure 5.15. 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.

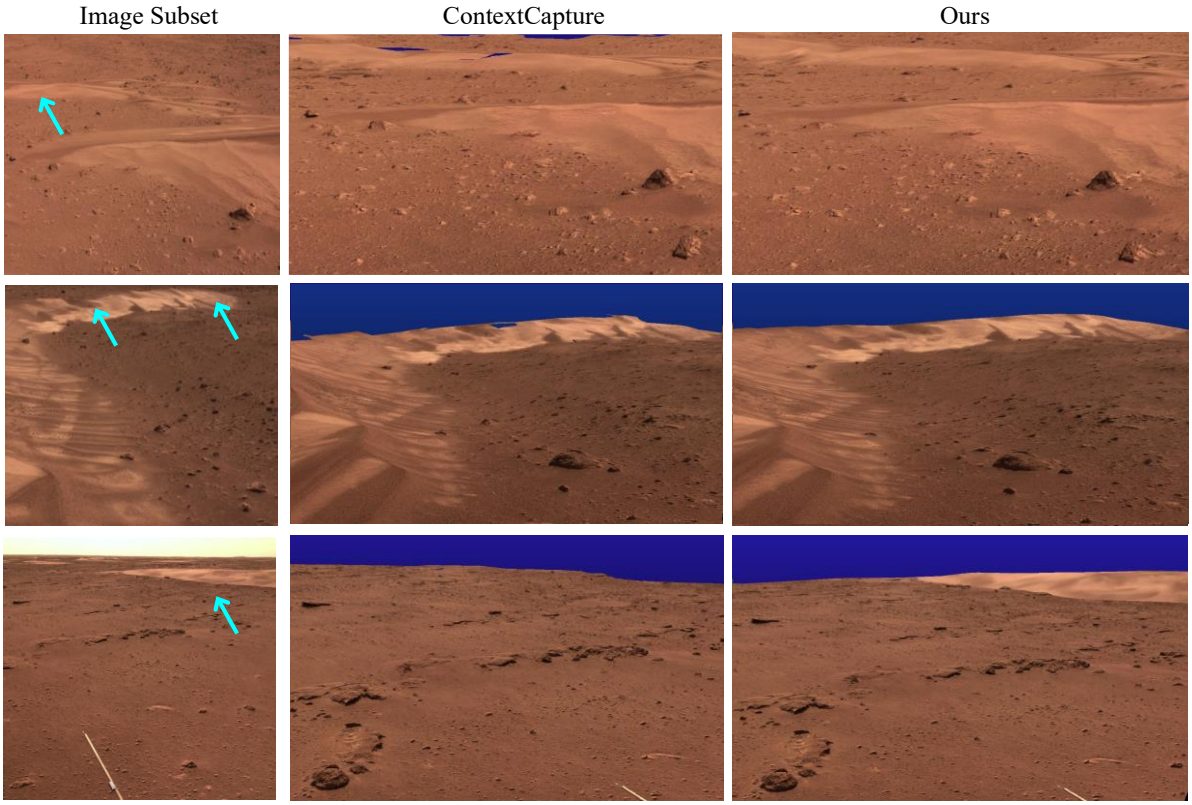


Figure 5.15 Comparison of the textured 3D mesh model for the 0303-24 dataset.

5.6 Summary

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.

Chapter 6 Conclusions and Discussion

This dissertation presents a series of approaches to the 3D reconstruction and semantic segmentation of planetary surfaces. Departing from traditional methods that process each component separately, this research endeavors to integrate these two tasks into a unified pipeline, thereby enriching the RGB information and yielding improved results. By bridging 3D reconstruction and semantic segmentation, this work aims to describe surface characterization in both geometric and semantic aspects, thereby revealing the underlying structural and compositional information embedded in planetary imagery. While the overall pipeline enables the generation of optimal 3D reconstruction results accompanied by semantic contextual information, the individual algorithms can also be employed separately, providing flexibility and versatility in various applications. The proposed algorithms are evaluated using a diverse set of images that cover various planetary surfaces, allowing for a comprehensive assessment of their performance in different environments. This chapter provides a summary of the concluding remarks drawn from this research, followed by an outline of potential future directions and plans for continued investigation.

6.1 Summary of the Research Work

(1) Multi-modal integration for 3D reconstruction of the planetary surfaces

Chapter 3 presents a multi-modal data integration approach for 3D reconstruction of planetary surfaces. This approach is divided into two primary stages. In the first stage, laser altimetry and photogrammetry data are integrated to generate a rigorous topographic product covering a large area. Initially, a rigorous polynomial camera model is established, which is then normalized using RPCs to generalize the subsequent process. A learning-based feature matching algorithm is incorporated into this pipeline, guided by nominal EO parameters to

retrieve more evenly-distributed tie-points, particularly in narrow textureless regions. With the improved and consistent EO parameters calculated from bundle adjustment, epipolar rectification is conducted. In the proposed pipeline, multiple CCD camera images are also fused into one image in this step to achieve better dense image matching results. Finally, space intersection is conducted to generate point clouds, which are later interpolated DEMs and used to rectify images into DOMs.

In the second stage, the generated low-resolution but robust DEM is refined through photoclinometry by integrating photometric data, thereby enhancing its accuracy and detail. To achieve this, four types of constraints are considered: (1) image similarity constraint, (2) low-resolution DEM constraint, (3) atmosphere constraint, and (4) albedo smoothness constraint. Each of these constraints examines different aspects of the parameters in the photometric equation, aiming to recover the real parameters of albedo, gradient, and atmosphere at the time when the image was captured. The optimization problem is formulated within the TensorFlow framework, leveraging the Adam optimizer and automatic parameter adaptation, to efficiently refine the DEM and produce a more accurate and pixel-wise representation of the terrain.

(2) Improved semantic segmentation facilitated by 3D information

Second, the integration of 3D information into semantic segmentation is thoroughly explored throughout the process. During the initial dataset construction stage, the 3D reconstruction results, which provide both the 3D position and pose of the camera and the terrain product, are utilized to facilitate semi-automatic dataset construction. This approach enables the creation of a robust dataset with minimal manual labeling effort, as a small set of labeled images can be augmented into a large number of images through 3D real-world roaming, thereby significantly reducing the need for extensive manual annotation. Furthermore, this approach also enables the generation of depth information, which in turn allows for the

utilization of depth-enhanced transformer-based neural networks to perform more detailed segmentation of the images.

Building upon this segmentation dataset, we propose a Siamese transformer-based semantic segmentation network that takes a pair of overlapping images as input. During training, the network not only evaluates the difference between the inferred segments and the corresponding labels, but also assesses the consistency among the overlapped images using cross-entropy loss based on tie-points. The network can thus be trained in a self-supervised manner after a few epochs, which benefits its robustness to incomplete or imperfect labels provided by humans.

(3) Optimal 3D reconstruction using the semantic cues assisted with frequency domain data

Based on consistent semantic segmentation, optimal 3D reconstruction can be achieved by enhancing both feature matching and dense image matching. With respect to feature matching, we propose to successively embed multiple levels of semantic cues into the state-of-the-art feature matching pipeline, SuperGlue, to enrich the original RGB information. This augmentation enhances the distinctiveness and robustness of the descriptor for each feature, leading to improved feature matching performance. Instead of leveraging the deep features extracted from a CNN, the Swin-transformer trained by the semantic segmentation task is leveraged to provide more global information. Meanwhile, the semantic labels are used to refine the attention weight in the aggregation stages to use distinct features to complement the descriptor, and the semantic probabilities are used to filter out false matches.

To enhance the performance of dense image matching, we integrate deep semantic features with frequency-domain distance calculations to construct the cost volume. Furthermore, we adaptively adjust the penalty term based on the boundaries of semantic segments to preserve

the discontinuities between landforms, thereby ensuring a more accurate and robust matching result.

The point clouds can be subsequently calculated from the disparity images, and then used to form a 3D mesh, which can be textured with the corresponding texture information.

6.2 Conclusions and Discussion

This study successfully accomplishes the stated objectives and leads to the following advances:

(1) A generic framework is developed for integrating laser altimetry and photogrammetry data to generate large-scale rigorous topographic products. This framework can process various types of satellite images, including along-track/cross-track stereo images with multiple CCD cameras or a single CCD camera. Experiments conducted with diverse satellite images validate the framework's capability and effectiveness.

(2) An atmosphere-aware framework for integrating photogrammetry and photoclinometry is developed. By comprehensively accounting for image, atmospheric, albedo, and gradient effects, this framework generates pixel-wise topographic products that preserve the precision of the original photogrammetric DEM.

(3) A semi-automatic dataset construction method is proposed, leveraging the 3D reconstruction results. By generating a semantic mesh model that is strictly aligned with the textured mesh model, semantic labels can be automatically obtained, along with depth and XYZ images, thereby enhancing the versatility of the dataset. Consequently, large datasets can be constructed with significantly reduced manual labor and more information. Depth-enhanced transformation can thus be performed, yielding more semantic labels with favorable accuracy.

(4) A Siamese transformer-based semantic segmentation network is designed, fully considering the consistency among photogrammetric methods, and can be trained in a self-supervised manner after only a few epochs. A consistency loss is proposed and established based on the extracted tie-points. Following training on the above dataset, the Siamese transformer demonstrates the ability to generate more semantic segments with transformation-invariant properties, indicating that pixel-wise stable features are obtained.

(5) Based on the semantic segmentation results, a semantic-aware feature matching algorithm is designed to match images with large variations in scale, coverage, perspective, and illumination. Experimental evaluation suggests that the semantic-aware algorithm can successfully establish abundant, evenly distributed tie-points, outperforming state-of-the-art algorithms.

(6) A semantic-aware dense image matching algorithm in the frequency domain is proposed, which integrates phase correlation with semantic similarity. Compared with off-the-shelf commercial software, the proposed semantic-aware dense image matching algorithm generates more favorable disparities, resulting in less noise and more accurate geometry that is closer to the real images.

Although the research objectives have been achieved, several limitations remain.

(1) There is a discrepancy in the experimental data used across chapters: Chapter 3 employs satellite imagery for evaluation, while Chapters 4 and 5 are based on rover imagery. This difference may initially appear inconsistent to the reader; however, from an algorithmic perspective, the methodologies are coherent and progressively integrated. Specifically, the EO-guided feature matching, object-based dense matching, and photoclinometric processing developed in Chapter 3 not only support the 3D

reconstruction framework in Chapter 4 but also provide the foundational geometric algorithms for the approach presented in Chapter 5.

Furthermore, despite the proposed semi-automatic method for training dataset construction, manual labeling of satellite imagery remains highly labor-intensive. To ensure feasibility and validate the core concepts efficiently, this study prioritized the use of rover images—which offer higher spatial resolution and simpler scene complexity—for initial development and testing. In future work, the framework will be extended to satellite-scale data, enabling broader planetary surface analysis.

(2) Second, while the three proposed approaches form a synergistic loop—where 3D reconstruction enhances semantic segmentation, and semantic outputs in turn refine the reconstruction—a natural question arises: when should this iterative process terminate? Based on comprehensive experimental evaluation, we find that the current three-stage pipeline is sufficient to achieve a stable and high-quality characterization of planetary surfaces in both geometric and semantic dimensions.

For semantic segmentation, the input of high-fidelity 3D reconstruction—featuring accurate geometry and pixel-level resolution—provides rich geometric priors that effectively support label augmentation in training dataset construction and enable geometry-aware supervision through tie-point constraints. These enhancements significantly improve segmentation accuracy without requiring further iterations.

Conversely, in the context of semantic-aware 3D reconstruction, the semantic cues derived from the current segmentation results exhibit sufficient discriminability, spatial consistency, and classification accuracy to guide the refinement of surface modeling—particularly in challenging regions such as shadows, low-texture areas, or complex terrains. Further feedback loops yield diminishing returns, with negligible improvements in topographic precision or structural coherence.

Therefore, we conclude that the three-stage integration strikes an optimal balance between performance gain and computational efficiency, achieving convergence in both geometric fidelity and semantic interpretability.

In summary, this study commences with the generation of high-quality 3D reconstruction products through the integration of multi-modal data. The semantic segmentation process leverages the results from the initial 3D reconstruction step to enrich the RGB information, and the improved results are subsequently fed back into the 3D reconstruction pipeline to produce an optimal 3D topographic product. Through the three main stages of this study, numerous detailed algorithms are proposed to address various specific issues, which can be applied to a wide range of applications extending beyond 3D reconstruction and semantic segmentation.

6.3 Future Works

Based on the accomplished experiments and proposed algorithms, there are many interesting topics worthy of further investigation and continued efforts.

(1) Semantic reconstruction from satellite images

Limited by manual labor, the experiments herein involving semantic parts were all conducted using frame rover images. However, the semantic segmentation of satellite images covering large regions can provide intuitive contextual information about the planetary surface, facilitating scientific analysis and the construction of large-scale landform databases. Besides the tremendous human labor required for labeling satellite images, identifying certain landforms is also quite challenging for non-experts, particularly when it comes to Mars, which has undergone various geological processes, such as atmospheric effects and volcanic eruptions, that have altered its surface. Even though some landforms (such as craters and volcanoes) may be relatively easy to identify, defining their boundaries remains a challenging task. This is

among the reasons why there are numerous landform detection studies, whereas segmentation-based approaches are relatively scarce. Hence, the downstream applications of semantic segmentation should be clearly defined before the organization and creation of a specific dataset. The SAM has been widely used to facilitate many segmentation tasks in the aspects of mask refinement, providing weak supervision and active learning. It can alleviate the workload of generating a large dataset, thereby facilitating the segmentation tasks of the planetary surface. Despite the advancements made, the boundary issue persists, thereby hindering the effective utilization of semantic cues to guide other tasks. Given the proper semantic segmentation results, the proposed semantic-aware image matching can be conducted on satellite images to achieve optimal 3D reconstruction results.

(2) Further incorporation of spectral images

The current segmentation method utilizes depth information to enrich the RGB information, whereas abundant spectral images are not used. Spectral images offer complementary information to traditional RGB images, facilitating a more comprehensive understanding of the scene. By capturing the reflectance properties of materials, spectral images enable the segmentation of objects based on their material composition, thereby providing a more in-depth representation of the environment. Furthermore, spectral images can detect subtle changes in material properties, leading to improved boundary detection and reduced noise in the segmentation output. The richer feature set provided by spectral images allows for more robust and informative feature extraction, which can enhance the overall performance of segmentation models. Additionally, spectral images are less susceptible to variations in lighting, shading, and atmospheric conditions, making them a valuable asset for segmentation pipelines operating in diverse environments.

(3) Learning-based dense image matching module

In the proposed 3D reconstruction or optimal 3D reconstruction pipeline, the feature matching component is replaced by a learning-based algorithm, whereas the dense image matching module still adheres to the conventional pipeline, and continues to be perturbed by inherent oversmoothness issues. One potential solution is to leverage the deep learning variants of traditional algorithms, such as MVSNet (Yao et al., 2018), which can potentially overcome some of the limitations of traditional methods. However, the dataset problem remains a significant obstacle. An alternative approach is to utilize emerging techniques like Gaussian splatting, which does not require supervised training stages, thereby bypassing the challenging task of constructing a large-scale training dataset. Although the current Gaussian Splatting algorithm relies on relative coordinates, our approach has already obtained accurate exterior orientation parameters, and the intersection process for generating dense point clouds is rigorously implemented. The application of Gaussian splatting can be reformulated as generating a more detailed and accurate depth map, which is subsequently fused with the precisely calculated disparity images. This fusion enables the guarantee of geometric accuracy while also enhancing the geometric appearance. As most learning-based algorithms are computed on GPUs, the large size of satellite images poses a significant challenge to the direct application of the current Gaussian Splatting algorithm. Furthermore, satellite images typically provide limited observations of a specific region, which may restrict the realization of the full potential of the Gaussian Splatting algorithm.

(4) Interaction between the rover and satellite images

While rover images offer more detailed and accurate information about the planetary surface, satellite images provide large-scale contextual information. A natural approach is to

integrate these two data sources to achieve a synergistic effect: enhancing the absolute geometric accuracy of topographic products generated from rover images, while also augmenting the detail and resolution of satellite images. By combining the strengths of both data sources, we can create more comprehensive and accurate representations of the planetary surface. The primary obstacle lies in the significant disparity between the appearance of the rover and satellite images, making it challenging to establish a connection between the two, particularly in textureless regions where visual cues are scarce. One potential solution is to incorporate images captured from the descending camera, but this approach is limited by the availability of images, which may only be feasible for research conducted in the immediate landing site region, where the lack of texture is even more pronounced. Another possible solution is to utilize images from unmanned aerial vehicles, but this option is currently only available for the Perseverance rover, which raises concerns about the algorithm's generalizability and applicability to other Mars exploration missions.

References

Abdulnabi, A.H., Wang, G., Lu, J., Jia, K., 2015. Multi-task CNN model for attribute prediction. IEEE Transactions on Multimedia 17, pp. 1949-1959.

Abdulwahab, S., Rashwan, H.A., García, M.Á., Jabreel, M., Chambon, S., Puig, D., 2020. Adversarial Learning for Depth and Viewpoint Estimation From a Single Image. IEEE Transactions on Circuits and Systems for Video Technology 30, pp. 2947-2958.

Acton, C., 1998. An overview of SPICE. Jet Propulsion Laboratory: Oak Grove, KY, USA.

Acton, C., Bachman, N., Semenov, B., Wright, E., 2016. Spice Tools Supporting Planetary Remote Sensing. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 41, pp. 357-359.

Agarwal, S., Mierle, K., Team, T., 2012. Ceres solver. < https://ceres-solver.org/>

Agarwal, S., Snavely, N., Simon, I., Seitz, S.M., Szeliski, R., 2009. Building Rome in a Day. International Conference on Computer Vision (ICCV), pp. 72-79.

Albee, A., 2002. The Mars Global Surveyor Mission: Description, Status, and Significant Results. Highlights of Astronomy 12, pp. 631-635.

Alexandrov, O., Beyer, R.A., 2018. Multiview shape-from-shading for planetary images. Earth and Space Science 5, pp. 652-666.

Anderson, J., Sides, S., Soltesz, D., Sucharski, T., Becker, K., 2004. Modernization of the integrated software for imagers and spectrometers, Lunar and planetary science conference, pp. 2039.

Anthony, W.Y., Krainak, M.A., Harding, D.J., Abshire, J.B., Sun, X., Cavanaugh, J., Valett, S., Ramos-Izquierdo, L., 2012. Multi-beam laser altimeter system simulator for the Lidar

Surface Topography (LIST) mission, CLEO: Applications and Technology. Optica Publishing Group, pp. ATu2G-6.

Barker, M.K., Mazarico, E., Neumann, G.A., Smith, D.E., Zuber, M.T., Head, J.W., 2021. Improved LOLA elevation maps for south pole landing sites: Error estimates and their impact on illumination conditions. Planetary and Space Science 203, 105119.

Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P., 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5855-5864.

Beyer, R.A., Alexandrov, O., McMichael, S., 2018. The Ames Stereo Pipeline: NASA's open-source software for deriving and processing terrain data. Earth and Space Science 5, pp. 537-548.

Bickel, V.T., Conway, S.J., Tesson, P.A., Manconi, A., Loew, S., Mall, U., 2020. Deep Learning-Driven Detection and Mapping of Rockfalls on Mars. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13, pp. 2831-2841.

Birchfield, S., Tomasi, C., 1998. A pixel dissimilarity measure that is insensitive to image sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, pp. 401-406.

Bostelmann, J., Heipke, C., 2011. Modeling Spacecraft Oscillations in Hrsc Images of Mars Express. ISPRS Hannover Workshop 2011: High-Resolution Earth Imaging for Geospatial Information 39-4, pp. 51-56.

Bostelmann, J., Heipke, C., 2014. Analyzing a block of HRSC image strips for a simultaneous bundle adjustment. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences; II-4 2, pp. 15-20.

Bottou, L., Cortes, C., Denker, J.S., Drucker, H., Guyon, I., Jackel, L.D., LeCun, Y., Muller, U.A., Sackinger, E., Simard, P., 1994. Comparison of classifier methods: a case study in

handwritten digit recognition, Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5). IEEE, pp. 77-82.

Bridle, J.S., 1990. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. Neurocomputing: Algorithms, architectures and applications, pp. 227-236.

Burns, D., Osfield, R., 2004. Tutorial: Open scene graph A: introduction tutorial: Open scene graph B: examples and applications, IEEE Virtual Reality 2004, pp. 265-265.

Chandler, J., Cooper, M., 1989. The extraction of positional data from historical photographs and their application to geomorphology. The Photogrammetric Record 13, pp. 69-78.

Chandraker, M., Agarwal, S., Kriegman, D., 2007. Shadowcuts: Photometric stereo with shadows. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-8.

Chandrasekhar, S., 1960. Radiative transfer.

Chen, H., Hu, X., Willner, K., Ye, Z., Damme, F., Gläser, P., Zheng, Y., Tong, X., Hußmann, H., Oberst, J., 2024a. Neural implicit shape modeling for small planetary bodies from multi-view images using a mask-based classification sampling strategy. ISPRS Journal of Photogrammetry and Remote Sensing 212, pp. 122-145.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 834-848.

Chen, P.Y., Liu, A.H., Liu, Y.C., Wang, Y.C.F., 2019. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation.

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2624-2632.

Chen, S., Wu, B., Li, H., Li, Z., Liu, Y., 2024b. Asteroid-NeRF: A deep-learning method for 3D surface reconstruction of asteroids. A&A 687, A278.

Chen, Z., Sun, X., Wang, L., Yu, Y., Huang, C., 2015. A deep visual correspondence embedding model for stereo matching costs. Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 972-980.

Chen, Z., Wu, B., Krasilnikov, S., Xun, W., Ma, Y., Liu, S., Li, Z., 2024c. A Global Database of Pitted Cones on Mars for Research on Martian Volcanism. Scientific Data 11, 942.

Chen, Z., Wu, B., Liu, W.C., 2021. Mars3DNet: CNN-Based High-Resolution 3D Reconstruction of the Martian Surface from Single Images. Remote Sensing 13, 839.

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1290-1299.

Cong, W., Liang, H., Fan, Z., Wang, P., Jiang, Y., Xu, D., Oztireli, A.C., Wang, Z., 2024. NeRF as Pretraining at Scale: Generalizable 3D-Aware Semantic Representation Learning from View Prediction, Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 2872-2882.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213-3223.

Couprie, C., Farabet, C., Najman, L., LeCun, Y., 2013. Indoor semantic segmentation using depth information. arXiv preprint arXiv:1301.3572.

Cuturi, M., 2013. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems 26.

Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5828-5839.

Daly, M., Barnouin, O., Dickinson, C., Seabrook, J., Johnson, C., Cunningham, G., Haltigin, T., Gaudreau, D., Brunet, C., Aslam, I., 2017. The OSIRIS-REx laser altimeter (OLA) investigation and instrument. Space Science Review 212, pp. 899-924.

Degnan, J., Wells, D., Machan, R., Leventhal, E., 2007. Second generation airborne 3D imaging lidars based on photon counting. Advanced Photon Counting Techniques II. SPIE, pp. 117-123.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248-255.

Deng, K., Liu, A., Zhu, J.-Y., Ramanan, D., 2022. Depth-supervised nerf: Fewer views and faster training for free. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12882-12891.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 224-236.

Dosovitskiy, A., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Eigen, D., Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2650-2658.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems 27.

Everingham, M., Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision 88, pp. 303–338.

Everingham, M., Zisserman, A., Williams, C.K., Van Gool, L., Allan, M., Bishop, C.M., Chapelle, O., Dalal, N., Deselaers, T., Dorkó, G., 2006. The 2005 Pascal visual object classes challenge, Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, pp. 117-176.

Fei-Fei, Li., Fergus, R., Perona, P., 2006. One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, pp. 594-611.

Flynn, J., Neulander, I., Philbin, J., Snavely, N., 2016. Deepstereo: Learning to predict new views from the world's imagery, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5515-5524.

Frankot, R.T., Chellappa, R., 1988. A method for enforcing integrability in shape from shading algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 10, pp. 439-451.

Furukawa, Y., Ponce, J., 2009. Accurate, dense, and robust multiview stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, pp. 1362-1376.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Garcia-Rodriguez, J., 2017. A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857.

Garg, R., Bg, V.K., Carneiro, G., Reid, I., 2016. Unsupervised cnn for single view depth estimation: Geometry to the rescue. Proceedings of the European Conference on Computer Vision (ECCV), Part VIII 14, pp. pp. 740-756.

Gaskell, R., Barnouin-Jha, O., Scheeres, D.J., Konopliv, A., Mukai, T., Abe, S., Saito, J., Ishiguro, M., Kubota, T., Hashimoto, T., 2008. Characterizing and navigating small bodies with imaging data. Meteoritics & Planetary Science 43, pp. 1049-1061.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The kitti dataset. The International Journal of Robotics Research 32, pp. 1231-1237.

Geng, X., Xu, Q., Xing, S., Lan, C.Z., 2020. A Generic Pushbroom Sensor Model for Planetary Photogrammetry. Earth and Space Science 7.

Gold, R.E., Solomon, S.C., McNutt Jr, R.L., Santo, A.G., Abshire, J.B., Acuña, M.H., Afzal, R.S., Anderson, B.J., Andrews, G.B., Bedini, P.D., 2001. The MESSENGER mission to Mercury: scientific payload. Planet Space Sci 49, pp. 1467-1479.

Golombek, M., Warner, N.H., Grant, J.A., Hauber, E., Ansan, V., Weitz, C.M., Williams, N., Charalambous, C., Wilson, S.A., DeMott, A., 2020. Geology of the InSight landing site on Mars. Nature Communications 11, 1014.

Gridin, I., 2022. Introduction to neural network intelligence, Automated Deep Learning Using Neural Network Intelligence: Develop and Design PyTorch and TensorFlow Models Using Python. Springer, pp. 1-30.

Grodecki, J., 2001. IKONOS stereo feature extraction-RPC approach, ASPRS annual conference St. Louis.

Grodecki, J., Dial, G., 2003. Block adjustment of high-resolution satellite images described by rational polynomials. Photogrammetric Engineering and Remote Sensing 69, pp. 59-68.

Grumpe, A., Wöhler, C., 2014. Recovery of elevation from estimated gradient fields constrained by digital elevation maps of lower lateral resolution. . ISPRS Journal of Photogrammetry and Remote Sensing 94, pp. 37-54.

Guizilini, V., Vasiljevic, I., Chen, D., Ambruș, R., Gaidon, A., 2023. Towards zero-shot scale-aware monocular depth estimation, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 9233-9243.

Gwinner, K., Jaumann, R., Hauber, E., Hoffmann, H., Heipke, C., Oberst, J., Neukum, G., Ansan, V., Bostelmann, J., Dumke, A., 2016. 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 126, pp. 93-138.

Gwinner, K., Scholten, F., Preusker, F., Elgner, S., Roatsch, T., Spiegel, M., Schmidt, R., Oberst, J., Jaumann, R., Heipke, C., 2010. Topography of Mars from global mapping by HRSC high-resolution digital terrain models and orthoimages: characteristics and performance. Earth and Planetary Science Letters 294.

Gwinner, K., Scholten, F., Spiegel, M., Schmidt, R., Giese, B., Oberst, J., Heipke, C., Jaumann, R., Neukum, G., 2009. Derivation and validation of high-resolution digital terrain models from Mars Express HRSC data. Photogrammetric Engineering & Remote Sensing 75, pp. 1127-1142.

Hadfield, R.H., 2009. Single-photon detectors for optical quantum information applications. Nature Photonics 3, pp. 696-705.

Hane, C., Zach, C., Cohen, A., Angst, R., Pollefeys, M., 2013. Joint 3D scene reconstruction and class segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 97-104.

Hapke, B., 1965. Effects of a simulated solar wind on the photometric properties of rocks and powders. Annals of the New York Academy of Sciences 123, pp. 711-721.

Hapke, B., 1981. Bidirectional reflectance spectroscopy: 1. Theory. Journal of Geophysical Research: Solid Earth 86, pp. 3039-3054.

Hapke, B., van Hoen, H., 1963. Photometric studies of complex surfaces, with applications to the Moon. Journal of Geophysical Research 68, pp. 4545-4570.

Harris, C., Stephens, M., 1988. A combined corner and edge detector, Alvey vision conference. Citeseer, pp. 147-152.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778.

Heipke, C., Piechullek, C., Ebner, H., 2001. Simulation studies and practical tests using multi-image shape from shading. ISPRS Journal of Photogrammetry and Remote Sensing 56, pp. 139-148.

Heymann, S., Muller, K., Smolic, A., Frolich, B., Wiegand, T., 2007. SIFT Implementation and Optimization for General-Purpose GPU. WSCG 2007 Full Papers Proceedings I and II, pp. 317-323.

Hirschmuller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 807-814.

Hoekzema, N., Garcia-Comas, M., Stenzel, O., Grieger, B., Markiewicz, W., Gwinner, K., Keller, H., 2010. Optical depth and its scale-height in Valles Marineris from HRSC stereo images. Earth and Planetary Science Letters 294, pp. 534-540.

Hoekzema, N.M., Markiewicz, W.J., Inada, A., Hviid, S.H., Keller, H.U., Gwinner, K., Hoffmann, H., Meima, J.A., Neukum, G., HRSC, Science, M., 2004. Atmospheric optical depths from HRSC stereo images of Gusev crater and elsewhere on Mars. AAS/Division for Planetary Sciences Meeting Abstracts 36, pp. 37-08.

Horn, B.K., 1990. Height and gradient from shading. International journal of computer vision 5, pp. 37-75.

Horn, B.K., Sjoberg, R.W., 1979. Calculating the reflectance map. Applied optics 18, 1770-1779.

Hu, H., Chen, C.T., Wu, B., Yang, X.X., Zhu, Q., Ding, Y.L., 2016. Texture-Aware Dense Image Matching Using Ternary Census Transform. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Science Volume 3, pp. 59-66.

Hu, H., Wu, B., 2018. Block adjustment and coupled epipolar rectification of LROC NAC images for precision lunar topographic mapping. Planet Space Sci 160, pp. 26-38.

Hu, H., Wu, B., 2019. Planetary3D: A photogrammetric tool for 3D topographic mapping of planetary bodies. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Science Volume IV-2/W5,2019.

Huynh, L., Nguyen-Ha, P., Matas, J., Rahtu, E., Heikkilä, J., 2020. Guiding monocular depth estimation using depth-attention. Proceedings of the European Conference on Computer Vision (ECCV), Part XXVI 16, pp. 581-597.

Jiang, C., Douté, S., Luo, B., Zhang, L., 2017. Fusion of photogrammetric and photoclinometric information for high-resolution DEMs from Mars in-orbit imagery. ISPRS Journal of Photogrammetry and Remote Sensing 130, pp. 418-430.

Jiang, Y., Hedman, P., Mildenhall, B., Xu, D., Barron, J.T., Wang, Z., Xue, T., 2023. AligNeRF: High-fidelity neural radiance fields via alignment-aware training. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 46-55.

Jonathan, E., 1986. NASA Sets Sensors for 1990 Return to Mars. Science News 129, p. 330.

Jung, J.-W., So, B.-C., Kang, J.-G., Lim, D.-W., Son, Y., 2019. Expanded Douglas–Peucker polygonal approximation and opposite angle-based exact cell decomposition for path planning with curvilinear obstacles. Applied Sciences 9, 638.

Kaula, W., Schubert, G., Lingenfelter, R., Sjogren, W., Wollenhaupt, W., 1973. Lunar topography from Apollo 15 and 16 laser altimetry, Proceedings of the Lunar Science Conference, vol. 4, p. 2811.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A., 2017. End-to-end learning of geometry and context for deep stereo regression. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 66-75.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graph. 42, 139:131-139:114.

King, M.D., Harshvardhan, 1986. Comparative Accuracy of Selected Multiple Scattering Approximations. Journal of Atmospheric Sciences 43, pp. 784-801.

Kingma, D.P., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., 2023. Segment anything. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4015-4026.

Kirk, R.L., 1987. Algorithm for Two-Dimensional Photoclinometry. California Institute of Technology.

Krähenbühl, P., Koltun, V., 2011. Efficient inference in fully connected CRFs with Gaussian edge potentials. Advances in neural information processing systems 24.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25.

Kundu, A., Li, Y., Dellaert, F., Li, F., Rehg, J.M., 2014. Joint semantic segmentation and 3D reconstruction from monocular video. Proceedings of the European Conference on Computer Vision (ECCV), Part VI 13, pp. 703-718.

LaChapelle, E., 1962. Assessing glacier mass budgets by reconnaissance aerial photography. Journal of Glaciology 4, pp. 290-297.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks. Fourth international conference on 3D vision (3DV). IEEE, pp. 239-248.

Lambert, J.H., 1760. Photometria sive de mensura et gradibus luminis, colorum et umbrae.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L., 1989. Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems 2.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D, 1989. Backpropagation applied to handwritten zip code recognition. Neural computation 1, pp. 541-551.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, pp. 2278-2324.

Ledebuhr, A.G., Kordas, J.F., Lewis, I.T., Richardson, M.J., Cameron, G.R., White III, W.T., Dobie, D.W., Strubhar, W.D., Tassinari, T.F., Sawyer, D.J., 1995. HiRes camera and lidar ranging system for the Clementine mission, Applied Laser Radar Technology II. SPIE, pp. 62-81.

Lenc, K., Vedaldi, A., 2016. Learning covariant feature detectors, Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. Springer, pp. 100-117.

Li, R.X., Hwangbo, J., Chen, Y.H., Di, K.C., 2011. Rigorous Photogrammetric Processing of HiRISE Stereo Imagery for Mars Topographic Mapping. IEEE Transactions on Geoscience and Remote Sensing 49, pp. 2558-2572.

Li, Y., Xiao, Z., Ma, C., Zeng, L., Zhang, W., Peng, M., Li, A., 2023a. Extraction and Analysis of Three-Dimensional Morphological Features of Centimeter-Scale Rocks in Zhurong Landing Region. Journal of Geophysical Research: Planets 128.

Li, Z., Snavely, N., 2018. Megadepth: Learning single-view depth prediction from internet photos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2041-2050.

Li, Z., Wu, B., Chen, Z., Ma, Y., 2023b. Transformer-Based Method for Semantic Segmentation and Reconstruction of the Martian Surface. Geospatial Week 2023, Vol. 48-1, pp. 1643-1649.

Li, Z., Wu, B., Li, Y., Chen, Z., 2023c. Fusion of aerial, MMS and backpack images and point clouds for optimized 3D mapping in urban areas. ISPRS Journal of Photogrammetry and Remote Sensing 202, pp. 463-478.

Li, Z., Wu, B., Liu, W.C., Chen, Z., 2021. Integrated Photogrammetric and Photoclinometric Processing of Multiple HRSC Images for Pixelwise 3-D Mapping on Mars. IEEE Transactions on Geoscience and Remote Sensing, pp. 1-13.

Light, D.L., 1972. Altimeter observations as orbital constraints. Photogrammetric Engineering and Remote Sensing 38, pp. 339-346.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2117-2125.

Liu, F., Shen, C., Lin, G., 2015. Deep convolutional neural fields for depth estimation from a single image. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5162-5170.

Liu, F., Zhang, C., Zheng, Y., Duan, Y., 2023. Semantic ray: Learning a generalizable semantic field with cross-reprojection attention. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17386-17396.

Liu, J., Li, C., Zhang, R., Rao, W., Cui, X., Geng, Y., Jia, Y., Huang, H., Ren, X., Yan, W., 2022. Geomorphic contexts and science focus of the Zhurong landing site on Mars. Nature Astronomy 6, pp. 65-71.

Liu, J., Li, H., Yang, Z., Wu, K., Liu, Y., Liu, R.W., 2019. Adaptive Douglas-Peucker algorithm with automatic thresholding for AIS-based vessel trajectory compression. IEEE Access 7, pp. 150677-150692.

Liu, W.C., Wu, B., 2020. An integrated photogrammetric and photoclinometric approach for illumination-invariant pixel-resolution 3D mapping of the lunar surface. ISPRS Journal of Photogrammetry and Remote Sensing 159, pp. 153-168.

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

Liu, W.C., Wu, B., Wöhler, C., 2018. Effects of illumination differences on photometric stereo shape-and-albedo-from-shading for precision lunar surface reconstruction. ISPRS Journal of Photogrammetry and Remote Sensing 136, pp. 58-72.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 10012-10022.

Lohse, V., Heipke, C., Kirk, R.L., 2006. Derivation of planetary topography using multi-image shape-from-shading. Planet Space Science 54, pp. 661-674.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Comput Vision 60, pp. 91-110.

Ma, C., Li, Y., Lv, J., Xiao, Z., Zhang, W., Mo, L., 2024. Automated Rock Detection From Mars Rover Image via Y-Shaped Dual-Task Network With Depth-Aware Spatial Attention Mechanism. IEEE Transactions on Geoscience and Remote Sensing 62, pp. 1-18.

Ma, J., Jiang, X., Fan, A., Jiang, J., Yan, J., 2021. Image Matching from Handcrafted to Deep Features: A Survey. International Journal of Computer Vision 129, pp. 23-79.

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5, pp. 115-133.

McEwen, A.S., 1986. Exogenic and endogenic albedo and color patterns on Europa. Journal of Geophysical Research: Solid Earth 91, pp. 8077-8097.

McEwen, A.S., 1991. Photometric functions for photoclinometry and other applications. Icarus 92, 298-311.

McEwen, A.S., 2021. Topography around the Zhurong Rover. < https://www.uahirise.org/ESP_069876_2055>

McEwen, A.S., Banks, M.E., Baugh, N., Becker, K., Boyd, A., Bergstrom, J.W., Beyer, R.A., Bortolini, E., Bridges, N.T., Byrne, S., 2010. The high resolution imaging science experiment (HiRISE) during MRO's primary science phase (PSP). Icarus 205, pp. 2-37.

McEwen, A.S., Eliason, E.M., Bergstrom, J.W., Bridges, N.T., Hansen, C.J., Delamere, W.A., Grant, J.A., Gulick, V.C., Herkenhoff, K.E., Keszthelyi, L., 2007. Mars reconnaissance orbiter's high resolution imaging science experiment (HiRISE). Journal of Geophysical Research: Planets 112.

Meng, Q., Wang, D., Wang, X., Li, W., Yang, X., Yan, D., Li, Y., Cao, Z., Ji, Q., Sun, T., 2021. High resolution imaging camera (HiRIC) on China's first Mars exploration Tianwen-1 mission. Space Science Review 217, pp. 1-29.

Meydenbauer, A., 1867. Die photometrographie. Wochenblatt des Architektenvereins zu Berlin 1, pp. 125-126.

Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM 65, pp. 99-106.

Mills, M.M., McEwen, A.S., Okubo, C.H., 2021. A preliminary regional geomorphologic map in Utopia Planitia of the Tianwen-1 Zhurong landing region. Geophysical Research Letters 48, e2021GL094629.

Montabone, L., Forget, F., Millour, E., Wilson, R.J., Lewis, S.R., Cantor, B., Kass, D., Kleinböhl, A., Lemmon, M.T., Smith, M.D., Wolff, M.J., 2015. Eight-year climatology of dust optical depth on Mars. Icarus 251, pp. 65-95.

Moravec, H., 1977. A non-synchronous orbital skyhook. Journal of the Astronautical Sciences 25, pp. 307-322.

Muja, M., Lowe, D.G., 2009. Fast approximate nearest neighbors with automatic algorithm configuration. VISAPP (1) 2, 2.

Mulverhill, C., Coops, N.C., Hermosilla, T., White, J.C., Wulder, M.A., 2022. Evaluating ICESat-2 for monitoring, modeling, and update of large area forest canopy height products. Remote Sensing of Environment 271, 112919.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted Boltzmann machines, Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807-814.

Neukum, G., Jaumann, R., 2004. HRSC: The high resolution stereo camera of Mars Express. In: Mars Express: the scientific payload. Ed. by Andrew Wilson, scientific coordination: Agustin Chicarro. ESA SP-1240, Noordwijk, Netherlands: ESA Publications Division, ISBN 92-9092-556-6, 2004, p. 17-35 1240, pp. 17-35.

Neumann, T.A., Martino, A.J., Markus, T., Bae, S., Bock, M.R., Brenner, A.C., Brunt, K.M., Cavanaugh, J., Fernandes, S.T., Hancock, D.W., 2019. The Ice, Cloud, and Land Elevation Satellite–2 Mission: A global geolocated photon product derived from the advanced topographic laser altimeter system. Remote sensing of environment 233, 111325.

Ni, J., Khan, Z., Wang, S., Wang, K., Haider, S.K., 2016. Automatic detection and counting of circular shaped overlapped objects using circular hough transform and contour detection, 2016 12th World Congress on Intelligent Control and Automation (WCICA). IEEE, pp. 2902-2906.

Oberst, J., Schwarz, G., Behnke, T., Hoffmann, H., Matz, K.D., Flohrer, J., Hirsch, H., Roatsch, T., Scholten, F., Hauber, E., Brinkmann, B., Jaumann, R., Williams, D., Kirk, R.L., Duxbury, T., Leu, C., Neukum, G., 2008. The imaging performance of the SRC on Mars Express. Planetary Space Science 56, 473-491.

Ohlhof, T., Montenbruck, O., Gill, E., 1994. New approach for combined bundle block adjustment and orbit determination based on Mars-94 three-line scanner imagery and radio-tracking data, ISPRS Commission III Symposium: Spatial Information from Digital Photogrammetry and Computer Vision. SPIE, pp. 630-639.

Ono, Y., Trulls, E., Fua, P., Yi, K.M., 2018. LF-Net: Learning local features from images. Advances in neural information processing systems 31.

Pentland, A.P., 1984. Local shading analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 170-187.

Peyré, G., Cuturi, M., 2019. Computational optimal transport: With applications to data science. Foundations and Trends in Machine Learning 11, 355-607.

Ping, J., Huang, Q., Yan, J., Cao, J., Tang, G., Shu, R., 2009. Lunar topographic model CLTM-s01 from Chang'E-1 laser altimeter. Science in China Series G: Physics, Mechanics and Astronomy 52, pp. 1105-1114.

Reddy, B.S., Chatterji, B.N., 1996. An FFT-based technique for translation, rotation, and scale-invariant image registration. IEEE Transactions on Image Processing 5, pp. 1266-1271.

Ren, X., Bo, L., Fox, D., 2012. Rgb-(d) scene labeling: Features and algorithms, 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2759-2766.

Robbins, S.J., Antonenko, I., Kirchoff, M.R., Chapman, C.R., Fassett, C.I., Herrick, R.R., Singer, K., Zanetti, M., Lehan, C., Huang, D., 2014. The variability of crater identification among expert and community crater analysts. Icarus 234, pp. 109-131.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer, pp. 234-241.

Rosten, E., Porter, R., Drummond, T., 2008. Faster and better: A machine learning approach to corner detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 105-119.

Rothermel, M., Wenzel, K., Fritsch, D., Haala, N., 2012. SURE: Photogrammetric surface reconstruction from imagery, Proceedings LC3D Workshop, Berlin.

Rothrock, B., Kennedy, R., Cunningham, C., Papon, J., Heverly, M., Ono, M., 2016. SPOC: Deep Learning-based Terrain Classification for Mars Rover Missions, AIAA SPACE 2016.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF, 2011 International conference on computer vision. Ieee, pp. 2564-2571.

Rudin, L.I., Osher, S., Fatemi, E., 1992. Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena 60, pp. 259-268.

Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T., 2008. LabelMe: a database and web-based tool for image annotation. Int J Comput Vision 77, pp. 157-173.

Sandhu, R., Dambreville, S., Yezzi, A., Tannenbaum, A., 2011. A Nonrigid Kernel-Based Framework for 2D-3D Pose Estimation and 2D Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, pp. 1098-1115.

Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. SuperGlue: Learning Feature Matching with Graph Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4937-4946.

Saxena, A., Sun, M., Ng, A.Y., 2008. Make3d: Learning 3d scene structure from a single still image. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, pp. 824-840.

Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision 47, pp. 7-42.

Schutz, B.E., Zwally, H.J., Shuman, C.A., Hancock, D., DiMarzio, J.P., 2005. Overview of the ICESat mission. Geophysical research letters 32.

Scherer, D., Schwatke, C., Dettmering, D., Seitz, F., 2022. ICESat-2 Based River Surface Slope and Its Impact on Water Level Time Series From Satellite Altimetry. Water Resources Research 58.

Shotton, J., Winn, J., Rother, C., Criminisi, A., 2006. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. Proceedings of the European conference on computer vision (ECCV), Part I 9. Springer, pp. 1-15.

Shvets, M., Zhao, D., Niethammer, M., Sengupta, R., Berg, A.C., 2024. Joint depth prediction and semantic segmentation with multi-view sam, Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pp. 1328-1338.

Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from rgbd images. Proceedings of the European conference on computer vision (ECCV), Part V 12. Springer, pp. 746-760.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sjogren, W., Wollenhaupt, W., 1973. Lunar shape via the Apollo laser altimeter. Science 179, pp. 275-278.

Smereka, M., Dulęba, I., 2008. Circular object detection using a modified Hough transform. International Journal of Applied Mathematics and Computer Science 18, pp. 85-91.

Smith, D., Zuber, M., Frey, H., Garvin, J., Head, J., Muhleman, D., Pettengill, G., Phillips, R., Solomon, S., Zwally, H., 1998. Topography of the northern hemisphere of Mars from the Mars Orbiter Laser Altimeter. Science 279, pp. 1686-1692.

Smith, D.E., Zuber, M.T., Frey, H.V., Garvin, J.B., Head, J.W., Muhleman, D.O., Pettengill, G.H., Phillips, R.J., Solomon, S.C., Zwally, H.J., 2001. Mars Orbiter Laser Altimeter: Experiment summary after the first year of global mapping of Mars. Journal of Geophysical Research: Planets 106, pp. 23689-23722.

Smith, D.E., Zuber, M.T., Neumann, G.A., Lemoine, F.G., 1997. Topography of the Moon from the Clementine lidar. Journal of Geophysical Research: Planets 102, pp. 1591-1611.

Smith, D.E., Zuber, M.T., Solomon, S.C., Phillips, R.J., Head, J.W., Garvin, J.B., Banerdt, W.B., Muhleman, D.O., Pettengill, G.H., Neumann, G.A., Lemoine, F.G., Abshire, J.B., Aharonson, O., Brown, C.D., Hauck, S.A., Ivanov, A.B., McGovern, P.J., Zwally, H.J., Duxbury, T.C., 1999. The global topography of Mars and implications for surface evolution. Science 284, pp. 1495-1503.

Snavely, N., Seitz, S.M., Szeliski, R., 2008. Modeling the world from internet photo collections. International Journal of Computer Vision 80, pp. 189-210.

Soderblom, J.M., Bell III, J.F., Hubbard, M.Y., Wolff, M.J., 2006. Martian phase function: Modeling the visible to near-infrared surface photometric function using HST-WFPC2 data. Icarus 184, pp. 401-423.

Song, C., Chen, Q., Li, F.W.B., Jiang, Z., Zheng, D., Shen, Y., Yang, B., 2024. Multi-feature fusion enhanced monocular depth estimation with boundary awareness. The Visual Computer 40, pp. 4955-4967.

Swan, R.M., Atha, D., Leopold, H.A., Gildner, M., Oij, S., Chiu, C., Ono, M., 2021. AI4MARS: A Dataset for Terrain-Aware Autonomous Driving on Mars. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1982-1991.

Tao, C.V., Hu, Y., 2002. 3D Reconstruction methods based on the rational function model. Photogramm Eng Rem S 68, pp. 705-714.

Tong, X., Ye, Z., Xu, Y., Gao, S., Xie, H., Du, Q., Liu, S., Xu, X., Liu, S., Luan, K., Stilla, U., 2019. Image Registration With Fourier-Based Image Correlation: A Comprehensive Review of Developments and Applications. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12, pp. 4062-4081.

Tong, X.H., Ye, Z., Xu, Y.S., Liu, S.J., Li, L.Y., Xie, H., Li, T.P., 2015. A Novel Subpixel Phase Correlation Method Using Singular Value Decomposition and Unified Random Sample Consensus. IEEE Transactions on Geoscience and Remote Sensing 53, pp. 4143-4156.

Torlegård, K., 1992. Sensors for photogrammetric mapping: review and prospects. ISPRS Journal of Photogrammetry and Remote Sensing 47, pp. 241-262.

Van Diggelen, J., 1951. A photometric investigation of the slopes and the heights of the ranges of hills in the Maria of the moon. Bulletin of the Astronomical Institutes of the Netherlands, Vol. 11, p. 283.

Vaswani, A., 2017. Attention is all you need. Advances in Neural Information Processing Systems (NIPS).

Vemulapalli, R., Tuzel, O., Liu, M.-Y., Chellapa, R., 2016. Gaussian conditional random field network for semantic segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3224-3233.

Verdie, Y., Yi, K., Fua, P., Lepetit, V., 2015. Tilde: A temporally invariant learned detector, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5279-5288.

Vu, H.-H., Labatut, P., Pons, J.-P., Keriven, R., 2011. High accuracy and visibility-consistent dense multiview stereo. IEEE Transactions on Pattern Analysis and Machine Ontelligence 34, 889-901.

Walter, S., Michael, G., Kneissl, T., 2015. Photometric Lambert correction for global mosaicking of HRSC image data, 46th Annual Lunar and Planetary Science Conference, p. 1434.

Wan, W., Wang, C., Li, C., Wei, Y., 2020. China's first mission to Mars. Nature Astronomy 4, p. 721-721.

Wan, X., Liu, J.G., Li, S., Yan, H., 2019. Phase Correlation Decomposition: The Impact of Illumination Variation for Robust Subpixel Remotely Sensed Image Matching. IEEE Transactions on Geoscience and Remote Sensing 57, pp. 6710-6725.

Wang, C., Reza, M.A., Vats, V., Ju, Y., Thakurdesai, N., Wang, Y., Crandall, D.J., Jung, S.-h., Seo, J., 2024. Deep learning-based 3D reconstruction from multiple images: A survey. Neurocomputing 597, 128018.

Wang, M., Hu, F., Li, J., 2011. Epipolar resampling of linear pushbroom satellite imagery by a new epipolarity model. ISPRS Journal of Photogrammetry and Remote Sensing 66, pp. 347-355.

Wang, P., Chen, X., Chen, T., Venugopalan, S., Wang, Z., 2022. Is Attention All That NeRF Needs? arXiv preprint arXiv:2207.13298.

Wang, X., Gong, P., Zhao, Y., Xu, Y., Cheng, X., Niu, Z., Luo, Z., Huang, H., Sun, F., Li, X., 2013. Water-level changes in China's large lakes determined from ICESat/GLAS data. Remote Sensing of Environment 132, pp. 131-144.

Wang, Y., Wu, B., 2019. Active machine learning approach for crater detection from planetary imagery and digital elevation models. IEEE Transactions on Geoscience and Remote Sensing 57, pp. 5777-5789.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13, pp. 600-612.

Weaver, W.R., Meador, W.E., 1977. An interpretation of photometric parameters of bright desert regions of Mars and their dependence on wave length. No. NASA-TN-D-8446.

Wehr, A., Lohr, U., 1999. Airborne laser scanning—an introduction and overview. ISPRS Journal of Photogrammetry and Remote Sensing 54, pp. 68-82.

Wöhler, C., 2004. Shape from shading under coplanar light sources, Joint Pattern Recognition Symposium. Springer, pp. 278-285.

Woodham, R., 1980. Photometric Method For Determining Surface Orientation From Multiple Images. Optical Engineering 19, 191139.

Wu, B., Dong, J., Wang, Y., Li, Z., Chen, Z., Liu, W.C., Zhu, J., Chen, L., Li, Y., Rao, W., 2021. Characterization of the Candidate Landing Region for Tianwen-1—China's First Mission to Mars. Earth and Space Science 8.

Wu, B., Dong, J., Wang, Y., Rao, W., Sun, Z., Li, Z., Tan, Z., Chen, Z., Wang, C., Liu, W.C., Chen, L., Zhu, J., Li, H., 2022. Landing Site Selection and Characterisation of Tianwen-1 (Zhurong Rover) on Mars. Journal of Geophysical Research-Planets.

Wu, B., Guo, J., Zhang, Y., King, B.A., Li, Z., Chen, Y., 2011. Integration of Chang'E-1 imagery and laser altimeter data for precision lunar topographic modeling. IEEE Transactions on Geoscience and Remote Sensing 49, pp. 4889-4903.

Wu, B., Hu, H., Guo, J., 2014. Integration of Chang'E-2 imagery and LRO laser altimeter data with a combined block adjustment for precision lunar topographic modeling. Earth and Planetary Science Letters 391, pp. 1-15.

Wu, B., Liu, W.C., Grumpe, A., Wöhler, C., 2018. Construction of pixel-level resolution DEMs from monocular images by shape and albedo from shading constrained with low-resolution DEM. . ISPRS Journal of Photogrammetry and Remote Sensing 140, pp. 3-19.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified perceptual parsing for scene understanding. Proceedings of the European Conference on Computer Vision (ECCV), pp. 418-434.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems 34, pp. 12077-12090.

Xie, H., Liu, X., Xu, Y., Ye, Z., Liu, S., Li, X., Li, B., Xu, Q., Guo, Y., Tong, X., 2022. Using Laser Altimetry to Finely Map the Permanently Shadowed Regions of the Lunar South Pole Using an Iterative Self-Constrained Adjustment Strategy. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15, pp. 9796-9808.

Xu, Q., Xu, Z., Philip, J., Bi, S., Shu, Z., Sunkavalli, K., Neumann, U., 2022. Point-nerf: Point-based neural radiance fields. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5438-5448.

Yang, X., 2000. Accuracy of rational function approximation in photogrammetry. Proceedings of ASPRS Annual Convention, Washington.

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. MVSNet: Depth inference for unstructured multi-view stereo. Proceedings of the European Conference on Computer Vision (ECCV), pp. 767-783.

Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L., 2019. Recurrent MVSNet for high-resolution multi-view stereo depth inference, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5525-5534.

Ye, Z., Xu, Y., Hoegner, L., Tong, X., Stilla, U., 2019. Precise disparity estimation for narrow baseline stereo based on multiscale superpixels and phase correlation. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W13, pp. 147-153.

Yuan, Y., Chen, X., Wang, J., 2020. Object-contextual representations for semantic segmentation. Proceedings of the European Conference on Computer Vision (ECCV), pp. 173-190.

Yue, Y., Fang, T., Li, W., Chen, M., Xu, B., Ge, X., Hu, H., Zhang, Z., 2023. Hierarchical Edge-Preserving Dense Matching by Exploiting Reliably Matched Line Segments. Remote Sensing.

Zhang, C., Wang, L., Yang, R., 2010. Semantic segmentation of urban scenes using dense depth maps. Proceedings of the European Conference on Computer Vision (ECCV), pp. 708-721.

Zhang, X., Srinivasan, P.P., Deng, B., Debevec, P., Freeman, W.T., Barron, J.T., 2021. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. ACM Transactions on Graphics (ToG) 40, pp. 1-18.

Zhang, Y., Xiong, C., Liu, J., Ye, X., Sun, G., 2023. Spatial-information guided adaptive context-aware network for efficient RGB-D semantic segmentation. IEEE Sensors Journal.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881-2890.

Zhao, W., Persello, C., Stein, A., 2023. Semantic-aware unsupervised domain adaptation for height estimation from single-view aerial images. ISPRS Journal of Photogrammetry and Remote Sensing 196, pp. 372-385.

Zhao, Z.-Q., Huang, D.-S., Sun, B.-Y., 2004. Human face recognition based on multi-features using neural networks committee. Pattern Recognition Letters 25, pp. 1351-1358.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H., 2015. Conditional random fields as recurrent neural networks. Proceedings of the IEEE International Conference on Computer Vision, pp. 1529-1537.

Zheng, T., Zhang, G., Han, L., Xu, L., Fang, L., 2022. BuildingFusion: Semantic-Aware Structural Building-Scale 3D Reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, pp. 2328-2345.

Zhi, S., Laidlow, T., Leutenegger, S., Davison, A.J., 2021. In-place scene labelling and understanding with implicit scene representation. Proceedings of the IEEE International Conference on Computer Vision (CVPR), pp. 15838-15847.

Zhou, H., Chen, Y., Hyyppä, J., Li, S., 2017. An overview of the laser ranging method of space laser altimeter. Infrared Physics & Technology 86, pp. 147-158.

Zhu, Q., Wang, Z., Hu, H., Xie, L., Ge, X., Zhang, Y., 2020. Leveraging photogrammetric mesh models for aerial-ground feature point matching toward integrated 3D reconstruction. ISPRS Journal of Photogrammetry and Remote Sensing 166, pp. 26-40.

Zuber, M.T., Smith, D.E., Phillips, R.J., Solomon, S.C., Neumann, G.A., Hauck, S.A., Peale, S.J., Barnouin, O.S., Head, J.W., Johnson, C.L., 2012. Topography of the northern hemisphere of Mercury from MESSENGER laser altimetry. Science 336, pp. 217-220.

Zuber, M.T., Smith, D.E., Solomon, S., Muhleman, D., Head, J., Garvin, J., Abshire, J., Bufton, J., 1992. The Mars Observer laser altimeter investigation. Journal of Geophysical Research: Planets 97, pp. 7781-7797.