Earth and Space Science

RESEARCH ARTICLE
10.1029/2024EA003812

DepthFormer: Depth-Enhanced Transformer Network for Semantic Segmentation of the Martian Surface From Rover Images

Yuan Ma[1], Zhaojin Li[1], Bo Wu[1], and Ran Duan[1]

[1]Department of Land Surveying & Geo-Informatics, Research Centre for Deep Space Explorations, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

Key Points:

- A new Mars image data set with labeled surface features and associated depth images for training semantic segmentation networks
- A new deep-learning transformer network incorporating four bands (red, green, blue, and depth) for more accurate semantic segmentation
- The proposed DepthFormer achieves an average accuracy of 98% for semantic segmentation of the Martian surface

Correspondence to:

Z. Li and B. Wu,
zhaojin.li@connect.polyu.hk;
bo.wu@polyu.edu.hk

Citation:

Ma, Y., Li, Z., Wu, B., & Duan, R. (2025). DepthFormer: Depth-enhanced transformer network for semantic segmentation of the Martian surface from rover images. *Earth and Space Science*, *12*, e2024EA003812. https://doi.org/10.1029/2024EA003812

Received 15 OCT 2024
Accepted 25 MAY 2025

Author Contributions:

Conceptualization: Bo Wu
Formal analysis: Bo Wu
Investigation: Zhaojin Li, Bo Wu
Methodology: Yuan Ma, Bo Wu
Resources: Bo Wu
Software: Yuan Ma
Validation: Yuan Ma, Zhaojin Li, Ran Duan
Writing – original draft: Yuan Ma, Bo Wu
Writing – review & editing: Bo Wu

Abstract The Martian surface, with its diverse landforms that reflect the planet's evolution, has attracted increasing scientific interest. While extensive data is needed for interpretation, identifying landform types is crucial. This semantic information reveals underlying features and patterns, offering valuable scientific insights. Advanced deep learning techniques, particularly Transformers, can enhance semantic segmentation and image interpretation, deepening our understanding of Martian surface features. However, current publicly available neural networks are trained in the context of Earth, rendering the direct use of the Martian surface impossible. Besides, the Martian surface features poorly texture and homogenous scenarios, leading to difficulty in segmenting the images into favorable semantic classes. In this paper, an innovative depth-enhanced Transformer network—DepthFormer is developed for the semantic segmentation of Martian surface images. The stereo images acquired by the Zhurong rover along its traverse are used for training and testing the DepthFormer network. Different from regular deep-learning networks only dealing with three bands (red, green and blue) of images, the DepthFormer incorporates the depth information available from the stereo images as the fourth band in the network to enable more accurate segmentation of various surface features. Experimental evaluations and comparisons using synthesized and actual Mars image data sets reveal that the DepthFormer achieves an average accuracy of 98%, superior to that of conventional segmentation methods. The proposed method is the first deep-learning model incorporating depth information for accurate semantic segmentation of the Martian surface, which is of significance for future Mars exploration missions and scientific studies.

1. Introduction

Mars, which was once a warm and wet planet, has been the focus of planetary exploration. Presently, over 30 dedicated Mars probes have conducted detailed examinations of the planet through remote sensing or in situ explorations, contributing to the enhancement of comprehensive understanding of the red planet (Shayler et al., 2005). Since NASA's Viking 1 achieved the first successful Mars landing in 1976 (Soffen & Snyder, 1976), multiple in situ exploration missions, including Viking 1 and 2 and the Pathfinder mission (Golombek, 1997), have collected extensive surface imagery that has enhanced our understanding of Martian landforms. The Mars Exploration Rovers (MER) mission, which included the Spirit and Opportunity rovers (Crisp et al., 2003), significantly advanced our knowledge of Mars's geomorphology and geology. Currently, the Curiosity and Perseverance rovers are exploring Gale Crater (Wray, 2013) and Jezero Crater (Pla-García et al., 2020), respectively, where they have discovered many surface features of scientific interest. China's Mars rover Zhurong, which landed on the southern Utopia Planitia, has also returned a tremendous amount of valuable image data of the Martian surface (Wu et al., 2022). Understanding the various types of surface features on Mars is crucial for exploration missions and scientific studies, of which semantic segmentation of the surface images is a key step. In recent years, deep-learning-based methods have been used for Martian image classification or segmentation (Iwashita et al., 2019; Liu et al., 2023; Rothrock et al., 2016; Taghanaki et al., 2021). However, due to the unstructured nature of the Martian surface—where the distribution of objects and terrains lacks explicit rules or fixed patterns compared to the structured environment on Earth (Liu et al., 2023)—and the presence of multiple objects at different scales and varying lighting conditions in the surface images, current segmentation methods face limitations when applied to the Martian environment for semantic segmentation (Petrovsky et al., 2022).

Early semantic segmentation methods grouped similar regions of an image into a class, assigning them the same color and a basic low level of textual information (Corso et al., 2008; Csurka, 2008; Wang et al., 2021). Deep

learning methods require less human involvement, provide sufficient data to achieve better results, and can continuously adjust themselves to improve their ability to predict results. Deep convolutional neural networks have gradually occupied a dominant position in semantic segmentation due to their compelling data analysis capability, learning ability, and the ability to model different scientific tasks (Garcia-Garcia et al., 2017; Holder & Shafique, 2022). Long et al. (2015) propose a new semantic image-processing method based on deep learning: a fully convolutional network (FCNet). Many researchers try to aggregate context information in different ways. In the FCN-based deep learning network architecture, Ronneberger et al. (2015) designed U-Net for biological image segmentation. Some semantic segmentation methods have utilized depth information to improve performance. For instance, Yan et al. (2021) presented a method that combines RGB and depth data to enhance image recognition accuracy through an attention mechanism. Zhang et al. (2023) introduced a lightweight network model with attention and context modules to effectively integrate RGB and depth data, thereby improving segmentation accuracy and computational efficiency compared to state-of-the-art methods.

Deep learning-based approaches have also been applied to Mars surface images for landform segmentation. Many of these approaches involve creating new Mars data sets or integrating existing ones for experimentation (Li et al., 2022; Liu et al., 2023; Swan et al., 2021; Zhang et al., 2022). Meanwhile, other approaches focus on developing lightweight networks to potentially enable automated Mars rover navigation in the future (Dai et al., 2022; Lv et al., 2023). Previous studies on Martian surface semantic segmentation mainly focused on basic feature identification with limited accuracy and adaptability across varying Martian terrains. Currently, Transformer-based neural networks have demonstrated superiority in extracting high-level features from images. The Vision Transformer (ViT) (Dosovitskiy, 2020) represents the first successful adaptation of the Transformer encoder for image classification tasks, showing superior performance over the widely used ResNet (He et al., 2016). However, due to its substantial computational requirements, a more efficient variant named the Hierarchical ViT using Shifted Windows (i.e., Swin Transformer) (Liu et al., 2021) was proposed. It outperforms similar network models such as EfficientNet (EffNet) (Tan & Le, 2019), Data-efficient Image Transformer (DeiT) (Touvron et al., 2021), Dual Attention Network (DANet) (Fu et al., 2019), DeepLabv3 (Chen et al., 2017), and Object-Contextual Representation Network (OCRNet) (Yuan et al., 2020). In this paper, we introduce the DepthFormer, an innovative depth-enhanced transformer model that improves segmentation performance by integrating depth perception and construct a data set based on the Mars surface imagery obtained by the Zhurong rover which simulates the rover's viewpoint through the derived 3D models to obtain the corresponding images and depth information to augment the data set. Afterward, the depth information and images are combined as four-dimensional inputs into the depth-enhanced semantic segmentation network for better segmentation of various types of surface features. The effectiveness of the proposed method is verified through quantitative and qualitative evaluation using synthesized and actual Mars image data sets.

2. Related Work

2.1. Deep-Learning-Based Semantic Segmentation for Mars Terrain

Many researchers have proposed deep learning-based methods for terrain classification tasks on Mars. Using a small number of samples provided by human experts, soil property and object classification (SPOC) uses a deep learning-based method, namely DeepLabV3 (Chen et al., 2017), which is based on convolutional neural network architecture, to identify the terrain types of the Martian surface (e.g., sand, rocks, outcrop) and Mars terrain features (e.g., scarps, ridges) (Rothrock et al., 2016). In addition, Rothrock et al. (2016) also proposed two applications for Mars rover missions, and one is the analysis of the traversability of the rover landing site area. The other is to find in previous Mars rover missions that the surface of Mars has excellent damage to the rover tires, and the rover slippage easily occurs on a relatively smooth surface, which is analyzed.

Aiming at the fact that the camera is easily affected by illumination changes when taking images, Iwashita et al. (2019) proposed two terrain classification methods based on deep learning, which can be well adapted to the change of illumination when trained on a data set that combines visible images and thermal images, that is, it is robust to illumination changes. Liu et al. (2023) believe that the current semantic segmentation method for the Martian terrain cannot be well applied to Mars because these standard methods are structured or unstructured scenes based on the segmentation scene. They propose a novel hybrid attention semantic segmentation (HASS) network that captures global attention and local attention to aid the final segmentation. In addition to this, they built a panoramic semantic segmentation data set (MarsScapes). For the consideration of safe landing, Claudet

et al. (2022) applied the efficient semantic segmentation algorithm to the safe landing of space aircraft, mainly through the identification and analysis of hazardous environments, which is trained on binary safety maps and HiRISE imagery data.

In addition to the methods described above, some researchers have tried to solve the problems with the Mars data set with semi-supervised frameworks. Wang et al. (2023) introduced a semi-supervised machine vision framework to Martian data for classification and segmentation. For the segmentation task, they used supervised contrastive learning in element-wise mode and introduced online pseudo labels for supervision on unlabeled areas. Goh et al. (2022) believe that the supervised classification method related to the segmentation of the Martian surface requires a large amount of data to achieve good results and efficient application. Moreover, for different segmentation tasks, the classification categories of the Martian terrain need to be different. They proposed a semi-supervised learning segmentation network framework and introduced a contrastive training method, and the final test results show that the method uses only 1% of the data set, 161 training images, and achieves a segmentation accuracy of 91.1%. Zhang et al. (2022) built a Martian terrain segmentation data set with high-quality labels with 6 K high-resolution images. Based on this data set, they designed a semi-supervised classification of multitasking mechanisms based on the idea of masked image modeling (MIM) to focus more attention on the textures of the images. Based on the encoding and decoding framework, Dai et al. (2022) introduced a mobile vision transformer (MViT) block to obtain local to global and multiscale information, introduced cross-scale feature modules (CFF) and compact feature aggregation module (CFA) the integrated context information and multi-level feature representation, they proposed a lightweight VIT-based Martian terrain segmentation network—SegMarsVit. Despite the advancements in Martian surface analysis, significant challenges such as handling shadows and lighting variations in images, and segmenting highly similar textural features, remain unresolved. The DepthFormer model addresses these challenges by utilizing a sophisticated depth-aware model that prioritizes underrepresented features, enhancing the model's ability to discern fine details in complex scenes.

2.2. Representative Mars Terrain Segmentation Data Sets

Deep learning requires extensive, sufficient, and accurate data to achieve good results. A number of large-scale image data sets have been established to help researchers on Mars terrain segmentation tasks. Swan et al. made the images that need to be annotated public through crowdsourcing to allow the public's aerospace enthusiasts to participate in the labeling of images (Swan et al., 2021). They built the first large-scale Mars data set, which is very helpful for the classification of Martian terrain and the assessment of traversability. The AI4Mars data set consists mainly of images taken by the Curiosity, Opportunity, and Spirit rovers. Contains four classes: Soil, bedrock, sand, and big rock. Li et al. have released the Mars-Seg data set with 5 K images containing nine classes: Martian Soil, Sand, Gravel, Bedrock, Rocks, Tracks, Shadows, Background, and Unknown (Li et al., 2022). Zhang et al. proposed a resolution, high-accuracy coefficient-labeled segmented data set divided into nine categories: Sky, Ridge, Soil, Sand, Bedrock, Rock, Rover, Trace, and Hole. The data set contains a total of 6,000 images of 1,200 × 1,200, taken from the camera aboard the Curiosity rover (Zhang et al., 2022).

In light of previous work, this paper highlights the following two main contributions:

1. To improve segmentation results, we incorporate 3D information (depth) into semantic segmentation, thereby enriching the original data.
2. We propose and introduce a method for generating the training data set in a semi-automatic manner. By leveraging readily available 3D information, we can establish a large and versatile data set based on real images.

3. DepthFormer for Semantic Segmentation of the Martian Surface

3.1. Overview of the Approach

The overview of the approach is illustrated in Figure 1. First, some images are manually labeled to attach semantic information to each label. Second, 3D reconstruction is conducted to generate the 3D scene based on the interior orientation (IO) parameters and exterior orientation (EO) parameters of the images. Third, the OpenGL pipeline (GLM, 2019), which ensures that each point in the original image is represented in normalized object space, is leveraged to render images from virtual viewpoints and generate the corresponding RGB and depth images. Besides, the traditional 2D augmentation techniques (i.e., translation, rotation, scaling, cropping) are also

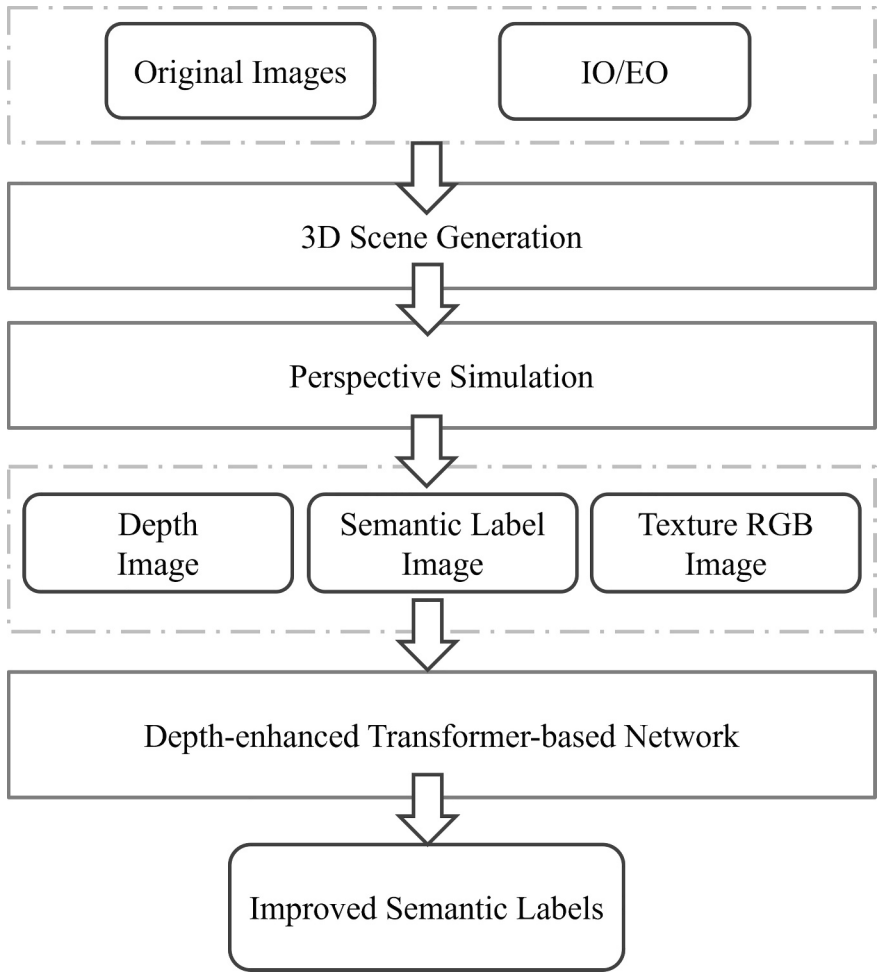

Figure 1. Overview of the approach (IO/EO indicates the interior and exterior orientation parameters of the images).

leveraged to further boost the volume of the data set. With the enriched depth information, the input to the neural network is configured as a combined representation of both textured and depth images, providing enhanced discriminative features for improved landform differentiation.

3.2. Data Set Construction

To provide a sufficient data set volume for training the neural network and generating essential depth data, we developed a new data set called DepthMars. We began with a rigorous 3D reconstruction using real rover images and then generated simulated images from the reconstructed 3D scene model. The raw Zhurong RGB images are collected by the Navigation and Terrain Camera (NaTeCam) onboard the Zhurong rover with a high resolution of 2,048 × 2,048. The categories of the original Mars imagery data set are based on the AI4Mars, including a total of nine categories: rover, sand, soil, rock, craters, wheel tracks, shadow, sky & far region, and others. The "Others" category specifically includes the rover's robotic arm and, in certain images, also encompasses the lander and parachute components. Six pairs of Zhurong images are shown in Figure 2, each pair consists of a raw image and a manually labeled label image. It is well known that Mars imagery data acquired along the rover's traverse shows a rich distribution of rocks, as well as the presence of large areas of sand. These two categories are, therefore, also heavily represented in this data set.

In this paper, depth images are incorporated to enhance image segmentation. Leveraging the raw Zhurong images and label data, a 3D model is created as a reference to simulate the Martian environment. By simulating the rover's perspective in image capture, depth images can be artificially produced. Furthermore, adjusting viewpoints and

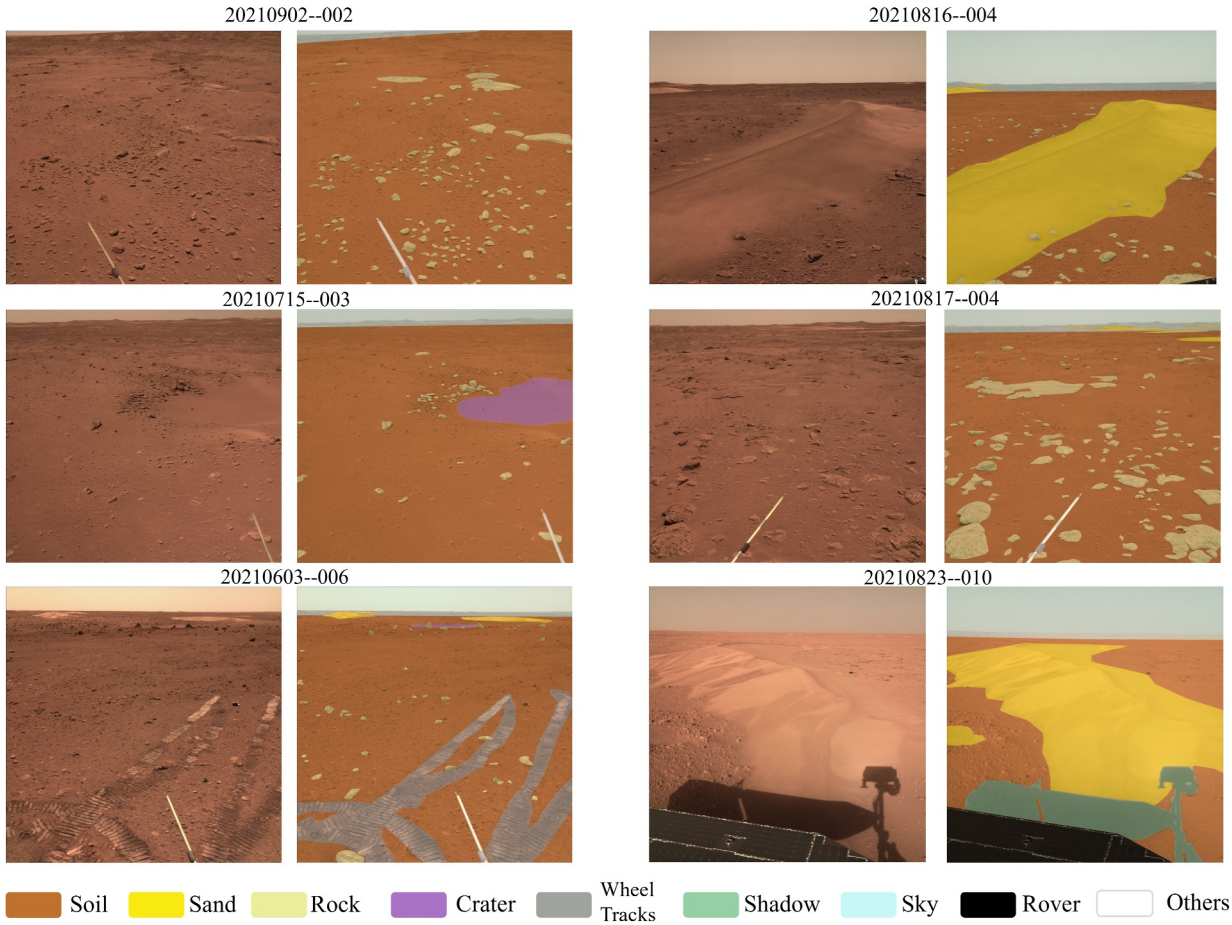

Figure 2. The landform features of the Martian surface and the labels. The date on which the images were captured, along with the image numbers at the stations, is listed above each subfigure.

positions allows for the generation of more Mars images and depth information, thereby enriching the data set. The data augmentation process is illustrated in Figure 3.

We utilized our in-house photogrammetric software (Liu & Wu, 2020; Wu, Dong, et al., 2021, Wu, Li, et al., 2021) to create high-resolution three-dimensional (3D) models from images captured by the Zhurong rover. The entire processing workflow follows the structure from motion (SfM) pipeline (Schonberger & Frahm, 2016), which begins with matching feature points across images to optimize camera parameters and generates sparse point clouds. Subsequently, the multi-view stereo (MVS) algorithm is employed to establish pixel-wise correspondences among images to generate dense point clouds, which are then interpolated and triangulated to form the 3D mesh model. Accordingly, the semantic segments annotated on the input images could be mapped onto the 3D mesh model based on the collinearity equation (Tang et al., 2016) with the optimized camera parameters.

Then, inspired by the approach of synthesizing ground images through the rendering of 3D meshes (Zhu et al., 2021), we generate 3D mesh models by leveraging MVS and SfM technologies. These models are then rendered and projected into 2D image planes from various viewpoints. The camera position and orientation information (IO/EO) for each view is then converted into the corresponding notations in OpenGL (GLM, 2019). OpenGL operates through a rendering pipeline that includes a model matrix for handling mesh positioning, a view matrix for processing camera parameters, and a projection matrix for implementing perspective projection. This pipeline ensures that each point in the original image is represented in normalized object space. During this simulation, the color image is undistorted using the Brown distortion model due to the neglecting of the camera's distortion parameters. Moreover, simulating the rover's viewpoint within the 3D scene resulted in the generation of color images and corresponding label images. By manipulating viewpoint selection, it is theoretically possible

10.1029/2024EA003812

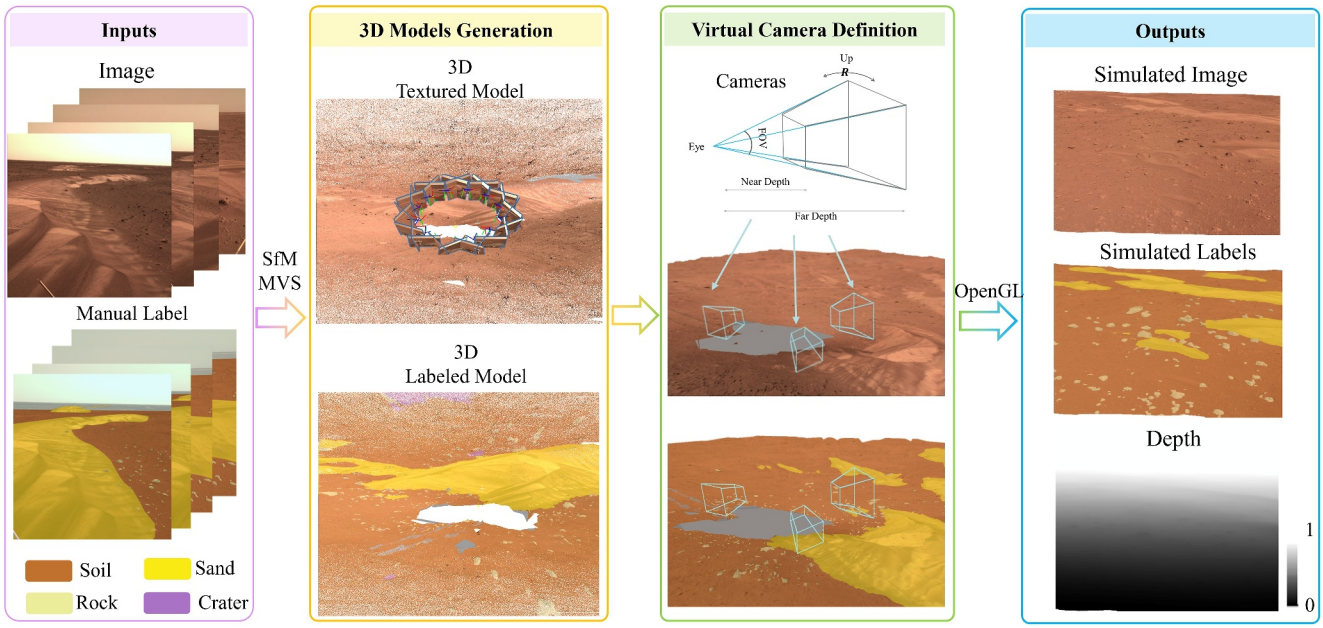

Figure 3. Illustration of the process for generating simulation results.

to create a vast number of simulated images. It is particularly noteworthy that the generated simulated images underwent meticulous manual screening to ensure data quality. This rigorous process eliminates images lacking sufficient information due to inappropriate viewpoint selection and prevents data set redundancy by filtering out excessively similar images with minimal variations in roll, pitch, and yaw angles.

In the development of the DepthMars data set, we included both real images captured by the Zhurong rover and virtually simulated images generated from our 3D models (Table 1). While this approach enriches the data set, it introduces a domain gap due to inherent differences between real-world imaging conditions and their simulated counterparts. These differences include, but are not limited to, variations in lighting, texture details, sensor noise, and atmospheric effects. In the paper, the construction of 3D models to simulate virtual images involves rigorous requirements for the images used to build the 3D models. The Zhurong Mars rover does not capture an extensive number of photos every day, except for particularly interesting targets. We therefore select raw images based on the adequacy of the images taken on a given date, typically 24 images around the rover, which are captured by the rover's stereo camera, to generate the surrounding 3D models of that location. Due to the high overlap between images and the short time differences between captures, we were able to significantly aid the precise generation of simulated images, effectively mitigating the effects of lighting and shadows. Moreover, we applied masking to the rover itself and its shadows to remove the constantly moving rover from the generated 3D models.

Table 1
Attributes of the Experimental Data Set

Data set	DepthMars
Type	Real + Synthetic
Classes	5
Real images	80
Synthetic images	543
Annotated images	623
Synthesized depth images	623
Image size	2,048 × 2,048

In terms of image textures and colors, we utilized high-resolution images of the Martian surface, and also applied masking to more distant locations to ensure more precise modeling. Throughout the experimental process, we continuously compared the generated images with the original images, making adjustments to the viewpoint and posture to ensure they closely match reality. Additionally, the rover captured images in a stable and safe environment, and the dates selected for generating the 3D models were close to the landing date, thus the performance of the rover remained largely unchanged.

In our ongoing research and future work, we plan to further improve the simulation process to achieve more accurate modeling. As more Martian images become available, we will enhance our data set accordingly. This rigorous approach ensures that the virtual images generated are a reliable basis for understanding and analyzing the Martian terrain, thus addressing the domain gap between real and simulated imagery. This careful curation and

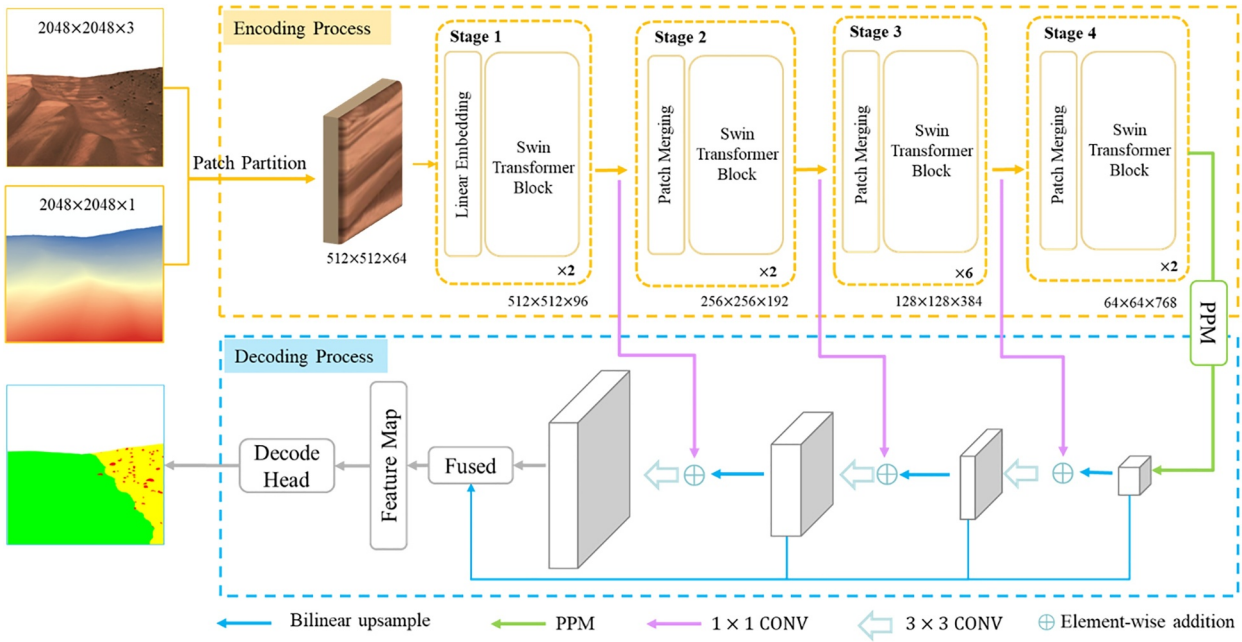

Figure 4. The architecture of the depth-enhanced Swin Transformer network—DepthFormer for Mars semantic segmentation.

processing of images helps to bridge the gap effectively, providing a robust data set for testing and developing image analysis algorithms.

In summary, we built a new DepthMars data set containing depth information to implement the Mars surface semantic segmentation experiments. The data set containing both raw and synthetic data from the Zhurong images and contains 623 images with a resolution of 2,048 × 2,048 pixels, of which 403 are used for training, 110 for validation, and 110 for testing. Depending on the traversal path taken by the rover and the degree of interest in the feature target, the data set is divided into five categories with varying sizes, textures, shapes, and scales. And the percentages of Soil, Rock, Sand, Craters, and Others are as follows: 45.01%, 1.23%, 12.67%, 0.01%, 41.97%. Notably, in the process of constructing 3D models, tie points are essential for aligning and reconstructing the images. Consequently, certain features, such as shadows, sky & far regions, wheel tracks, and the rover itself, are not reconstructed in the 3D models. This is because they either lack sufficient tie points or are absent from the images used for reconstruction. As a result, these features are missing in the subsequent simulated images generated from the 3D models.

3.3. Network Architecture

The overall architecture of the proposed network model DepthMars is shown in Figure 4. In the encoding process, the images are sliced into individual patches through the Patch Partition module before feeding the images into the block. The main function of the module is to convert the 2,048 × 2,048 × 4 tensor, synthesized from RGB and depth maps, into a 512 × 512 × 64 tensor. We subsequently adopt the Swin Transformer as the backbone network, owing to its exceptional performance in image segmentation and classification. The Swin Transformer is a hierarchical transformer architecture, which uses shifted windows for efficient computation and captures long-range dependencies with local attention within each transformer block. Each transformer block employs window multi-head self-attention (W-MSA) and shifted window multi-head self-attention (SW-MSA) to enable interaction between local and neighboring features, which greatly reduces the computational cost compared to global attention (Liu et al., 2021). This backbone network is composed of four stages, designed to extract image features at various semantic levels. These multi-level features are then fused in the decoding process using the pyramid pooling module (PPM), which captures contextual information at multiple scales through parallel pooling operations with different window sizes, to generate the final segmentation results.

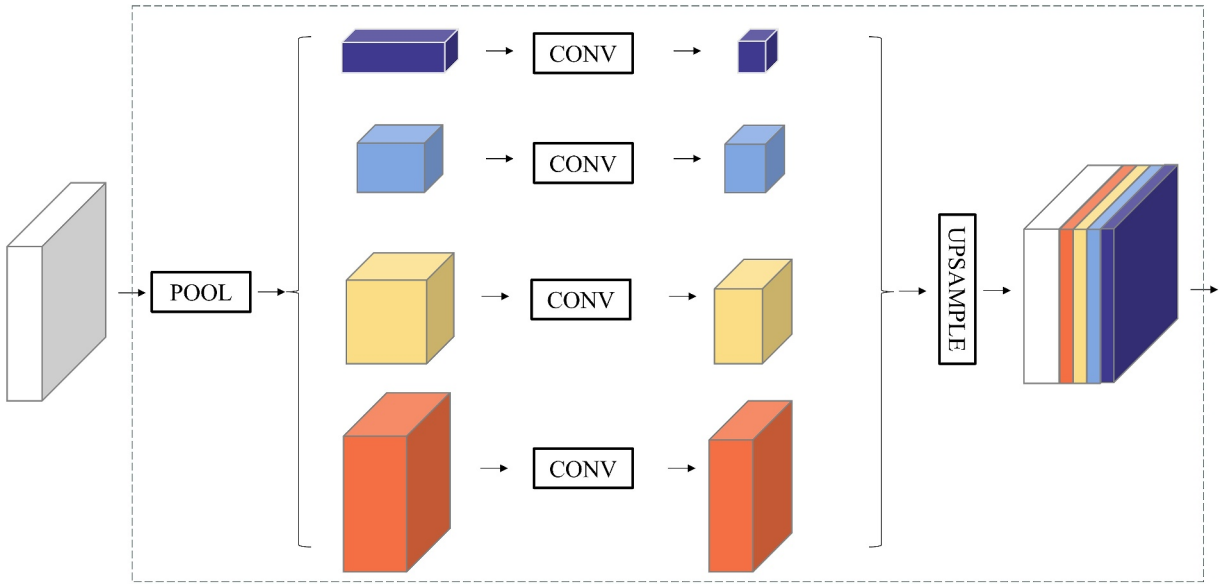

Figure 5. The architecture of the Pyramid Pooling Module.

3.3.1. Encoder

The encoder of this model is sequentially composed of a Patch Partition module and a backbone module made up of an efficient hierarchical transformer structure. When the RGB image (size H × W × 3) and the depth image (size H × W × 1) are entered in the Patch Partition module, the images are partitioned into 4 × 4 image patch tokens and become 64 feature dimension data, with a resolution of H/4 × W/4 for reducing computational complexity and improving model efficiency, for the following Swin Transformer block. In a Transformer block, the regular multi-head self-attention module is replaced by a self-attention module that alternates between W-MSA and SW-MSA. The window-based attention module greatly reduces the computational effort of the network by dividing the input image evenly into non-overlapping windows and limiting the attention calculation to each window. Additionally, the use of shifted windows-based modules can improve the information interaction between windows. In the first stage, the Swin Transformer block is repeated twice, and its output is subsequently fed into the following three consecutive stages, each beginning with Patch Merging to enable hierarchical feature extraction. Specifically, it reduces the spatial resolution of feature maps while increasing channel dimensions, mimicking pooling operations in CNNs. Given an input feature map of shape $[B, H, W, C]$, it groups 2×2 patches, flattens them into vectors of length $4C$, and concatenates them, reducing the spatial size to $\left[\frac{H}{2}, \frac{W}{2}\right]$, with $4C$ hannels. A linear layer then projects these to $2C$ channels, yielding a feature map of shape $\left[B, \frac{H}{2}, \frac{W}{2}, 2C\right]$. This process enables efficient multi-scale feature learning, balancing computational cost and expressive power. In Stage 2, 3, and 4, the Swin Transformer block is repeated two, six, and two times, respectively.

3.3.2. Decoder

The design of the Decoder draws inspiration from UPerNet (Unified Perceptual Parsing Network) (Xiao et al., 2018). The architecture of UPerNet is based on the Feature Pyramid Networks (FPN) (Lin et al., 2017), it can express the multi-layer features in the pyramid hierarchy, with a top-down structure, which can fuse semantic information from high-level to low-level. On this basis, the pyramid pooling module (PPM) from PSPNet (Zhao et al., 2017) is introduced in the last layer of the backbone network before entering the FPN top-down branch.

As depicted in Figure 5, the PPM is specifically designed to tackle the challenge of capturing contextual information at multiple scales. The primary function of pooling in this module is to distill the essential features from the feature maps by reducing their spatial dimensions while preserving vital information. This is accomplished through a hierarchical pooling approach, where features are extracted from subregions of the input feature map at various scales. The extracted features are then aggregated to form a comprehensive global representation that seamlessly integrates both local and global contextual information.

Figure 5 illustrates the architecture of the PPM, which comprises multiple branches, each corresponding to a distinct level of pooling. Specifically, the module employs average pooling with progressively increasing kernel sizes (e.g., 1×1, 2×2, 3×3, etc.), which reduces the spatial resolution of the input feature map while retaining the most salient information. The resulting pooled features are then refined through convolutional layers, which align their dimensionality and enhance their representational capacity. Finally, the features from all scales are upsampled to match the original resolution and concatenated along the channel dimension, yielding the pyramid-pooled global features.

This module plays a crucial role in enhancing the model's capacity to adapt to objects of varying scales in the image. By pooling features at different scales, it aggregates contextual information from local details to global patterns. This process is particularly important for Mars surface image analysis, where targets exhibit significant variations in size and appearance. The pyramid-pooled features provide a comprehensive representation of the image, improving the accuracy and robustness of semantic segmentation by capturing multi-scale contextual relationships.

The outcomes are returned to the highly compatible top-down FPN architecture, providing adequate global prior knowledge, and increasing the receptive field and the global information. Finally, the extracted features of different levels are fused into the feature map. The head consists of the convolutional layer and a classifier to analyze the fused feature map and ultimately output a prediction result of the same size as the original image.

3.3.3. Loss Function

We implemented the weighted cross-entropy (WCE) loss function to evaluate the prediction results in this paper. The WCE loss is a refined extension of the prevalent cross-entropy approach (Panchapagesan et al., 2016; Sudre et al., 2017). It allocates distinct weights to various classes rather than uniform values to focus more on intricate pixels that are challenging to discern, endowing them with higher weight, particularly evident in the boundary pixels of landforms classes. The loss function is as follows.

$$\mathcal{L}_{\mathrm{WCE}} = -\sum_{j} \sum_{n_1}^{n} \theta_k A_k^{(j,n_1)} \log\left(B_k^{(j,n_1)}\right) \tag{1}$$

Where n is the number of classes, the θ_k, $A_k^{(j,n_1)}$, $B_k^{(j,n_1)}$ are the allocated weights to pixel j in an image I_k. θ_k is the weight assigned to each pixel j in class n_1 in the image k, $A_k^{(j,n_1)}$ represents the ground truth probability for pixel j in class n_1, $B_k^{(j,n_1)}$ represents the predicted probability for pixel j being in class n_1. log() measures the logarithmic difference between the predicted and true probabilities. It ensures that different classes or pixel types (e.g., boundary pixels, specific landform features) are given different importance during the loss computation. The calculation of the weight for pixel j can be referenced in the following equation.

$$\theta_k^{(j)} = 1 + a\ddot{I}\left(\left|\nabla A_k^j\right| > 0\right) + b\ddot{I}\left(A_k^j = \mathtt{C}\right) \tag{2}$$

Where a and b are constant hyperparameters that control the impact of boundary pixels (at the edges of different classes) and underground class pixels (refers to a specific class in the data set, possibly representing a background or less prominent feature), respectively. $\ddot{I}()$ is the function to indicate the one and zero for the true and other values The $\mathtt{C}$, ∇ represents the underground class and the divergent, respectively.

4. Experimental Evaluation

4.1. Implementation and Evaluation Metrics

By scrutinizing images obtained by the Zhurong rover and Martian scenes, this experimental data set DepthMars encompasses classes such as sand, rock, and craters. Emphasis is placed on harmonizing the distribution of landforms across different locales to enrich the data set for training the transformer-based model. It is worth noting that in forthcoming experiments, analogous features can be subdivided, and the data set can be augmented by amplifying the instances of particular classes in accordance with scientific inquiry. Therefore, the data set employed for training, validation, and testing emphasizes the selection of images across various scales and

Table 2
Comparison of Quantitative Results Before and After Using the Weighted Cross-Entropy Loss Function Method in This Paper

Methods	Soil	Rock	Sand	Craters	Others	aAcc	mIoU
DepthFormer without WCE	**94.15**	39.82	90.9	34.8	**96.9**	96.35	71.31
DepthFormer	93.82	**60.25**	**92.66**	**38.45**	94.78	**98.28**	**75.99**

Note. The best results are highlighted in bold.

perspectives, the depth is also considered to evaluate the impact of feature depth to test and affirm the efficacy of depth information.

The model implementation for this study was completed through secondary development based on the open-source MMSegmentation (Contributors, 2020) framework, which is a modular and extensible semantic segmentation library designed to facilitate the training and evaluation of deep learning models for image segmentation. MMSegmentation provides a unified interface for various backbones (e.g., ResNet, Swin Transformer), decoders (e.g., U-Net, DeepLab), and loss functions, enabling rapid experimentation and customization. It supports multi-scale training, data augmentation, and a variety of optimizers and learning rate (LR) scheduling policies, making it highly adaptable for different segmentation tasks.

The model was trained on an NVIDIA Geforce RTX2080Ti with GPU acceleration with a batch size of 16. As the base Transformer framework, pre-training is done at ade20k based on the tiny version, and the window size is set to 7. Particularly, the image is cropped to 512×512 during the training process. The optimizer chosen is AdamW, and the iteration is set to 160 K. The learning rate policy chosen is "poly," which controls the step size at each iteration during optimization, and the initial value is set to an initial value of 1×10^{-6} to ensure stable convergence and avoids overshooting the loss landscape.

The evaluation metrics used in image segmentation are usually intersection over union (IoU) and accuracy. The IoU indicator treats the real and predicted labels as two sets, which means calculating the ratio of the intersection and union of the predicted label and the real label on each category. The calculation of the IoU is

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{3}$$

where the number of predicted pixels for true positives, false positives, true negatives, and false negatives is denoted by the TP, FP, TN, and FN, respectively. mIoU is usually calculated based on classes, where the IOU of each class is calculated and then added together before averaging to get an overall evaluation. The average accuracy (aAcc) is calculated as the proportion of the predicted correct quantity to the total quantity for each class.

4.2. Experimental Results

4.2.1. Evaluation on the Performance of the Weighted Cross-Entropy Loss Function

To address the data set's category imbalance, we employed the WCE loss function. Specifically, we assigned weighting coefficients of 0.1, 0.3, 8, 3, and 5 to the Others, Soil, Rock, Sand, Craters. This approach allows the model to prioritize minority categories and features of particular interest. The selection of these weights is typically influenced by the severity of category imbalance, which can be addressed through methods such as inverse proportional, inverse proportional square root, and proportion-based approaches.

We applied these methods to derive the weights, tailoring them to the unique characteristics of Martian surface scenes and the specific categories under investigation. Through a series of experiments, we fine-tuned these weights to optimize model performance. Consequently, our refined approach yielded improved results, which are detailed both quantitatively and qualitatively in Table 2 and Figure 6. Our analysis reveals a notable improvement in rock recognition, which can be attributed to the fact that rocks, despite occupying a relatively small pixel proportion in the data set, are the most numerous in terms of annotated instances. However, their prominent presence in the images does not translate to effective segmentation performance when compared to using the WCE loss function. To address this, we assign a larger weight to rocks to enhance their segmentation accuracy. In

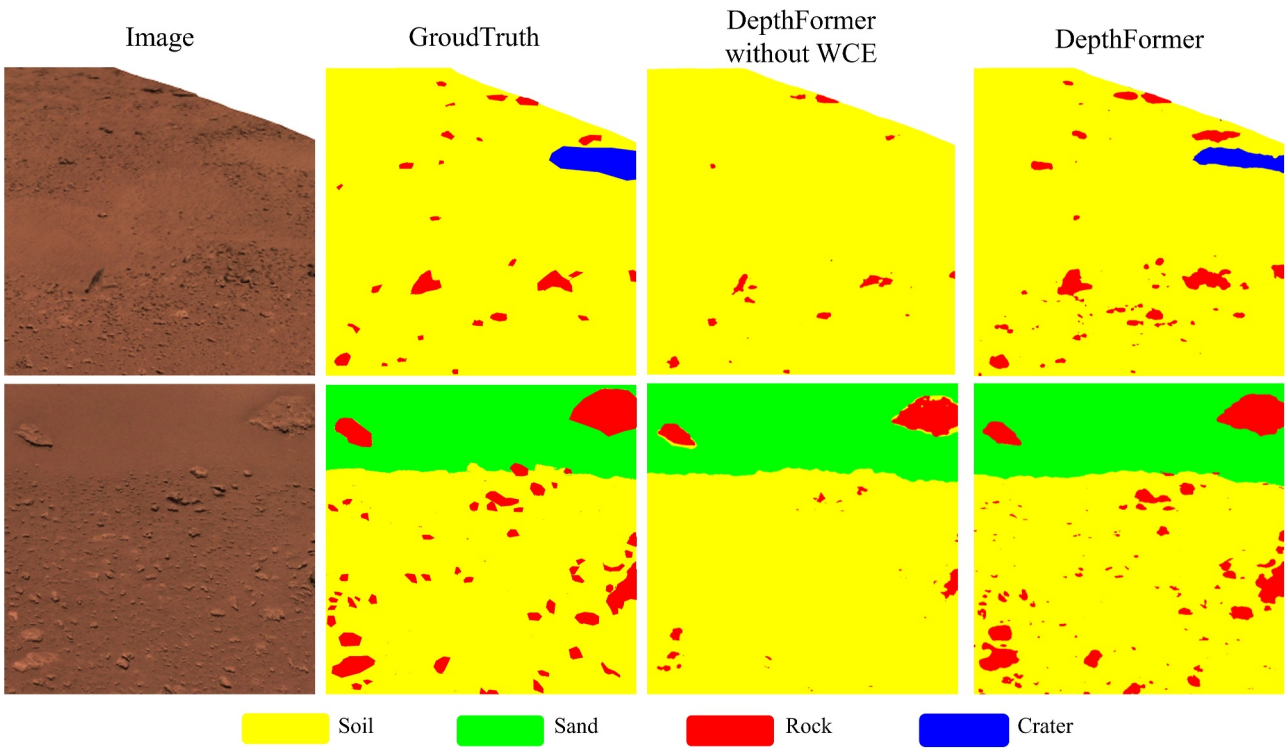

Figure 6. Qualitative comparison of the segmentation results before and after using weighted cross-entropy loss function method.

contrast, sand segmentation challenges are mostly limited to edge cases, which have a minimal impact on the overall results. Craters, on the other hand, are severely underrepresented in the data set, leading to suboptimal training outcomes. Nevertheless, assigning an excessively high weight to craters could result in a surge of erroneous segmentation, highlighting the need for a balanced approach.

4.2.2. Comparison Between DepthFormer and Other Methods

In order to evaluate the effectiveness of our proposed depth-enhanced method, we conducted extensive experiments on the DepthMars using a number of efficient, state-of-the-art semantic segmentation methods. These models represent the latest segmentation technology and provide a robust benchmark for evaluating our model. These models are developed using the MMSegmentation (Contributors, 2020) framework, leveraging pre-trained weights from large-scale data sets as a foundational starting point. These models undergo a comprehensive fine-

Table 3
Comparison With SOTA Semantic Segmentation Methods on the DepthMars Data Set

Methods	Soil	Rock	Sand	Craters	Others	aAcc	mIoU
OCRNet (Yuan et al., 2020)	6.62	47.45	11.26	0.0	67.25	71.17	26.52
DeepLabV3 (Chen et al., 2017)	74.81	**99.00**	69.20	17.85	**98.82**	95.17	71.94
FCN (Long et al., 2015)	74.22	98.98	69.82	31.99	98.80	96.15	74.76
SegFormer (Xie et al., 2021)	96.31	42.96	93.26	3.86	98.55	97.91	66.99
Mask2Former (Cheng et al., 2022)	96.17	57.21	92.46	34.75	98.59	97.87	75.84
SGACNet (Zhang et al., 2023)	95.70	34.83	87.81	0.0	96.76	96.87	63.02
Swin transformer (Liu et al., 2021)	92.67	54.09	91.47	34.52	95.14	95.85	73.58
DepthFormer	**93.82**	60.25	**92.66**	**38.45**	94.78	**98.28**	**75.99**

Note. The best results are highlighted in bold.

tuning process on our specific data set, which includes rigorous training and validation phases. To maintain consistency and ensure comparability of results across all experiments, we adhere to a standardized protocol throughout our investigation. As shown in Table 3, the best results are indicated in bold. From the experimental results, our proposed method has the best aAcc, and mIoU, which are 98.28%, and 75.99%, respectively, and outperforms the second-best result by 2.41% according to mIoU. Furthermore, our method achieves the best recognition accuracy in terms of IoU for all feature classes except rock and others. We speculate that this is owing to the fact that on the Martian surface, rocks differ in size, scale, and even shaded areas, which causes other methods to mistakenly classify other features as rocks except DeepLabV3. We speculate that this disparity arises because rocks on the Martian surface exhibit significant variability in size, scale, and shading, which leads to misclassification by other methods. In contrast, DeepLabV3 outperforms DepthFormer, primarily because DepthFormer tends to misclassify certain areas of soil as rock. The low accuracy of the crater class can be attributed to two main reasons. Firstly, the rover's oblique perspective significantly distorts the shape of craters, which confuses the network. Secondly, the limited number of crater training samples is due to safety considerations when selecting the Zhurong landing site (Wu, Dong, et al., 2021, Wu, Li, et al., 2021; Wu et al., 2022). In contrast to other methods, which exhibit poor performance, our approach achieves a notable accuracy of 38.45%. In addition, with the wide distribution of sand on Mars and its potential scientific value, this method may be of great help in identifying sand.

Figure 7 shows the segmentation results of the selected methods, along with the corresponding authentic images of the surface. Compared to other methods (i.e., OCRNet, FCN, DeepLabV3, SegFormer, Mask2Former, SGACNet, Swin Transformer), our proposed method is more capable of recognizing rocks at small scales and distinguishing more precise boundaries between different features. As a whole, FCN, DeepLabV3, and Swin Transformer are able to extract the main Martian surface features. However, our approach works better for large-scale sand extraction as well as for smaller-scale rocks, as shown in the enlarged views in Figure 7.

4.2.3. Effectiveness of Involving Depth Information in DepthFormer

To further demonstrate the impact of depth information on the semantic segmentation of the Martian surface, we select three image sets to visualize the segmentation results, as shown in Figure 8. A comparison is made with the results obtained using the Swin Transformer, which has previously demonstrated its superior performance in both quantitative and qualitative analyses.

In the first image row, the Swin Transformer faced challenges in distinguishing between rocks and Soil due to their similar colors, the abundance of small rocks, and the irregularity of rock edges. Our DepthFormer method significantly improved this segmentation. In the second image row, DepthFormer successfully segmented the rocks at both the top and bottom of the image. The third row presented a large sand area, with a blue-boxed region exhibiting a similar color and texture to the soil representation, making it challenging to differentiate. Nevertheless, DepthFormer effectively segmented the sand area and detailed the boundary between sand and Soil. Our proposed method thus demonstrates its ability to extract key features of the Martian surface while accurately identifying features of varying scales, similar colors, and textures.

4.3. Discussions

The proposed method, which includes data set generation and semantic segmentation algorithms, is designed to be applicable to any rover imagery data set. Although our experiments focused on Zhurong rover images as a proof of concept, the methodology can be extended to images collected by other Mars rovers, such as Curiosity and Perseverance, and potentially to lunar rovers as well (Li & Wu, 2018). The method's adaptability is mainly constrained by the manual effort needed for initial data set annotation and validation. It is worth noting that even the publicly available AI4MARS data set, which comprises labeled data from other rover images, has two main limitations: (a) the labels are not pixel-wise precise and appear relatively coarse, and (b) the images are not systematically organized by traversal stations, which is crucial for our 3D reconstruction workflow. To address these gaps, we plan to extend our labeling efforts to these publicly available data sets from the Planetary Data System (PDS) archive. This will enable us to construct 3D semantic models for different Mars exploration sites and further validate the generalizability of our algorithm.

Regarding the computational cost of the proposed method, generating the 3D models is relatively modest. However, it's important to note that instances of suboptimal image quality or insufficient image overlap may

10.1029/2024EA003812

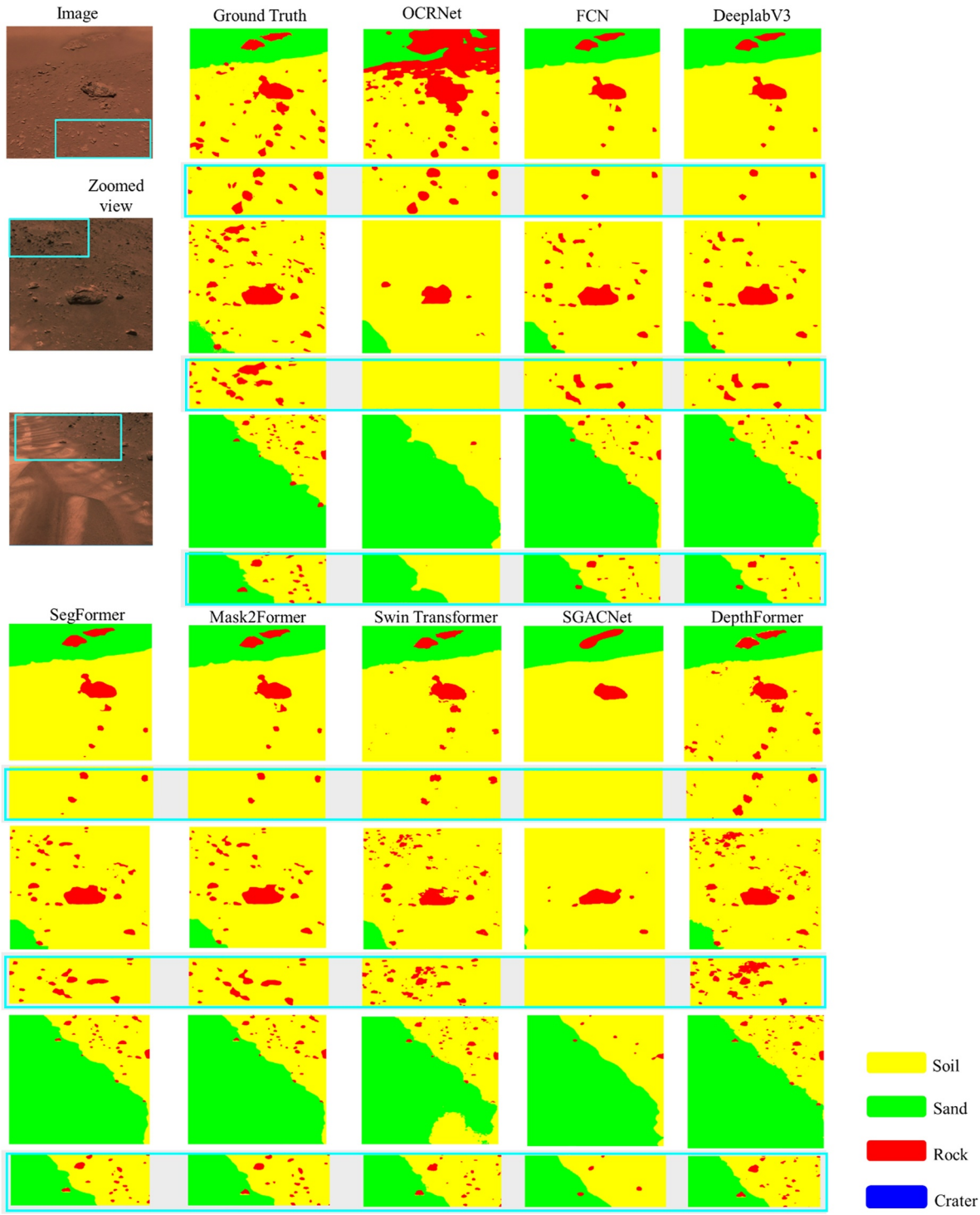

Figure 7. Qualitative comparison of the segmentation results from the OCRNet, FCN, DeepLabV3, SegFormer, Mask2Former, SGACNet, Swin Transformer, and DepthFormer. The enlarged views of the highlighted bounding boxes from the original images are also displayed below each result.

necessitate additional efforts for manual adjustments and the reselection of images, which may require more time and resources.

The results of semantic segmentation can enhance geomorphological and geological studies on Mars. For example, previous research has investigated the planet's geological evolution (Chen et al., 2022; Ye et al., 2021)

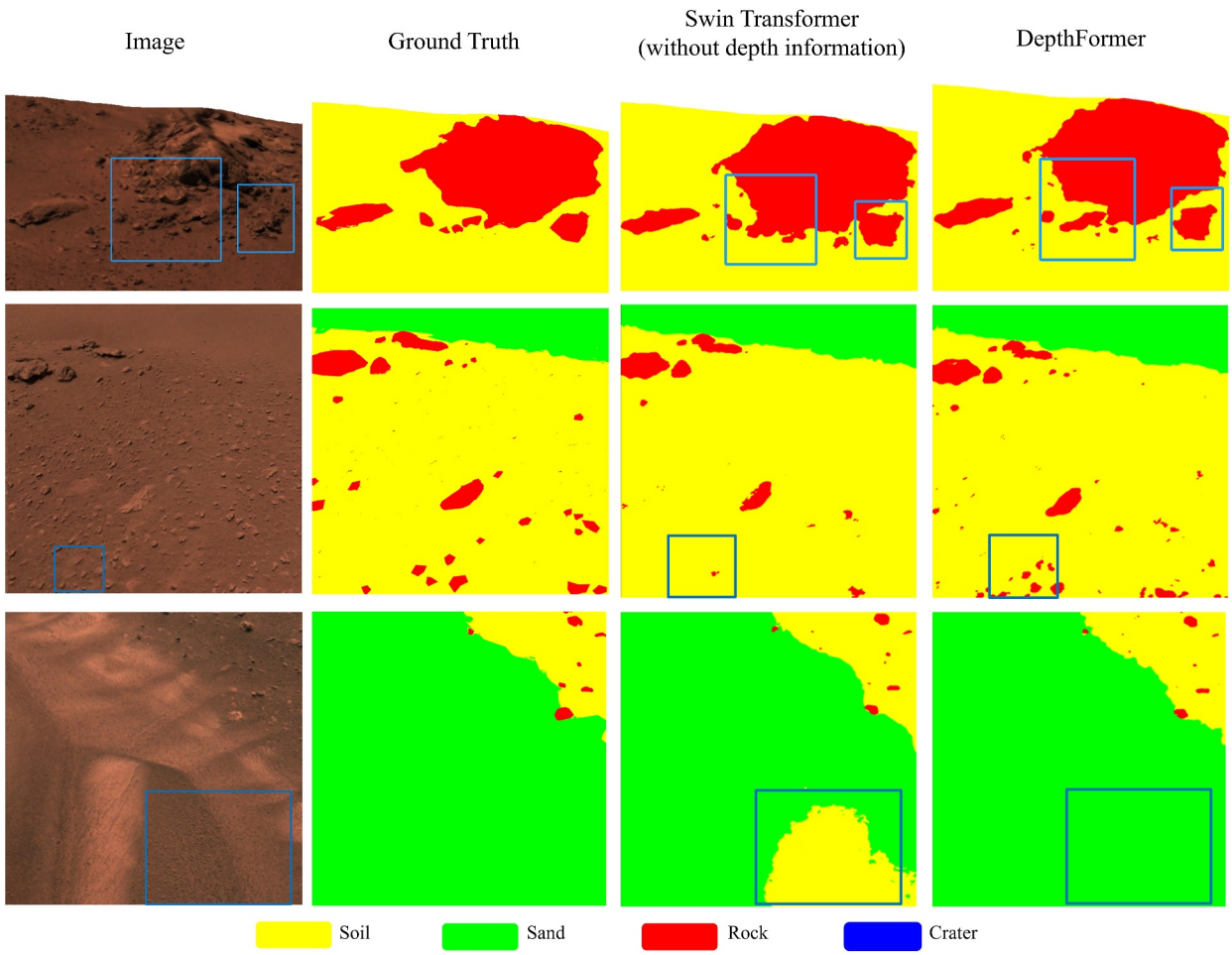

Figure 8. Enlarged views of the results show the effectiveness of involving depth information in semantic segmentation.

and hydrological activities (Ding et al., 2022; Liu et al., 2022) by analyzing the distribution of rocks, craters, and sand dunes at the Zhurong landing site. Similar studies have delved into various aspects of Martian terrain, contributing to our understanding of its complex landscape. Our future research aims to refine these algorithms for application to other in situ rover missions. This advancement would facilitate the automatic identification of diverse landforms across different regions, thereby offering a deeper insight into the geomorphological and geological implications suggested by these features.

5. Conclusions

In this paper, we proposed DepthFormer, a depth-enhanced transformer model for semantic segmentation on Mars, where the environment is complex and unstructured. In order to compensate for the lack of the Mars data set and simulate the depth images, a set of images is generated based on the high-resolution images returned by the Zhurong rover to simulate the Martian surface images and the corresponding depth images, DepthMars. It is proved through experiments that the introduction of depth information allows the DepthFormer model to aggregate the contextual information better and achieve an average accuracy of 98% for semantic segmentation of the surface images of Mars, superior to that of regular segmentation methods. The proposed DepthFormer can also finely delineate the boundaries between features. The limitation of this model lies in its dependency on the quality and quantity of available Martian surface data, which includes variations in lighting, terrain features, and atmospheric conditions for training. The model's performance is heavily influenced by the diversity and richness of the training data. The generalizability of the model to different Martian terrains and environmental conditions remains to be explored in future experiments and in real-world scenarios.

The proposed method advances automatic image segmentation, enhancing image analysis and scene interpretation. Its potential lies in its ability to interpret planetary surfaces more comprehensively from a semantic perspective, thereby improving surface operations of exploration rovers and revealing geomorphological features and patterns for more effective scientific research.

Data Availability Statement

The Martian surface images collected by the Zhurong rover are available at Lunar and Planetary Data Release System (https://moon.bao.ac.cn/web/enmanager/home). The derived semantic segmentation data set is available at Zenodo (https://doi.org/10.5281/zenodo.10939837 (Ma et al., 2024)).

Acknowledgments
This work was supported by grants from the Research Grants Council of Hong Kong (Project PolyU 15210520, Project PolyU 15215822, Project PolyU 15236524, RIF Project R5043-19, CRF Project C7004-21GF). The authors would like to thank all those who worked on the archive of the data sets to make them publicly available.

References

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, Z., Wu, B., Wang, Y., Liu, S., Li, Z., Yang, C., et al. (2022). Rock abundance and erosion rate at the Zhurong landing site in southern Utopia Planitia on Mars. *Earth and Space Science*, *9*(8), e2022EA002252. https://doi.org/10.1029/2022ea002252

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1280–1289. https://doi.org/10.1109/cvpr52688.2022.00135

Claudet, T., Tomita, K., & Ho, K. (2022). Benchmark analysis of semantic segmentation algorithms for safe planetary landing site selection. *IEEE Access*, *10*, 41766–41775. https://doi.org/10.1109/access.2022.3167763

Contributors, M. (2020). MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation

Corso, J. J., Yuille, A., & Zhuowen, T. (2008). Graph-shifts: Natural image labeling by dynamic hierarchical computing. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. https://doi.org/10.1109/cvpr.2008.4587490

Crisp, J. A., Adler, M., Matijevic, J. R., Squyres, S. W., Arvidson, R. E., & Kass, D. M. (2003). Mars exploration rover mission. *Journal of Geophysical Research*, *108*(E12). https://doi.org/10.1029/2002je002038

Csurka, G., & Perronnin, F. (2008). A simple high performance approach to semantic segmentation. *Procedings of the British Machine Vision Conference 2008*, 22.1–22.10. https://doi.org/10.5244/C.22.22

Dai, Y., Zheng, T., Xue, C., & Zhou, L. (2022). SegMarsViT: Lightweight mars terrain segmentation network for autonomous driving in planetary exploration. *Remote Sensing*, *14*(24), 6297. https://doi.org/10.3390/rs14246297

Ding, L., Zhou, R., Yu, T., Gao, H., Yang, H., Li, J., et al. (2022). Surface characteristics of the Zhurong Mars rover traverse at Utopia Planitia. *Nature Geoscience*, *15*(3), 171–176. https://doi.org/10.1038/s41561-022-00905-6

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

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3141–3149. https://doi.org/10.1109/cvpr.2019.00326

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*.

GLM. (2019). Opengl mathematics. Retrieved from https://glm.g-truc.net

Goh, E., Chen, J., & Wilson, B. (2022). Mars terrain segmentation with less labels. *2022 IEEE Aerospace Conference (AERO)*, 1–10. https://doi.org/10.1109/aero53065.2022.9843245

Golombek, M. P. (1997). The mars pathfinder mission. *Journal of Geophysical Research*, *102*(E2), 3953–3965. https://doi.org/10.1029/96je02805

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*.

Holder, C. J., & Shafique, M. (2022). On efficient real-time semantic segmentation: A survey. *arXiv preprint arXiv:2206.08605*.

Iwashita, Y., Nakashima, K., Stoica, A., & Kurazume, R. (2019). Tu-net and tdeeplab: Deep learning-based terrain classification robust to illumination changes, combining visible and thermal imagery. *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 280–285. https://doi.org/10.1109/mipr.2019.00057

Li, J., Zi, S., Song, R., Li, Y., Hu, Y., & Du, Q. (2022). A stepwise domain adaptive segmentation network with covariate shift alleviation for remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–15. https://doi.org/10.1109/tgrs.2022.3152587

Li, Y., & Wu, B. (2018). Analysis of rock abundance on lunar surface from orbital and descent images using automatic rock detection. *Journal of Geophysical Research: Planets*, *123*(5), 1061–1088. https://doi.org/10.1029/2017je005496

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*.

Liu, H., Yao, M., Xiao, X., & Cui, H. (2023). A hybrid attention semantic segmentation network for unstructured terrain on Mars. *Acta Astronautica*, *204*, 492–499. https://doi.org/10.1016/j.actaastro.2022.08.002

Liu, J., Li, C., Zhang, R., Rao, W., Cui, X., Geng, Y., et al. (2022). Geomorphic contexts and science focus of the Zhurong landing site on Mars. *Nature Astronomy*, *6*(1), 65–71. https://doi.org/10.1038/s41550-021-01519-5

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*, 153–168. https://doi.org/10.1016/j.isprsjprs.2019.11.017

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the 2021 IEEE/CVF international conference on computer vision*, 9992–10002. https://doi.org/10.1109/iccv48922.2021.00986

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the 2015 IEEE conference on computer vision and pattern recognition*.

Lv, W., Wei, L., Zheng, D., Liu, Y., & Wang, Y. (2023). MarsNet: Automated rock segmentation with transformers for Tianwen-1 mission. *IEEE Geoscience and Remote Sensing Letters*, *20*, 1–5. https://doi.org/10.1109/lgrs.2022.3227338

Ma, Y., Wu, B., Li, Z., & Duan, R. (2024). DepthMars dataset for semantic segmentation of the Martian surface from rover images [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.10939837

Panchapagesan, S., Sun, M., Khare, A., Matsoukas, S., Mandal, A., Hoffmeister, B., & Vitaladevuni, S. (2016). Multi-task learning and weighted cross-entropy for DNN-based keyword spotting. *Proc. Interspeech*, *2016*, 760–764.

Petrovsky, A., Kalinov, I., Karpyshev, P., Tsetserukou, D., Ivanov, A., & Golkar, A. (2022). The two-wheeled robotic swarm concept for Mars exploration. *Acta Astronautica*, *194*, 1–8. https://doi.org/10.1016/j.actaastro.2022.01.025

Pla-García, J., Rafkin, S. C., Martinez, G., Vicente-Retortillo, Á., Newman, C., Savijärvi, H., et al. (2020). Meteorological predictions for Mars 2020 perseverance rover landing site at Jezero Crater. *Space Science Reviews*, *216*(8), 148. https://doi.org/10.1007/s11214-020-00763-x

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*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Rothrock, B., Kennedy, R., Cunningham, C., Papon, J., Heverly, M., & Ono, M. (2016). SPOC: Deep learning-based terrain classification for mars rover missions. In *AIAA space 2016* (p. 5539).

Schonberger, J. L., & Frahm, J.-M. (2016). Structure-from-motion revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Shayler, D. J., Salmon, A., & Shayler, M. D. (2005). *Marswalk one: First steps on a new planet* (pp. 19–42). Springer London.

Soffen, G. A., & Snyder, C. W. (1976). The first Viking mission to Mars. *Science*, *193*(4255), 759–766. https://doi.org/10.1126/science.193.4255.759

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Jorge Cardoso, M. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017*, 240–248. https://doi.org/10.1007/978-3-319-67558-9_28

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. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1982–1991. https://doi.org/10.1109/cvprw53098.2021.00226

Taghanaki, S. A., Abhishek, K., Cohen, J. P., Cohen-Adad, J., & Hamarneh, G. (2021). Deep semantic segmentation of natural and medical images: A review. *Artificial Intelligence Review*, *54*(1), 137–178. https://doi.org/10.1007/s10462-020-09854-1

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning, 2019 PMLR*, *97*, 6105–6114.

Tang, S., Wu, B., & Zhu, Q. (2016). Combined adjustment of multi-resolution satellite imagery for improved geo-positioning accuracy. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 125–136. https://doi.org/10.1016/j.isprsjprs.2016.02.003

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jegou, H. (2021). Training data-Efficient image transformers & distillation through attention proceedings of the 38th international conference on machine learning. *Proceedings of Machine Learning Research*. https://proceedings.mlr.press/v139/touvron21a.html

Wang, W., Lin, L., Fan, Z., & Liu, J. (2023). Semi-supervised learning for mars imagery classification and segmentation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *19*(4), 1–23. https://doi.org/10.1145/3572916

Wang, Y., Wu, B., Xue, H., Li, X., & Ma, J. (2021). An improved global catalog of lunar impact craters (≥1 km) with 3D morphometric information and updates on global crater analysis. *Journal of Geophysical Research: Planets*, *126*(9), e2020JE006728. https://doi.org/10.1029/2020je006728

Wray, J. J. (2013). Gale crater: The mars science laboratory/curiosity rover landing site. *International Journal of Astrobiology*, *12*(1), 25–38. https://doi.org/10.1017/s1473550412000328

Wu, B., Dong, J., Wang, Y., Li, Z., Chen, Z., Liu, W. C., et al. (2021). Characterization of the candidate landing region for Tianwen-1–China's first mission to Mars. *Earth and Space Science*, *8*(6), e2021EA001670. https://doi.org/10.1029/2021ea001670

Wu, B., Dong, J., Wang, Y., Rao, W., Sun, Z., Li, Z., et al. (2022). Landing site selection and characterization of Tianwen-1 (Zhurong Rover) on Mars. *Journal of Geophysical Research: Planets*, *127*(4), e2021JE007137. https://doi.org/10.1029/2021je007137

Wu, B., Li, Y., Liu, W. C., Wang, Y., Li, F., Zhao, Y., & Zhang, H. (2021). Centimeter-resolution topographic modeling and fine-scale analysis of craters and rocks at the Chang'E-4 landing site. *Earth and Planetary Science Letters*, *553*, 116666. https://doi.org/10.1016/j.epsl.2020.116666

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)*, 432–448. https://doi.org/10.1007/978-3-030-01228-1_26

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*, 12077–12090.

Yan, X., Hou, S., Karim, A., & Jia, W. (2021). RAFNet: RGB-D attention feature fusion network for indoor semantic segmentation. *Displays*, *70*, 102082. https://doi.org/10.1016/j.displa.2021.102082

Ye, B., Qian, Y., Xiao, L., Michalski, J. R., Li, Y., Wu, B., & Qiao, L. (2021). Geomorphologic exploration targets at the Zhurong landing site in the southern Utopia Planitia of Mars. *Earth and Planetary Science Letters*, *576*, 117199. https://doi.org/10.1016/j.epsl.2021.117199

Yuan, Y., Chen, X., & Wang, J. (2020). Object-contextual representations for semantic segmentation. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 173–190. https://doi.org/10.1007/978-3-030-58539-6_11

Zhang, J., Lin, L., Fan, Z., Wang, W., & Liu, J. (2022). S^5Mars: Self-Supervised and semi-supervised learning for mars segmentation. *arXiv preprint arXiv:2207.01200*.

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*, *23*(19), 23512–23521. https://doi.org/10.1109/jsen.2023.3304637

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*.

Zhu, Q., Huang, S., Hu, H., Li, H., Chen, M., & Zhong, R. (2021). Depth-enhanced feature pyramid network for occlusion-aware verification of buildings from oblique images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *174*, 105–116. https://doi.org/10.1016/j.isprsjprs.2021.01.025