S^{5}Mars: Self-Supervised and Semi-Supervised Learning for Mars Segmentation

Jiahang Zhang*, Lilang Lin*, Zejia Fan, Wenjing Wang, Jiaying Liu, *Senior Member, IEEE,*

***Abstract*—Deep learning has become a powerful tool for Mars exploration. Mars terrain segmentation is an important Martian vision task, which is the base of rover autonomous planning and safe driving. However, existing deep-learning-based terrain segmentation methods face two problems: one is the lack of sufficient detailed and high-confidence annotations, and the other is the over-reliance of models on annotated training data. In this paper, we address these two problems from the perspective of joint data and method design. We first present a new Mars terrain segmentation dataset which contains 6K high-resolution images and is sparsely annotated based on confidence, ensuring the high quality of labels. Then to learn from this sparse data, we propose a representation-learning-based framework for Mars terrain segmentation, including a self-supervised learning stage (for pre-training) and a semi-supervised learning stage (for fine-tuning). Specifically, for self-supervised learning, we design a multi-task mechanism based on the masked image modeling (MIM) concept to emphasize the texture information of images. For semi-supervised learning, since our dataset is sparsely annotated, we encourage the model to excavate the information of unlabeled area in each image by generating and utilizing pseudo-labels online. We name our dataset and method Self-Supervised and Semi-Supervised Segmentation for Mars (S^{5}Mars). Experimental results show that our method can outperform state-of-the-art approaches and improve terrain segmentation performance by a large margin. Our project is available at https://jhang2020.github.io/S5Mars.github.io/.**

***Index Terms*—Mars vision tasks, terrain segmentation, image segmentation, self-supervised learning, semi-supervised learning.**

I. Introduction

HUMANS have shown great enthusiasm for Mars. The history of human research on Mars can date back to the 1960s. So far, more than 30 rovers have been dispatched to the red planet, and the increasing amount of available data promotes the application and development of deep learning algorithms. Deep-learning-based methods have already assisted in prioritizing data selection [1], collecting data, and analyzing data [2–4]. This paper explores the task of Mars terrain segmentation, which aims to identify the drivable areas and the specific terrains. It is of great significance to obstacle avoidance, traversability estimation, data collection, and path planning [5, 6], ensuring the safety and productivity of ongoing and future missions to Mars.

Mars terrain segmentation faces problems from both data and method design. First, the lack of satisfactory and available data hinders the development of deep learning methods to some extent. On the one hand, because of the high cost of Mars rovers, limited bandwidth and data transmission loss from Mars to Earth, collecting Martian data is very expensive. On the other hand, due to the complexity and similarity of the terrain, data annotation is highly specialized and time-consuming. Accordingly, previous datasets [7, 8] are not satisfactory because of the low-quality annotations or the roughly defined categories. AI4Mars [8], a newly published Mars terrain segmentation dataset, only defines four simple categories which are difficult to meet actual requirements of complex terrain identification. Besides, some datasets [7, 8] collected through crowdsourcing often do not have satisfactory annotation quality due to the inconsistent standards.

From a methodological point of view, the existing methods rely too much on large amounts of training data and lack targeted and effective design. Early works generally defined the set of terrain categories in advance and then directly applied a certain machine learning algorithm such as Support Vector Machines (SVM) [6]. With the rapid development of deep learning, methods based on deep neural networks including but not limited to Convolutional Neural Networks (CNN) are proposed to solve the terrain segmentation task [5, 8, 9], greatly improving the segmentation performance. However, most existing deep-learning-based methods still rely on standard supervised learning pipelines that require a lot of high-quality labeled data, which is often difficult to achieve. Besides, some works directly migrate the segmentation frameworks designed for Earth without sufficient consideration of the characteristics of Martian data. Goh *et al*. [10] alleviate the data dependency problem by applying an existing contrastive-learning-based method [11]. However, Goh *et al*. do not take into account the similarity between different terrains in Martian images, which makes the contrastive learning framework less effective.

In summary, there are two main challenges in the Mars terrain segmentation task: 1) the lack of data with sufficient detailed and high-confidence annotations, 2) the over-reliance of existing methods on annotated training data. We solve the above problems from the perspective of *both data and method design*, which are named **S**elf-**S**upervised and **S**emi-**S**upervised **S**egmentation for **Mars** (**S^{5}Mars**). We first create a new dataset to provide a high-quality and fine-grained labeled dataset for Mars terrain segmentation. Our dataset contains 6K high-resolution images captured on the surface of Mars, each of which is annotated by a professional team. There are 9 categories defined in our dataset, including sky, ridge, soil, sand, bedrock, rock, rover, trace, and hole, respectively. To improve the quality of labels, the annotation of the dataset adopt a sparse labeling style. Only the area with high human

* Equal contribution. The authors are with the Wangxuan Institute of Computer Technology, Peking University, Beijing, 100080, China, e-mail: {zjh2020, linlilang, zejia, daooshee, liujiaying}@pku.edu.cn.

confidence is annotated.

To learn from this sparse data, we propose a representation learning-based Martian semantic segmentation framework. In general, we first pre-train the model on the pre-designed auxiliary prediction tasks (known as pretext tasks) to obtain strong feature representations, and then fine-tune the model on the downstream task in a semi-supervised manner to make full use of the unlabeled area in the dataset. In this way, we reduce the dependence of model training on a large amount of annotated data.

There is a large literature on self-supervised learning. However, most existing methods are not designed for Mars image data and ignore to consider its special nature. Moreover, many approaches target at solving instance-level prediction problems like image classification, which is sub-optimal for dense prediction tasks like segmentation. Therefore, we design a pixel-level pretext task to bridge this gap. Specifically, our method is based on masked image modeling (MIM), which is inspired by masked language modeling [12, 13] in natural language processing. MIM allows the model to learn a visual representation by predicting the raw pixels [14–16] or designed features [17–19] of the masked image. However, because of the similar colors of images on Mars, predicting raw pixels alone will cause the network to tend to output the average of surrounding pixels and ignore the high-frequency textures, making the pretext task easier and less effective. It is observed that texture information plays an important role in Mars image segmentation. For example, the sky, soil, and sand all have similar color and spatial arrangement, while they can be distinguished from different textures, such as the relative proportion of grain sizes, roughness, and bumpiness. Therefore, we introduce a new pretext task, guiding the network to jointly model the low-frequency (color) and high-frequency (texture) information through a multi-task mechanism.

In the fine-tuning stage, since our dataset is sparsely annotated and the areas hard to distinguish in the image are not labeled, we propose a semi-supervised learning method based on uncertainty to further improve performance. A pseudo-label-based approach is designed to take full advantage of the information in unlabeled areas. After our self-supervised pre-training, we fine-tune the model with ground-truth labels on labeled areas and pseudo-labels generated online on unlabeled areas. Furthermore, we exploit the task uncertainty to improve the quality of pseudo-labels. Experimental results demonstrate that our method can improve the performance to a large extent for Mars imagery segmentation.

Our contributions can be summed up as follows:

- We collect a new fine-grained labeled Mars dataset for terrain semantic segmentation, which contains a large amount of Martian geomorphological data. Our dataset is sparsely annotated by a professional team under multiple rounds of inspection rework. The high-quality dataset can provide accurate and rich segmentation guidance.
- We propose a self-supervised multi-task learning approach to improve the performance of Mars imagery segmentation, in which the network can learn strong representations by explicitly modeling both the low-frequency color feature and high-frequency texture feature of the input image. It enables self-supervised learning to extract more useful and comprehensive information, providing better initialization for downstream tasks.
- An uncertainty-based semi-supervised training strategy is introduced to take full use of unlabeled Mars image areas. We exploit task uncertainty to generate confident pseudo-labels, reducing noise in labeled areas. Semi-supervision is applied to employ more data to improve the generalization ability of the model without compromising the feature representation ability by introducing noise.

The rest of this article is organized as follows. In Section II, we provide a detailed survey on Martian datasets and a brief review of deep learning for Mars. Section III introduces our Mars segmentation dataset. Section IV and Section V describe our framework for Mars semantic segmentation. Experimental results are shown in Section VI and Section VII. The conclusion is finally given in Section VIII.

II. Related Works

A. Deep Learning for Mars

With the increasing amount of available data and the rapid development of computing power, deep learning is playing an increasingly important role in Mars exploration.

For many reasons such as limited computing resources, existing deep learning methods are usually ex-situ (Earth edge). For terrain identification, Deep Mars [21] trains an AlexNet to classify engineering-focused rover images (*e.g.*, those of rover wheels and drill holes) and orbital images. However, it can only recognize one object in a single image. The Soil Property and Object Classification (SPOC) [9] proposes to segment the Mars terrains in an image by using a fully convolutional neural network. Swan *et al*. [8] collect a terrain segmentation dataset and evaluate the performances using DeepLabv3+ [32]. Considering the dependence of existing methods on large amounts of data, [10] utilizes a self-supervised method and trains the model on less labeled images. For other tasks, Zhang *et al*. [33] deal with Mars visual navigation problem by utilizing a deep neural network, which can find the optimal path to the target point directly from the global Martian environment.

Meanwhile, intrigued by the vision of autonomous probes that rely on deep learning even without human-in-the-loop requirements, scientists are studying the potential of implementing in-situ (Mars edge) deep learning algorithms using high-performance chips [34]. For example, the Scientific Captioning Of Terrain Images (SCOTI) [1] model automatically creates captions for pictures of the Martian surface based on LSTM, which helps selectively transfer more valuable data within downlink bandwidth limitations. For energy-optimal driving, Higa *et al*. [29] propose to predict energy consumption from images based on a PNASNet-5 [35].

It is foreseeable that deep learning will play an irreplaceable role in future Mars exploration. However, it also requires a lot of annotated training data, which is often hard to obtain. In the paper, we present a powerful self- and semi-supervised learning framework, which can learn a good visual representation from large amounts of unlabeled data, to resolve the terrain segmentation task.

TABLE I
SUMMARY OF MARS TERRAIN-AWARE DATASETS.

Type	Source	Dataset	Scale	Classes	Description
Real	Curiosity rover	[9]	5k	-	Wheel slip and slope angles prediction
			700	6	Terrain segmentation
		[5]	300	3	Terrain classification
		[20]	620	4	Terrain classification
		[21]	6k	24	Terrain classification
		[22]	405	-	Rock detection
		[1]	1k	-	Image description
		[23]	310k	-	Compressed image quality evaluation with automatic labeling
	Opportunity, Spirit rovers	[24]	117	-	Rock detection
	Curiosity, Opportunity, Spirit rovers	[25]	46	2	Terrain segmentation
		[8]	35k	4	Terrain segmentation
		[26]	5k	9	Terrain segmentation
		[7]	5k	6 (17 sub)	Terrain segmentation
Real + Synthetic	Curiosity rover	[27]	30k	5	Terrain classification
Synthetic	ROAMS rover simulator	[24]	55	-	Rock detection
Simulation field	Atacama Desert Zoë rover prototype	[28]	30	-	Rock detection
	JPL Mars Yard FIDO rover Platform	[24]	35	-	Rock detection
	JPL Mars Yard Athena rover Platform	[29]	91k	-	Rover energy consumption
	Devon Island	[30]	400	-	Rock detection
Real + Simulation field	Opportunity, Spirit rovers	[31]	36	2	Terrain segmentation

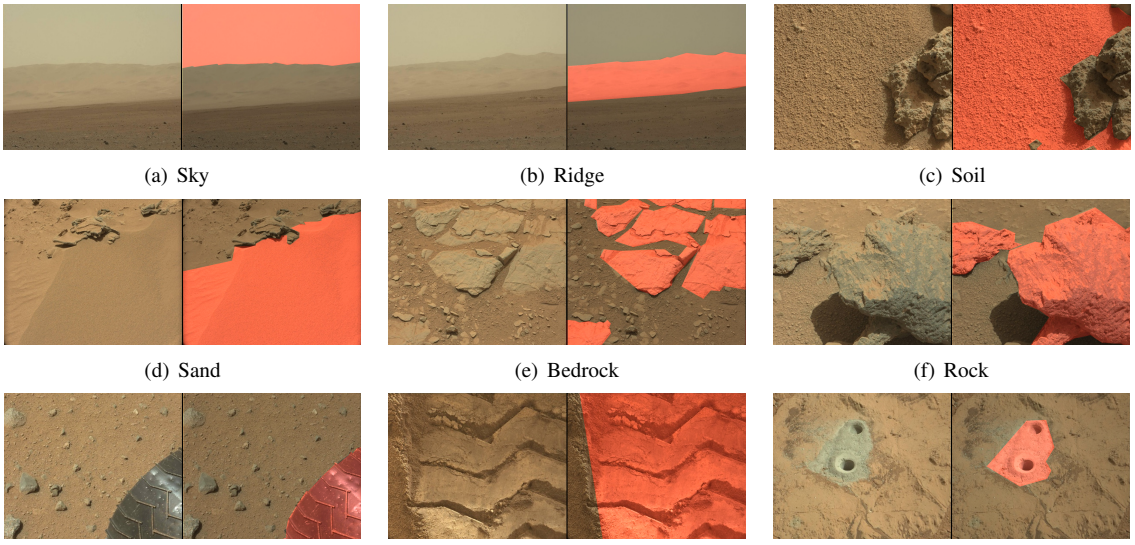
B. Datasets for Mars Vision

Datasets are the basis for intelligent algorithms development. At present, there are various datasets of planetary surfaces, such as digital simulation Lunar landscape segmentation dataset ALLD and Mars Satellite image dataset DoMars16k [27]. As for Mars, the commonly used terrain-aware datasets can be divided into three categories: rover shooting real data, artificial synthetic data, and earth simulation field shooting data. The rover shooting data are captured by devices of rovers that land on Mars. The number of rovers sent to Mars will gradually increase along with the progress of space research. However, the amount of data available now is still relatively limited. Synthesizing Mars datasets by means of digital modeling simulation or adversarial learning is an important data supplement, but can differ greatly from the real Mars data. Earth simulation field shooting way requires building a simulation platform or finding a similar landscape on Earth to Mars, which is difficult to implement. The current Mars terrain-aware datasets are shown in Table I. A large proportion of them have an image quantity of less than 1000, which can not meet the training needs of the machine learning models. The richness of Mars terrain-aware datasets still needs to be strengthened.

C. Self-Supervised and Semi-Supervised Learning

Self-supervised learning aims to learn a robust feature space from unlabeled visual data by pretext tasks, usually constructed by image operations [14, 36–38] or spatiotemporal operations [39–41]. It alleviates the need for expensive annotations and enables representation learning from unlabeled data. A common approach is to pre-train the network on a large dataset such as ImageNet [42] using a standard classification task and fine-tune it on a small dataset for certain downstream task [21]. However, this method will suffer from domain shift caused by the differences to some extent in image properties between the pre-training data and fine-tuning data. Another popular method lately is contrastive learning [11, 43–46], which capitalizes on augmentation invariance. By extending the distance between negative samples while narrowing the distance between positive samples, the model learns a more separable feature space. However, most of the above methods are designed for instance-level classification and can be sub-optimal for dense prediction tasks like segmentation. Recently, the masked image modeling [14–17, 37, 47] has achieved surprising results in self-supervised learning, which provides a natural pixel-level pretext task. We explore the potential in Mars image semantic segmentation by masked image modeling in this paper.

Semi-supervised learning [48, 49] utilizes the manifold structure of unlabeled data to assist learning with labeled data. The pseudo-label method [50] assigns pseudo-labels to unlabeled data through a classifier trained on supervised data. The semantic information of unlabeled data is extracted by this method. To reduce the noise of pseudo-labels and improve the quality, Zou *et al*. [51] use label regularization to filter labels with high confidence. To correct its own mistakes of pseudo-label, Mugnai *et al*. [52] propose Gradient Reversal Layer (GRL) for fine-graining the labels. However, the main



(a) Sky (b) Ridge (c) Soil

(d) Sand (e) Bedrock (f) Rock

(g) Rover (h) Trace (i) Hole

Fig. 1. Examples for each label category (highlighted in red).

downside of such methods is the low quality of pseudo-label, which will seriously interfere with network training. We address this issue by introducing uncertainty estimation to the pseudo-label selection. Only the pseudo-labels with high confidence are retained for training.

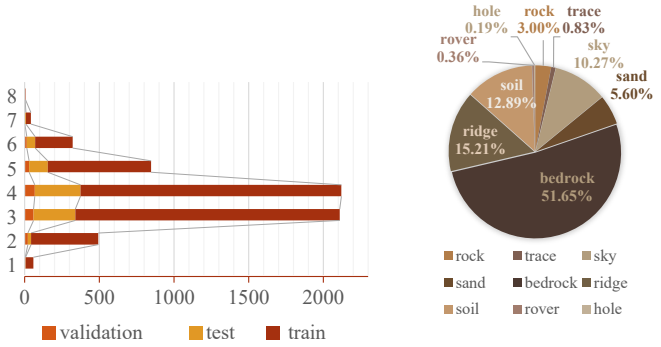
III. Proposed Mars Imagery Segmentation Dataset

In order to solve the problem of scarce available training data for deep learning, we create a fine-grained labeled Mars dataset for the exploration on Mars surface, which can guide the rovers and support space research missions. In the following, we mainly use S^{5}Mars to represent our dataset for clarity. The dataset includes 6,000 high-resolution images taken on the surface of Mars, by color mast camera (Mastcam) from Curiosity (MSL). The spatial resolution of RGB images in this dataset is 1200 $\times$ 1200. We divide the dataset into training, validation, and test sets randomly. The training set contains 5,000 images, while the validation set and the test set contains 200 and 800 images, respectively.

A. Labeling Process

We annotate the dataset at pixel level in a deterministic sparse labeling style. There are 9 label categories, sky, ridge, soil, sand, bedrock, rock, rover, trace, and hole, respectively. Examples of each category are shown in Fig. 1. The labeling criteria are as follows:

- **Sky.** The Martian sky, often at the top of a distant image, bounded by the upper edge of a mountain or horizon.
- **Ridge.** The distant peaks bounded by the sky above and the horizon below.
- **Soil.** Unconsolidated or poorly consolidated weathered material on the surface of Mars, with larger and coarse-grained grains containing small stones.
- **Sand.** Granular material, more fluid, less viscous, some with windward and leeward sides, most of the time with sand ridges.

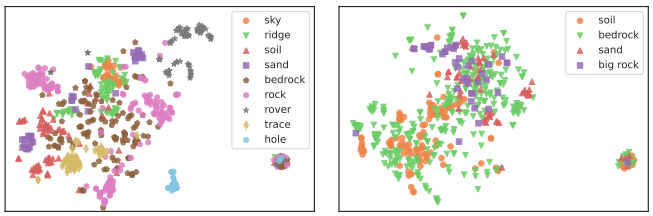


(a) Dist. of the number of labels (b) Dist. of label area

Fig. 2. Numerical statistics on our S^{5}Mars dataset. The figures show the richness of the categories contained in the image from two aspects: distribution of the number of labels and distribution of label area.

- **Bedrock.** Partially covered by the soil and buried at varying depths.
- **Rock.** A stone that is completely exposed to the ground and is roughly lumpy or oval in shape, usually with distinct shadows.
- **Rover.** The rover itself.
- **Trace.** The trace left by the rover when it passed over the ground.
- **Hole.** The hole left by the rover during its sampling operation on Mars, contains the surrounding soil of different colors.

Martian surface condition is complicated due to the harsh and volatile Martian environment. The terrain types can mix and overlap with each other and it becomes hard for humans to distinguish the correct categories clearly. Considering the situation, we apply sparse labeling, which ensures that only the pixels with enough human confidence are labeled. The overall annotation priority is in a coarse-to-fine manner, which means we label each image in order of object size. In addition, the



(a) Visualization on S^{5}Mars (b) Visualization on AI4Mars

Fig. 3. Visualization of pixel-level feature distribution on the S^{5}Mars and the AI4Mars. Features are extracted by Swin pre-trained backbone and are visualized with t-SNE.

trace left by rover will be assigned a higher priority since its appearance is relatively infrequent.

As for the annotating process, the annotation rules are discussed more than ten times and take each annotator's feedback into account to keep consistency and preciseness. Each annotation result passes more than two turns of quality inspections. Annotation work is carried out by a professional team, 90% of the annotators have been engaged in such annotation work more than six times, and the annotators are experienced. The age distribution of the annotators ranges from 18 to 37 years old, with an average age of about 24 years old. The annotation time of each terrain image is about 30 minutes.

B. Comparison and Analysis

We make a statistical analysis on the semantic labels in the dataset, as shown in Fig. 2. We show the distribution of the number of labeled categories contained in each image in Fig. 2(a). Most images are relatively complex with three or four annotations in one scene. This distribution on training, validation and test sets keeps in good consistency.

We make statistics of the distribution of label area of each category, as shown in Fig. 2(b). There are 2,730 images of distant views in the dataset, which contains the sky and the surface division line. Bedrock is the label of the largest annotation area, ridge the second. Rocks appear in most of the images in the dataset, but the total area is small. The artificial impact, *e.g.*, rover, trace, and hole, accounts for few portions of the labeled area, but they have a greater variety of shapes and are crucial to the observation and judgment system for intelligence research on Mars. In that sense, the distribution of the dataset offers new challenges for future research, since it tests the generalization and robustness of the research methods for long-tail data.

We provide the visualization of pixel-level feature distribution on the S^{5}Mars, while comparing with the same type labeled dataset AI4Mars [8], as in Fig. 3. The features are randomly extracted with Swin [53] segmentation model. We take the features from the second to last layer of the pre-trained backbone and visualize them with t-SNE data dimension reduction algorithm. The Swin backbone is not finetuned on any Mars data for the fairness of the comparison. It can be observed that S^{5}Mars shows a more separable feature distribution. The categories like the sky, trace, hole, and rover, are highly distinguishable from other data. Categories like soil, rock, and sand also show a significant aggregation. AI4Mars contains 4 categories of labels, and the separability of each type of label is poor, especially that between bedrock and soil. Since AI4Mars is a crowdsourcing project, though the number of submissions is large, the annotators may have inconsistent understandings of labeling standards. This reflects the importance of establishing clear labeling criteria and professional training for annotators.

Mars-Seg [26] is also a public Mars terrain segmentation dataset. The dataset has 1,064 high-resolution grayscale images and 4,184 RGB images with a spatial resolution of 560×500, while S^{5}Mars is composed of high-resolution RGB images, which offers more accurate and more abundant high-frequency information and texture details for detection and segmentation tasks. Two datasets differ in labelling process. Mars-Seg divides the terrain into 9 categories and labels every pixel. However, categories like gravel, sand, and rocks mix up with each other and it is inaccurate to determine the terrain scene into any one category. S^{5}Mars applies a high-confidence sparse-labeled manner, that only regions with high confidence to judge the terrain type will be labeled. This way we guarantee the labels are strongly representative in each category and reduce the label noise introduced in the labeling work. Considering this, the dataset offers an ideal scenario for self-supervised and semi-supervised learning research.

IV. Self-Supervised Multi-Task Learning for Mars Imagery Segmentation

In this section, we will introduce our approach based on self-supervised learning for the Mars imagery semantic segmentation task. To enrich the visual representation of the network, the model is first pre-trained on the pretext task using unlabeled data in a multi-tasking manner. In the following subsections, we give an overview first and then go into more details about our method.

A. Motivation and Overview

Considering that our downstream task semantic segmentation is a dense predictive task, a pixel-level pretext task rather than an instance-level one is expected in the self-supervised learning phase. Compared to other pretext tasks such as jigsaw puzzles and rotation prediction, image inpainting, which aims to restore the masked image in RGB space, provides a natural pixel-level pretext task. It is an efficient representation learning method and has achieved remarkable results recently [15, 16]. However, due to the optimization of mean distance, the output is usually a blurry averaged image. This is attributed to the fact that the L2 (or L1) loss often prefers a blurry and smooth solution, over highly precise textures [14]. Therefore, for Mars data with similar colors and unclear object contours, it is difficult for the network to learn distinguishing features only by predicting low-frequency information of the image. High-frequency information like texture, which plays an important role in terrain identification, is needed to assist the model to obtain a more strong representation ability.

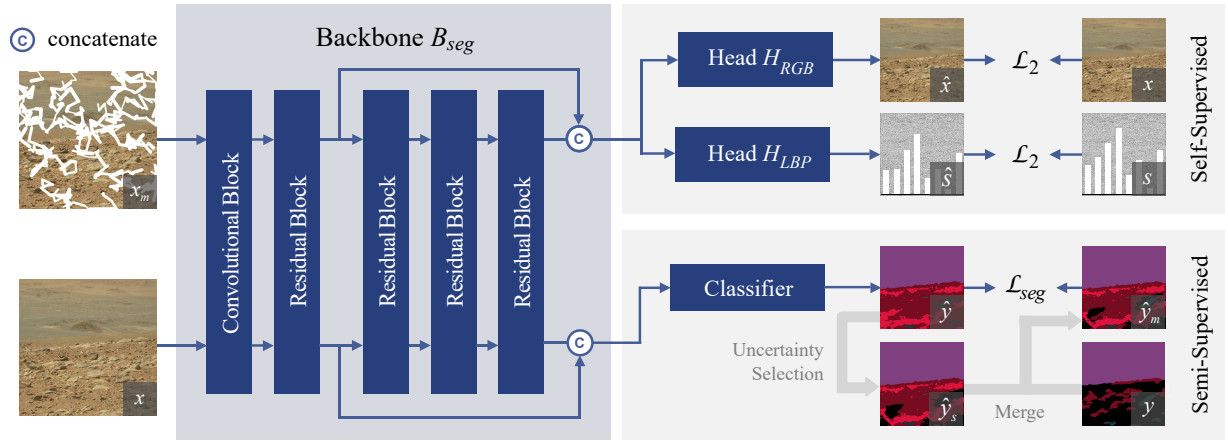


Fig. 4. The framework of our method for Mars image segmentation. The framework can be divided into two stages as a whole, namely, the pre-training stage for representation learning guided by the self-supervised pretext task, and the semi-supervised fine-tuning stage based on pseudo-label fusion. In the self-supervised stage, we utilize the raw pixel value prediction and texture feature prediction on the masked image area to make the network learn the effective feature representation. In the semi-supervised fine-tuning stage, we introduce task uncertainty to generate and select high-quality pseudo-labels, making full use of the supervised information in the unlabeled area of the data.

Therefore, we propose a multi-task mechanism in the pre-training stage, to explicitly guide the model to focus on both the low-frequency color feature and high-frequency texture feature of the image. An overview of our self-supervised architecture is shown in the upper of Fig. 4. In our method, the random masked image is fed into the network which is constrained to predict the original image under texture constraints. Different from the previous works [14, 15, 17], in which the image data usually has distinct colors, clear and mostly regular object contours, we model the masked image in a task scenario with stronger texture correlation.

In the following, a general description of the multi-task modeling algorithm and several options for implementation will be given and discussed.

B. Multi-Task and Joint Learning

Multi-task. Inspired by [17], we explore the potential of introducing traditional operators into existing deep learning frameworks. In our method, a simple yet efficient handcrafted texture feature descriptor, Local Binary Pattern (LBP) [54], is adopted as an additional prediction target to better guide the network to learn texture features. The network is required to predict the raw pixels of masked regions in the input image, and simultaneously optimize the task of LBP prediction.

LBP [54] is a popular texture operator with discriminative power and computational simplicity. It labels the pixels of a grayscale image by thresholding the neighborhood of each pixel and considers the result as a binary number. The values of the pixels in the thresholded neighborhood are multiplied by different weights and summed to obtain the final LBP value. LBP can be seen as a simple Bag-of-Words (BoW) descriptor [55], in which each value corresponds to a word. It can detect the microstructures (*e.g.*, edges, lines, spots, flat areas) whose underlying distribution is estimated by the computed occurrence histogram [56]. Due to its simplicity and efficiency, LBP is widely used in solving computer vision problems, such as texture analysis and face recognition [57, 58].

There are many extensions of LBP [56, 59–63] proposed to improve its performance. In our method, an improved version introducing the "uniform pattern" (ULBP) [56] is used. On the one hand, it drastically reduces feature dimensionality by grouping the "nonuniform" patterns under one label, making it efficient to calculate the loss. On the other hand, a few useful properties are provided like rotation and grayscale invariance, which make the feature more robust to some common variations. To verify the effectiveness of the operator, we conduct a simple experiment on the S^5Mars dataset. The features extracted by different descriptors are fed into an SVM classifier and the output is the terrain category of the corresponding region. We consider the accuracy which reflects the representation ability, and throughput which reflects the computation efficiency, respectively. The results are shown in the Fig. 5. As we can see, ULBP (hereinafter referred to as LBP) performs well and maintains high throughput at the same time. MR8 (Maximum Response Filters) [64], which is also a texture operator based on the filter bank, achieves the highest accuracy among these descriptors. However, the high computational complexity is not satisfactory. We adopt LBP operator to assist the model in modeling texture information.

In the implementation, we first calculate the LBP map of the original image, and then we divide the LBP map into different patches and compute the occurrence histograms separately. After normalized to unit length, LBP histogram $\mathbf{s} \in \mathbb{R}^{C_p \times H \times W}$ is obtained as one of our targets, where C_p is the number of different LBP patterns we defined. We constrain the network to predict the LBP histogram and RGB value of the image mask region simultaneously.

Joint Learning. After being encoded by the backbone network, the feature map of the image is fed into two decoders respectively to predict the corresponding targets. Note that the loss is calculated only on the masked region and the reconstruction of the visible area will not be involved in the calculation. We use L2 loss to minimize the distance between the prediction and the ground truth.

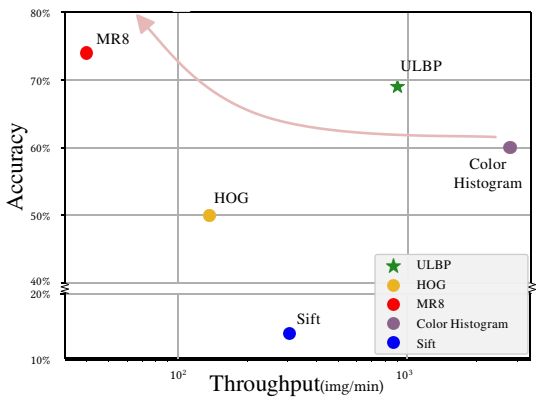


Fig. 5. The results of different descriptors on the S^{5}Mars dataset. Features extracted from the input image are mapped to terrain categories through a SVM classifier. MR8 [64], ULBP [56], Sift [65], HOG [66], and Color histogram are adopted and we consider their recognition accuracy and throughput. The pink arrow represents the tendency towards higher accuracy and longer computation time.

Specifically, given an input image $\mathbf{x}$, the corresponding LBP histogram $\mathbf{s}$, and binary mask $\mathbf{M}$ in which 1 represents the invalid pixel, the overall loss function $\mathcal{L}_{pre-train}$ can be defined as:

$$\mathcal{L}_{inp} = \left\|\left(\mathbf{g}\left(\mathbf{f}\left(\mathbf{x}\odot\left(\mathbf{1}-\mathbf{M}\right)\right)\right)-\mathbf{x}\right)\odot\mathbf{M}\right\|_2, \tag{1}$$

$$\mathcal{L}_{lbp} = \left\|\left(\mathbf{h}\left(\mathbf{f}\left(\mathbf{x}\odot\left(\mathbf{1}-\mathbf{M}\right)\right)\right)-\mathbf{s}\right)\odot\mathbf{M}\right\|_2, \tag{2}$$

$$\mathcal{L}_{pre-train} = \lambda_{inp}\mathcal{L}_{inp} + \lambda_{lbp}\mathcal{L}_{lbp}, \tag{3}$$

where $\mathbf{f}(\cdot)$ is the encoder, and $\mathbf{h}(\cdot)$, $\mathbf{g}(\cdot)$ represent the two decoders respectively. $\mathbf{1}$ stands for an all-ones matrix and $\odot$ is the element-wise multiplication operator.

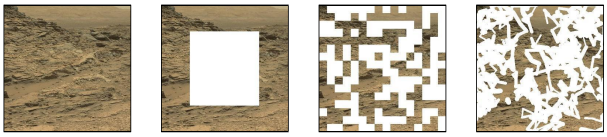
C. Masking Strategy

We apply the masking operation, as shown in Fig. 6, to the original image to generate the input. For masking strategy, we consider these options:

Rectangular Random Masking. A very simple implementation is to generate a rectangular mask directly at the center of the image [14]. However, the features learned by the network are usually difficult to generalize to downstream tasks due to the lack of randomness. Hence we extend the condition to generate a rectangular mask at a random position.

Patch-wise Random Masking. We also follow the recent works based on Transformer [15–17] to divide the whole image into different non-overlapping patches. Patches to be masked are randomly sampled in a certain proportion. In our experiment, the patch size is 32 × 32.

Free-form Random Masking. [67] introduces an efficient and controllable algorithm to generate free-form masks like Fig. 6. It can flexibly generate a variety of masks in the form of lines by controlling the thickness and the number of generated lines. The diversity of masks alleviates the over-fitting problem to a certain extent.



(a) (b) (c) (d)

Fig. 6. Different mask strategies. From left to right: (a) Original image, (b) Rectangular random mask, (c) Patch-wise random mask, (d) Free-form random mask.

V. Semi-Supervised Learning for Mars Imagery Segmentation

The previous chapter mentions that our dataset has two properties: (1) Sparse annotation. (2) High-confidence annotations. To exploit the two properties of the dataset to aid training, we propose an uncertainty-based semi-supervised training strategy. By generating pseudo-labels for unlabeled area in each image, we propose a semi-supervised learning algorithm to exploit the information of unlabeled pixels. Besides, we propose a selection mechanism to select pixels with lower uncertainty to be added to the training set.

A. Task-Uncertainty-Based Pseudo Labeling

In this section, we describe how to estimate uncertainty in data, *i.e.*, task uncertainty. Task uncertainty is the noise introduced during labeling, which is related to the downstream tasks. For segmentation tasks, data on object boundaries have higher task uncertainty because pixels on the boundaries are more difficult to predict segmentation labels. Similar objects, such as rocks and bedrock, are easily confused in labeling. Therefore, there will also be higher task uncertainty. To reduce task uncertainty, in the labeling process, we employ high-confidence labeling. Our annotations avoid unclear boundaries and discard annotations for uncertain objects. This allows us to apply the annotation information to estimate the task confidence for different pixels. Since the labeled data are more confident, while unlabeled data are difficult to annotate, the predicted pseudo labels are noisy and inaccurate. Therefore, we train a discriminator to estimate the task uncertainty by judging whether a pixel is labeled, and add pixels with low task uncertainty to the training data.

B. Estimating Task Uncertainty via the Discriminator

To measure task uncertainty, we train a discriminator to predict the uncertainty of the data. Specifically, we assume that the input image is $\mathbf{x}$, and its annotated pixel data is $\mathbf{y}$. If -1 means unlabeled, we generate labels by $\mathbf{q} = \mathbb{I}[\mathbf{y} \neq -1]$, where $\mathbb{I}[\cdot]$ is the indicator function. Because labeled data and unlabeled data are unbalanced, and most areas of an image are often unlabeled, we employ dice loss to train the discriminator $\mathbf{d}(\cdot)$, which takes the output of the encoder $\mathbf{f}(\cdot)$ as input. Let $\mathbf{p}_{h,w} = \mathbf{d}(\mathbf{f}(\mathbf{x}_{h,w}))$ be the predicted label probability of the pixel located at (h, w), which represents the degree of certainty of the pixel. We formalize the loss function as:

$$\mathcal{L}_{dice} = 1 - \frac{2\sum \mathbf{p}_{h,w}\mathbf{q}_{h,w}}{\sum \mathbf{p}_{h,w} + \sum \mathbf{q}_{h,w}}. \tag{4}$$

C. Reducing Task Uncertainty of Pseudo Labeling

In pseudo-label-based semi-supervised learning, we utilize the encoder $\mathbf{f}(\cdot)$ and classifier $\phi(\cdot)$ trained on labeled areas of data to predict their classes on unlabeled areas as pseudo-labels to aid in training. Moreover, we employ the discriminator $\mathbf{d}(\cdot)$ to pick out the data $\hat{\mathbf{y}}_s$ with high confidence (*i.e.*, $\mathbf{p} > t$, where t is the threshold) in the predicted label $\hat{\mathbf{y}}$. Then, we merge the high-confidence predictions $\hat{\mathbf{y}}_s$ with the ground truth $\mathbf{y}$ to obtain pseudo-labels $\hat{\mathbf{y}}_m$. Given a sample $\mathbf{x}$ in the training dataset $\mathbf{X}$, pseudo-label semi-supervised learning based on task uncertainty can be written as:

$$\mathcal{L}_{pseudo} = -\sum_{\mathbf{x}\in\mathbf{X}}\sum_{\mathbf{f}_{h,w}\in\phi(\mathbf{f}(\mathbf{x}))}\left[\log\frac{\exp(\mathbf{f}_{h,w}^{\hat{c}_i})}{\sum_{c_j=1}^{C}\exp(\mathbf{f}_{h,w}^{c_j})}\right], \quad (5)$$

where $\mathbf{f}_{h,w}^{c_j}$ represents the prediction of this pixel at (h, w) belonging to c_j, and $\hat{c}_i$ is the pseudo-label $\hat{\mathbf{y}}_m$. C is the number of different categories defined in the dataset.

D. Full Model

The whole model architecture is shown in Fig. 4. In the pre-training stage, the two tasks, RGB value prediction, and LBP histogram prediction, are jointly optimized. We first pre-train the model as described in Section IV-B, and obtain the pre-trained model which provides initialization for the fine-tuning stage. In fine-tuning, the optimization objective of segmentation task is cross-entropy loss $\mathcal{L}_{ce}$, which can be defined as:

$$\mathcal{L}_{ce} = -\sum_{\mathbf{x}\in\mathbf{X}}\sum_{\mathbf{f}_{h,w}\in\phi(\mathbf{f}(\mathbf{x}))}\left[\log\frac{\exp(\mathbf{f}_{h,w}^{c_i})}{\sum_{c_j=1}^{C}\exp(\mathbf{f}_{h,w}^{c_j})}\right], \quad (6)$$

where $\mathbf{f}_{h,w}^{c_j}$ represents the prediction of this pixel at (h, w) belonging to c_j, and c_i is the ground truth label $\mathbf{y}$.

In supervised fine-tuning, we optimize the following classification loss:

$$\mathcal{L}_{sup} = \lambda_{ce}\mathcal{L}_{ce}. \quad (7)$$

In semi-supervised learning, the overall loss for the semi-supervised learning stage can be formulated as:

$$\mathcal{L}_{semi} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{pseudo}\mathcal{L}_{pseudo}. \quad (8)$$

VI. Experiments for Self-Supervised Learning

In this section, we evaluate the performance of our proposed self-supervised learning method for terrain semantic segmentation. Note that the fine-tuning process is consistent across all the experiments.

A. Experiment Setup

Our model is based on DeepLabV3+ [32], adopting a ResNet-101 [68] as the backbone. We evaluate the method by fine-tuning the whole model on the semantic segmentation task after self-supervised pre-training. The batch size is set to 12 when pre-training and 16 in the fine-tuning stage. λ_{inp} is set to 0.5 and λ_{lbp} is set to 0.5 for self-supervised learning. The numbers of iterations for pre-training and fine-tuning are 15,000, 50,000 for the S^{5}Mars dataset, and 30,000, 60,000 for AI4MARS dataset respectively. Unless otherwise noted, the image size for training is 512×512.

TABLE II
Segmentation performance on the S^{5}Mars dataset. † indicates the change of backbone.

Method	ACC (%)	MACC (%)	FWIoU (%)	mIoU (%)
Baseline	92.13±0.21	77.97±0.96	85.74±0.34	71.78±0.67
DenseCL [69]	92.22±0.46	79.41±1.08	86.05±0.57	72.88±1.58
PixPro [70]	92.31±0.23	79.87±0.22	86.09±0.38	73.74±0.27
MAE† [15]	92.10±0.19	78.99±1.53	85.79±0.28	72.36±1.21
MaskFeat† [17]	91.91±0.67	79.62±1.18	85.77±0.64	72.89±1.74
Ours	**92.46**±0.19	**81.13**±0.81	**86.41**±0.31	**74.36**±1.33

TABLE III
Segmentation performance on the AI4Mars dataset.

Method	ACC (%)	MACC (%)	FWIoU (%)	mIoU (%)
Baseline	93.70±0.13	73.95±0.19	88.18±0.26	69.02±0.11
MAE† [15]	93.63±0.18	74.91±0.39	88.07±0.33	69.75±0.43
MaskFeat† [17]	93.61±0.12	74.02±0.11	88.16±0.21	68.84±0.12
Ours	**93.82**±0.15	**74.92**±0.31	**88.42**±0.27	**69.83**±0.35

We evaluate performance using the following metrics:

- **Pixel Accuracy** (ACC): It simply computes a ratio between the amount of properly classfied pixels and the total number of pixels.
- **Mean Pixel Accuracy** (MACC): It computes the ACC metrics in a per-class basis and then averages them over all semantic classes.
- **Mean Intersection over Union** (mIoU): This is the widely used metric for semantic segmentation. IoU computes the ratio between the intersection and the union of the ground truth label and the predicted label on a per-class basis. Then IoU is averaged over all semantic classes to get mIoU.
- **Frequency Weighted Intersection over Union** (FWIoU): It is an improved version over the mIoU which weights each class importance depending on their appearance frequency. The weighted sum of the IoU of each class is given as the results.

B. Comparison Results

We compare our model with state-of-the-art self-supervised learning methods, including the MIM-based methods and contrastive learning methods. In our experiment, we report the mean and variance of the experimental results of 5 runs with the same backbone for the S^{5}Mars dataset and 3 runs for the AI4MARS [8] dataset, which has been introduced in Section III. As shown in Table II and Table III, our method achieves the best results on both our dataset and AI4MARS dataset. (Note that we do not use official implementation for [15] and [17] because of the difference of the backbone. Here we transfer them to our backbone and only focus on methodological differences, illustrating them by attaching special notation †.)

TABLE IV
EFFECT OF MASKING STRATEGY WITH CROPPED SIZE OF 256×256.

Mask type	Mask ratio	ACC (%)	MACC (%)	FWIoU (%)	mIoU (%)
Rectangular	30%	**91.06**±0.28	**71.13**±1.07	**83.71**±0.43	**65.10**±1.23
	40%	90.85±0.17	70.28±0.26	83.47±0.14	64.61±0.34
	50%	91.03±0.13	70.29±1.87	83.46±0.30	64.13±0.66
	60%	90.95±0.21	70.24±0.35	83.10±0.30	64.36±0.32
Patch-wise	30%	90.85±0.24	70.74±1.78	83.42±0.41	64.77±1.79
	40%	**91.05**±0.23	70.80±1.42	**83.71**±0.40	**65.13**±1.37
	50%	90.96±0.19	**71.15**±1.08	83.53±0.25	65.10±1.49
	60%	90.95±0.21	70.30±0.36	83.53±0.30	64.56±0.32
Free-form	50%	90.54±0.06	70.58±0.62	83.15±0.27	64.60±0.43
	60%	**91.07**±0.27	**71.29**±0.33	83.59±0.24	**65.27**±0.18
	70%	91.04±0.11	70.68±1.16	**83.66**±0.26	64.75±1.36

TABLE V
ABLATION STUDIES ON MULTI-TASKING.

$\mathcal{L}_{inp}$	$\mathcal{L}_{lbp}$	FWIoU(%)	mIoU(%)
✓		85.79±0.28	72.36±1.21
	✓	85.84±0.61	74.01±1.99
✓	✓	**86.41**±0.31	**74.36**±1.33

TABLE VI
ABLATION STUDIES ON DATA AUGMENTATION. RC: RANDOM CROP. RF: RANDOM FLIP. CJ: COLOR JITTER.

RC	RF	CJ	ACC (%)	MACC (%)	FWIoU (%)	mIoU (%)
✓			92.21±0.25	**81.64**±0.60	86.21±0.19	74.34±0.46
✓	✓		**92.46**±0.19	81.13±0.81	**86.41**±0.31	**74.36**±1.33
✓	✓	✓	92.38±0.14	80.31±1.04	86.25±0.20	73.77±1.78

Compared to [15], which uses the raw pixels as predicted targets, our method introduces an extra prediction target to assist the network to extract identifiable features, thus, achieving better results. [17] uses the HOG feature as the prediction target. The gradient information it focuses on is mainly concentrated at the contours. However, for Mars image data, the contours of many objects are not clear and small objects like broken rocks are often masked completely, which makes the representation learning of the network less effective.

We also test the performance of the contrastive learning methods designed for downstream tasks of dense prediction. [69] proposes a method for dense prediction task, which optimizes a pairwise contrastive loss at both global image level and local pixel. [70] encourages the corresponding pixels of the two different views of an image to be consistent. However, these contrastive learning methods fail to achieve a good performance, because of the lack of consideration of the characteristics of Mars images. There is a high degree of similarity between different terrains on Mars, and data augmentation cannot expand data distribution as expected. It makes the regional contrastive learning that relies only on data transformation without label guidance less effective.

Our method aims to guide the model to extract both the low-frequency signal (*e.g.* the color feature) and high-frequency signal (*e.g.* the texture feature). The multi-task mechanism makes the features of different categories more discriminative and obtains better performance on the Mars dataset. In addition, our method has higher training efficiency than the contrastive learning methods which require a large memory bank to store negative pairs.

C. Ablation Studies

In the following, we give the ablation studies on the S^5Mars dataset. We report the mean and variance of experimental results of at least 3 runs to get a more reliable analysis of our method.

Masking Strategy. We study the effect of the different masking strategies in Table IV. Specifically, we consider fine-tuning performances under different mask types and mask ratios.

We can see that our method is robust for different mask types. One thing to be noted is that different mask types correspond to different optimal mask rates. For the free-form mask, a mask ratio of about 60% is recommended from the results while a relatively low mask ratio of 40% - 50% results in better performance when using patch-wise masks. Meanwhile, a much lower mask ratio of 30% is suitable for rectangular masks. This is because different mask types have different local information loss rates, resulting in different predicted distances [16]. In our experiment, we adopt a free-form masking strategy with a mask ratio of 0.6 by default, which gives the best performance in the experiment.

Multi-tasking. Table V ablates the effect of different combination of task. As we can see, multi-tasking is better than optimizing any single task. Optimization by the single inpainting task often results in insufficient diversity of feature space. Nevertheless, the LBP prediction task focuses on the high-frequency signals of the image, and optimization only by it will lead to unstable training to some extent.

Data Augmentation. We consider the following common augmentations: random crop and flip, and color jittering. As shown in Table VI, unlike most contrastive learning methods, our method does not rely on a large number of data transformations to produce a wide data distribution. The model can still achieve good performance when only the random crop

TABLE VII
ABLATION STUDIES ON LBP IMPLEMENTATION. P, R REPRESENT THE THE ANGULAR RESOLUTION AND SPATIAL RESOLUTION OF THE OPERATOR RESPECTIVELY.

(P, R)	ACC (%)	MACC (%)	FWIoU (%)	mIoU (%)
(16, 2)	91.84±0.58	80.38±0.34	85.84±0.49	73.50±0.63
(24, 3)	**92.46** ±0.19	**81.13**±0.81	86.41±0.31	**74.36**±1.33
(32, 4)	92.36±0.27	80.15±1.05	**86.51**±0.37	73.96±0.68

TABLE VIII
ABLATION STUDIES ON LOSS REGION. $Whole$ MEANS THAT LOSS IS CALCULATED ON THE WHOLE IMAGE, AND $Masked$ MEANS THAT LOSS IS CALCULATED ONLY ON THE MASKED AREA.

Loss region	ACC (%)	MACC (%)	FWIoU (%)	mIoU (%)
$Whole$	92.42±0.17	80.31±0.35	**86.41**±0.26	74.00±1.03
$Masked$	**92.46**±0.19	**81.13**±0.81	**86.41**±0.31	**74.36**±1.33

is applied. However, we observe that color jittering results in performance degradation. We infer that this is probably caused by the concentrated color distribution of the Mars images, and the color jittering leads to a significant difference in color distribution between training data and testing data.

LBP Implementation. We study the effect of different LBP implementations. Specifically, we set different (P, R) values which denote the angular resolution (quantization of the angular space) and spatial resolution of the operator. Following the default setting in [56], we set (P, R) as (16, 2), (24, 3), and (32, 4) respectively in our experiment. The results are summarized in Table VII. The values of P and R directly determine the number of pattern types (the dimension of the LBP histogram to be predicted) extracted by the operator, which is closely related to the representation ability of the operator. When the (P, R) is set too small, the representation power is reduced due to the insufficient number of quantization modes. But when (P, R) is too large, the model performance will degrade in a way due to the introduction of more noise. We finally set the (P, R) to (24, 3) in our experiment.

Loss Region. Table VIII shows the effect of different loss regions in pre-training stage. It is better to calculate loss only in the mask area than on the whole image. Similar results are found in [16]. It indicates that for masked image modeling, the prediction task can obtain a better feature representation than the reconstruction task, which may be because the reconstruction of unmasked areas is relatively simple and is not consistent with the prediction task of masked areas. Features learned by the network are less effective for downstream task fine-tuning due to the wasted part of network capacity [16].

VII. EXPERIMENTS FOR SEMI-SUPERVISED LEARNING

Our approach to semantic segmentation is evaluated in this section as a semi-supervised learning framework. Because our dataset is sparsely labeled, unlabeled areas of the data can be used to further improve the generalization ability and performance of the model.

A. Experiment Setup

As in the setting of supervised fine-tuning, in semi-supervised learning, we adopt DeepLabV3+ [32] as our backbone. The feature dimension of the output for per pixel is 256. We simultaneously pre-train our encoder with a self-supervised task, and then fine-tune our network with two loss functions, segmentation loss and pseudo-label loss. Among them, the segmentation loss is trained for 50,000 iterations. The pseudo-label loss is applied after training for 30,000 iterations.

B. Comparison Results

In Table IX, the results of our method and other semi-supervised segmentation methods are shown. It can be seen that our method always outperforms other methods and achieves better performance.

Mean Teacher utilizes a teacher network to generate pseudo-labels, and its teacher network utilizes an offline momentum method to update parameters. However, such an approach may still produce noisy labels that can harm the training process. Whereas our method reduces task uncertainty in pseudo-labels, guaranteeing small conditional entropy $H(\mathbf{Y}|\mathbf{Z})$. This enhances the inter-class separability in the feature space. ReCo [75] applies contrastive learning to extract the information of unlabeled data, and assigns pseudo-labels to unlabeled data to bring the pixel features of the same category closer and push the pixel features of different categories farther. However, contrastive learning requires numerous positive and negative samples to learn, and the computational cost of pixel-level contrastive learning is huge. Our method employs pseudo-labels to assist training and does not require many positive and negative samples. Compared with some semi-supervised methods based on data augmentation, our method also has advantages. Because of the similarity between different terrains on Mars, traditional data augmentation cannot effectively enhance the distribution of the training set. Our method applies the task uncertainty to reduce the noise in the label data, making the training more effective.

C. Ablation Studies

In this subsection, we conduct experiments on various parts of the semi-supervised method to illustrate its effect and role.

Different Modules. Table X shows the effect of different tasks for semi-supervised learning. We can notice that simply using pseudo-labels does not yield a satisfactory performance boost. This is because the generated pseudo-labels contain a lot of noise and errors, which cannot effectively improve the separability and compactness of features. Better performance is obtained by applying task uncertainty as a constraint, because this removes the uncertainty and the selected pseudo-labels are more accurate. Combining these losses together can achieve the best performance, where the network can better utilize the information from the initialization obtained from self-supervised learning and unlabeled data to obtain stronger generalization performance.

Certainty Thresholds. Table XI shows the effect of different thresholds for semi-supervised learning. We consider the labels of these selected data to be more credible and accurate.

TABLE IX
SEGMENTATION PERFORMANCE ON THE S^{5}MARS AND THE AI4MARS DATASET. † INDICATES IMAGENET PRE-TRAINING.

Method	S^{5}Mars		AI4Mars	
	FWIoU (%)	mIoU (%)	FWIoU (%)	mIoU (%)
Mean Teacher [71] + CutOut [72]	62.98	63.05	75.31	58.01
Mean Teacher [71] + CutMix [73]	62.56	62.62	74.81	57.70
Mean Teacher [71] + ClassMix [74]	62.84	62.89	70.67	53.84
ReCo† [75] + CutOut [72]	76.47	76.38	83.23	68.73
ReCo† [75] + CutMix [73]	76.28	76.16	83.11	68.78
ReCo† [75] + ClassMix [74]	76.70	76.59	83.14	68.77
Ours	**87.18**	**77.20**	**88.82**	**70.64**

TABLE X
ABLATION STUDIES ON DIFFERENT MODULES WITH CROPPED SIZE OF 256 × 256 ON THE S^{5}MARS DATASET.

Modules	ACC	MACC	FWIoU	mIoU
w/o Semi-Supervised Learning	91.07%	71.29%	83.59%	65.27%
Pseudo Label	90.74%	69.57%	83.15%	63.77%
Pseudo Label + Task Uncertainty	**91.73%**	**74.73%**	**84.78%**	**69.38%**

TABLE XI
ABLATION STUDIES ON DIFFERENT CERTAINTY THRESHOLDS WITH CROPPED SIZE OF 256 × 256 ON THE S^{5}MARS DATASET.

Thresholds	ACC (%)	MACC (%)	FWIoU (%)	mIoU (%)
0.3	90.77	71.52	83.41	63.83
0.5	91.16	72.55	83.93	66.48
0.7	91.17	72.64	84.04	66.50
0.9	**91.73**	**74.73**	**84.78**	**69.38**
0.99	91.29	73.06	84.14	65.88
0.999	90.98	71.15	83.78	65.13

As the threshold increases, the segmentation accuracy first increases and then decreases. Because the data selected when the threshold is small introduces too much noise, the pseudo-labels are not accurate enough. When the threshold is too large, less data is selected, which cannot effectively enhance the dataset and improve the generalization ability.

D. Visualizations and Interpretability

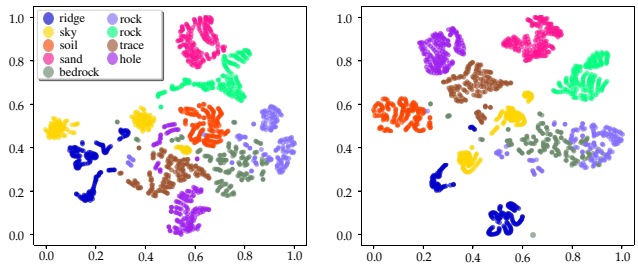
In this part, we give the visualization results to further prove the validity of our method.

Inpainting Results. We give an example of recovered images by the model in the self-supervised pre-training stage. As shown in Fig. 7, the model learns the capability to restore the image, predicting the pixels of masked regions. However, the model prefers to give a blurry solution as we mentioned before, and the edges cannot be well recovered. It is because different images of Mars usually have similar colors and unclear contours, making it difficult for the model to learn good semantic information. Therefore, we introduce the high-frequency texture feature as our extra target to assist the representation learning of the model.

Feature Visualization. The visualization of the output of the last layer in the encoder $\mathbf{f}(\cdot)$ is shown in Fig. 8. Our method

(a) Original image (b) Masked image (c) Restored image

Fig. 7. The inpainting results in self-supervised learning stage.



(a) Baseline (b) Ours

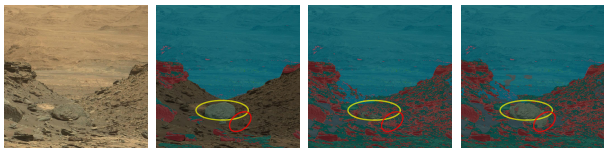
Fig. 8. Visualization of the extracted features for each pixel on the S^{5}Mars dataset. Compared with the baseline model, our method generates a better feature distribution.

enables features learned by the network to be more compact and separable than the baseline algorithm. In our method, self-supervised pre-training provides a good initialization for the fine-tuning stage, which improves the quality of pseudo-labels to some extent. The fine-tuning stage in a semi-supervised manner provides more supervised signals and results in a more separable feature space learned by the network.

Subjective Segmentation Results. We give subjective segmentation results in Fig. 9. The segmentation results of our method is more accurate on the rock in the middle of the presented example. Moreover, benefiting from the semi-supervised approach based on uncertainty estimation, our model has a more reasonable result in the unlabeled area.

E. Failure Case Study and Future Work

Errors occur mainly in confusion between rock and bedrock, while the former is completely exposed to the ground and the latter is often partially covered. It is difficult to distinguish them only by texture information, which is some of the challenging aspects of our dataset. In terms of future work, one direction is to carry out specific designs for difficult categories,



(a) Input (b) Ground Truth (c) Baseline (d) Ours

Fig. 9. Subjective segmentation results compared with the baseline. Compared with the baseline, our method can give more reasonable prediction results in both the labeled region (yellow circle) and the unlabeled region (red circle).

for example, a certain category classifier can be trained to improve the performance.

VIII. Conclusion

We propose a fine-grained annotated dataset for Martian terrain segmentation. This dataset provides sparse and high-confidence labeled data, which effectively assists the subsequent Mars exploration work. To fully utilize this dataset, we propose a self-supervised pre-training, semi-supervised fine-tuning learning mechanism. In the pre-training stage, a multi-tasking mechanism is used to extract features such as shape and texture in the landform, which provides meaningful information for downstream segmentation tasks. In the fine-tuning stage, we adopt a pseudo-label semi-supervised method based on confidence selection, which makes full use of the unlabeled areas in the dataset to improve the generalization ability of the model. Our method finally outperforms previous segmentation algorithms with satisfactory performance gains.

References

[1] D. Qiu, B. Rothrock, T. Islam, A. K. Didier, V. Z. Sun, C. A. Mattmann, and M. Ono, "Scoti: Science captioning of terrain images for data prioritization and local image search," *Planetary and Space Science*, 2020.

[2] F. Goesmann, W. B. Brinckerhoff, F. Raulin, W. Goetz, R. M. Danell, S. A. Getty, S. Siljeström, H. Mißbach, H. Steininger, R. D. Arevalo Jr *et al.*, "The mars organic molecule analyzer (MOMA) instrument: characterization of organic material in martian sediments," *Astrobiology*, 2017.

[3] V. DaPoian, E. Lyness, W. Brinckerhoff, R. Danell, X. Li, and M. Trainer, "Science autonomy and the exomars mission: Machine learning to help find life on mars," *Computer*, 2021.

[4] I. Priyadarshini and V. Puri, "Mars weather data analysis using machine learning techniques," *Earth Science Informatics*, 2021.

[5] R. Gonzalez and K. Iagnemma, "Deepterramechanics: Terrain classification and slip estimation for ground robots via deep learning," *arXiv*, 2018.

[6] M. Dimastrogiovanni, F. Cordes, and G. Reina, "Terrain estimation for planetary exploration robots," *Applied Sciences*, 2020.

[7] S. Schwenzer, M. Woods, S. Karachalios, N. Phan, and L. Joudrier, "Labelmars: Creating an extremely large martian image dataset through machine learning," in *Lunar and Planetary Science Conference*, 2019.

[8] R. M. Swan, D. Atha, H. A. Leopold, M. Gildner, S. Oij, C. Chiu, and M. Ono, "AI4MARS: A dataset for terrain-aware autonomous driving on mars," in *IEEE CVPR Workshops*, 2021.

[9] B. Rothrock, R. Kennedy, C. Cunningham, J. Papon, M. Heverly, and M. Ono, "Spoc: Deep learning-based terrain classification for mars rover missions," in *AIAA SPACE*, 2016.

[10] E. Goh, J. Chen, and B. Wilson, "Mars terrain segmentation with less labels," *arXiv*, 2022.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv*, 2018.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv*, 2019.

[14] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *IEEE CVPR*, 2016.

[15] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *arXiv*, 2021.

[16] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," *arXiv*, 2021.

[17] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," *arXiv*, 2021.

[18] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv*, 2021.

[19] H. Tan, J. Lei, T. Wolf, and M. Bansal, "Vimpac: Video pre-training via masked token prediction and contrastive learning," *arXiv*, 2021.

[20] J. Li, L. Zhang, Z. Wu, Z. Ling, X. Cao, K. Guo, and F. Yan, "Autonomous martian rock image classification based on transfer deep learning methods," *Earth Science Informatics*, 2020.

[21] K. L. Wagstaff, Y. Lu, A. Stanboli, K. Grimes, T. Gowda, and J. Padams, "Deep mars: CNN classification of mars imagery for the PDS imaging atlas," in *IAAI*, 2018.

[22] X. Xiao, M. Yao, H. Liu, J. Wang, L. Zhang, and Y. Fu, "A kernel-based multi-featured rock modeling and detection framework for a mars rover," *IEEE TNNLS*, 2021.

[23] H. R. Kerner, J. F. Bell III, and H. B. Amor, "Context-dependent image quality assessment of jpeg compressed mars science laboratory mastcam images using convolutional neural networks," *Computers & Geosciences*, 2018.

[24] D. R. Thompson and R. Castano, "Performance comparison of rock detection algorithms for autonomous planetary geology," in *2007 IEEE Aerospace Conference*, 2007.

[25] D. R. Thompson, W. Abbey, A. Allwood, D. Bekker, B. Bornstein, N. A. Cabrol, R. Castano, T. Estlin, T. Fuchs, and K. L. Wagstaff, "Smart cameras for remote science survey," 2012.

[26] J. Li, S. Zi, R. Song, Y. Li, Y. Hu, and Q. Du, "A stepwise domain adaptive segmentation network with covariate shift alleviation for remote sensing imagery," *IEEE TGRS*, 2022.

[27] T. Wilhelm, M. Geis, J. Püttschneider, T. Sievernich, T. Weber, K. Wohlfarth, and C. Wöhler, "Domars16k: A diverse dataset for weakly supervised geomorphologic analysis on mars," *Remote Sensing*, 2020.

[28] S. Niekum, "Reliable rock detection and classification for autonomous science," *C. Thesis*, 2005.

[29] S. Higa, Y. Iwashita, K. Otsu, M. Ono, O. Lamarre, A. Didier, and M. Hoffmann, "Vision-based estimation of driving energy for planetary rovers using deep learning and terramechanics," *IEEE Robotics and Automation Letters*, 2019.

[30] F. Furlán, E. Rubio, H. Sossa, and V. Ponce, "Rock detection in a mars-like environment using a cnn," in *Mexican Conference on Pattern Recognition*, 2019.

[31] X. Xiao, H. Cui, M. Yao, and Y. Tian, "Autonomous rock detection on mars through region contrast," *Advances in Space Research*, 2017.

[32] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[33] J. Zhang, Y. Xia, and G. Shen, "A novel deep neural network

architecture for mars visual navigation," *arXiv*, 2018.
[34] M. Ono, B. Rothrock, K. Otsu, S. Higa, Y. Iwashita, A. Didier, T. Islam, C. Laporte, V. Sun, K. Stack *et al.*, "Maars: machine learning-based analytics for automated rover systems," in *2020 IEEE Aerospace Conference*, 2020.
[35] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy, "Progressive neural architecture search," in *ECCV*, 2018.
[36] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *ECCV*, 2018.
[37] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *IEEE ICCV*, 2015.
[38] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *ECCV*, 2016.
[39] B. Fernando, H. Bilen, E. Gavves, and S. Gould, "Self-supervised video representation learning with odd-one-out networks," in *IEEE CVPR*, 2017.
[40] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: unsupervised learning using temporal order verification," in *ECCV*, 2016.
[41] D. Pathak, R. Girshick, P. Doll'ar, T. Darrell, and B. Hariharan, "Learning features by watching objects move," in *IEEE CVPR*, 2017.
[42] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009.
[43] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE TPAMI*, 2015.
[44] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *IEEE CVPR*, 2020.
[45] X. Chen, H. Fan, R. B. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv*, 2020.
[46] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *NeurIPS*, 2020.
[47] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *ICML*, 2020.
[48] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," *arXiv*, 2020.
[49] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, 2020.
[50] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *IEEE CVPR*, 2020.
[51] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, "Confidence regularized self-training," in *IEEE ICCV*, 2019.
[52] D. Mugnai, F. Pernici, F. Turchini, and A. Del Bimbo, "Fine-grained adversarial semi-supervised learning," *ACM TOMM*, 2022.
[53] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE CVPR*, 2021.
[54] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, 1996.
[55] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikäinen, "From bow to cnn: Two decades of texture representation for texture classification," *IJCV*, 2019.
[56] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE TPAMI*, 2002.
[57] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE TPAMI*, 2006.
[58] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *ECCV*, 2004.
[59] M. Pietikäinen, T. Ojala, and Z. Xu, "Rotation-invariant texture classification using feature distributions," *Pattern Recognition*, 2000.
[60] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition," in *IEEE ICCV*, 2005.
[61] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE TPAMI*, 2007.
[62] V. Ojansivu, E. Rahtu, and J. Heikkila, "Rotation invariant local phase quantization for blur insensitive texture analysis," in *IEEE ICPR*, 2008.
[63] L. Liu, S. Lao, P. W. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Median robust extended local binary pattern for texture classification," *IEEE TIP*, 2016.
[64] E. Hayman, B. Caputo, M. Fritz, and J.-O. Eklundh, "On the significance of real-world conditions for material classification," in *ECCV*, 2004.
[65] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.
[66] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE CVPR*, 2005.
[67] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *IEEE ICCV*, 2019.
[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.
[69] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *IEEE CVPR*, 2021.
[70] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *IEEE CVPR*, 2021.
[71] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NeurIPS*, 2017.
[72] T. Devries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv*, 2017.
[73] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *IEEE ICCV*, 2019.
[74] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *IEEE WACV*, 2021.
[75] S. Liu, S. Zhi, E. Johns, and A. J. Davison, "Bootstrapping semantic segmentation with regional contrast," *arXiv*, 2021.