

S⁵Mars: Semi-Supervised Learning for Mars Semantic 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 semantic segmentation is an important Martian vision task, which is the base of rover autonomous planning and safe driving. However, there is a lack of sufficient detailed and high-confidence data annotations, which are exactly required by most deep learning methods to obtain a good model. To address this problem, we propose our solution from the perspective of joint data and method design. We first present a new dataset S⁵Mars for Semi-Supervised learning on Mars Semantic Segmentation, 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 semi-supervised learning (SSL) framework for Mars image semantic segmentation, to learn representations from limited labeled data. Different from the existing SSL methods which are mostly targeted at the Earth image data, our method takes into account Mars data characteristics. Specifically, we first investigate the impact of current widely used natural image augmentations on Mars images. Based on the analysis, we then proposed two novel and effective augmentations for SSL of Mars segmentation, *AugIN* and *SAM-Mix*, which serve as strong augmentations to boost the model performance. Meanwhile, to fully leverage the unlabeled data, we introduce a soft-to-hard consistency learning strategy, learning from different targets based on prediction confidence. Experimental results show that our method can outperform state-of-the-art SSL approaches remarkably. Our proposed dataset is available at <https://jhang2020.github.io/S5Mars.github.io/>.

Index Terms—Mars vision tasks, terrain segmentation, image semantic segmentation, 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 semantic segmentation, which aims to identify the drivable areas and the specific terrains from images. 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 semantic segmentation faces problems from both data and method design. First, the lack of satisfactory and available

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2018AAA0102702, and in part by the National Natural Science Foundation of China under Grant 62172020. (*Corresponding author: Jiaying Liu*)

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

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, delicate and dense pixel-level labeling 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 the actual requirements of complex terrain identification. Besides, some datasets [7, 8] collected through crowdsourcing often do not have satisfactory annotation quality due to inconsistent standards.

From a methodological point of view, the existing methods heavily rely on large amounts of training data and lack targeted and effective design. Early works directly applied a certain machine learning algorithm such as Support Vector Machines (SVM) [6]. With the rapid development of deep learning, the terrain segmentation performance is greatly improved by methods based on deep neural networks [5, 8, 9]. However, they still rely on fully supervised learning pipelines that require a lot of high-quality labeled data, which is often difficult to achieve. To this end, semi-supervised learning (SSL) has attracted lots of attention, which learns representations from limited labeled data as well as the amounts of unlabeled data. However, most existing SSL methods are designed for Earth image data and cannot be directly transferred to Mars image segmentation tasks, due to the properties of Mars images. First, the color of Mars images is less diverse. Traditional color augmentations, which are crucial and widely used in SSL works [10–12], can cause over-distortion problem [13] in the form of color distribution shift for the Mars images, and fail to improve the performance as shown in Fig. 5. Note that color distribution shift can arise in different data domains with some similar properties as the Mars images, *e.g.*, less diverse color distributions, when applying the traditional color augmentations. Nevertheless, it is still less explored in previous SSL works, especially from the perspective of data augmentations. Besides, the objects in Mars images are often with irregular contours and obvious occlusions, *e.g.*, between rocks and soil/sand. As a result, the high background complexity makes the model suffer from greater uncertainty in the consistency learning of unlabeled data, leading to the sub-optimal representations. Moreover, some categories are more confusing between each other, *e.g.*, rocks and bedrocks, soil and sand, which require more fine-grained representations to

distinguish.

In summary, there are two main challenges in the Mars terrain segmentation task: 1) the lack of data with adequate detailed and high-confidence annotations, 2) insufficient studies targeted at SSL on Mars image data. We solve the above problems from the perspective of *both data and method design*, which are named **Semi-Supervised Semantic Segmentation for Mars (S⁵Mars)**. We first create a new dataset to provide a high-quality and fine-grained labeled data 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 adopts a sparse labeling style, *i.e.*, only the area with high human confidence is annotated.

To learn from this sparse data, we propose a new semi-supervised framework for Mars image terrain segmentation. Our method is based on the recently popular consistency regularization-based methods, which utilize weak-to-strong augmentations to generate the perturbation while pursuing the perturbation consistency. Specifically, we first investigate the impact of widely used Earth image augmentations on Mars data and are surprised to find their adverse effects on the SSL of Mars segmentation. Based on this analysis, we further propose two novel and effective augmentations, *AugIN* and *SAM-Mix*. *AugIN* exchange statistics between images to generate new data views while avoiding drastic color distribution shift. *SAM-Mix* utilizes the pretrained Segment-Anything Model (SAM) [14] to generate high-quality object masks, reducing the uncertainty of the mixed images. These two data augmentations lead to better consistency learning and improve the performance remarkably. Finally, we introduce the soft-to-hard consistency learning strategy, which utilizes the soft pseudo-labels in low-confidence regions, while using the hard pseudo-labels in high-confidence regions, fully taking advantage of the unlabeled data. Extensive experiments and ablation studies verify the effectiveness of the proposed method.

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 systematically study the data augmentations used in current mainstream SSL methods and find their detrimental impact on Mars image segmentation, especially the traditional color augmentations. We analyze this problem and further propose two new and effective augmentations, *SAM-Mix* and *AugIN*, boosting the performance of SSL methods for Mars image segmentation.
- To fully take advantage of the unlabeled data, a soft-to-hard consistency learning strategy is introduced. The model is constrained to learn consistency by the hard pseudo-labels in high-confidence regions as well as the soft pseudo-labels in low-confidence regions, further improving the consistency.

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 proposed Mars segmentation dataset. Then we present our framework for Mars semantic segmentation in Section IV. Experimental results and analysis are shown in Section V. The conclusion is finally given in Section VI.

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 [16] 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+ [28]. Considering the dependence of existing methods on large amounts of data, [29] utilizes a self-supervised method and trains the model on less labeled images. Recently, Transformer-based network is studied [18, 30] for Martian rock segmentation task. For other tasks, Zhang *et al.* [31] 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 [32]. 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.* [25] propose to predict energy consumption from images based on a PNASNet-5 [33].

However, many existing works still directly transfer the technology designed for the Earth scene to the Mars task, which can be sub-optimal due to the properties of Mars data. Meanwhile, due to the significant bandwidth and computational resource limitations, the model is expected to be lightweight and efficient, and hence the large models are unsuitable to employ. Most importantly, most of these methods require a lot of annotated training data, which is expensive and hard to obtain. Although some domain adaptation methods [22, 34] also can learn the target domain knowledge without many labels, they still suffer from taxonomy inconsistencies in segmentation detail as discussed in [29, 35]. To this end, in this paper we present a powerful semi-supervised learning framework designed for the Mars images.

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
		[15]	620	4	Terrain classification
		[16]	6k	24	Terrain classification
		[17]	405	-	Rock detection
		[18]	8k	-	Rock detection
		[1]	1k	-	Image description
		[19]	310k	-	Compressed image quality evaluation with automatic labeling
		[20]	117	-	Rock detection
Real + Synthetic	Curiosity, Opportunity, Spirit rovers	[21]	46	2	Terrain segmentation
		[8]	35k	4	Terrain segmentation
		[22]	5k	9	Terrain segmentation
		[7]	5k	6 (17 sub)	Terrain segmentation
		[23]	30k	5	Terrain classification
Synthetic	ROAMS rover simulator	[20]	55	-	Rock detection
Simulation field	Atacama Desert Zoë rover prototype	[24]	30	-	Rock detection
	JPL Mars Yard FIDO rover Platform	[20]	35	-	Rock detection
	JPL Mars Yard Athena rover Platform	[25]	91k	-	Rover energy consumption
	Devon Island	[26]	400	-	Rock detection
Real + Simulation field	Opportunity, Spirit rovers	[27]	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. 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, which are shot by the Mars rovers. 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. Semi-Supervised Learning

Semi-supervised learning [36] utilizes the manifold structure of unlabeled data to assist learning with labeled data. The key issue is how to exploit the information of unlabeled data.

Generally, the cross-entropy loss is optimized by the ground-truth label on the labeled data, while a regularization term is applied to the model *w.r.t* the unlabeled data. For example, the pseudo-label method [37] assigns pseudo-labels to unlabeled data through a classifier trained on supervised data, which is typical in entropy minimization methods.

Regarding the utilization of unlabeled data, many researchers have conducted extensive studies, covering unsupervised contrastive learning [38, 39], uncertainty attention mechanism [39, 40], and extra correcting networks [41]. However, these methods improve the performance at the cost of increasing space and computational complexity. Recently, consistency regularization-based methods have attracted lots of attention, due to their simplicity and effectiveness. They rely on various perturbation techniques (augmentations) to generate different data patterns, which maintain similar semantic information as the original data. Then, the consistency regularization objective is performed to guide the learning of the unlabeled data. MixMatch [42] mixes labeled and unlabeled data using MixUp [43] and performs consistency regularization utilizing low-entropy labels. FixMatch [10] further assigns pseudo-labels, which are the predictions by teacher model on weakly augmented data, to the corresponding strongly augmented data. Inheriting from FixMatch, FlexMatch [44] and FreeMatch [45] propose to learn the threshold for different classes adaptively to filter the low-confidence pseudo-labels. Zhao *et al.* [11] proposed a series of strong data augmentations to enhance the augmented space. UniMatch [12] utilizes both

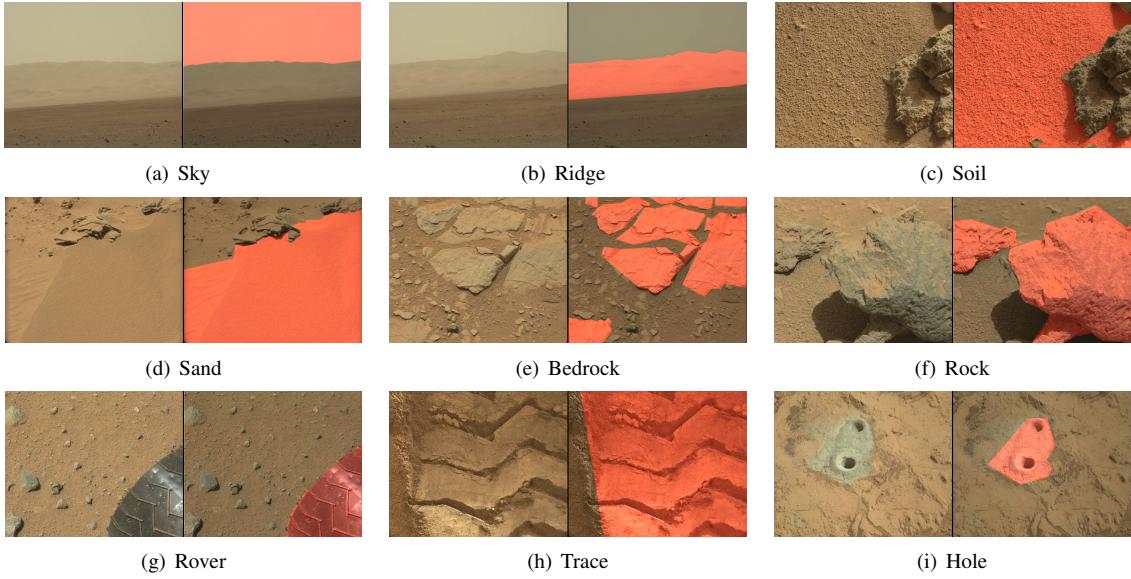
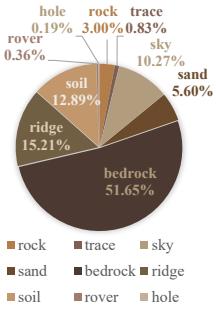


Fig. 1. Examples for each label category (highlighted in red).



(a) The number of images with n categories appearing simultaneously.



(b) The distribution of different label area.

Fig. 2. Numerical statistics on our S^5 Mars dataset. The figures show the richness of the categories from two aspects: distribution of the number of different labels in each image and distribution of label area. Note no image contains 9 labels simultaneously in its annotation, so it is omitted in (a).

the data-level and feature-level augmentations to constrain the consistency learning.

In these consistency regularization methods, the augmentations, *i.e.*, the perturbation techniques, are crucial for the semantic segmentation. Many techniques, *e.g.*, geometric-based, color-based, mixing-based, and feature perturbation-based methods have been studied. Furthermore, some random auto-augment modules are developed [46] to further expose data patterns. However, these methods are less suitable and effective for the Mars semantic segmentation, due to the special properties of Mars images as we discussed in Section I. Therefore, it is significant and critical for the study of augmentations for Mars image data. In this paper, we analyze the characteristics of Mars images, and study the performance of existing common augmentations on Mars data. Meanwhile, we propose two effective new augmentations to boost SSL for Mars image segmentation.

III. PROPOSED MARS IMAGERY SEGMENTATION DATASET

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, namely, S^5 Mars. Our dataset includes 6,000 high-resolution images taken on the surface of Mars, by color mast camera (Mastcam) from Curiosity (MSL), with the spatial resolution of 1200×1200 . The dataset is divided in a roughly stratified sampling manner to make the label distribution similar among different splits, yielding a training set of 5000 images, a validation set of 200 images and a test set of 800 images.

A. Labeling Process

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.
- **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.

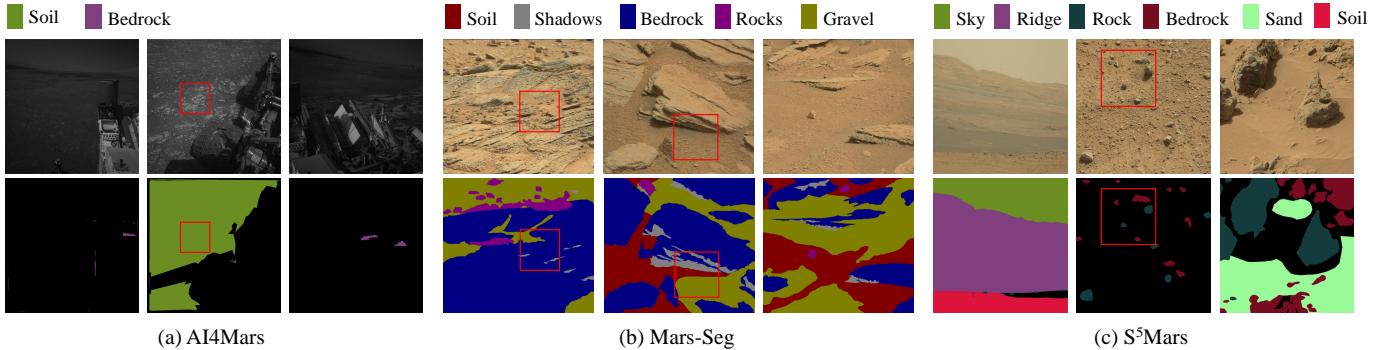


Fig. 3. Some image-label examples in different datasets: (a) AI4Mars [8]. Due to the few defined categories, the annotation diversity and adequacy are insufficient. Meanwhile, there are some cases of mislabeling (red box). (b) Mars-Seg [22], which gives a complete pixel-level labeling. However, the label can be misleading when different categories mix up with each other (red box). (c) Our dataset S⁵Mars, which provides accurate labeling for regions with high confidence.

- **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*, i.e., 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, and the total pixel annotation ratio is 48.9%. As for the annotating process, the annotation rules are discussed more than ten times 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, where 90% of the annotators have been engaged in such annotation work more than six times. 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 different 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). The total pixel-wise label ratio is 49%. For the labeled regions, 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.

AI4Mars contains 4 categories with gray-scale images available solely, which can only provide limited task knowledge. Moreover, since AI4Mars is a crowdsourcing project, though the number of submissions is large, the annotators may have

inconsistent understandings of labeling standards, which can lead mislabeling in the annotations as shown in Fig. 3(a). In contrast, our dataset is equipped with high-resolution RGB images including 9 semantic categories. Meanwhile, we establish clear labeling criteria and provide professional training to annotators, making the proposed dataset more reliable.

Mars-Seg [22] 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 S⁵Mars is composed of high-resolution RGB images, which offers more accurate and more abundant semantic information for detection and segmentation tasks. Meanwhile, categories in Mars-Seg like gravel, sand, and rocks mix up with each other, making it hard to determine the terrain scene into any one category, as shown in Fig. 3(b). Instead, S⁵Mars applies confidence-based sparse-labeled manner. This way we guarantee the labels are strongly representative in each category and reduce the label noise introduced in the labeling work, as shown in Fig. 3(c).

IV. THE PROPOSED METHOD

In this section, we introduce the proposed method for Mars image semantic segmentation. The overview and motivations are first provided in Section IV-A. Then, we systematically investigate the augmentations for Mars images in Section IV-B, and propose two effective augmentation techniques based on the analysis. Finally, in Section IV-C, we introduce the soft-to-hard consistency learning strategy and present the full model.

A. Preliminaries and Motivation

1) Overview: As introduced in the previous sections, our proposed dataset is annotated in a sparse style, *i.e.*, some areas of an image are annotated and some are not. For clarity, we no longer distinguish between unlabeled images and unlabeled areas in an image, which can be aligned with a few minor changes. Following the dominant consistency regularization semi-supervised methods [10, 36], the model is trained on both labeled and unlabeled images simultaneously. Given a batch of labeled images $\mathcal{B}_l = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{B}_l|}$ and a batch of unlabeled images $\mathcal{B}_u = \{(\mathbf{u}_i)\}_{i=1}^{|\mathcal{B}_u|}$, the goal of SSL is to train a model

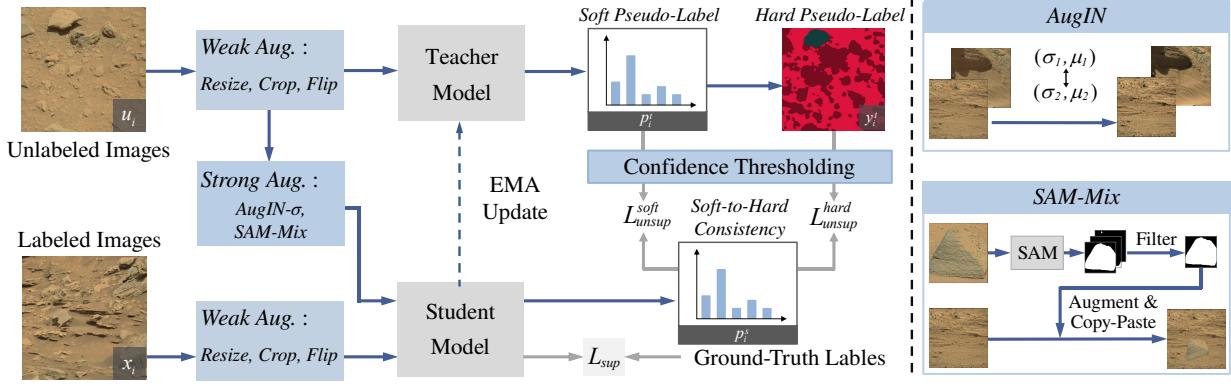


Fig. 4. The overview of the proposed framework for semi-supervised Mars semantic segmentation. We adopt a two-branch teacher-student architecture. Two novel augmentations are proposed as strong augmentations, *AugIN* and *SAM-Mix*. *AugIN* exchanges the statistics of the two samples, *i.e.*, mean and standard deviation. *SAM-Mix* utilizes an off-the-shelf SAM to obtain the object binary masks to perform copy-paste operation, reducing the uncertainty of the augmented images. Finally, the model is optimized according to a soft-to-hard consistency learning strategy, utilizing both the soft labels \mathbf{p}_i^t and the hard labels \mathbf{y}_i^t based on the confidence.

$f(\cdot; \theta)$ with good representations by optimizing the following objective \mathcal{L} :

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_u \mathcal{L}_{unsup}, \quad (1)$$

where \mathcal{L}_{sup} is the supervised loss on the labeled images, *i.e.*, the cross-entropy loss, and the \mathcal{L}_{unsup} is the unsupervised loss for unlabeled images. λ_u controls the weight of unsupervised term.

Our method is based on the recent popular consistency regularization-based SSL method, FixMatch [10]. Specifically, a two-branch network is adopted, consisting of a teacher model $f(\cdot; \theta_t)$ and a student model $f(\cdot; \theta_s)$. The teacher model $f(\cdot; \theta_t)$ can be identical to the student model sharing the same weights. Alternatively, it can be updated gradually via the exponential moving averaging (EMA) of the student model weights:

$$\theta_t \leftarrow m\theta_t + (1 - m)\theta_s, \quad (2)$$

where $m \in [0, 1]$ is the momentum coefficient. We follow the EMA setting to update the teacher model, which is also recommended in Mean-Teacher [47]. The student model is optimized via the backward gradients.

The core implementation in FixMatch is the weak-to-strong augmentation strategy, which serves as the perturbations and generate different augmented data views. Specifically, given the weak augmentations \mathcal{T}_w and strong augmentations \mathcal{T}_s , the augmented views $\mathbf{u}_i^w = \mathcal{T}_w(\mathbf{u}_i)$, $\mathbf{u}_i^s = \mathcal{T}_s(\mathbf{u}_i)$ are constructed and fed into the teacher and student model to encode, respectively. The teacher model assigns the pseudo-labels for weakly augmented images, which are then utilized in the learning of student model for strongly augmented images. Concretely, the unsupervised consistency loss can be formulated as follows:

$$\mathcal{L}_{unsup}^{ce} = \frac{1}{|\mathcal{B}_u|} \sum_{i=1}^{|\mathcal{B}_u|} \frac{1}{H \times W} \sum_{j=1}^{H \times W} \mathcal{L}_{ce}(\mathbf{p}_i^s(j), \mathbf{y}_i^t(j)), \quad (3)$$

$$\mathbf{y}_i^t(j) = \mathbb{1}(argmax(\mathbf{p}_i^t(j))), \quad (4)$$

where $\mathbf{p}_i^s(j)/\mathbf{p}_i^t(j)$ is the predicted scores output by the student/teacher model after softmax layer corresponding to the

j th pixel of the i th unlabeled image \mathbf{u}_i . $\mathbf{y}_i^t(j)$ is one-hot encoding of the pseudo-label generated from the teacher model and $\mathbb{1}$ is the one-hot indicator function. H and W are the height and width of the image. \mathcal{L}_{ce} is the cross-entropy loss function.

2) Motivation: For SSL in Mars image semantic segmentation, there are two main challenges to be solved: (a) Previous augmentations for the natural images on Earth can be ineffective due to the different properties of Mars images. (b) The unlabeled regions of the Mars images tend to be with high uncertainty, making the pseudo-labels less reliable for training. These problems affect the performance of the existing SSL frameworks for Mars image segmentation. To overcome these challenges, we propose a simple yet effective SSL framework, as shown in Fig. 4, which adopts effective augmentations and learns semantic representations by exploring soft-to-hard consistency, which will be introduced in the following parts.

B. Augmentations for Mars Images

As pointed in previous works [10–12], the augmentation module plays an important role in SSL, encouraging the model to learn the consistency in the perturbations. Generally, the common augmentations adopted for SSL methods can be divided into the following categories:

- **Geometrical Augmentation.** It utilizes some geometrical transformations, *e.g.* *Flip* and *Translate*, to generate new data views. These augmentations often serve as the basic augmentations, *i.e.*, the weak augmentations, due to their efficiency and stability.
- **Noise-Based Augmentation.** Different augmented views can be obtained by simply injecting random noise into the original image, *e.g.*, *Gaussian Noise*, and *Random Mask*.
- **Color-Based Augmentation.** A series of color transformations are introduced to further enlarge the data distributions, *e.g.*, *Gaussian Blur*, *Equalize*, and *Sharpness*. More details can be found in [46]. These transformations facilitate the model to learn the intrinsic semantic consistency by perturbing the color distribution of images.

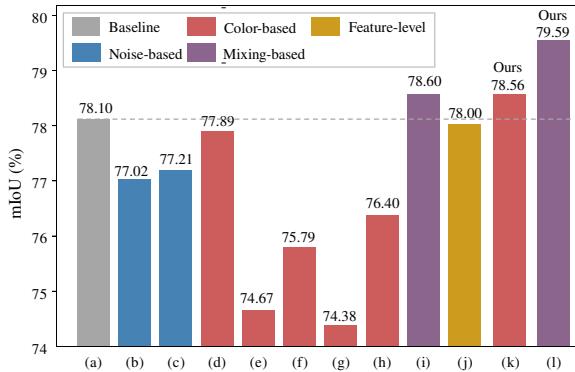


Fig. 5. Comparison of different augmentations on SSL for Mars segmentation. (a): Identity, (b): *Gaussian Noise*, (c): *CutOut*, (d): *Gaussian Blur*, (e): *Hue*, (f): *Contrast*, (g): *Equalize*, (h): *Brightness*, (i): *CutMix*, (j): *Dropout*. (k) and (l) are the proposed *AugIN* and *SAM-Mix*.

TABLE II

COMPARISON OF THE STATISTICAL INFORMATION BETWEEN THE MARS IMAGES AND EARTH IMAGES. LOWER VALUES OF THE METRICS INDICATE LESS DISPERSION OF THE DATA DISTRIBUTION.

Dataset	(R, G, B)	
	Standard Deviation	Variable Coefficient
S ⁵ Mars	(0.134, 0.121, 0.099)	(0.214, 0.233, 0.273)
ImageNet [52]	(0.229, 0.224, 0.225)	(0.472, 0.491, 0.554)

- **Mixing-Based Augmentation.** Mixing methods have been proven effective for SSL scenarios. They mix the two samples via the interpolation (*Mixup* [43]) or cut-paste (*CutMix* [48]) operations. Some advanced mixing methods are further developed for SSL such as *CowMix* [49] and *ClassMix* [50].
- **Feature-Level Augmentation.** The most common augmentation in feature-level is *Dropout* [51] operation, which can also be regarded as a kind of model perturbation. It is often utilized as strong augmentations in conjunction with other augmentations.

We focus on the latter four, which serve as strong augmentations and have a significant impact on model performance. Following the recent work [11], we adopt *Resize*, *Crop*, and *Flip* as the weak augmentations. In addition, we choose different augmentations, which are commonly used and found beneficial for the learning of Earth images, as strong augmentations to demonstrate their impact separately. The results are shown in Fig. 5. As we can see, unlike natural images on Earth, the noise-, color-based and feature augmentations cannot bring a boost compared with the “identity” baseline. To further understand this phenomenon, we analyze the data from a statistical perspective and present the comparison of standard deviation and coefficient of variation between Mars and Earth images. As shown in Table II, the dispersion of RGB values in the Mars image is much less than that of Earth natural image, which indicates that the color distribution of the Mars image is more concentrated. It is in line with our observation that there is a high similarity within and between Mars images. Based on this conclusion, we argue that the traditional color-based per-



Fig. 6. Examples of color-based augmented images. More details of the augmentations can be found in [46].

turbations lead to the color distribution shift of Mars images, causing the over-distortion problem [13] as shown in Fig. 6, which is not conducive to the model segmentation learning. Note that this is not trivial in the context of SSL because most previous SSL works adopt the color augmentations as a strong technique by default and lack specific consideration on the Mars images. Meanwhile, we empirically find that the feature perturbation *Dropout* also fails to improve the performance, because it does not generate new input samples and cannot help the model learn richer semantic information. Besides, due to the irregular objects with occlusions and unclear contours, the model can face more serious uncertainty and consistency learning difficulty under the noise-based and random mixing-based augmentations, which will be discussed in the following.

To this end, we propose two effective augmentations designed for Mars images, *AugIN* and *SAM-Mix*, and employ them in our method to boost the SSL performance.

1) AugIN. To avoid drastic changes in image color distribution caused by direct perturbation, we propose *AugIN* (Augment Instance Normalization), which generates augmented data views by exchanging statistics of different images, *i.e.*, the mean and standard deviation. This is inspired by the successful practice of style transfer [53]. Specifically, given a image \mathbf{u}_i and a randomly sampled image \mathbf{u}_j , we exchange the mean and standard deviation as follows:

$$\text{AugIN}(\mathbf{u}_i, \mathbf{u}_j) = \sigma(\mathbf{u}_j) \left(\frac{\mathbf{u}_i - \mu(\mathbf{u}_i)}{\sigma(\mathbf{u}_i)} \right) + \mu(\mathbf{u}_j), \quad (5)$$

where the $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation functions. Meanwhile, we can spontaneously obtain the two variants, *AugIN- μ* and *AugIN- σ* , which only exchange the mean or standard deviation between two samples. In the implementation, we exchange the image statistics within the same batch following a randomly generated permutation. Note that the operation in our method that exchanges the statistics of images within the same batch does not change the statistics of the entire batch, which can be theoretically verified easily. This stabilizes the color distribution after augmentation and generates more reasonable augmented data. In contrast, traditional color augmentations change the statics directly without considering the whole color distribution, making the model suffer from the potential color distribution shift problem.

2) SAM-Mix. As shown in Fig. 5, *CutMix* achieves a modest performance gain over the baseline, failing to meet the expected level of improvement. This is because there are many fragmentary objects with unclear edges in Mars images, and random cut-pasting manner may lead to high uncertainty, limiting the model performance. To this end, we propose *SAM-Mix*, which is formulated as a generalization of *CutMix* using



Fig. 7. Examples of image-mask pairs. We show the filtered masks of each image with high predicted confidence. Note these masks are output by SAM in an instance-wise manner, and we illustrate them in different colors.

binary masks output by an off-the-shelf Segment-Anything Model (SAM) [14].

SAM attracts lots of attention recently, which can produce binary masks for the objects in an image from input or randomly generated prompts. We utilize an off-the-shelf SAM to produce a mask of the target object and paste it into the source image. Compared with random rectangular mask generation, SAM can generate high-quality masks to segment specific objects as shown in Fig. 7. Specifically, given an image, a list of binary masks with the corresponding confidence score is output by SAM. These masks are first filtered so that 1) the size of the mask is limited to a certain range and 2) the confidence of the mask is above a certain threshold. If there is no qualified mask, a random rectangular mask will be directly generated. Then a Gaussian filter is applied to the masks to eliminate possible noise. Subsequently, we randomly select a qualified mask and further transform the masked object, *i.e.*, *Rotation*, *Flip*, and *Rescaling*. The pasting position will not be adjusted, that is, it will generally pasted corresponding to the position of the original image, to avoid some unreasonable cases, *e.g.*, the sky appearing in the bottom half of the image. The corresponding segmentation labels are also generated in the same way, which are used to train the model as previous work [54].

We note that *SAM-Mix* shares similarities with other segmentation-based mixing augmentation strategies [50, 55–58], which develop the binary mask generation in an instance-wise or class-wise manner. However, in contrast to the above mixing methods, *SAM-Mix* gets rid of the reliance on ground-truth labels, making it possible for the augmentation on unlabeled images. Furthermore, SAM’s strong generalization ability enables us to produce high-quality masks for individual objects efficiently, which is compatible with images of Mars that contain multiple objects simultaneously. *SAM-Mix* reduces the uncertainty caused by random mixing and further improves the performance of the model.

C. Soft-to-Hard Consistency Learning

As mentioned in the previous section, the Mars images are with more confusing categories, such as sand and soil, rock and bedrock, which require a more fine-grained representation learning target, especially for the unlabeled regions with high uncertainty in our dataset. Meanwhile, for the data collection and annotation, it is more difficult to obtain large-scale and high-quality annotated Mars images than natural Earth images, due to the complexity of the Mars terrain, the required expert knowledge, and the limited transmission bandwidth. Therefore, previous works using only unlabeled regions with

high confidence for training can be sub-optimal in the Mars SSL context.

To this end, we propose a soft-to-hard consistency learning strategy, which utilizes both the soft and hard pseudo-labels according to a confidence thresholding policy. The hard pseudo-label is the one-hot label representation $y_i^t(j)$ in Eq. 3, which is obtained by the *argmax*(\cdot) operation. The soft label is represented as the model prediction scores $\mathbf{p}_i^s(j)$, which denotes the probability distribution over different semantic categories. Specifically, the optimization objective for the soft pseudo-label can be formulated as:

$$\mathcal{L}_{unsup}^{soft} = -\frac{1}{|\mathcal{B}_u|} \sum_{i=1}^{|\mathcal{B}_u|} \frac{1}{H \times W} \sum_{j=1}^{H \times W} \mathbf{p}_i^t(j) \log(\mathbf{p}_i^s(j)). \quad (6)$$

Intuitively, Eq. 6 optimizes the similarity of the two distributions, *i.e.*, p_i^t and p_i^u , which indicate the predicted class probability of the teacher and student model. Based on this objective, we can further find that:

- When $\max(\mathbf{p}_i^t(j)) \approx 1$, the teacher model assigns the pseudo-labels with high confidence, and the Eq. 6 degenerates to be almost equivalent to Eq. 3.
- When $\max(\mathbf{p}_i^t(j)) < 1 - t$ where t is a positive constant, the predictions of the teacher model is less confident. This objective encourages the student model to learn the consistency measured by the relevance of current features to different prototype anchors. This can be seen as a more fine-grained smooth label of the unknown regions in Mars images, which can belong to a new class or the old class with high uncertainty.

Therefore, the hard label provides a confident target to force the model to predict a distribution with low entropy, learning the explicit semantic mapping in images. In contrast, the soft label objective encourages the model to learn the consistency in a more gentle way, which can be viewed as performing the self-distillation [59] of relational knowledge, modeled as the feature similarity to the prototype features stored in the weights of the classification head. This allows the model to make better use of unlabeled data to improve the representation consistency learning in an unsupervised manner, achieving a better representation space.

Based on the above analysis, we propose a confidence-based thresholding policy to integrate the two objective functions organically. We utilize the hard pseudo-labels in high-confidence regions while using soft pseudo-labels in low confidence regions, fully taking advantage of the training signals from unlabeled data. Specifically, we first obtain the confidence score of teacher model predictions as $\max(\mathbf{p}_i^t(j))$. Then, the student model is optimized as follows (t_{hard} and t_{soft} are the threshold hyper-parameters):

- 1) If $\max(\mathbf{p}_i^t(j)) > t_{hard}$, Eq. 3 is applied to optimize the model with the highly confident one-hot pseudo-label;
- 2) If $\max(\mathbf{p}_i^t(j)) < t_{soft}$, the soft label objective is optimized, to avoid noisy signals from other prototype features in the high confidence region.

Finally, the model is optimized in an end-to-end manner using the objective in Eq. 1. The supervised term \mathcal{L}_{sup} is

TABLE III
SEGMENTATION PERFORMANCE ON THE S⁵MARS USING THE RESNET-50 AS THE BACKBONE.

Method	20% data		50% data		100% data	
	mAcc (%)	mIoU (%)	mAcc (%)	mIoU (%)	mAcc (%)	mIoU (%)
Supervised	76.71	66.26	80.17	72.65	82.12	75.04
MT [47]	76.89	70.92	81.95	75.93	82.32	77.87
MT [47] + ClassMix [50]	77.38	71.54	82.34	76.42	83.21	78.04
FixMatch [10]	77.80	71.08	79.47	72.98	80.75	73.69
RanPaste [61]	76.36	66.66	80.51	74.08	82.28	75.78
U ² PL [38]	78.03	72.32	82.29	77.56	83.41	78.60
AugSeg [11]	78.28	72.38	81.52	75.40	81.82	76.90
UniMatch [12]	76.92	69.83	78.53	71.56	79.82	72.60
Ours	82.85	76.49	83.56	78.66	84.73	80.15

the cross-entropy loss on the labeled images. The whole consistency regularization term \mathcal{L}_{unsup} is:

$$\mathcal{L}_{unsup} = \mathcal{L}_{unsup}^{hard} + \lambda_s \mathcal{L}_{unsup}^{soft}, \quad (7)$$

where $\mathcal{L}_{unsup}^{hard}$ is exactly the \mathcal{L}_{unsup}^{ce} in Eq. 3 and λ_s is the weight coefficient.

V. EXPERIMENTS AND RESULTS

A. Dataset

We use the proposed S⁵Mars dataset and AI4Mars dataset, which are introduced in the Section III. For semi-supervised learning evaluation, we adopt a stratified sampling strategy to extract different proportions of data from the dataset as labeled data and the rest as unlabeled data, to jointly train our model. Note that all methods are evaluated under the same data partition lists.

B. Implementation Details and Metrics

Our model is based on DeepLabV3+ [28], adopting a ResNet-50 [60] pre-trained on Image-Net [52] as the segmentation backbone. We use an output stride of 16 by default. The batch size is set to 8. An SGD optimizer with a momentum of 0.9. A polynomial learning-rate decay with an initial value of 0.01 are adopted to train the student model. Specifically, the learning rate is scaled by $(1 - iter/max_iter)^{0.9}$. The EMA momentum coefficient m is set as $\min(1 - 1/(iter + 1), 0.996)$ following the [11]. λ_r and λ_{unsup} are set to 1.0 and 2.0 by default. The model is trained for 240 epochs by default and the teacher model is used for the evaluation. The images for training are cropped to the size of 512×512. The test images are center-cropped to 1024×1024 size. We train our model on a single NVIDIA RTX 3090 GPU.

We evaluate performance using **Mean Pixel Accuracy** (mAcc) and **Mean Intersection over Union** (mIoU) as the metrics.

C. Comparison Results

Compared with the SSL Methods. We compare our model with state-of-the-art semi-supervised learning methods, Mean Teacher [47] (MT), ClassMix [50], FixMatch [10],

TABLE IV
SEGMENTATION PERFORMANCE ON THE AI4MARS.

Method	20% data		100% data	
	mAcc	mIoU	mAcc	mIoU
Supervised	71.68	66.14	74.43	68.34
MT [47]	75.44	70.13	77.82	71.98
MT [47]+ClassMix [50]	76.86	70.49	78.56	72.67
FixMatch [10]	76.27	70.36	77.37	71.90
RanPaste [61]	75.16	70.37	77.39	70.59
U ² PL [38]	77.11	70.89	78.62	72.41
AugSeg [11]	76.88	70.15	77.45	72.34
UniMatch [12]	75.60	70.24	77.21	71.36
Ours	77.60	71.79	80.33	74.68

RanPaste [61], U²PL [38], AugSeg [11], and UniMatch [12], covering the latest consistency regularization-based and contrastive learning-based methods as well as the naive supervised training results, which only utilize the labeled data. We conduct these compared methods with their official implementations (except for the MT which we adopt a better training hyper-parameters for fair), using the same backbone ResNet-50. As shown in Table V-B and Table IV, our method achieves the best performance across different labeled data ratios and datasets. MT [47] uses *Gaussian noise* and *Dropout* as augmentations for both teacher and student branches. However, earlier works do not use weak-to-strong augmentation strategy, which makes them sub-optimal. We also employ the ClassMix [50] augmentation with MT. But the quality of the generated mixed image strongly depends on the quality of the pseudo-labels, which cannot be guaranteed on the Mars images with unclear object contours. FixMatch [10] and UniMatch [12] perform even worse than the supervised baseline when the more labeled data are available. This is mainly because they employ a shared encoder instead of the teacher-student architecture, which we find is less effective in our setting. Besides, AugSeg [11] achieves state-of-the-art performance on Earth image benchmarks, while it is not satisfying in Mars semantic segmentation task due to the adopted various color augmentations. As for the contrastive learning-based methods, U²PL uses filtered pseudo-labels to perform the pixel-wise contrastive learning. However, the training cost of such methods is generally high, which will

TABLE V
SEGMENTATION PERFORMANCE OF DIFFERENT CLASSES ON THE S^5 MARS USING ALL LABELED TRAINING DATA.

Method	Class IoU (%)									mIoU
	Sky	Ridge	Soil	Sand	Bedrock	Rock	Rover	Trace	Hole	
FixMatch [10]	88.97	90.06	83.05	74.19	92.09	20.31	52.54	88.75	73.26	73.69
U ² PL [38]	94.86	93.25	85.80	81.02	92.18	23.55	93.77	78.06	64.96	78.60
AugSeg [11]	94.75	93.00	84.78	77.36	92.02	10.33	83.08	81.00	75.84	76.90
Ours	95.64	94.20	87.18	83.18	92.42	22.89	87.17	85.60	74.10	80.15

TABLE VI
COMPARISON WITH THE ZERO-SHOT MODELS ON S^5 MARS DATASET.

Methods	mAcc (%)	mIoU (%)	Inference Time
SAM-CLS [14]	36.38	28.53	691 ms/img
SAN [62]	14.70	3.28	102 ms/img
SAN-FT [62]	11.60	10.54	102 ms/img
Ours	84.73	80.15	19 ms/img

be discussed in Section V-D.

We also present the segmentation performance of each class in Table V. Note that the optimal data augmentations and learning strategy may differ across different categories. Remarkably, our method can achieve the best results on the head classes. Meanwhile, the performance on the tail classes, of which the sample number is small, is also comparable with other methods.

Compared with Zero-Shot General Models. Recently, the general large models for segmentation achieves a great success, which can deal with the unseen data in training with the help of massive amounts of training data or the help of the vision-language model, *e.g.*, CLIP [63]. Here we highlight that current zero-shot learning methods or open-vocabulary methods are still unable to handle the Mars image semantic segmentation task well due to the fine-grained feature classification and required expert knowledge.

SAM [14] can produce the high-quality object binary masks. To obtain the corresponding label, we apply a classification head subsequent to the image encoder. Specifically, the pre-trained image encoder in SAM is fixed and we train a classification decoder, which takes the encoded feature as input and outputs the pixel-wise semantic label, denoted by SAM-CLS. We can find that the extracted features by the encoder are not discriminative for Mars semantic segmentation, as shown in Table VI. As for vision-language models, we compare with state-of-the-art method, SAN [62]. We evaluate the model with both official model weights and fine-tuned weights on the target dataset, denoted as SAN-FT. However, poor performance is observed as shown in Table VI. This is mainly due to two aspects: 1) the Martian terrain category is relatively rare in the corpus, 2) domain-specific expert knowledge is required for the fine-grained classification, which lead to the difficulty of feature space alignment for the Mars segmentation dataset. Moreover, their high training and inference overhead makes them sub-optimal for resource-constrained extraterrestrial tasks.

Overall, our method achieves remarkable performance,

TABLE VII
ABLATION STUDIES ON *AugIN*.
IDENTITY DENOTES THE METHOD WITHOUT *AugIN*.

Method	mAcc	mIoU
Identity	84.20	79.59
EFDM [64]	83.97	78.78
WCT [65]	83.38	78.83
FDA [58]	84.20	78.70
<i>AugIN</i> - μ	83.64	78.88
<i>AugIN</i> - σ	84.73	80.15
<i>AugIN</i>	83.58	78.69

TABLE VIII
ABLATION STUDIES ON DIFFERENT MIXING METHODS. IDENTITY DENOTES THE METHOD WITHOUT ANY MIXING AUGMENTATION.

Method	mAcc	mIoU
Identity	83.33	78.56
<i>CutMix</i> [48]	83.85	78.98
<i>ClassMix</i> [50]	83.81	78.89
DACS [66]	83.80	78.69
<i>SAM-Mix</i> (Ours)	84.73	80.15

TABLE IX
ABLATION STUDIES ON THE OBJECT AUGMENTATIONS IN *SAM-Mix*.

Flip	Rotate	Rescaling	mAcc (%)	mIoU (%)
✓			84.57	79.46
✓	✓		84.81	79.71
✓	✓	✓	84.80	79.88
			84.73	80.15

which verifies the effectiveness of the proposed method.

D. Ablation Studies

In the following, we conduct a series of ablations studies on the S^5 Mars dataset using full data by default. ResNet-50 is adopted as the backbone.

Effect of AugIN. Recall that *AugIN* exchanges the mean or standard deviation of different samples as shown in Fig. 8. First, we give the ablation studies on the different variants of the *AugIN* in Table VII. *AugIN*- μ , *AugIN*- σ and *AugIN* denote swapping the mean, standard deviation, and both, respectively. As we can see, *AugIN*- σ can bring boost to the model performance, while swapping mean shows adverse effect. By comparing the performance of different categories, we find that the main performance degradation comes from the hole, rover, and rock. We argue that this is because the mean of a image corresponds to the brightness. Exchanging the mean can cause the inappropriate brightness changes of objects. For example, the brightness of rover is obviously different from that of rock, and corrupting this information is not conducive to distinguish objects. In contrast, the standard deviation mainly affects the degree of dispersion of the data while maintaining the overall brightness, which is mainly reflected in the clarity of the object edge. This helps the model to produce better prediction results at the object edge. We finally choose *AugIN*- σ in implementation.

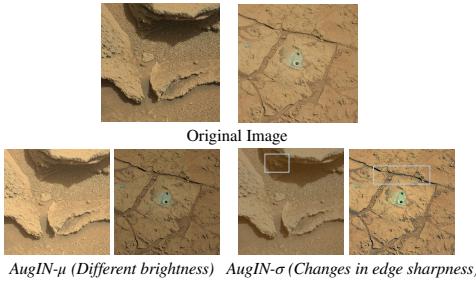


Fig. 8. Examples of augmented results by *AugIN*. *AugIN*- μ mainly affects the image brightness while *AugIN*- σ changes the sharpness of the object edges.

TABLE X
ABLATION STUDIES ON SOFT AND HARD PSEUDO-LABELS.

Soft Label	Hard Label	mAcc (%)	mIoU (%)
✓		83.36	78.12
	✓	84.28	79.70
✓	✓	84.73	80.15

Besides, to give a comprehensive evaluation, we provide the results of replacing *AugIN* with other methods [58, 64, 65] that exchange inter-image information. FDA [58] exchanges low-frequency information which is similar to *AugIN*- μ . Other style-transfer-based methods, *i.e.*, EFDM [64] and WCT [65], also fail to improve the performance because they still cannot avoid the caused color distribution shift.

Effect of SAM-Mix. Table VIII gives the analysis of the *SAM-Mix* augmentation. The main difference between these approaches lies in the way the masks are generated. However, these generated masks have high uncertainty, making it difficult for the model to learn consistency in SSL tasks. *CutMix* randomly generates a rectangular mask from the beta distribution. *ClassMix* takes the predicted region of a certain category as the mask through the generated pseudo-labels, while *DACS* uses the ground-truth labels to mix images. However, these methods are performed in class-wise instead of instance-wise. For example, all rocks would be cut and pasted to another image, which can cause serious occlusions and increase the difficulty of consistency learning. Meanwhile, *ClassMix*, which relies on pseudo-labels, still cannot provide good guidance in the early stage of training, while *DACS*, which is based on ground-truth labels, heavily relies on the number of labels and can only generate limited mixed samples. In contrast, we utilize the masks output by SAM to locate the objects and filter out the mask with higher confidence, achieving better performance.

Meanwhile, we present the effect of the object-wise augmentations, *i.e.*, *Rotation*, *Flip*, and *Rescaling* for the cut out object. As shown in the Tabel IX, these simple available geometrical object augmentations can bring further improvement slightly by promoting the mask diversity.

Soft-to-Hard Pseudo-Label. To depict the semantic features in a more fine-grained manner, we utilize the soft pseudo-labels in addition to the hard labels for the unlabeled images. Table X shows the effect of the soft and hard label, respectively. As we can see, the model gives the best results

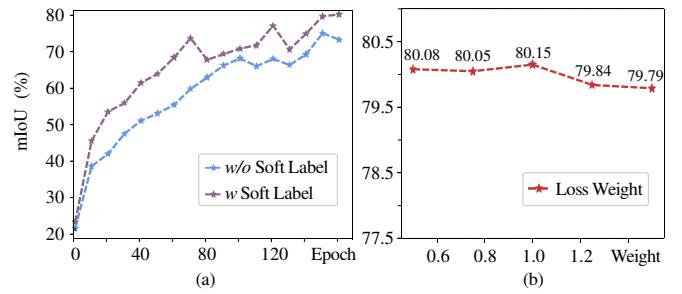


Fig. 9. (a): The effect of soft pseudo-labels on performance of different epochs. Faster convergence is observed when equipped with soft labels. (b) Ablation study results on the loss weight of soft pseudo-labels.

TABLE XI
ABLATION STUDY ON THE THRESHOLD t_{hard} .

t_{hard}	mIoU (%)
0.9	78.82
0.7	79.31
0.6	79.87
0.5	80.15
0.4	80.00

TABLE XII
ABLATION STUDY ON THE THRESHOLD t_{soft} .

t_{soft}	mIoU (%)
1.0	79.83
0.95	79.75
0.9	80.15
0.8	80.04
0.6	79.58

when both kinds of labels are utilized. We note that only employing the soft labels can not yield a good performance, which illustrates that the hard label is important for its low entropy constraints on the model output predictions. On the other hand, soft labels serve as a useful complement to hard labels, especially by making full use of the supervised signals in low-confidence regions. It promotes consistency learning by using the correlation between features and prototypes of different categories as the objective, which can be deemed as a relational knowledge distillation process.

Meanwhile, we note that the dual label optimization strategy can significantly improve the convergence speed of the model as shown in Fig. 9 (a). This is because of the extra supervisory signal provided by those low-confidence regions, which would be discarded using the hard labels.

Confidence-based Thresholding Strategy. We utilize hard pseudo-labels in highly confident regions ($> t_{hard}$) while using soft labels in regions with low confidence ($< t_{soft}$). Table XI gives the results of different thresholds. For the hard labels, a too-high threshold (> 0.7) can significantly reduce the number of training labels, limiting the effect of pseudo-labels. Meanwhile, a too-low threshold can degrade the performance slightly, due to the additional introduced uncertain pseudo-labels in model consistency learning. However, this effect is relatively weak because only few pixels are involved. For the soft labels, we discard the high confidence region. Because these regions can be considered reliable, and the model should give predictions with high confidence. In this case, correlations generated with other semantic prototypes are often noise, which is not conducive to consistency learning as shown in Table XII. However, a too-low threshold degrades the performance significantly, which is because these learning targets with high uncertainty lead to an unstable learning process.

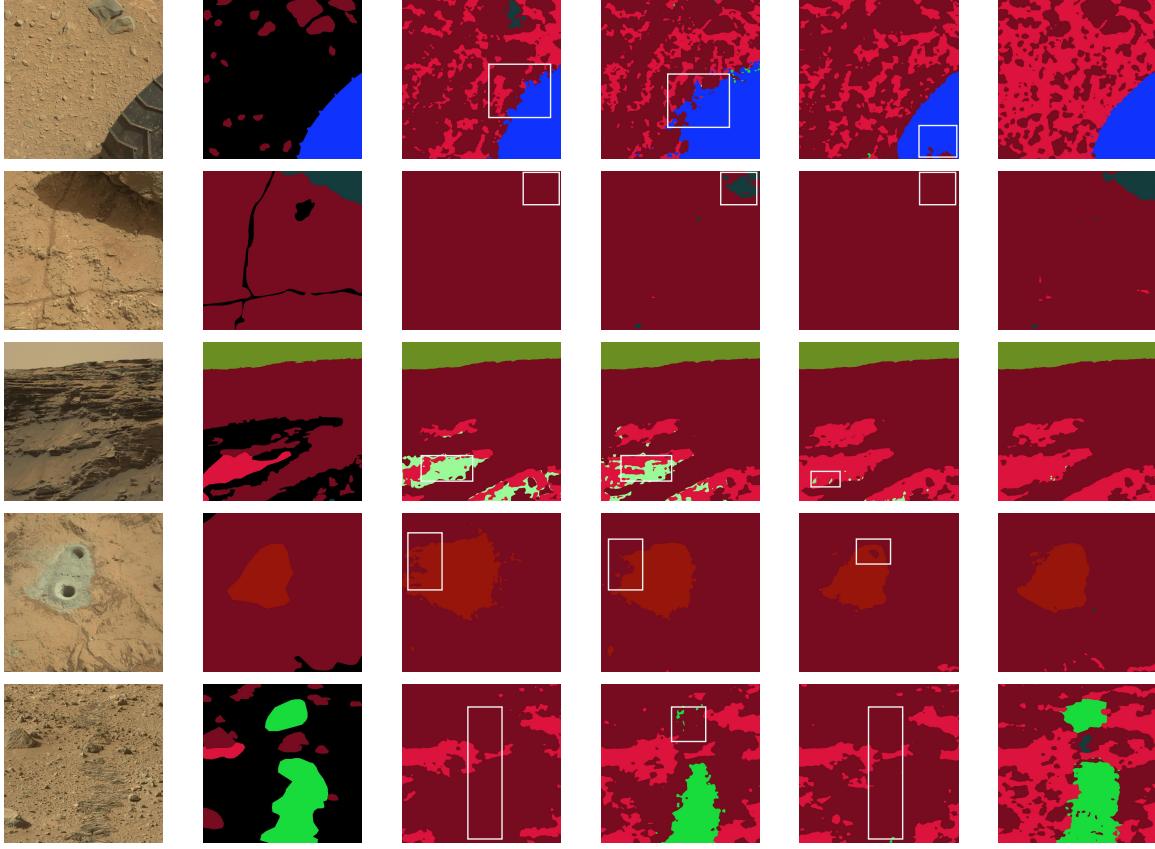


Fig. 10. Qualitative results on S^5 Mars dataset with full labels. Columns from left to right denote the original images, the ground-truth, the supervised results, the MT [47]+ClassMix [50] results, AugSeg [11] results, and our method results, respectively.

TABLE XIII
COMPLEXITY ANALYSIS ON S^5 MARS DATASET.

Methods	Training Time	Params	mIoU (%)
U ² PL [38]	98.7h	196.15M	78.60
FixMatch [10]	34.3h	40.47M	73.69
AugSeg [11]	35.9h	80.94M	76.90
Ours	42.8h	80.94M	80.15

Loss Weight. We further conduct the ablation studies on the loss weight λ_s of the soft pseudo-labels $\mathcal{L}_{unsup}^{soft}$ in Fig. 9 (b). As we can see, the model performance degrades when the loss weight is too large. This indicates the hard pseudo-label signals of the high-confidence regions are important to the model, which constrain the model predictions to be with low entropy. For the loss weight λ_{unsup} , we set it to be 2.0 following the previous works [10, 11] to make our method more general.

Complexity Analysis. We first present the results of U²PL [38] based on contrastive learning in Table XIII. These methods have high complexity because they need to maintain a large number of samples in the memory bank and perform instance discrimination tasks with multiple negative samples. As we can see, our method based on consistency regularization has an obvious complexity advantage, which shows the advantage of this type of method for extraterrestrial missions. Compared with other consistency-regularization methods [10, 11], our method has a longer training time due to the additional

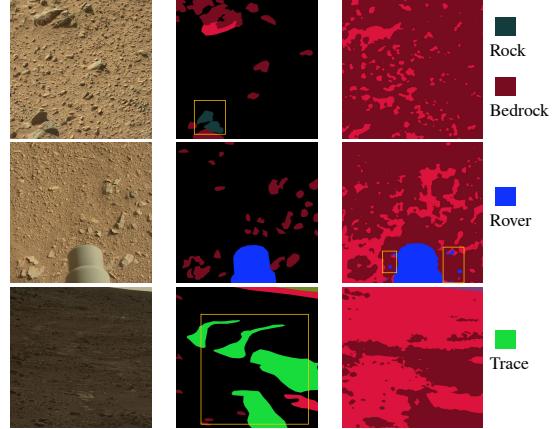


Fig. 11. Some failure cases of our method. From left to right are the original image, the ground-truth label, and the predicted result.

back-propagation process for soft pseudo-label optimization and the data augmentations. Since the same backbone network is used, the inference time is the same across these methods. As for the model parameters, the difference lies in whether the teacher and student models share parameters. Despite the cost incurred in terms of training time and space parameters, we argue that they are acceptable where a significant performance gain is observed.

E. Qualitative Results

We present subjective segmentation results in Fig. 10. The compared methods employ the same ResNet50 backbone. As we can see, the segmentation results of our method are more accurate than other methods, which is reflected in clearer object contours (first and fourth rows), more sensitive and accurate object detection (second and fifth rows), and less category mixing in segmentation map (third row). Compared with AugSeg [11] which is also based on two-branch architecture but with many color augmentations, our method can generate better results, demonstrating the effectiveness of the proposed augmentations for SSL Mars segmentation task.

F. Limitations and Discussions

Our method is efficiently designed in terms of data augmentation and pseudo-label optimization for semi-supervised Mars segmentation task. Notably, it scales well to the existing techniques, *e.g.*, augmentation anchoring and distribution alignment in [67] and is simple to implement, making it a strong model to provide the basis for future work.

We present some failure cases in Fig. 11. As we can see, the limitations mainly lie in two aspects: 1) There is confusion between similar categories as shown in the first case. For example, to distinguish the rock and bedrock, the model needs to classify whether the rock is exposed to the ground, which can be difficult. One direction is to carry out specific designs, *e.g.*, training a independent classifier, for these difficult categories. 2) The images directly taken by the rovers on Mars often have a long-tailed label distribution, which may affect the reliability of the pseudo-labels in semi-supervised learning, causing the noisy prediction or misclassification in the tail classes as shown in the second and third case. We point out these observed problems as future work, in the hope that more meaningful works will emerge.

VI. CONCLUSION

In this paper, we address the SSL for Mars semantic segmentation problem from both data and method perspective. First, we propose a fine-grained annotated dataset S⁵Mars for Martian terrain segmentation. This dataset provides sparse and high-confidence labeled data, which effectively assists the subsequent Mars exploration work. Then, we propose a simple yet effective SSL framework. Specifically, we analyze the effect of current used augmentations for Mars image segmentation. And two effective augmentations, *AugIN* and *SAM-Mix* are further proposed to improve the model performance. Meanwhile, a soft-to-hard consistency learning strategy is introduced to fully utilize the unlabeled data in a confidence-based manner. Extensive comparison and ablation experiments demonstrate the effectiveness of our method.

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*, vol. 188, p. 104943, 2020.
- [2] F. Goessmann, 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*, vol. 17, no. 6-7, pp. 655–685, 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*, vol. 54, no. 10, pp. 69–77, 2021.
- [4] I. Priyadarshini and V. Puri, “Mars weather data analysis using machine learning techniques,” *Earth Science Informatics*, vol. 14, no. 4, pp. 1885–1898, 2021.
- [5] R. Gonzalez and K. Iagnemma, “Deepterramechanics: Terrain classification and slip estimation for ground robots via deep learning,” *arXiv:1806.07379*, 2018.
- [6] M. Dimastrogiovanni, F. Cordes, and G. Reina, “Terrain estimation for planetary exploration robots,” *Applied Sciences*, vol. 10, no. 17, p. 6044, 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition 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] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, “Fixmatch: Simplifying semi-supervised learning with consistency and confidence,” *Proc. Advances in Neural Information Processing Systems*, 2020.
- [11] Z. Zhao, L. Yang, S. Long, J. Pi, L. Zhou, and J. Wang, “Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2023.
- [12] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, “Revisiting weak-to-strong consistency in semi-supervised semantic segmentation,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2023.
- [13] J. Yuan, Y. Liu, C. Shen, Z. Wang, and H. Li, “A simple baseline for semi-supervised semantic segmentation with strong data augmentation,” in *Proc. Int'l Conference on Computer Vision*, 2021.
- [14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” *arXiv:2304.02643*, 2023.
- [15] 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*, vol. 13, no. 3, pp. 951–963, 2020.
- [16] 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 *Proc. Innovative Applications of Artificial Intelligence*, 2018.
- [17] 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 Trans. on Neural Networks and Learning Systems*, vol. 34, no. 7, pp. 3335–3344, 2021.
- [18] H. Liu, M. Yao, X. Xiao, and Y. Xiong, “Rockformer: A u-shaped transformer network for martian rock segmentation,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.
- [19] 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*, vol. 118, pp. 109–121, 2018.

- [20] D. R. Thompson and R. Castano, "Performance comparison of rock detection algorithms for autonomous planetary geology," in *2007 IEEE Aerospace Conference*, 2007.
- [21] 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.
- [22] 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*, vol. 60, pp. 1–15, 2022.
- [23] T. Wilhelm, M. Geis, J. Pütschneider, T. Sievernich, T. Weber, K. Wohlfarth, and C. Wöhler, "Domars16k: A diverse dataset for weakly supervised geomorphologic analysis on mars," *Remote Sensing*, vol. 12, no. 23, p. 3981, 2020.
- [24] S. Niekum, "Reliable rock detection and classification for autonomous science," *C. Thesis*, 2005.
- [25] 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*, vol. 4, no. 4, pp. 3876–3883, 2019.
- [26] 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.
- [27] X. Xiao, H. Cui, M. Yao, and Y. Tian, "Autonomous rock detection on mars through region contrast," *Advances in Space Research*, vol. 60, no. 3, pp. 626–635, 2017.
- [28] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. European Conference on Computer Vision*, 2018.
- [29] E. Goh, J. Chen, and B. Wilson, "Mars terrain segmentation with less labels," *arXiv:2202.00791*, 2022.
- [30] Y. Xiong, X. Xiao, M. Yao, H. Liu, H. Yang, and Y. Fu, "Marsformer: Martian rock semantic segmentation with transformer," *IEEE Trans. on Geoscience and Remote Sensing*, 2023.
- [31] J. Zhang, Y. Xia, and G. Shen, "A novel deep neural network architecture for mars visual navigation," *arXiv*, 2018.
- [32] M. Ono, B. Rothrock, K. Otsu, S. Higa, Y. Iwashita, A. Didier, T. Islam, C. Laporte, V. Sun, K. Stack *et al.*, "Maars: machine learning-based analytics for automated rover systems," in *2020 IEEE Aerospace Conference*, 2020, pp. 1–17.
- [33] 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 *Proc. European Conference on Computer Vision*, 2018.
- [34] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proc. European Conference on Computer Vision Workshops*, 2016.
- [35] J. Lambert, Z. Liu, O. Sener, J. Hays, and V. Koltun, "Mseg: A composite dataset for multi-domain semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [36] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," *arXiv:2006.05278*, 2020.
- [37] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [38] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2022.
- [39] Y. Zhong, B. Yuan, H. Wu, Z. Yuan, J. Peng, and Y.-X. Wang, "Pixel contrastive-consistent semi-supervised semantic segmentation," in *Proc. Int'l Conference on Computer Vision*, 2021.
- [40] L. Hu, J. Li, X. Peng, J. Xiao, B. Zhan, C. Zu, X. Wu, J. Zhou, and Y. Wang, "Semi-supervised npc segmentation with uncertainty and attention guided consistency," *Knowledge Based Systems*, vol. 239, p. 108021, 2022.
- [41] R. Mendel, L. A. De Souza, D. Rauber, J. P. Papa, and C. Palm, "Semi-supervised segmentation based on error-correcting supervision," in *Proc. European Conference on Computer Vision*, 2020.
- [42] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Proc. Advances in Neural Information Processing Systems*, 2019.
- [43] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int'l Conference on Learning Representations*, 2018.
- [44] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinohaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Proc. Advances in Neural Information Processing Systems*, 2021.
- [45] Y. Wang, H. Chen, Q. Heng, W. Hou, M. Savvides, T. Shinohaki, B. Raj, Z. Wu, and J. Wang, "Freematch: Self-adaptive thresholding for semi-supervised learning," *arXiv:2205.07246*, 2022.
- [46] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020.
- [47] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Advances in Neural Information Processing Systems*, 2017.
- [48] 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 *Proc. Int'l Conference on Computer Vision*, 2019.
- [49] G. French, A. Oliver, and T. Salimans, "Milking cowmask for semi-supervised image classification," *arXiv:2003.12022*, 2020.
- [50] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.
- [51] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [52] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [53] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. Int'l Conference on Computer Vision*, 2017.
- [54] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," *arXiv:1906.01916*, 2019.
- [55] D. Dwibedi, I. Misra, and M. Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in *Proc. Int'l Conference on Computer Vision*, 2017.
- [56] S. Tripathi, S. Chandra, A. Agrawal, A. Tyagi, J. M. Rehg, and V. Chari, "Learning to generate synthetic data via compositing," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [57] H.-S. Fang, J. Sun, R. Wang, M. Gou, Y.-L. Li, and C. Lu, "Instaboost: Boosting instance segmentation via probability map guided copy-pasting," in *Proc. Int'l Conference on Computer Vision*, 2019.
- [58] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [59] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning

- for image recognition,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [61] J.-X. Wang, S.-B. Chen, C. H. Ding, J. Tang, and B. Luo, “Ran-paste: Paste consistency and pseudo label for semisupervised remote sensing image semantic segmentation,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.
- [62] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, “Side adapter network for open-vocabulary semantic segmentation,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2023.
- [63] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *Proc. Int'l Conference for Machine Learning*, 2021.
- [64] Y. Zhang, M. Li, R. Li, K. Jia, and L. Zhang, “Exact feature distribution matching for arbitrary style transfer and domain generalization,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2022.
- [65] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, “Universal style transfer via feature transforms,” *Proc. Advances in Neural Information Processing Systems*, 2017.
- [66] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, “Dacs: Domain adaptation via cross-domain mixed sampling,” in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.
- [67] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, “Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring,” *arXiv:1911.09785*, 2019.



Zejia Fan received the B.S. degree in Computer Science in 2021 from Peking University, Beijing, China, where she is currently working toward the Ph.D degree with the Wangxuan Institute of Computer Technology. Her current research interests include image enhancement and deep learning.



Wenjing Wang (Student Member, IEEE) received the B.S. degree in data science from Peking University, Beijing, China, in 2019, where she is currently pursuing a doctoral degree with the Wangxuan Institute of Computer Technology, Peking University. She has authored over 20 technical articles in refereed journals and proceedings, and she holds five granted patents. Her current research interests include image enhancement, image synthesis, and deep learning.



Jiahang Zhang received the B.S. degree in computer science from Peking University, Beijing, China, in 2023, where he is currently pursuing the Ph.D. degree with the Wangxuan Institute of Computer Technology. His current research interests include action recognition and self-supervised learning.



Jiaying Liu (Senior Member, IEEE) received the PhD degree (Hons.) in computer science from Peking University, Beijing China, 2010. She is currently an Associate Professor, Boya Young Fellow with the Wangxuan Institute of Computer Technology, Peking University, China. She has authored more than 100 technical articles in refereed journals and proceedings, and holds 70 granted patents. Her current research interests include multimedia signal processing, compression, and computer vision. She is a senior member of IEEE/CSIG, and a distinguished member of CCF. She was a visiting scholar with the University of Southern California, Los Angeles, California, from 2007 to 2008. She was a visiting researcher with Microsoft Research Asia, in 2015 supported by the Star Track Young Faculties Award. Dr. Liu has served as a member of Multimedia Systems and Applications Technical Committee (MSA TC), and Visual Signal Processing and Communications Technical Committee (VSPC TC) in IEEE Circuits and Systems Society. She received the IEEE ICME 2020 Best Paper Award and IEEE MMSP 2015 Top10% Paper Award. She has also served as the Associate Editor of the IEEE Trans. on Image Processing, the IEEE Trans. on Circuits Systems for Video Technology and Journal of Visual Communication and Image Representation, the Technical Program Chair of ACM MM Asia-2023/IEEE ICME-2021/ACM ICMR-2021/IEEE VCIP-2019, the Area Chair of CVPR-2021/ECCV-2020/ICCV-2019, ACM ICMR Steering Committee member and the CAS Representative at the ICME Steering Committee. She was the APSIPA Distinguished Lecturer (2016-2017).



Lilang Lin (Student Member, IEEE) received the B.S. degree in data science from Peking University, Beijing, China, in 2021, where he is currently pursuing the Ph.D. degree with the Wangxuan Institute of Computer Technology. His current research interests include action recognition, self-supervised learning, and unsupervised learning.