Contents lists available at ScienceDirect

Acta Astronautica

journal homepage: www.elsevier.com/locate/actaastro

A hybrid attention semantic segmentation network for unstructured terrain on Mars

Haiqiang Liu [a], Meibao Yao [a,*], Xueming Xiao [b], Hutao Cui [c]

[a] School of Artificial Intelligence, Jilin University, Changchun, 130012, China
[b] CVIR Lab, Changchun University of Science and Technology, Changchun, 130013, China
[c] Deep Space Exploration Research Center, Harbin Institute of Technology, Harbin, 150006, China

ARTICLE INFO

Keywords:
Semantic segmentation
Unstructured environment
Martian terrain
Fully convolution network
Context aggregation

ABSTRACT

Semantic segmentation of Martian terrain is crucial for the route planning and autonomous navigation of rovers on Mars. However, existing methods are restricted to structured or semi-structured scenes, performing poorly on Mars that is a completely unstructured environment. Therefore, we propose a novel hybrid attention semantic segmentation (HASS) network, which contains a global intra-class attention branch, a local inter-class attention branch and a representation merging module. Specifically, the global attention branch draws the consistencies of all homogeneous pixels in the whole image, and the local attention branch models the relationships between specific heterogeneous pixels with the supervision of elaborately designed loss function. The merging module aggregates the contexts from the two branches for the final segmentation. Furthermore, we establish a panorama semantic segmentation dataset of Martian landforms, named MarsScapes, which provides fine-grained annotations for eight semantic categories. Extensive experiments on our MarsScapes and the public AI4Mars datasets show the superiority of the proposed method.

1. Introduction

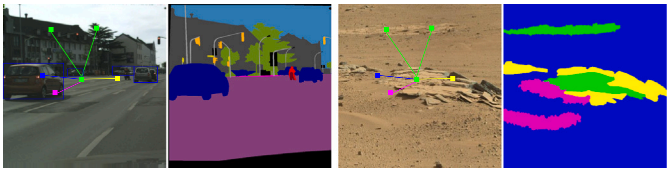
Semantic segmentation is a challenging issue in computer vision, whose goal is to assign a class label to each pixel in a given image. It has been applied in autonomous driving [1,2], human–computer interaction [3,4] and medical diagnosis [5,6]. Although numerous methods have achieved significant progress in the segmentation of structured and semi-structured scenes [7–11], they are not fully applicable to a completely unstructured environment like Mars [12]. This is mainly because the definitions of unstructured terrains are quite different from those of objects in a structured environment. In a remote sensing image of unstructured environment, the classification of an area depends not only on its consistencies with all intra-class pixels in homogeneous terrains, but also on its relationships with designated inter-class pixels in the neighboring heterogeneous terrains. Taking Fig. 1 for example, a box with multiple wheels is usually regarded as a car, no matter where it appears on a structured scene of street. On an unstructured Mars scene, however, a block is considered as a big rock (in yellow) if it is much higher than the surrounding landforms, but is seen as a bedrock (in green) if it is embedded in the ground.

To facilitate the semantic segmentation for Unstructured terrains on Mars, we propose a novel contextual aggregation network called hybrid attention semantic segmentation (HASS), which contains a global intra-class branch to capture the representations of all pixels in homogeneous terrains and a local inter-class branch to explore the relationships between pixels specified in neighboring heterogeneous terrains. Specifically, we adopt a local diversity loss (LDLoss) in the local intra-class branch. For each pixel in an input image, LDLoss supervises a local attention map to highlight the heterogeneous pixels located in adjacent landforms. After multiplying the local attention map by the original features, we get the designated local features, which is a significant supplement to features extracted by existing global attention methods for unstructured terrain segmentation.

Deep learning-based methods for semantic segmentation require a large amount of annotated images. To acquire enough unstructured terrain images, there are two main ways. (1) Air-view methods, some of which use UAVs [13] and others resort to satellites [14,15]. However, they are unavailable in some harsh environments (caves, forests, exoplanets, etc.), and are not fine-grained enough for the off-road autonomous navigation of mobile robots. (2) Ground-view methods, usually implemented by mobile mapping systems in ground robots [16,17], are appropriate for more environments. But the collected images are limited by the viewing angle [18], which cannot help the robot plan

[*] Corresponding author.
E-mail address: meibaoyao@jlu.edu.cn (M. Yao).

https://doi.org/10.1016/j.actaastro.2022.08.002
Received 24 April 2022; Received in revised form 25 July 2022; Accepted 4 August 2022
Available online 8 August 2022



(a) Structured scene (b) Unstructured scene

Fig. 1. An illustration of the difference between structured scene and unstructured scene understanding. In a structured scene (a), The definition of cars remains unchanged, no matter where they appear on the street. On Mars (b), however, the segmentation of an bedrock depends not only on its consistencies with other bedrocks (denoted as green lines), but also on its relationships with neighboring heterogeneous terrains (represented as lines of other colors). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

its route in advance. Moreover, these images are taken in the earth environments, where many terrains can be easily distinguished by color, which ignores the difficulties of unstructured scene understanding.

To this end, we establish MarsScapes, the first panoramic image dataset of Martian landforms for semantic and instance segmentation, aiming at promoting the research of unstructured terrain understanding and serving for the intelligent exploration of robots on Mars. The dataset contains detailed annotations of navigable terrains, such as soil and bedrock, as well as less navigable terrains, such as sand, gravel, big rocks and steep slopes. MarsScapes provides panorama images composited from a large number of local images collected by the Curiosity rover [19]. We define two benchmark tasks on the dataset, semantic segmentation of all terrains (MarsScapes-ALL) and prediction of navigable terrains (MarsScapes-Navigability). Referring to the processing methods of [20], we obtain 18460 samples from the original panoramas, which are sufficient to support the training of the two tasks. Abundant experiments on MarsScapes and another public AI4Mars dataset show the superiority of our method in understanding unstructured Martian terrain.

In summary, our main contributions are as follows:

- We propose a novel hybrid attention semantic segmentation method with a dual-branch network, which integrates the specific global and local contexts of unstructured terrains with the supervision of elaborately designed loss functions.
- We establish a panorama dataset of Martian landforms with detailed annotations and define two benchmark tasks, aiming at promoting the development of methods for unstructured environments understanding.
- We demonstrate the proposed method outperforms existing approaches through extensive experiments on our MarsScapes and the public AI4Mars dataset.

2. Related work

2.1. Context aggregation methods

Context aggregation can augment feature representations by capturing long-range dependencies, which has been applied in many tasks [21–23]. In the early years, aggregating context was mainly realized by generating multi-scale feature maps. PSPNet [7] adopts a pyramid pooling module to aggregate effective spatial information. DeepLab v2 [24] and DeepLab v3 [25] employ parallel dilated convolutions with different dilation rates to explore contextual dependencies. DenseASPP [26] brings dense connections into ASPP to capture the contexts of different scales. However, they failed to exploit the relationships between objects in a global perspective, which is crucial to scene understanding.

To address this, attention-based methods have become popular for context aggregation. DAN [8] designs two parallel attention modules, one is used to model long-range dependencies in the spatial dimension, and the other is used to learn channel relationships in the channel dimension. CCNet [9] introduces a novel criss-cross attention module that explores contextual dependencies from the full image in an efficient way. CPNet [10] obtains the global intra-class and inter-class contexts based on a Context Prior map supervised by Affinity Loss. Nevertheless, these self-attention mechanisms focus more on exploring global contexts from all pixels regardless of their positions, ignoring the significant role of local heterogeneous pixels in unstructured terrain understanding.

In this work, we propose a novel attention-based aggregation network, which integrates the dependencies of all intra-class pixels and the relationships between adjacent inter-class pixels. Qualitative and quantitative comparisons with other models verify the effectiveness of our method.

2.2. Unstructured terrain segmentation datasets

Presently, many unstructured landform datasets are composed of air-level images taken by satellites or UAVs [27]. For example, GID [14] is composed of 150 high-resolution images taken by GF-2 from more than 60 cities in China. DeepGlobe 2018 [15] contains three datasets aiming to solve three different satellite image understanding tasks. Semantic Drone [13] consists of 400 available urban images taken at an altitude of 5 to 30 m. However, these collection methods are not applicable to some harsh environments such as forests and caves, and the collected images are not fine-grained enough for the autonomous navigation of mobile robots [28,29].

This problem can be solved by using mobile mapping systems in ground robots. Angelova et al. [30] introduce a small-scale off-road terrain dataset collected by a wheeled ground robot. Freiburg Forest [16] comprises 366 annotated images taken by a Viona autonomous robot platform in forested environments. RUGD [17] is a large-scale dataset collected by a small unmanned mobile robot from creek, park, trail and village. RELLIS-3D [31] contains annotations of 6235 images taken by a Clearpath Robotics Warthog platform on the Rellis Campus of Texas A&M University. Existing ground-view unstructured terrain segmentation datasets, however, are limited by the angle of view, which is not conducive to the route planning of robots in a large area. Besides, since these datasets are collected in the earth environments, many terrains can be easily distinguished by color, which limits their utility in Mars environments that mostly share similar color. Although some datasets [32–34] have been collected from Mars-like landforms that are simulated on Earth, they make limited achievements on the Martian scene understanding due to the geological differences between Mars and Earth caused by volcanic activity, wind and liquid erosion.

To this end, some researchers have contributed themselves to creating semantic segmentation datasets of Mars environment. Xiao et al. [35] build a small-scale dataset named MarsData to separate Martian rocks from the complex background, which is pivotal for planetary vehicles to avoid obstacles and discover scientific goals. NASA releases AI4Mars [36], a large-scale Martian terrain dataset, which contains coarse-grained annotations for four basic terrain categories. Based on AI4Mars, Li et al. [37] propose Mars-Seg that appends the annotations of four additional terrain types, reaching a total of eight categories. However, these datasets do not provide panoramas and only contain semantic annotations (without instance annotations) of Mars environment. In contrast, our MarsScapes is more targeted and versatile for the future exploration on Mars with advanced techniques in scene understanding.

3. Proposed method

In this section, we first illustrate the overall framework of HASS, and then introduce the local heterogeneous matrix and local diversity loss used for local inter-class attention. Finally, we present how to merge the two branches and design a total loss function for the whole network.

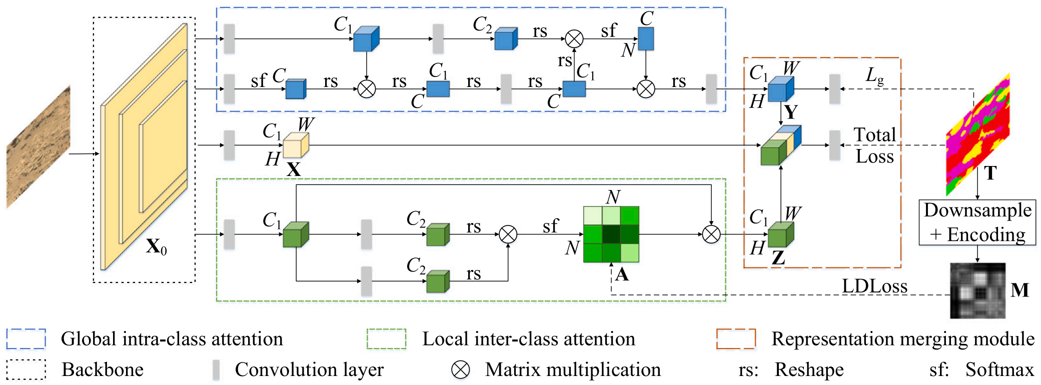


Fig. 2. The framework of HASS, which contains a backbone, a global intra-class attention branch, a local inter-class attention branch, and a representation merging module.

3.1. HASS architecture

Existing attention-based networks focus more on exploring global dependencies from homogeneous and heterogeneous pixels regardless of their distance. However, as previously analyzed, the judgment of an unstructured terrain depends not only on global information, but also on its relative relationships with adjacent terrains. Based on this, we propose HASS to integrate the two contexts for unstructured terrain segmentation.

HASS is a fully convolution network, composed of a backbone network, a global intra-class attention branch, a local inter-class attention branch, and a representation merging module, as shown in Fig. 2. We choose HRNet-W48 [38] with output stride 4 as the backbone. Then, the features $\mathbf{X}_0$ yielded by HRNet-W48 are fed into two parallel attention modules. We use OCR [11] to draw the consistencies of all homogeneous pixels in the global branch, yielding the feature map $\mathbf{Y}$. Note that it can be replaced with other attention-based methods, such as DANet [8] and CPNet [10]. For the local inter-class attention branch, we first encode a local heterogeneous matrix $\mathbf{M}$ from ground truth and utilize it to train a local inter-class attention map $\mathbf{A}$ with the supervision of the local diversity loss. Then, we capture designated features by $\mathbf{Z} = \mathbf{A}\mathbf{X}$. The concatenated features $\mathbf{F} = [\mathbf{X}, \mathbf{Y}, \mathbf{Z}]$ in channel dimension are used for the final semantic classification.

3.2. Local inter-class attention

Given an image, the semantic label of each pixel is shown in ground truth. It is difficult for a model to know a prior relationship between any two pixels until the segmentation result is obtained. To fulfill this, Yuan et al. [11] obtain a soft segmentation map computed from a backbone network in advance and refine the segmentation result through a coarse-to-fine scheme. However, they cannot distinguish between the contexts extracted from inter-class pixels and the ones extracted from intra-class pixels. Yu et al. [10] distinguish the two contextual dependencies by training an affinity map. But they model the relationships between inter-class pixels in a global view (i.e., compute the link weights of all inter-class pixels regardless of distance), which not only hurdles the understanding of unstructured terrains, but leads to high computational complexity.

To model the relationships between specific heterogeneous pixels, we introduce the local diversity loss (LDLoss), which guides the network to learn a local inter-class attention map $\mathbf{A}$, where the conjunction of two adjacent inter-class pixels is highlighted and others are weakened. For the ground truth of an image, we can find out which category the pixels belong to and whether the two pixels are located in adjacent landforms. We encode these information into a local heterogeneous matrix $\mathbf{M}$ to supervise the training of $\mathbf{A}$.

Given an image $\mathbf{I}$ and the ground truth $\mathbf{T}$ with size $H_0 \times W_0$, we acquire a feature map $\mathbf{X}_0$ with size $H \times W$ from the backbone. As shown in Fig. 3, we down-sample the ground truth $\mathbf{T}$ to the same shape as $\mathbf{X}_0$, obtaining a smaller ground truth $\mathbf{T}_d$ with size $H \times W$. $\mathbf{T}_d$ is encoded into the local heterogeneous matrix $\mathbf{M}$ in three steps as displayed in Algorithm 1. First, we recursively search 4-adjacent neighbors of each pixel and put the intra-class pixels into the same instance, so that $\mathbf{T}_d$ is divided into n instances: $I_1, I_2, \dots, I_n$. For each pixel $I_i(j)$ in instance I_i, we then traverse other pixels around it in four directions (i.e., left, up, right, and down) until four closest heterogeneous pixels (i.e., p_l, p_u, p_r, and p_d) have been found, and we find out four instances (i.e., I_l, I_u, I_r, and I_d) that the four pixels belong to. Finally, we construct the local heterogeneous matrix M, where $M(j,k)$ is set to 1 if the jth pixel belongs to instance I_i and the kth pixel belongs to any of the above searched instances, otherwise $M(j,k)$ is set to 0.

Algorithm 1 Encoding for the local heterogeneous matrix

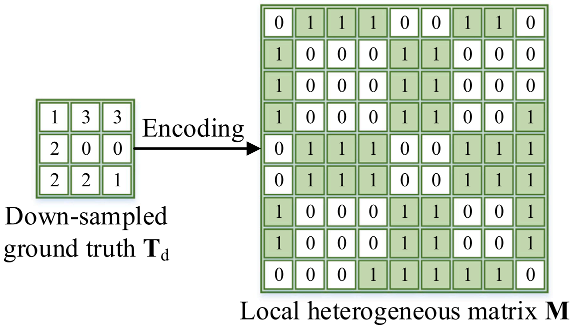
Input: Down-sampled ground truth $\mathbf{T}_d \in \mathbb{R}^{W\times H}$, Class number C
Output: The local heterogeneous matrix $\mathbf{M} \in \mathbb{R}^{N\times N}$

```
Instance id: $n \leftarrow 1$.
while $\mathbf{T}_d \neq \varnothing$ do
    $q \leftarrow Queue()$.
    $q.put(\mathbf{T}_d(0))$.
    while $q \neq \varnothing$ do
        $p \leftarrow q.get()$.
        Search the 4-adjacent pixels of $p$, and save homogeneous pixels in $I_n$.

        $q.put(I_n)$.
    Remove $I_n$ from $\mathbf{T}_d$.
    $n \leftarrow n + 1$.
for $i \leftarrow 1$ to $n$ do
    Count the number of pixels in $I_i$: $N_i$.
    for $j \leftarrow 1$ to $N_i$ do
        Traverse the pixels around $I_i(j)$ in four directions (left, up, right, and
        down) until four heterogeneous pixels $p_l$, $p_u$, $p_r$, and $p_d$ are found.
        Find out four instances $I_l$, $I_u$, $I_r$, and $I_d$ that $p_l$, $p_u$, $p_r$, and $p_d$ belong
        to.
    for $j \leftarrow 1$ to $N_i$ do
        $N = W \times H$
        for $k \leftarrow 1$ to $N$ do
            if $p_k \in \{I_l, I_u, I_r, I_d\}$ then
                $M(j,k) \leftarrow 1$.
            else
                $M(j,k) \leftarrow 0$.
return $\mathbf{M} \in \mathbb{R}^{N\times N}$
```

In this way, the training of $\mathbf{A}$ can be formulated as a binary classification problem of each pixel. A common objective function that is applicable to this problem is the binary Cross Entropy Loss:

$$L = -\frac{1}{N^2}\sum_{i=1}^{N^2}\left(m_i \log a_i + \left(1 - m_i\right)\log\left(1 - a_i\right)\right), \tag{1}$$


Encoding
Down-sampled ground truth $\mathbf{T}_d$
Local heterogeneous matrix $\mathbf{M}$

1	3	3
2	0	0
2	2	1

0	1	1	1	0	0	1	1	0
1	0	0	0	1	1	0	0	0
1	0	0	0	1	1	0	0	0
1	0	0	0	1	1	0	0	1
0	1	1	1	0	0	1	1	1
0	1	1	1	0	0	1	1	1
1	0	0	0	1	1	0	0	1
1	0	0	0	1	1	0	0	1
0	0	0	1	1	1	1	1	0

Fig. 3. Illustration of the encoding of local heterogeneous matrix.

where N^2 is the number of pixels in $\mathbf{A}$ (i.e., $N^2 = H \times W \times H \times W$), $m_i \in \mathbf{M}$, and $a_i \in \mathbf{A}$. However, such a loss function only uses m_i to guide the attention learning at the same position in $\mathbf{A}$ (i.e., a_i), ignoring the correlations with nearby pixels. For example, if pixel p_i and pixel p_j constitute a local inter-class pair, namely $M(i,j) = 1$, the pixel on the right of p_i (i.e., p_{i+1}) and the pixel below p_i (i.e., $p_{i+1\times W}$) are quite likely to form a local inter-class pair with p_j. To fully utilize this contextual correlation, we put forward LDLoss by revising Eq. (1) as follows:

$$L_{ld} = -\frac{1}{N^2}\sum_{i=1}^{N^2}\Big(\left(m_i + R_i\right)\log a_i + \left(1 - m_i + \tilde{R}_i\right)\log\left(1 - a_i\right)\Big), \tag{2}$$

$$R_i = \frac{1}{N_R}\sum_{k=1}^{N_R}\left(\alpha a_{i+k} + \beta a_{i+kW}\right), \tag{3}$$

$$\tilde{R}_i = \frac{1}{N_R}\sum_{k=1}^{N_R}\left(\alpha\left(1 - a_{i+k}\right) + \beta\left(1 - a_{i+kW}\right)\right). \tag{4}$$

In the first item of Eq. (2), we add an additional reward R_i if the next N_R pixels on the right of a_i or the next N_R pixels below a_i are also predicted as 1. On the contrary, if they are all predicted as 0, the first item of LDLoss is equivalent to that of Eq. (1). α and β are weights in two directions, which are both set to 0.5 empirically. The additional reward $\tilde{R}_i$ in the second item of Eq. (2) follows the same design rules. N_R will be discussed in the next section. It is noteworthy that $\mathbf{M}$ is symmetric, so we do not have to add rewards in the left and up directions.

3.3. Representation merging module

The feature maps generated by the global and local branches are diverse due to their differences in design purpose and loss function, which leads to a limited effect if only one of them is used for semantic segmentation. Therefore, we propose a representation merging module, where a merged feature map is derived from the combination of the two branches. We employ three conventional merging strategies, element plus, channel concatenation, and element multiply. As shown in the next section, they are all useful for our tasks, but the channel connection works best.

Moreover, in order to balance the influence of two branches on the final classification, we design a total loss function to supervise the training of the whole network. It is defined as follows:

$$L_t = \lambda_{ld} L_{ld} + \lambda_g L_g + \lambda_m L_m, \tag{5}$$

$$L_g = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log p_{ij}^{g}, \tag{6}$$

$$L_m = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log p_{ij}^{m}, \tag{7}$$

where L_t is the total loss, L_g is the semantic loss of global attention branch, and L_m is the final semantic loss of representation merging module. λ_{ld}, λ_g, and λ_m are hyper-parameters to balance the losses of different branches and module. We empirically set them as: $\lambda_{ld} = 0.8$, $\lambda_g = 1$, and $\lambda_m = 1$. L_g and L_m are both semantic loss based on cross-entropy, where $y_{ij} = 1$ if the semantic label of the ith pixel is j, and $y_{ij} = 0$ otherwise. p_{ij}^{g} and p_{ij}^{m} are predicted semantic labels in the representation merging module and global attention branch, respectively.

4. Experiments

4.1. Datasets

We conduct semantic segmentation experiments on AI4Mars [36] that is partially public at present. As a supplement, we also create a challenging Martian landform segmentation dataset with more classes, called "MarsScapes", which helps to better evaluate the robustness of various methods.

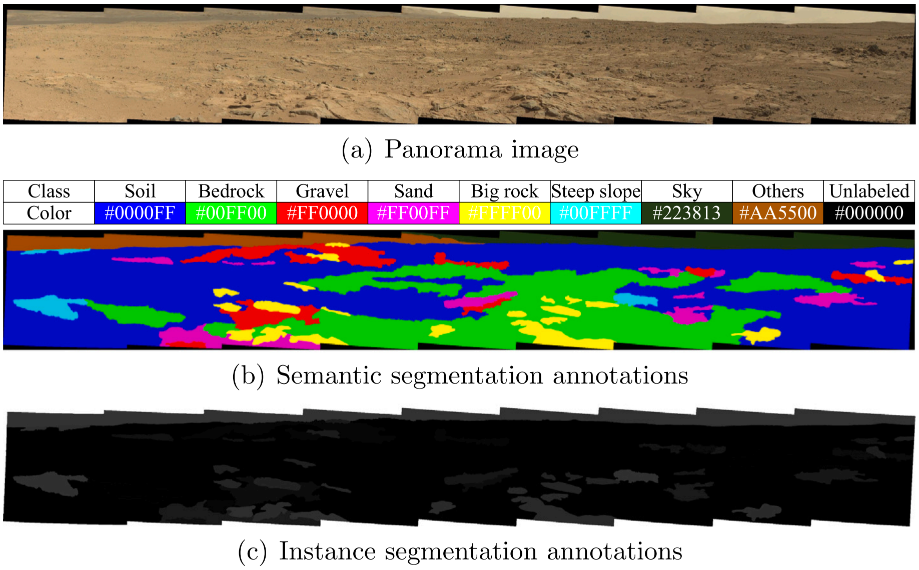
4.1.1. Ai4mars

AI4Mars is the first large-scale semantic segmentation dataset of Martian terrain, proposed by NASA in 2021. The dataset contains 35 K images collected by Curiosity, Opportunity and Spirit rovers. The pixels of each image were grouped into four terrain categories (soil, bedrock, sand and big rock), implemented by multiple volunteers, rover planners and scientists to ensure the quality of semantic annotation. Currently, the Mars Science Laboratory part of AI4Mars (AI4Mars-MSL) has been released but the Mars Exploration Rovers part (AI4Mars-MER) has remained inaccessible. We use AI4Mars-MSL for our evaluation, which contains 16064 annotated images in the training set and 322 in the test set.

4.1.2. MarsScapes

The raw RGB images of MarsScapes are from the official website of NASA [19], which are collected by a mast camera on the Curiosity rover. To integrate abundant global geomorphic information, we concatenate a group of spatially continuous images into a panorama. After removing the images with extreme shooting distance and severe light distortion, 195 samples were produced, whose width ranges from 1230 to 12062 pixels and height from 472 to 1649 pixels. Fig. 4 displays a sample of MarsScapes, including a panorama image in (a), semantic annotation for each terrain category in (b) and instance annotation for each terrain in (c). Although the previously released AI4Mars contains four types of terrains, it is insufficient for a higher-level semantic understanding of the Mars. After analyzing the high-risk accidents that may be encountered during the exploration process of rovers [39,40], we divide the Martian environment into eight classes in MarsScapes. The dataset not only contains the four categories that exist in AI4Mars, but also introduces the classes of gravel, steep slope, sky and others to enhance the Martian scene understanding for navigation. Besides, to relieve the influence of perspective effect on ground-level images (i.e., remote terrains look smaller while nearby terrains look larger from the perspective of the camera), we divide the maps vertically into three regions and employ different annotation standards for each region, especially for big rocks and steep slopes.

Table 1 shows quantitative comparison of MarsScapes against other ground-level unstructured terrain segmentation datasets, including three datasets collected on the earth and three datasets acquired from Mars. These datasets have a similar number of unstructured classes, but differ greatly in the quantity of annotated pixels, which is determined by the size and amount of images. Although MarsScapes offers less annotated samples than RUGD [17], it contains the largest annotation scope with widths varying from 1230 to 12062 and heights from 472 to 1649. Moreover, MarsScapes provides annotations for instance segmentation and is the first panorama dataset for extraterrestrial environment.



(a) Panorama image

Class	Soil	Bedrock	Gravel	Sand	Big rock	Steep slope	Sky	Others	Unlabeled
Color	#0000FF	#00FF00	#FF0000	#FF00FF	#FFFF00	#00FFFF	#223813	#AA5500	#000000

(b) Semantic segmentation annotations

(c) Instance segmentation annotations

Fig. 4. MarsScapes contains 195 panorama images (a sample on the 412-th Sol is shown in (a)) with fine-grained annotations for eight semantic categories (in (b)), including soil, bedrock, sand, gravel, big rock, steep slope, sky and others. In addition, instance annotation for each terrain is shown in (c).

Table 1
Statistics of MarsScapes and other ground-level unstructured terrain segmentation datasets.

	Unstructured Classes	Annotated images	Image size	Panorama	Instance
Angelova et al. [30]	6	150	AVG 258×258	×	×
Freiburg Forest [16]	6	366	1024 × 768	×	×
RUGD [17]	8	7456	1376×1110	×	×
AI4Mars-MSL [36]	4	16064	1024 × 1024	×	×
MER-Seg [37]	8	1024	1024 × 1024	×	×
MSL-Seg [37]	8	4155	560 × 500	×	×
MarsScapes	8	195	Width:1230~12062Height: 472~1649	✓	✓

To conduct sufficient evaluations on MarsScapes, we define two tasks: MarsScapes-ALL and MarsScapes-Navigability. The former is a standard semantic segmentation of all classes on Mars. The latter predicts the traversability of terrains, which divides the Martian landforms into three base categories: navigable (soil and bedrock), less navigable (sand, gravel, big rock and steep slope) and unknown (sky and others). Both tasks are helpful in promoting the development of unstructured terrain understanding and rover navigation on Mars. Referring to SkyScapes (another panorama dataset of structured environment) [20] before training, we pre-process the panoramas by cropping them into 256 × 512 patches with 50% overlap between adjacent sub-images in both horizontal and vertical directions. After flipping vertically, we obtain 18460 samples and divide them into 12976 for training, 4184 for validation, and 3640 for testing.

Table 2
Comparison with state-of-the-art methods on the test set of AI4Mars. The proposed HASS outperforms existing methods with dominant advantage.

Method	mIoU	pixAcc
PSPNet [7]	79.08	95.96
DeepLab v3 [25]	82.10	96.12
EncNet [43]	84.75	96.34
GCN [44]	82.54	96.45
DANet [8]	86.90	96.69
CCNet [9]	86.23	96.68
CPNet [10]	87.18	96.79
OCR [11]	88.29	96.98
HASS	**90.76**	**97.10**

4.2. Implementation details

We carry out experiments based on Pytorch with two NVIDIA GeForce RTX 3090 GPUs. The batch size for per GPU is set to 16 for MarsScapes and 8 for AI4Mars. We train the network using the stochastic gradient descent (SGD) algorithm [41] with 0.9 momentum. To reduce network oscillation, we adopt the polynomial learning rate strategy $\gamma = \gamma_0 \times \left(1 - \frac{n_i}{n_{total}}\right)^p$, where total iteration number n_{total} is set to 200, the base learning rate γ_0 is set to 0.01 and p is set to 0.9 for both datasets. For data augmentation, we randomly scale samples with the range of [0.5,2] and perform brightness disturbance within the range of [−5,5] in the training stage. Considering the encoding computation of local heterogeneous matrix $\mathbf{M}$, we obtain $\mathbf{T}_d$ by down-sampling 1/4 of ground truth for MarsScapes (i.e., 128 × 64) and 1/8 for AI4Mars (i.e., 128 × 128). For AI4Mars and MarsScapes-ALL, the output channels of different layers are set as follows: $C_1 = 512$, $C_2 = 256$, and $C = 8$. For MarsScapes-Navigability on MarsScapes, the output channels are set as follows: $C_1 = 256$, $C_2 = 128$, and $C = 3$. To ensure fairness of the comparison, we uniformly choose HRNet-W48 [38] with output stride 4 as the backbone of all networks and initialize it using the model pre-trained on ImageNet [42].

4.3. Experiment results

4.3.1. Experiments on ai4mars

Table 2 presents the IoU, mean IoU (mIoU) and pixel accuracy (pixAcc) indicators of state-of-the-art approaches on AI4Mars, including multi-scale-based methods (PSPNet [7], DeepLab v3 [25], EncNet [43]) and GCN [44]) and attention-based methods (DANet [8], CCNet [9],

Table 3
Results on the MarsScapes-All task of MarsScapes. The proposed HASS exhibits competitive performance over other state-of-the-art methods.

Method	MarsScapes-All								
	Soil	Bedrock	Gravel	Sand	Big rock	Steep slope	Sky	Others	mIoU
PSPNet [7]	55.78	57.32	34.87	60.16	32.59	6.02	55.56	42.30	43.08
DeepLab v3 [25]	63.53	58.02	36.25	61.58	45.30	7.49	58.52	41.51	46.52
GCN [44]	65.88	66.15	45.99	69.30	42.77	5.11	60.05	46.59	50.23
EncNet [43]	70.75	68.00	45.95	66.89	40.17	7.03	61.96	53.12	51.73
DANet [8]	70.88	71.90	46.82	72.72	47.39	**7.53**	68.37	53.95	54.95
CCNet [9]	67.00	65.91	42.09	71.66	40.00	1.64	64.46	49.50	50.28
CPNet [10]	74.45	74.38	44.33	75.73	51.12	0.79	69.74	56.81	55.92
OCR [11]	75.36	72.85	48.42	74.98	47.68	4.95	70.54	54.92	56.21
HASS	**77.22**	**77.67**	**52.00**	**79.08**	**53.33**	3.07	**71.33**	**62.46**	**59.52**

Table 4
Results on the MarsScapes-Navigability task of MarsScapes. Our HASS outperforms existing methods with significant advantage.

Method	MarsScapes-Navigability			
	Navigable	Less navigable	Unknown	mIoU
PSPNet [7]	60.38	56.97	70.60	62.65
DeepLab v3 [25]	68.54	59.78	79.43	69.25
GCN [44]	70.24	65.12	82.00	72.45
EncNet [43]	70.76	63.09	81.20	71.68
DANet [8]	79.18	66.55	88.61	78.11
CCNet [9]	75.37	64.21	85.86	75.15
CPNet [10]	80.58	68.83	87.71	79.04
OCR [11]	79.27	**71.50**	90.49	80.42
HASS	**84.80**	69.10	**92.41**	**82.10**

Table 5
Ablation studies on the test set of MarsScapes-All task to verify the justifiability of each branch and module in HASS. GB represents the global intra-class attention branch, LB represents the local inter-class attention branch, A_m, A_p, and A_c denote the merging strategies of element multiply, element plus, and channel concatenation, respectively.

Branch/Module	Method	mIoU (%)
GB	CCNet	50.28
	DANet	54.95
	CPNet	55.92
	OCR	**56.21**
LB	Ours	40.54
GB + LB + A_m	CCNet + LB + A_m	51.60
	DANet + LB + A_m	55.68
	CPNet + LB + A_m	56.65
	OCR + LB + A_m	**57.22**
GB + LB + A_p	CCNet + LB + A_p	52.12
	DANet + LB + A_p	56.05
	CPNet + LB + A_p	56.98
	OCR + LB + A_p	**57.83**
GB + LB + A_c	CCNet + LB + A_c	55.78
	DANet + LB + A_c	58.38
	CPNet + LB + A_c	59.05
	OCR + LB + A_c (HASS)	**59.52**

CPNet [10], and OCR [11]). It can be seen that attention-based methods are more effective than multi-scale-based methods for unstructured terrain segmentation on Mars. For example, DANet, CCNet, CPNet and OCR exceed DeepLab v3 in mIoU with a large margin of 4.80%, 4.13%, 5.08% and 6.19%, respectively. Further, our HASS achieves 90.76% mIoU and 97.10% pixAcc, which exhibits promising advantage over existing methods and even surpasses OCR that performs prominently in structured scenes segmentation.

4.3.2. Experiments on marsscapes

To further evaluate the effectiveness of our network, we carry out semantic segmentation experiments on the MarsScapes-All and MarsScapes-Navigability tasks in Tables 3 and 4. The first and second best results are highlighted with bold and underline, respectively. For the MarsScapes-All task, we can learn that HASS is 3.31% higher than the second best method OCR [11] and performs the best prediction for each single class except steep slope. We speculate that the exception on steep slope is due to its small proportion compared with other categories, so the IoU of steep slope is less capable to evaluate the network performance. For the MarsScapes-Navigability task, our approach exhibits the first or the second best IoU in each class and increases the mIOU from 80.42% to 82.10%, which confirms the superiority of our HASS over existing state-of-the-art methods in predicting navigable terrains. Moreover, the results of these approaches on MarsScapes-Navigability are much higher than MarsScapes-All. It is because the features of each class become more distinguishable after merging eight terrains into three base classes, which makes it easier to differentiate each category.

Fig. 5 illustrates the visual comparison of our HASS with two representative approaches, including a multi-scale-based method PSPNet and an attention-based model DANet. The results of HASS are the most similar to ground truth, but it occasionally gives some wrong labels, such as recognizing bedrocks as big rocks. DANet often predicts less or more terrains and there are a few flaws in the map. PSPNet sometimes has large areas of mislabeling and predicts with more flaws. Besides, there are some common problems. For example, because the edge of bedrock is quite similar to that of big rock (the details of specific distinguishing method is provided in the supplementary material), it is difficult for the three models to distinguish them clearly.

Furthermore, we have visualized the local inter-class attention maps for better understanding of HASS in the sixth column of Fig. 5, where m represents the terrain to be segmented and n_i shows one of heterogeneous terrains adjacent to m. The whiter n_i is, the closer its relationship to m. It can be observed that the local attention branch of HASS could clearly capture the relationships between m and most of its neighboring inter-class terrains n_i. In the first sample, a big rock m is marked and its local inter-class attention map highlights areas of neighboring gravel n_1 and bedrock n_3. In the second sample, when soil m is marked, HASS could pay more attention to the neighboring sand n_2 and two big rocks $n_1 + n_3$ than to other terrains. In the third sample, the soil n_2 and sand n_3 are highlighted around the selected bedrock m. Although there are some exceptions (the relationship between m and n_2 in the first sample and that between m and n_1 in the third sample are unobvious), HASS achieves promising performance with the help of the local inter-class attention branch.

4.4. Ablation studies

To verify the justifiability of HASS, we conduct ablation experiments on the test set of MarsScapes-All task with different settings in Table 5. We first only use the global attention branch (i.e., GB) for segmentation, where four advanced methods: CCNet [9], DANet [8], CPNet [10], and OCR [11] are taken into consideration. Then, we only employ our local attention branch (i.e., LB) for features exploration. Finally, we integrate the two branches with the representation merging module, where A_m, A_p, and A_c denote the merging strategies of element multiply,

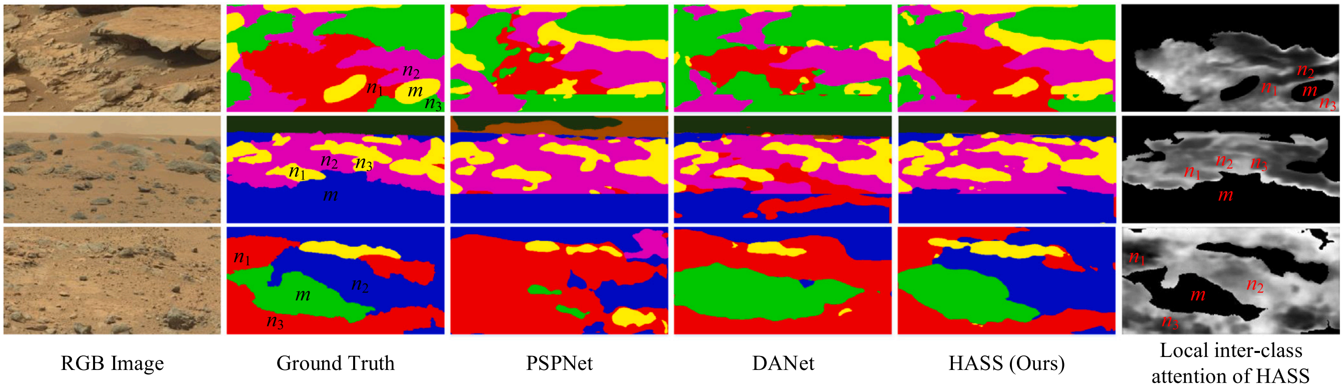


Fig. 5. Visualization results on the test set of MarsScapes-All task.

Table 6
Performance comparison between different loss function settings on the test set of MarsScapes-All task.

Method	mIoU (%)
L_m	56.24
$L_m + L_g$	57.39
$L_m + L_g + L_{ld}$ $(N_R = 0)$	58.45
$L_m + L_g + L_{ld}$ $(N_R = 5)$	59.11
$L_a + L_g + L_{ld}$ $(N_R = 10)$	**59.52**
$L_m + L_g + L_{ld}$ $(N_R = 15)$	58.98
$L_m + L_g + L_{ld}$ $(N_R = 20)$	58.23
$L_m + L_g + L_{ld}$ $(N_R = 25)$	57.06

element plus, and channel concatenation, respectively. Although the mIoU of our local attention branch is much lower than that of the global branch, the performance is remarkably improved when they are combined, which suggests that the features extracted by the two branches are highly complementary for unstructured terrain segmentation. Meanwhile, the merging module with channel concatenation achieves competitive advantage over other strategies. Taking the dual-branch of OCR + LB for example, A_c is 1.69% higher than A_p and 2.30% higher than A_m.

Furthermore, we demonstrate the effectiveness of proposed loss functions on the test set of MarsScapes-All task in Table 6. Compared with the base loss L_m of representation merging module, an additional calculation of the semantic loss of the global attention branch (i.e., $L_m + L_g$) yields a result of 57.39% in mIoU, acquiring a 1.15% improvement. To evaluate the role played by the local diversity loss, we add it to the total loss. It can be seen that setting suitable value of N_R contributes to performance improving. When $N_R = 10$, the total loss gain the highest score of 59.52%, which outperforms $L_m + L_g$ by 2.13%. However, the performance decreases with the unrestricted increase of N_R. When $N_R = 25$, the mIoU is 57.06%, even lower than the loss without L_{ld}.

5. Conclusion

In this paper, we propose a novel attention-based network called HASS, which utilizes both local inter-class and global intra-class contexts for the unstructured terrain segmentation on Mars. Moreover, we establish MarsScapes, the first panoramic image dataset for Martian terrain segmentation. Through quantitative and qualitative experiments on MarsScapes and AI4Mars, we find that the proposed method can integrate specific regional relationships with long-range consistencies, which reports significant improvements and achieves new state-of-the-art performances on the MarsScapes and AI4Mars datasets. In the future researches, we would focus on making the framework more lightweight and introducing domain adaptation to the semantic segmentation of Martian terrain.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was sponsored by the National Natural Science Foundation of China (NSFC) through grants No. 62103163 and No. 62003055, Natural Science Foundation of Jilin Province, China through grant No. YDZJ202101ZYTS033, and Chinese University Industry-Academia-Research Innovation Fund from Science and Technology Development Center of the Ministry of Education through grant No. 2020HYA09001. We thank the above mentioned funds for their financial support.

References

[1] M. Siam, M. Gamal, M. Abdel-Razek, S. Yogamani, M. Jagersand, H. Zhang, A comparative study of real-time semantic segmentation for autonomous driving, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 587–597.

[2] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, K. Dietmayer, Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges, IEEE Trans. Intell. Transp. Syst. 22 (3) (2020) 1341–1360.

[3] Y. Guo, Y. Liu, T. Georgiou, M.S. Lew, A review of semantic segmentation using deep neural networks, Int. J. Multimedia Inf. Retr. 7 (2) (2018) 87–93.

[4] D. Rozenberszki, G. Sörös, S. Szeier, A. Lőrincz, 3D semantic label transfer in human-robot collaboration, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2602–2611.

[5] H.J. Lee, J.U. Kim, S. Lee, H.G. Kim, Y.M. Ro, Structure boundary preserving segmentation for medical image with ambiguous boundary, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4817–4826.

[6] A. Raju, C.-T. Cheng, Y. Huo, J. Cai, J. Huang, J. Xiao, L. Lu, C. Liao, A.P. Harrison, Co-heterogeneous and adaptive segmentation from multi-source and multi-phase CT imaging data: a study on pathological liver and lesion segmentation, in: European Conference on Computer Vision, Springer, 2020, pp. 448–465.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.

[8] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.

[9] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 603–612.

[10] C. Yu, J. Wang, C. Gao, G. Yu, C. Shen, N. Sang, Context prior for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12416–12425.

[11] Y. Yuan, X. Chen, J. Wang, Object-contextual representations for semantic segmentation, 2019, arXiv preprint arXiv:1909.11065.

[12] A. Petrovsky, I. Kalinov, P. Karpyshev, D. Tsetserukou, A. Ivanov, A. Golkar, The two-wheeled robotic swarm concept for Mars exploration, Acta Astronaut. 194 (2022) 1–8.

[13] M. Christian, M. Michael, H. Nikolaus, P.P. Jesus, F. Friedrich, Semantic drone dataset, 2019, https://www.tugraz.at/index.php?id=22387. (Accessed 25 January 2019).

[14] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, L. Zhang, Learning transferable deep models for land-use classification with high-resolution remote sensing images, 2018, arXiv preprint arXiv:1807.05713.

[15] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, R. Raskar, Deepglobe 2018: A challenge to parse the earth through satellite images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 172–181.

[16] A. Valada, G.L. Oliveira, T. Brox, W. Burgard, Deep multispectral semantic scene understanding of forested environments using multimodal fusion, in: International Symposium on Experimental Robotics, Springer, 2016, pp. 465–477.

[17] M. Wigness, S. Eum, J.G. Rogers, D. Han, H. Kwon, A RUGD dataset for autonomous navigation and visual perception in unstructured outdoor environments, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2019, pp. 5000–5007, http://dx.doi.org/10.1109/IROS40897.2019.8968283.

[18] M. Yu, H. Cui, S. Li, Y. Tian, Database construction for vision aided navigation in planetary landing, Acta Astronaut. 140 (2017) 235–246.

[19] M.E. Program, Mars curiosity rover, 2015, https://mars.nasa.gov/msl/multimedia/raw-images. (Accessed 1 April 2021).

[20] S.M. Azimi, C. Henry, L. Sommer, A. Schumann, E. Vig, Skyscapes fine-grained semantic understanding of aerial scenes, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7393–7403.

[21] G. Lin, C. Shen, A. Van Den Hengel, I. Reid, Efficient piecewise training of deep structured models for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3194–3203.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[23] H. Liu, C. Peng, C. Yu, J. Wang, X. Liu, G. Yu, W. Jiang, An end-to-end network for panoptic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6172–6181.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.

[25] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, 2017, arXiv preprint arXiv:1706.05587.

[26] M. Yang, K. Yu, C. Zhang, Z. Li, K. Yang, Denseaspp for semantic segmentation in street scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3684–3692.

[27] X. Liu, P. Lu, B. Pan, Survey of convex optimization for aerospace applications, Astrodynamics 1 (1) (2017) 23–40.

[28] D. Izzo, M. Märtens, B. Pan, A survey on artificial intelligence trends in spacecraft guidance dynamics and control, Astrodynamics 3 (4) (2019) 287–299.

[29] M. Yu, S. Li, X. Huang, S. Wang, A novel inertial-aided feature detection model for autonomous navigation in planetary landing, Acta Astronaut. 152 (2018) 667–681.

[30] A. Angelova, L. Matthies, D. Helmick, P. Perona, Slip Prediction Using Visual Information, MIT Press, 2007.

[31] P. Jiang, P. Osteen, M. Wigness, S. Saripalli, Rellis-3d dataset: Data, benchmarks and analysis, 2020, arXiv preprint arXiv:2011.12954.

[32] P. Furgale, P. Carle, J. Enright, T.D. Barfoot, The devon island rover navigation dataset, Int. J. Robot. Res. 31 (6) (2012) 707–713.

[33] C.H. Tong, D. Gingras, K. Larose, T.D. Barfoot, E. Dupuis, The Canadian planetary emulation terrain 3D mapping dataset, Int. J. Robot. Res. 32 (4) (2013) 389–395.

[34] L. Meyer, M. Smíšek, A. Fontan Villacampa, L. Oliva Maza, D. Medina, M.J. Schuster, F. Steidle, M. Vayugundla, M.G. Müller, B. Rebele, et al., The MADMAX data set for visual-inertial rover navigation on Mars, J. Field Robotics 38 (6) (2021) 833–853.

[35] X. Xiao, M. Yao, H. Liu, J. Wang, L. Zhang, Y. Fu, A kernel-based multi-featured rock modeling and detection framework for a Mars rover, IEEE Trans. Neural Netw. Learn. Syst. (2021).

[36] R.M. Swan, D. Atha, H.A. Leopold, M. Gildner, S. Oij, C. Chiu, M. Ono, Ai4mars: A dataset for terrain-aware autonomous driving on mars, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1982–1991.

[37] J. Li, S. Zi, R. Song, Y. Li, Y. Hu, Q. Du, A stepwise domain adaptive segmentation network with covariate shift alleviation for remote sensing imagery, IEEE Trans. Geosci. Remote Sens. (2022).

[38] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, J. Wang, High-resolution representations for labeling pixels and regions, 2019, arXiv preprint arXiv:1904.04514.

[39] J. Carsten, A. Rankin, D. Ferguson, A. Stentz, Global path planning on board the mars exploration rovers, in: 2007 IEEE Aerospace Conference, IEEE, 2007, pp. 1–11.

[40] R. Francis, T. Estlin, D. Gaines, B. Bornstein, S. Schaffer, V. Verma, R. Anderson, M. Burl, S. Chu, R. Castaño, et al., AEGIS autonomous targeting for the Curiosity rover's ChemCam instrument, in: 2015 IEEE Applied Imagery Pattern Recognition Workshop, AIPR, IEEE, 2015, pp. 1–5.

[41] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 1097–1105.

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

[43] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, Context encoding for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7151–7160.

[44] C. Peng, X. Zhang, G. Yu, G. Luo, J. Sun, Large kernel matters–improve semantic segmentation by global convolutional network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4353–4361.