Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SD-SHAP: A Semantic Distinction-Based SHAP Explanation Method for Mars Terrain Segmentation Models

jinshuo qu
Nanjing Tech University
mujia shi
Nanjing Tech University
ziyang jia
Southeast University
lihang Feng
lfeng8@njtech.edu.cn

Nanjing Tech University
dong wang
Southeast University
yong Hu
Beijing Institute of Technology
jiantao shi
Nanjing Tech University

Research Article

Keywords: Explainable artificial intelligence, Semantic segmentation, SHAP, Black-box explanation, Mars terrain perception, Region-level attribution

Posted Date: May 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9554905/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SD-SHAP: A Semantic Distinction-Based SHAP Explanation Method for Mars Terrain Segmentation Models

Jinshuo Qu · Mujia Shi · Ziyang Jia · Lihang Feng · Dong Wang · Yong Hu · Jiantao Shi

Corresponding author: Lihang Feng
Email: lfeng8@njtech.edu.cn

Abstract

Artificial intelligence is expected to play an important role in Mars rover vision systems, yet the limited transparency of deep neural networks remains a concern for reliable deployment. Although many explainability methods have been developed for image-level prediction tasks, methods tailored to semantic segmentation in Mars terrain analysis are still limited. To address this issue, this paper presents Semantic Distinction-Based Shapley Additive Explanations, an extension of Shapley Additive Explanations (SHAP) for Mars terrain segmentation models. The proposed method constructs semantically coherent feature units and introduces a segmentation-oriented contribution formulation, enabling SHAP to better reflect region-level decision behavior in dense prediction tasks. The method is post hoc and does not require access to internal model parameters, while its current form assumes the availability of semantic priors for the analyzed sample. Experiments on the AI4Mars dataset show that, under the proposed region-completeness-based evaluation setting, SD-SHAP consistently outperforms baselines based on SHAP and Local Interpretable Model-agnostic Explanations (LIME) in contribution direction consistency, alignment with output variation, and stability.

Keywords: Explainable artificial intelligence; Semantic segmentation; SHAP; Black-box explanation; Mars terrain perception; Region-level attribution

1 Introduction

Over the past decades, multiple unmanned rovers have been launched to the surface of Mars[1][2]. These rovers have achieved great success in helping humans understand the planet. However, due to the presence of numerous hazardous terrains, rover mobility still faces significant challenges. For example, the well-known Mars rover Spirit became permanently immobilized after being trapped in soft sand[3][4].

At present, most rover navigation plans still rely on images transmitted back to Earth for manual assessment and path planning[5]. With the continuous development of deep learning, semantic segmentation-based vision models are expected to replace traditional machine vision systems, thereby enabling autonomous path planning and navigation for Mars rovers[6]. The key to achieving this goal lies in whether the decisions of vision models can be trusted.

Recent studies on Mars image segmentation have moved toward more task-specific model design, and this trend also makes it increasingly important to examine how model decisions should be interpreted and validated in practical use[7]. There are two main reasons for this. On the one hand, a single poor decision may place the rover in a dangerous situation and lead to costly human and material losses[8]. On the other hand, the "black-box" nature of vision models deployed on rovers makes it difficult for researchers to understand their decision-making logic. In this setting, explainability is useful not only for transparency, but also for failure diagnosis and offline verification, because acceptable segmentation accuracy does not by itself guarantee that the model relies on stable and meaningful visual evidence. This concern is consistent with broader findings in computer vision and Explainable Artificial Intelligence (XAI), where deep models are known to be affected by texture bias and spurious correlations, while explanation and evaluation methods for semantic segmentation remain comparatively limited[9][10].

To address the problem of decision opacity in image network models, many explainability methods have emerged in the field of explainable artificial intelligence. The DeConvNets (Deconvolutional Networks) method proposed by Zeiler and Fergus [11] reconstructs the forward process of Convolutional Neural Networks (CNNs) through

Corresponding author: Lihang Feng
lfeng8@njtech.edu.cn

Jinshuo Qu
qujinshuo@njtech.edu.cn

Mujia Shi
202461106034@njtech.edu.cn

Ziyang Jia
220243813@seu.edu.cn

Dong Wang
kingeast16@seu.edu.cn

Yong Hu
nhhy@hotmail.com

Jiantao Shi
shijt@njtech.edu.cn

1 College of Control Science and Engineering, Nanjing Tech University, Nanjing 211816, China

2 School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China

3 Science and Technology on Space Intelligent Control Laboratory, Beijing Institute of Control Engineering, Beijing 100190, China

deconvolution and unpooling, mapping high-level abstract features back to the pixel space, and it was the first to achieve pixel-level visualization of CNN features. The CAM (Class Activation Mapping) method proposed by Zhou et al. [12] uses global average pooling and the weights of the fully connected layer to associate each channel of the final convolutional feature map with the importance of the target class, generating a heatmap. The Grad-CAM (Gradient-weighted Class Activation Mapping) method proposed by Selvaraju et al. [13] replaces the weights in CAM with the gradients of the target class with respect to the final convolutional feature maps, thereby calculating the importance of each channel for class decisions. In addition, the Attention Rollout method proposed by Abnar and Zuidema [14] propagates attention weights recursively to quantify the information flow path in the model.

However, many of the aforementioned methods rely on internal gradients, feature maps, or architecture-specific information, which limits their use in black-box deployment settings. Moreover, for terrain segmentation models, an effective explanation method should provide richer information, such as whether a feature contributes positively or negatively, in order to facilitate a deeper understanding of model decisions. To address this issue, this paper proposes an improved SHapley Additive exPlanations (SHAP) method, namely SD-SHAP (A Semantic Distinction-Based SHAP Explanation Method), an extension of KernelSHAP for Mars terrain segmentation[15]. The method is post-hoc and does not require access to internal model parameters, but it assumes a semantic prior map for the sample under analysis. The main contributions of this work are as follows:

- We propose a semantic-consistent construction strategy for feature patches. During the construction process, the semantic segmentation label map of the target class is incorporated, while the model prediction probability map is introduced to emphasize the regions of primary interest to the model. This results in feature patches that not only represent the local characteristics of the class but are also closely aligned with the semantic information of that class.
- We design a contribution calculation mechanism specifically tailored to terrain segmentation tasks. In contrast to the mainstream mechanisms designed for image classification, the proposed mechanism is customized for segmentation, including the selection of the baseline value and the construction of the feature matrix.

The remainder of this paper is organized as follows. Section 2 reviews the progress of related explainability methods. Section 3 presents the framework and implementation of the proposed improved SHAP method based on semantic distinction. Section 4 validates the effectiveness of the method through experiments. Finally, Section 5 concludes the paper.

2 Related work

2.1 Segmentation Models for Terrain Perception

In recent years, deep learning has been widely applied to image understanding tasks, including semantic segmentation and robot perception. In dense prediction, the DeepLab series remains a representative framework because of its strong multi-scale feature extraction capability[16-19]. Beyond such backbone architectures, the field of image segmentation has continued to improve prediction quality through stronger attention mechanisms, feature fusion, and spatial-context modeling. Representative examples include Lv et al.'s hybrid-attention semantic segmentation network for remote sensing interpretation [20], Hu et al.'s ULAF-Net for real-time semantic segmentation [21], and Li et al.'s SCFI-ESeg for efficient Transformer-based segmentation [22]. In Mars exploration scenarios, representative studies include Mobile-DeepRFB proposed by Feng et al. [23], Light4Mars proposed by Xiong et al. [24], and LisseMars proposed by Lin et al. [25]. Although these studies have substantially improved terrain perception performance, they mainly focus on prediction quality itself rather than post hoc explanation of region-level decisions. This leaves open the question of how such models arrive at a specific terrain prediction and which local regions support or suppress that decision.

2.2 Post-hoc Explainability Methods for Image Models

Recent studies have classified post-hoc explainability methods into two categories: model-agnostic and model-specific [26]. Among model-specific methods, the most well-known series is CAM, which has further derived Grad-CAM, Grad-CAM++ proposed by Chattopadhyay et al. [27], and Score-CAM proposed by Wang et al. [28]. Among them, the most widely used is Grad-CAM, whose basic principle is to trace the backpropagation process from the prediction output to the input, thereby attributing the prediction results to the input features, and analyzing the contribution of each feature to the final decision. This method has the advantages of high computational efficiency and class-discriminative visualization.

In model-agnostic methods, a typical representative is the LIME (Local Interpretable Model-Agnostic Explanations) method proposed by Ribeiro et al. [29]. The core idea of this method is to generate a series of new samples in the neighborhood of the target instance by perturbing the input, and then use the original black-box model to obtain the predictions for these samples. A simple and interpretable surrogate model (such as a linear model) is then trained on these samples, aiming to locally approximate the decision boundary of the complex model. The process can be formalized as a local loss minimization problem:

$$\underset{g \in G}{argmin}\, \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{1}$$

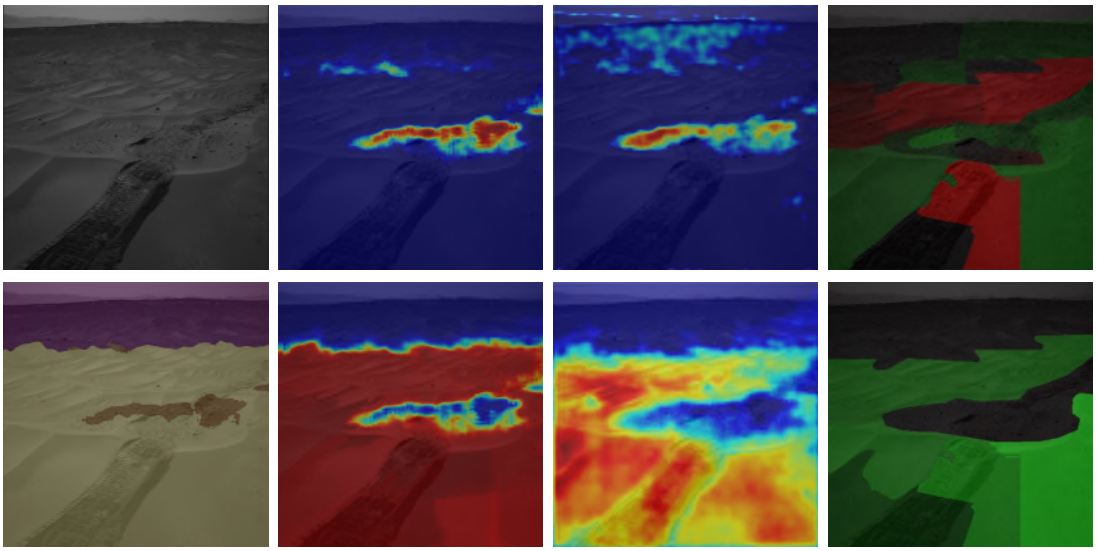
where f denotes the complex black-box model to be explained, g is a simple model (e.g., a linear model) selected from the family of interpretable models $\mathcal{G}$, $\mathcal{L}$ is a loss function that measures the local approximation accuracy of g to f, π_x is a weight distribution over neighborhood samples around the instance x, and $\Omega(g)$ is a regularization term that controls the model complexity.

Another well-known method is RISE (Randomized Input Sampling for Explanation), proposed by Petsiuk et al. [30]. It generates a large number of random binary masks to create perturbed images, and records the class score of each mask while predicting the perturbed images, thereby producing a pixel-level importance map. This random sampling strategy reduces the potential bias caused by fixed partitioning. The core idea of RISE is to use the weighted sum of the model responses to masked images to evaluate the importance of each pixel in the original image. Its core formula is expressed as:

$$S_I(x,y) \approx \frac{1}{N \cdot E[M]} \sum_{i=1}^{N} f\,(I \odot M_i) \cdot M_i(x,y) \qquad (2)$$

where $S_I(x,y)$ is the importance score of pixel (x,y), I is the original image, M_i is the i-th random mask, N is the total number of masks, $f(I \odot M_i)$ is the model prediction score for the masked image, and $E[M]$ is the expected value of the masks.

Up to now, most existing explanation methods mainly focus on image-level prediction tasks such as image classification or image captioning, where the core goal is to explain why an image belongs to a certain category [31], [32]. However, research on why a specific image region is assigned to a certain category remains much more limited [33]. This limitation becomes more evident in semantic segmentation, where the target of explanation is a structured region-level output rather than a single class label. For Martian terrain segmentation, an explanation method should preferably operate in a post-hoc manner under black-box access, retain a clear mathematical basis, and provide region-level contribution information rather than only coarse attention maps. These requirements motivate the development of segmentation-oriented perturbation-based explanation methods.



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

Fig. 1.
(a) Input image (top) and model output (bottom)
(b) Explanation generated by the RISE method;
(c) Explanation generated by the Grad-CAM method
(d) Explanation generated by SD-SHAP

For example, as shown in Fig. 1, the saliency maps generated by the model-agnostic method RISE and the model-specific method Grad-CAM are visually similar to segmentation maps, which means that they can well reveal the decision basis of image classification tasks because their heatmaps can show the regions of interest to the model. However, when facing image segmentation tasks, the information they provide is relatively limited, while SD-SHAP can provide more detailed information, such as the contribution direction (green for positive contributions and red for negative contributions) of the feature patches around the segmentation regions to the model decision.

In addition, among model-agnostic interpretation methods, the feature units used to provide positive and negative contribution measures are usually constructed via masks. Commonly used mask construction approaches include superpixel-based segmentation [34] and rule-based shape generation [35], which can rapidly generate interpretable feature units for global prediction tasks [36]. However, in terrain segmentation tasks, these feature units may suffer from semantic ambiguity, since a single unit may contain multiple categories. In such cases, the interpretation results cannot be attributed to a single category, which in turn reduces interpretability. As shown in Fig. 2, three sample patches generated by superpixel segmentation are presented. The feature patch labeled 10 contains two categories: rock and soil. This semantic mixing makes it difficult to distinguish the specific contributing category when calculating the contribution of this patch. The other two images illustrate the semantic fragmentation caused by texture that may occur in superpixel segmentation, where the rock category is not distinguished as a single integral patch. A similar problem exists with rule-based patches. On this basis, we choose to use feature patches based on semantic segmentation to avoid such situations as much as possible.



(a) (b) (c)

Fig. 2. Examples of semantic ambiguity and fragmentation caused by superpixel segmentation on Mars terrain images

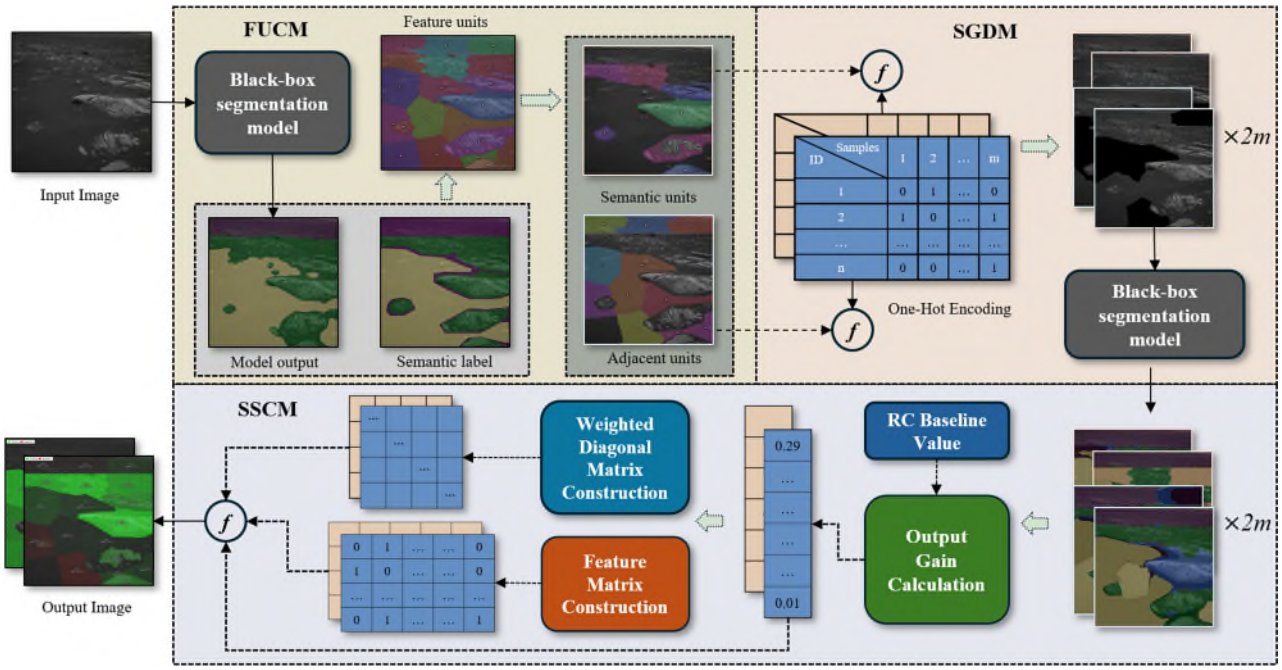


Fig. 3. The structural framework of SD-SHAP

3 Method

3.1 The Method Framework

To explain the output of Mars terrain segmentation models under black-box access, this paper proposes SD-SHAP, a post-hoc explanation framework. The current form of the framework does not require gradients or internal parameters, but it assumes a semantic prior map for the sample under analysis. As shown in Fig. 3, the framework consists of FUCM (Feature Unit Construction Module), SGDM (Sampling-Generated Disturbance Module), and SSCM (Segmentation-Based SHAP Calculation Module).

In FUCM, to avoid semantic ambiguity in feature units, semantic label maps are used to assist in constructing class masks. At the same time, to capture high-probability regions of the model, the probability map of the model output is introduced. Finally, considering both semantic ranges and high-probability regions, superpixel segmentation is applied to generate feature units with semantic consistency, which are further divided into semantic groups and adjacent groups according to spatial semantic relationships. Compared with traditional superpixel segmentation and rule-based shape segmentation, the semantic feature units obtained in this way provide better explainability and are more consistent with human cognition.

In SGDM, different local sampling strategies are applied to semantic groups and adjacent groups according to the number of feature units. All sampling subsets are then used to generate disturbance samples by randomly replacing pixels in the regions corresponding to the same mask ID. To further improve both stability and robustness of the results, a perturbation enhancement mechanism is introduced, where multiple disturbed images are generated for each disturbance sample, and the final output result considers all disturbed images.

In SSCM, the RC (Region Completeness) baseline value and sampling weights are calculated to measure the model gain. Then, the disturbed samples are fed into the model for prediction, and the output gains are computed. Finally, a feature unit matrix is constructed based on the activation state of each feature unit within the subsets (activated = 1, not activated = 0). Using the feature unit matrix, the output gains, and the sampling weights, a linear system is established, and the explanation results for each feature unit are obtained by solving this system.

3.2 Semantic-consistent Feature Units

In image semantic segmentation tasks, the model outputs the class probability for each pixel. However, attempting to explain the model behavior at the pixel level requires a large amount of computation and time, and the explanation results are also difficult for humans to understand. A more practical approach is to aggregate pixel-level fragmented information into semantically meaningful high-level feature units, which are referred to as feature patches in image processing tasks. Using feature patches as the minimum explanatory unit can not only significantly reduce computational complexity but also facilitate human understanding.

At present, the traditional construction methods of feature patches include superpixel segmentation and rule-based shape segmentation. These segmentation methods can provide fast and convenient feature patches when handling global prediction tasks (such as image classification). However, when dealing with terrain segmentation tasks, their segmentation approaches are prone to semantic ambiguity. To address this issue, this paper proposes a segmentation method that ensures semantic consistency of feature patches.

First, a semantic prior map of the image to be explained is introduced. In the experiments of this paper, this map is taken from the ground-truth label map. Let $L(x,y)$ denote the semantic label (x,y) at coordinate of the input image, and let C be the number of semantic categories. For each category $c \in 1,2,\dots,C$, the corresponding semantic mask is defined as:

$$\mathcal{M}_c^L = \{(x,y) \mid L(x,y) = c\} \tag{3}$$

To distinguish the importance of each feature patch, the high-confidence regions of the model prediction should be considered in the feature patch construction process. For this purpose, the model output probability map $P(x,y,c)$ is introduced, which represents the predicted probability that pixel (x,y) belongs to class c. Based on a probability threshold α, the high-confidence region H_c and the low-confidence region L_c of class c are defined as:

$$\begin{aligned} H_c &= \{(x,y) \mid P(x,y,c) \geq \alpha\} \\ L_c &= \{(x,y) \mid P(x,y,c) < \alpha\} \end{aligned}, \alpha \in (0,1) \tag{4}$$

Then, superpixel segmentation is applied to regions H_c and L_c respectively. The superpixel set obtained from H_c is denoted as $\{S_{c,i}^h\}$, and the superpixel set obtained from L_c is denoted as $\{S_{c,j}^l\}$. These superpixels serve as the candidate feature patches for class c. To control computational complexity, the total number of feature patches for each class is restricted within a fixed range. When the number of generated superpixels exceeds this threshold, they are merged in ascending order according to their area and contact area until the number of units falls below the threshold.

3.3 Sampling-Generated Disturbance

Explanation methods such as LIME and SHAP, which fit a local model to approximate a complex model, require a large number of sampled instances to ensure accuracy. This inevitably increases computational time and resource consumption. Therefore, this paper introduces a targeted optimization of the sampling strategy. For all feature blocks of the object to be explained, we divide them into two groups according to semantic spatial logic and then perform sampling separately. This effectively reduces the number of feature blocks involved in computation, thereby lowering the demand for resources and time. The sampling strategy requires that the feature blocks in each group have low semantic correlation, so that no strong interaction occurs between them.

For a target semantic category c, let the corresponding feature block set be $\mathcal{U}_c$. We divide $\mathcal{U}_c$ into two groups: the semantic group G_{sem} and the adjacency group G_{adj}. The semantic group G_{sem} consists of feature blocks overlapping with the ground-truth region of category c, that is, any $u \in \mathcal{U}_c$ satisfying $u \cap \mathcal{M}_c^L \neq \emptyset$. The adjacency group G_{adj} consists of the feature blocks not included in G_{sem}, but spatially adjacent to at least one block in G_{sem}, i.e., sharing a boundary. Alternatively, a block also belongs to G_{adj} if it lies within the high-confidence radiation range of category c and has a strong influence on it. Since the subsequent treatment of the two groups is the same, we collectively refer to them as the feature group $\mathcal{G}_c$. Perturbed samples are generated by removing feature blocks from $\mathcal{G}_c$. Let $N = |\mathcal{G}_c|$ denote the number of feature blocks in the group. Different sampling strategies are adopted depending on the value of N.

When the number of feature blocks is small (e.g., $N \leq 12$), enumeration can be used to further improve the explanation accuracy. All possible subsets of the feature group $\mathcal{G}_c$, including the empty set and the full set, are enumerated. In total there are $M = 2^N$ subsets, denoted as the set $\mathcal{S}$. When the number of feature blocks is large (e.g., $N > 12$), stratified sampling is adopted, specifically stratified Monte Carlo sampling. A total sampling budget M is specified. The power set of the feature group $\mathcal{G}_c$ is stratified according to the subset size $k = 1,2,\dots,N-1$. Each stratum contains $\binom{N}{k}$ subsets. The sampling budget M is allocated to each stratum in proportion to $\binom{N}{k}$, and within each stratum, M_k subsets are selected by Monte Carlo random sampling. The sampled subset collection is denoted as S:

$$M_k = \frac{\binom{N}{k}}{\sum_{j=1}^{N-1}\binom{N}{j}} \cdot M \tag{5}$$

For each subset $S \in \mathcal{S}$, a perturbed image is generated by removing the feature blocks in S. Specifically, a perturbed image x_S is constructed by replacing the pixels of each feature block u in S with the pixels at the same locations from another image x',where x' is sampled from the same dataset. That is, for all $(x,y) \in u$,let $x_S(x,y) = x'(x,y)$.Moreover, to reduce the bias caused by random image selection, R independent perturbed images are generated for each subset S The corresponding perturbed image set is denoted as $\{x_S^{(1)},\dots,x_S^{(R)}\}$. For each perturbed set, the images are input into the model and the outputs are recorded. Combined with the one-hot encoding of subset S, these prediction results are subsequently used to calculate the contribution values of the feature blocks.

3.4 Segmentation-Based SHAP Calculation

Based on the feature patches and perturbed samples generated above, our goal is to approximate the Shapley values through weighted linear model fitting. Considering that the explanation task targets terrain segmentation models, this paper conducts targeted optimizations on SHAP, and these improvements help enhance the explanation accuracy.

RC (Region Completeness) baseline value: When applying the SHAP method to explain global prediction tasks, the model output on a plain image can be chosen as the baseline. A better approach is to simulate the global output of the model without any feature input. For terrain segmentation tasks, we expect the baseline value to reflect the region completeness of the model prediction within the actual area of the target class c in the test image. Let $\mathcal{M}_c^L$ denote the semantic region of class c. The baseline is defined as the proportion of pixels within $\mathcal{M}_c^L$ predicted as class c:

$$baseline = \frac{1}{|\mathcal{M}_c^L|}\sum_{(x,y)\in\mathcal{M}_c^L} 1\left(\operatorname{argmax}_k f(x,y) = c\right) \tag{6}$$

where $f(x,y)$ denotes the model's predicted label at pixel

(x, y), and $1(\bullet)$ is the indicator function. This baseline reflects the model's inherent tendency to predict class c for this region without perturbation.

Model output gain: For each subset S, we input the corresponding perturbed set $\{x_S^{(1)}, \ldots, x_S^{(R)}\}$ into the model and obtain the predictions. Let p_S denote the mean proportion of pixels predicted as class c within $\mathcal{M}_c^L$ over the R perturbed images:

$$p_S = \frac{1}{R}\sum_{r=1}^{R} \frac{1}{|\mathcal{M}_c^L|} \sum_{(x,y) \in \mathcal{M}_c^L} 1\left(\underset{k}{\operatorname{argmax}}\, f\left(x_S^{(r)}, (x, y)\right) = c\right) \quad (7)$$

the prediction gain is then defined as:

$$y_S = p_S - baseline \quad (8)$$

where y_S indicates the change, relative to the RC baseline, in the proportion of pixels within the region $\mathcal{M}_c^L$ predicted as class c after removing the feature blocks in subset S.

Sample weights: According to the possible sampling strategies described in Section 3.3, different weighting schemes are used. When the number of feature blocks is small and the enumeration strategy is adopted, the sampling weights are calculated exactly. For any feature subset $S \subset \mathcal{G}_c$, the weight is defined as:

$$\omega(S) = \begin{cases} \frac{|S|!\cdot(N-|S|-1)!}{N!}, & S \subset \mathcal{G}_c \setminus \{i\}, \forall i \\ \frac{1}{N}, & S = \emptyset \ \vee \ S = \mathcal{G}_c \end{cases} \quad (9)$$

When the number of feature blocks is large and stratified Monte Carlo sampling is applied, the kernel weight for a subset in the k-th layer S_k is defined as:

$$\omega(S) = \frac{(N-1)}{\binom{N}{|S_k|}\cdot|S_k|\cdot(N-|S_k|)\cdot M_k} \quad (10)$$

Linear system construction: For any subset $S \subset \mathcal{G}_c$, it is encoded into an N-dimensional one-hot vector $x_s \in \{0,1\}^N$, where $x_{s,i} = 1$ indicates that the i-th feature block is included in S, and 0 otherwise. All subset vectors are stacked row by row to form a matrix $X \in \mathbb{R}^{M\times N}$, where M is the number of subsets and N is the number of feature blocks. Let $y \in \mathbb{R}^M$ denote the vector of model outputs corresponding to all subsets, and let $W \in \mathbb{R}^{M\times M}$ be a diagonal weight matrix with diagonal elements $W_{ss} = \omega(S)$, the weight of subset S. The computation of SHAP values is then formulated as a weighted least-squares regression problem:

$$\min_{\phi} \sum_{S\in\mathcal{S}} \omega\,(S) \cdot (y_S - x_S^T \phi)^2 \quad (11)$$

In matrix form, this can be written as:

$$\min_{\phi} (X\phi - y)^T W (X\phi - y) \quad (12)$$

Taking the derivative of the weighted least-squares objective with respect to ϕ and setting the gradient to zero, the closed-form solution is obtained as:

$$\phi = (X^\top W X)^{-1} X^\top W y \quad (13)$$

Here, the solution vector $\phi \in \mathbb{R}^N$ contains the Shapley values for each feature block. The i-th element ϕ_i represents the marginal contribution of the i-th feature block to the prediction of class c within the target semantic region.

4 Experiment

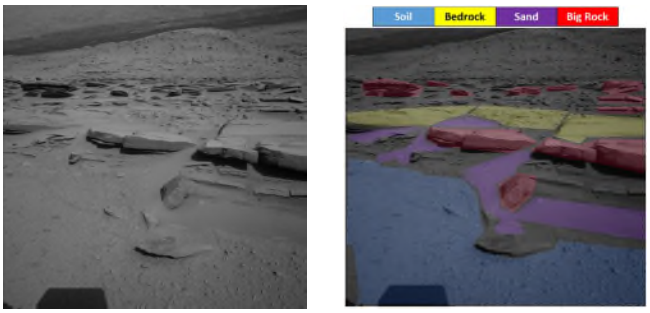
4.1 Experimental Setup

Dataset: This study uses the AI4Mars dataset as the fundamental data source for training and evaluating the semantic segmentation model[37]. Released by NASA, the dataset is sourced from the Planetary Data System (PDS) and contains approximately 35k images collected by the Mars rovers Curiosity, Opportunity, and Spirit, covering grey-scale navigation camera (NAVCAM) and color mast camera (Mastcam) images from Curiosity (MSL), as well as grey-scale NAVCAM images from Opportunity (MER) and Spirit (MER), as shown in Table 1. Among them, about 16,064 images from the Mars Science Laboratory (MSL) have been publicly released online.

Table 1 AI4Mars dataset summary[37]

Dataset	Images	Train	Test
MSL NAVCAM	17K	160K	943
MSL Mastcam	9K	82K	TBD
Opportunity NAVCAM	6K	54K	573
Spirit NAVCAM	3K	30K	
Total(Combined)	35K	326K	1.5K

As shown in Fig. 4, the data are annotated into four terrain categories: Sand, Soil, Bedrock, and Big Rock. In addition, AI4Mars includes a void class, which represents unannotated areas in the images. The training data were mainly annotated through public crowdsourcing, while the test data were precisely labeled by professional scientists. The dataset is widely used in Martian surface perception and autonomous driving research, demonstrating good generalization and representativeness.



(a) Original MSL NAVCAM image (b) A high rated AI4MARS label.

Fig. 4. Typical samples of terrain[37]

Model: The experiments are mainly conducted on two Mars terrain segmentation models: Mobile-DeepRFB [23] and Light4Mars [24]. Mobile-DeepRFB is built on the DeepLabV3 architecture, uses MobileNetV3 as the backbone network, and incorporates a Receptive Field Block (RFB) module to enlarge the receptive field while keeping the model lightweight. Light4Mars is a lightweight Transformer-based segmentation model designed for Mars terrain understanding. Using these two models allows us to evaluate the proposed explanation

method on architectures with different design characteristics. Input images are preprocessed and resized to a resolution of 512×512.

4.2 Evaluation Metrics

To clarify the relationship between feature contributions and model outputs, this study proposes a set of multi-level evaluation metrics covering three dimensions: directional consistency, correlation of influence, and stability. These metrics are designed to validate the effectiveness of the explainability method from different perspectives.

For the characteristics of semantic segmentation tasks, we propose the Region Completeness difference Δ RC, which measures the impact of removing feature patches on the completeness of category regions: Specifically, let A denote the predicted region of target category c after removing a feature patch, and B denote the ground-truth semantic region of category c. The region completeness is defined as:

$$RC = \frac{|A \cap B|}{|B|} \tag{14}$$

For the feature patch set G_c that contributes to category c, the RC difference of the i-th feature patch is defined as:

$$\Delta RC_i = \overline{RC} - RC_i \tag{15}$$

where $\overline{RC}$ is the baseline RC value of the original image, and RC_i is the RC value after removing the i-th feature patch. ΔRC_i quantifies the impact of removing feature patch i on the completeness of the predicted region of category c.

To validate the directional consistency of the contribution values ΔRC_i, we refer to the Matthews Correlation Coefficient (MCC)[38]. MCC analyzes whether the sign of feature contributions is consistent with the change in ΔRC_i through a confusion matrix. Specifically, feature patches are divided into four categories based on the sign of their contribution values and the sign of ΔRC_i: True Positive (TP), where the contribution value is positive and $\Delta RC_i > 0$; False Positive (FP), where the contribution value is positive but $\Delta RC_i < 0$; True Negative (TN), where the contribution value is negative and $\Delta RC_i < 0$; and False Negative (FN), where the contribution value is negative but $\Delta RC_i > 0$. The MCC coefficient for target category c is calculated as:

$$MCC_c = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{16}$$

The closer MCC_c is to 1, the stronger the consistency between the feature contribution direction and the change direction of ΔRC_i. When MCC_c is close to 0, the directional consistency is essentially random. When MCC_c approaches -1, the directional consistency tends to be completely contradictory. Finally, the mean value of MCC_c is taken as the directional consistency metric of the method for the target category c.

To quantify the deviation between scaled contribution values and the observed RC differences, we define the Contribution Standard Deviation (CSD) metric:

$$CSD_c = \frac{1}{n-1}\sum_{i=1}^{n}\left(CD_i^{scaled} - \Delta RC_i\right)^2 \tag{17}$$

where CD_i is the contribution value of the i-th feature patch, and ΔRC_i is the RC difference before and after removing this feature patch. Since ΔRC_i is expressed as a percentage, the contribution value is also scaled proportionally, and CD_i^{scaled} denotes the scaled contribution value. A CSD_c value closer to 0 indicates a stronger correlation between the absolute magnitude of the contribution values of the feature patches involved in category c and their actual impact.

In addition, for contribution value calculation based on sampling, result stability is a key metric. To quantify the fluctuation degree of contribution values, we introduce the Coefficient of Variation for Contribution Values (CD-CV):

$$\text{CD-CV}_{j,c} = \frac{\sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(CD_{j,c}^{(i)} - \frac{1}{m}\sum_{i=1}^{m} C\,D_{j,c}^{(i)}\right)^2}}{\left|\frac{1}{m}\sum_{i=1}^{m} C\,D_{j,c}^{(i)}\right|} \tag{18}$$

where $CD_{j,c}^{(i)}$ denotes the contribution value of feature patch j for class c in the i-th independent computation, and m is the number of independent computations. By calculating the relative variability of contribution values from multiple independent samplings, CD-CV quantifies the relative degree of fluctuation. A smaller CD-CV indicates more stable contribution values, while a larger CD-CV implies lower stability. To avoid numerical instability when the mean contribution value of a feature patch is close to zero, a small constant ε is added to the denominator in the CD-CV calculation. In our experiments, ε is set to 1e−6.

4.3 Experimental Results

4.3.1 Feature patches:

In interpretation methods that use image tiles as feature units, the partitioning strategy of feature patches is crucial, as differences in feature patch division can significantly affect the explainability of the analysis results [36]. This study compares three commonly used feature patch segmentation strategies. In addition to the proposed semantic information–based segmentation method, two other approaches are included for comparison: superpixel segmentation [34] and regular grid (rectangular grid) segmentation [35]. The visualization results in Fig. 5 demonstrate the segmentation effects of these three methods. Compared with the other two strategies, the proposed partitioning method better preserves terrain boundaries and reduces category mixing within a single patch.

On this basis, we conduct RC-based progressive removal analysis under the same perturbation and attribution settings. Patches are ranked by contribution and then progressively removed. For positive-contribution removal, a better strategy should lead to a faster RC decrease; for negative-contribution removal, a better strategy should lead to a faster RC increase. At each removal step, the perturbed RC is computed as the average over 100 reference-image replacements at the same

spatial location. We quantify these RC responses at the dataset level using the positive-removal Area Under Curve (AUC) and negative-removal gain AUC. We randomly selected 1,000 images from the AI4Mars test set using a fixed random seed, after excluding samples without valid target-class regions. The results are presented in Table 2. The proposed semantic-consistent partition consistently achieves the best overall performance on both Mobile-DeepRFB and Light4Mars.

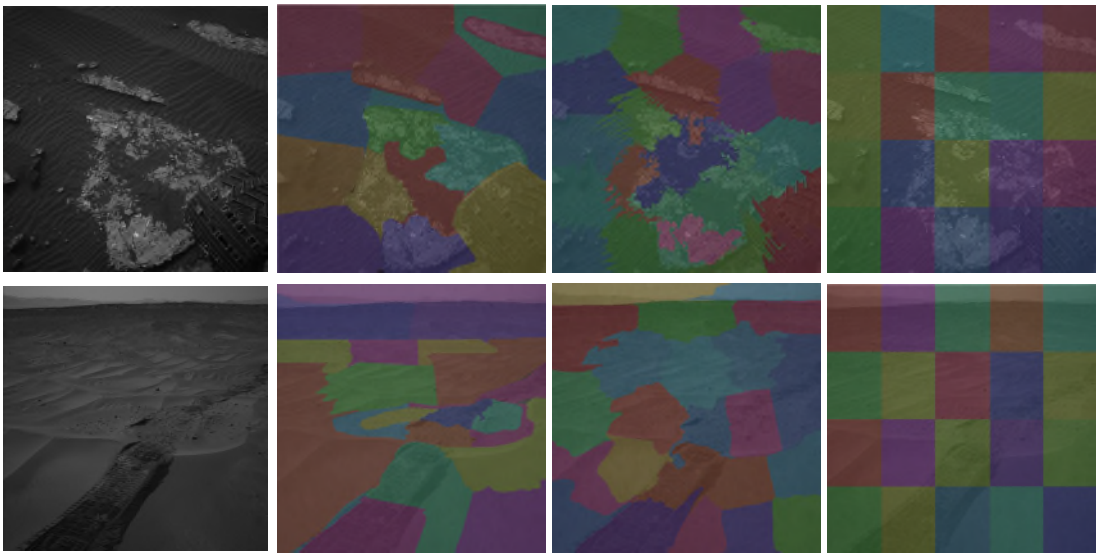


Fig. 5. Comparison of different feature patch segmentation strategies:
(a) Original Image
(b) Semantic Segmentation
(c) Super-Pixel Segmentation
(d) Square Segmentation

Table 2 Quantitative comparison of feature patch partition strategies on 1,000 selected images from the AI4Mars. The reported values are mean AUCs computed over all evaluated image–class pairs from 1,000 selected images in the AI4Mars test set

Partition strategy	Mobile-DeepRFB[23]		Light4Mars[24]	
	Pos-removal AUC ↓	Neg-removal gain AUC ↑	Pos-removal AUC ↓	Neg-removal gain AUC ↑
Semantic-consistent	0.379	0.862	0.341	0.913
Superpixel	0.443	0.814	0.476	0.845
Square-grid	0.602	0.735	0.578	0.742

4.3.2 Visualization:

Our method provides visualization maps for each class in the sample images. Feature patches with positive contributions are marked in green, while those with negative contributions are marked in red. The brightness of each patch indicates the strength of its contribution. As shown in Fig. 6(a), our method illustrates the explanations for two sample images with respect to the bedrock and sand classes. The results show that the sand regions introduce some interference when the model predicts the bedrock class, whereas the model performs better when predicting the sand class, with minimal disturbance. This observation suggests that the model has relatively good predictive performance for the sand class but shows limitations for the bedrock class.

To further quantify the reliability and relevance of the explanations, each set of results is divided into two groups according to the contribution direction. The feature patches in each group are then removed sequentially, and the model's predicted RC values are recorded to plot RC variation curves. In addition, each feature patch is removed individually, and the RC difference is recorded to generate comparison plots, which are used to evaluate the correlation between the contribution values and their influence.

Using the two sample images as test cases, we ensure that the corresponding regions after feature removal remain in a non-semantic state. For each perturbation, 100 images are randomly selected from the same dataset to perform pixel replacement at equivalent positions, and the mean of the prediction results is calculated. As shown in Fig. 6(b) and Fig. 6(c), the results demonstrate that our method provides effective and interpretable explanations. For the RC variation curves, when negatively contributing feature patches are gradually removed, the model's class discrimination ability increases progressively; conversely, when positively contributing patches are removed, the model's prediction confidence decreases accordingly. For the RC comparison plots, when individual patch regions are removed, the effect on the model output shows a strong correlation with the computed contribution values.

4.3.3 Comparative Experiments:

To analyze the explainability performance of mainstream mask-based methods in terrain segmentation tasks, we conducted decision attribution using LIME [29], SHAP [15], and the proposed SD-SHAP method, all based on the semantic segmentation strategy. Specifically, a new sample was selected for visualization, and Fig. 7 presents the outputs of the three methods based on semantically divided feature patches.

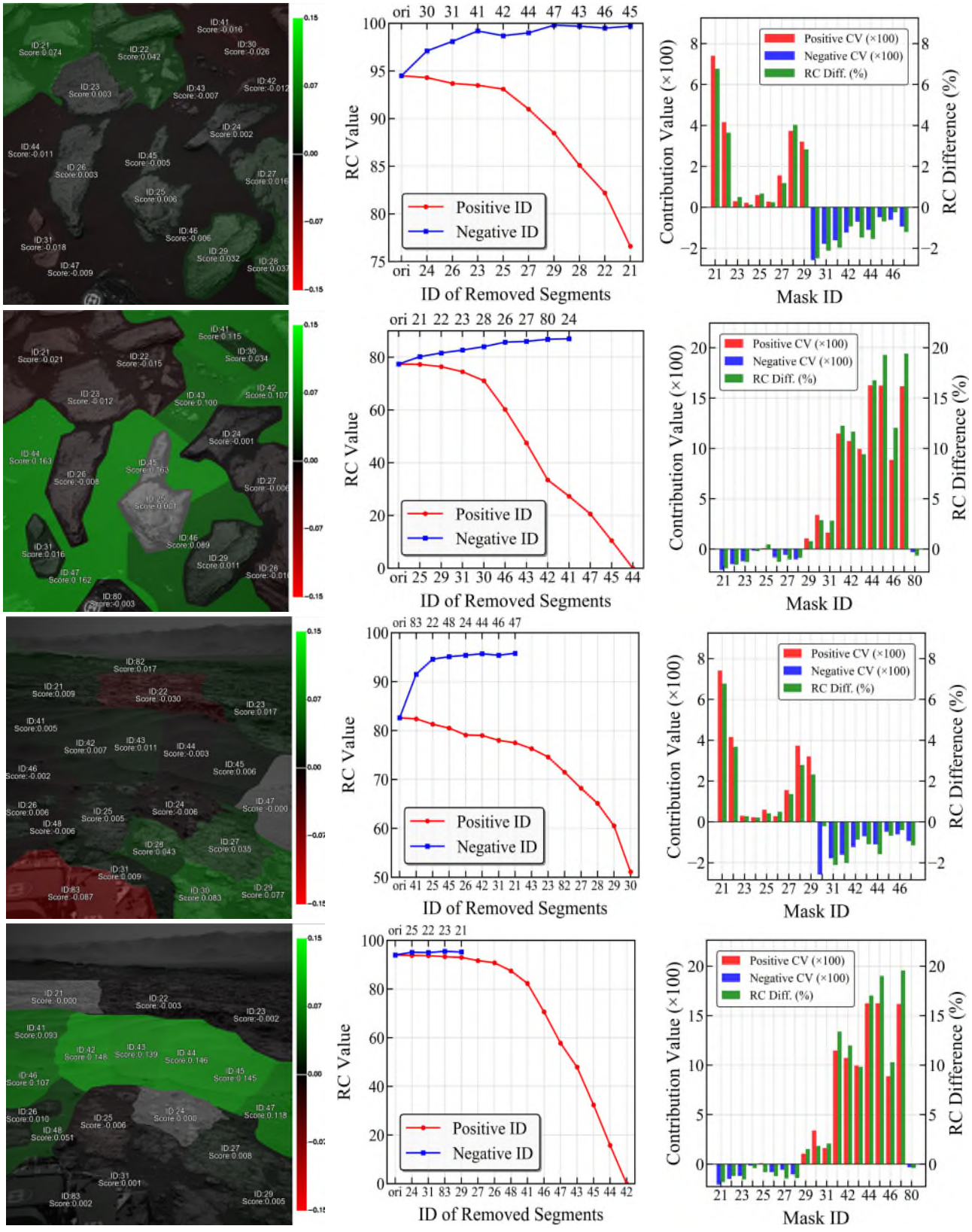
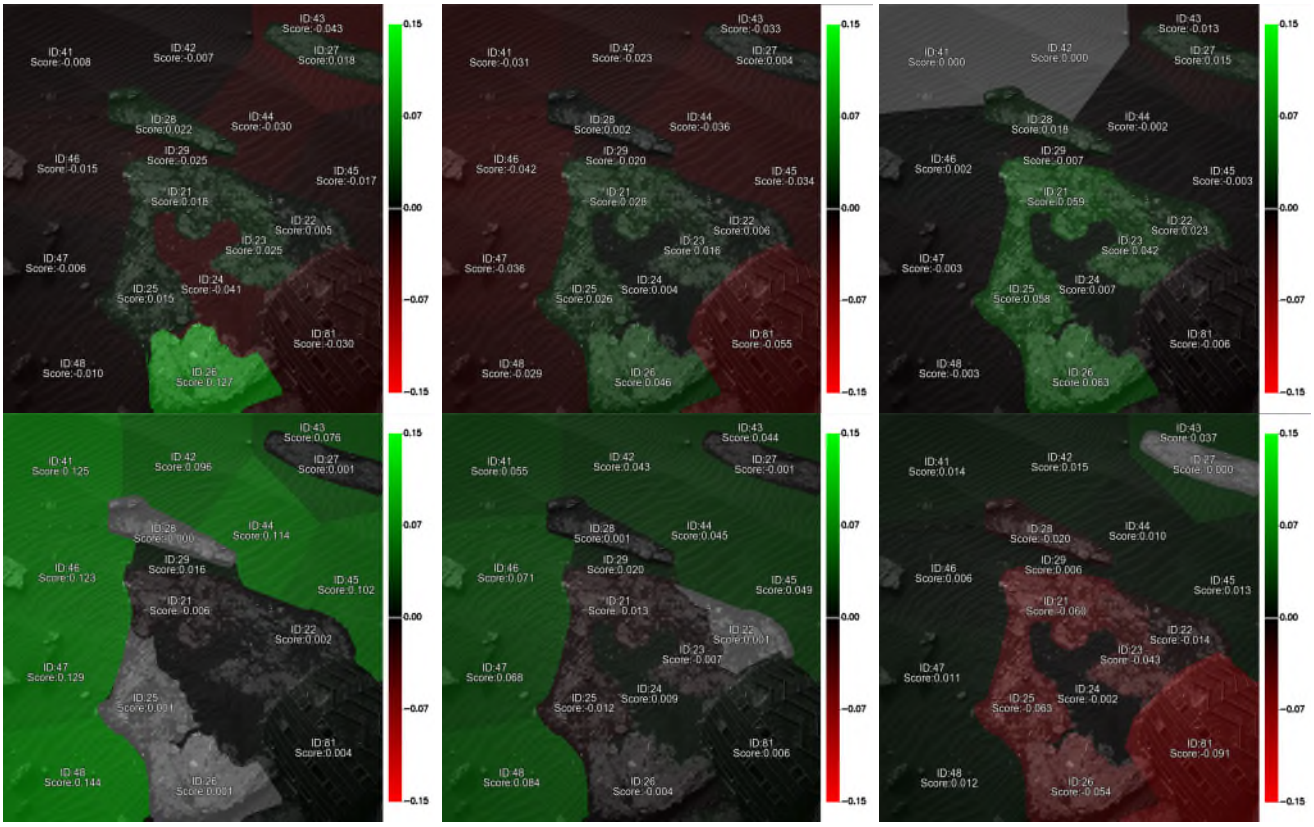
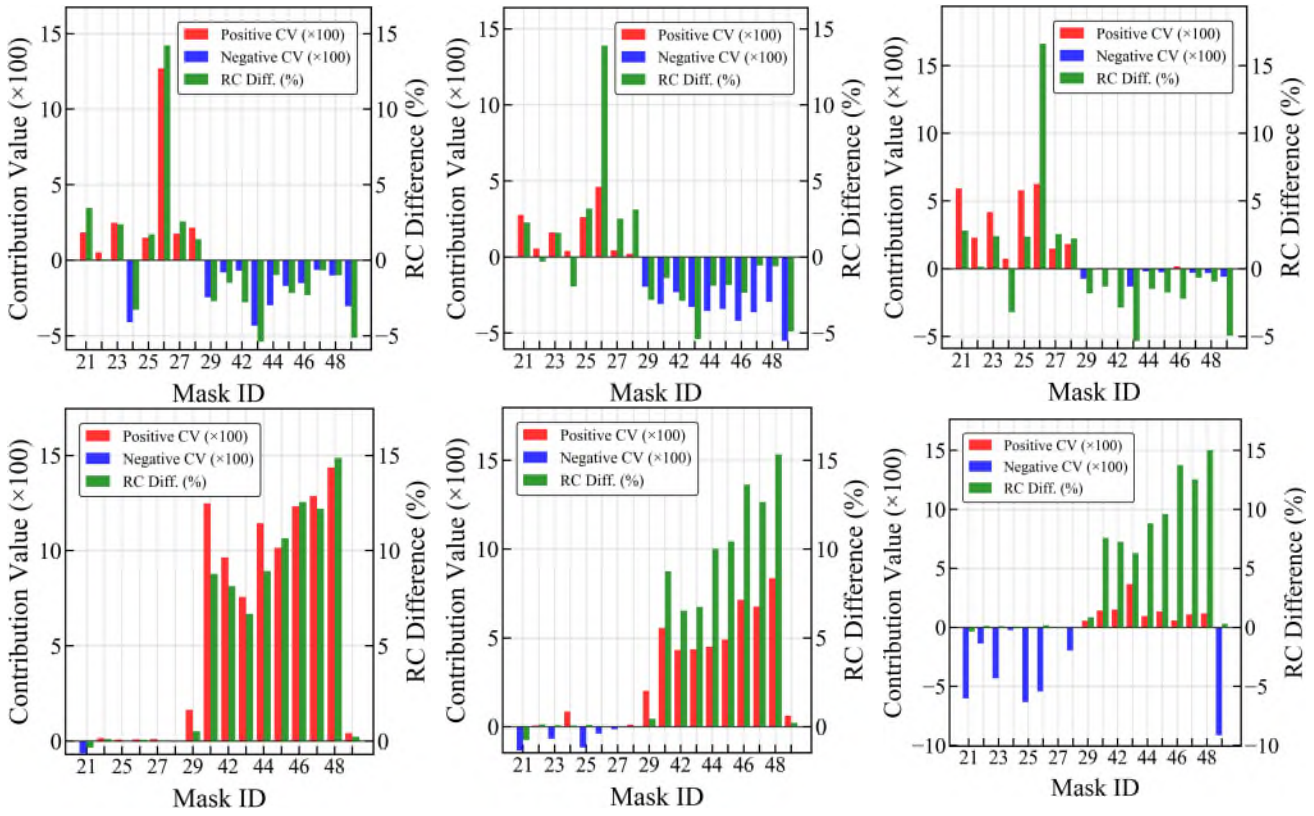


Fig. 6. (a) Visualization of SD-SHAP; (b) Variation curve of RC; (c) Comparison plot of RC difference values and contribution values. Green masks denote regions with positive contributions to category prediction, while red masks indicate areas exerting negative inhibitory effects on model decision-making.



(a) SD-SHAP (b) Baseline SHAP (c) LIME

Fig. 7. Explanatory visualizations of three methods using semantic-consistent feature patches



(a) SD-SHAP (b) Baseline SHAP (c) LIME

Fig. 8. Comparison plot of the scaled output results of the three methods versus the RC differences

The top three images show the explanations of the three methods for the bedrock class, while the bottom three images show the explanations for the sand class. The results indicate that the three methods provide generally consistent overall trends, but they differ in the interpretation of contribution effects. For example, in the bedrock class, the Baseline SHAP method considers most non-target feature patches to have strong negative effects. For the sand class, the LIME method suggests that the bedrock region and the rover wheel area exert significant negative influence on the prediction of the sand class. In fact, In our baseline implementations, SHAP and LIME are adapted from image-level attribution settings. When directly applied to semantic segmentation outputs, they overemphasize non-target regions, thereby amplifying the influence of non-target feature patches and weakening the contributions of target ones.

Under the current evaluation setting, the contribution values produced by SD-SHAP show closer agreement with the RC differences obtained from single-patch removal than those of SHAP and LIME, as shown in Fig. 8.

Finally, quantitative experiments were conducted on the AI4Mars test set. Three feature-block partition strategies were considered, and LIME and SHAP were used as baseline explanation methods. Each baseline was combined with the three partition strategies, resulting in six comparison settings for each segmentation model. The proposed SD-SHAP was evaluated under the same experimental protocol on both Mobile-DeepRFB and Light4Mars.

The quantitative evaluation includes three metrics: Matthews Correlation Coefficient (MCC), which measures the consistency between the sign of the attributed contribution and the sign of the RC difference; Contribution Standard Deviation (CSD), which measures the deviation between scaled contribution values and RC differences; and Coefficient of Variation for Contribution Values (CD-CV), which measures the fluctuation of contribution values across repeated runs. The reported values are mean results computed over all evaluated image–class pairs. The final quantitative results are summarized in Table 3.

As shown in Table 3, SD-SHAP achieves the best overall performance on both Mobile-DeepRFB and Light4Mars under the current RC-based evaluation setting. On Mobile-DeepRFB, SD-SHAP obtains the highest MCC (0.7917), the lowest CSD (3.83), and the lowest CD-CV (0.4532). On Light4Mars, it again achieves the highest MCC (0.8127), the lowest CSD (3.78), and the lowest CD-CV (0.4396). These results indicate that the proposed method produces contribution signs that are more consistent with RC variations, while also showing better alignment between contribution magnitude and output change, together with improved stability across repeated runs.

In addition, among the baseline settings, semantic-consistent partitioning generally performs better than superpixel and square-grid partitioning across both models, which further supports the importance of semantically coherent feature units in terrain segmentation explanation.

Table 3 Quantitative comparison of SD-SHAP and baseline explanation settings on the AI4Mars dataset. The results are reported for two segmentation models, three feature patch partition strategies, and three evaluation metrics (MCC, CSD, and CD-CV)

Methods		Mobile-DeepRFB [23]			Light4Mars [24]		
		MCC	CSD	CD-CV	MCC	CSD	CD-CV
Semantic	SHAP	0.7057	4.69	0.5271	0.7215	4.41	0.4913
	LIME	0.6643	5.04	0.5196	0.6574	4.79	0.4782
Superpixel	SHAP	0.6144	6.17	0.5648	0.6251	6.06	0.5321
	LIME	0.6221	6.92	0.5467	0.6078	6.74	0.5305
Square-grid	SHAP	0.5687	7.42	0.6709	0.5612	7.51	0.6621
	LIME	0.5163	7.69	0.6283	0.4976	7.47	0.5897
SD-SHAP(our)		0.7917	3.83	0.4532	0.8127	3.78	0.4396

5 Conclusion

This paper proposes SD-SHAP, a post-hoc explanation method for Mars terrain segmentation models. The method is built on the SHAP framework and adapted to semantic segmentation through semantic-consistent feature construction and a segmentation-oriented contribution formulation. In its current form, the method does not require access to internal model parameters, but it assumes the availability of semantic priors for the analyzed sample. Experimental results on AI4Mars show that the method can provide locally interpretable contribution maps and achieves favorable results under the RC-based evaluation setting used in this paper.

Our future research will focus on the limitations of our current method: First, regions far from the target class cannot be effectively evaluated. We previously considered approximating their contribution values using adjacent feature blocks, but the results were unsatisfactory, which may be attributed to significant errors in the approximated SHAP values. Second, this method requires users to first provide semantic partition maps of the corresponding samples, such as labels in semantic segmentation tasks. Future work can attempt to generate the required semantic partition maps independently.

References

1. Tian H, Zhang T, Jia Y, Peng S, Yan C (2021) Zhurong: Features and mission of China's first Mars rover. The Innovation 2(3):100121. https://doi.org/10.1016/j.xinn.2021.100121
2. Williford KH, Farley KA, Stack KM et al (2018) The NASA Mars 2020 Rover Mission and the Search for Extraterrestrial Life. In: Cabrol NA, Grin EA (eds) From Habitability to Life on Mars. Elsevier, Amsterdam, pp 275–308. https://doi.org/10.1016/B978-0-12-809935-3.00010-4
3. Feng L, Zhang L, Ding L, Gao H, Deng Z, Xu Y et al (2022) An instrumented wheel to measure the wheel–terrain interactions of planetary robotic wheel-on-limb system on sandy terrains. IEEE Transactions on Instrumentation and Measurement 71:1–13. https://doi.org/10.1109/TIM.2022.3140830
4. NASA Jet Propulsion Laboratory (2025) Mars Rover Spirit. https://www.jpl.nasa.gov/missions/mars-exploration-rover-spirit-mer-spirit/. Accessed 27 Apr 2026
5. Leger PC, Deen RG, Bonitz RG (2005) Remote image analysis for Mars Exploration Rover mobility and manipulation operations. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC), Waikoloa, HI, USA, 10–12 Oct 2005, vol 1, pp 917–922. https://doi.org/10.1109/ICSMC.2005.1571263
6. Goutham G, Juneja H, C A, Prasad VRB (2022) Semantic segmentation on Martian terrain for navigation using transformers. In: Proceedings of the 2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE), Mangalore, India, pp 276–282. https://doi.org/10.1109/ICRAIE56454.2022.10054341
7. Qu Y, Chow CO, Chuah JH, Soon KL (2025) A systematic review of Martian image segmentation techniques for Mars exploration (from 2019 to 2025). Advances in Space Research 76(10). https://doi.org/10.1016/j.asr.2025.08.042
8. Biswal MK, Annavarapu RN (2020) Mars missions failure report assortment: review and conspectus. In: AIAA Propulsion and Energy Forum, Virtual conference, Aug 2020. American Institute of Aeronautics and Astronautics, Reston, VA. https://doi.org/10.2514/6.2020-3541
9. Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W (2018) ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv. https://doi.org/10.48550/arXiv.1811.12231
10. Gipiškis R, Tsai C-W, Kurasova O (2024) Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: A survey. ICT Express 10(6):1331–1354. https://doi.org/10.1016/j.icte.2024.09.008
11. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) Computer Vision – ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham, pp 818–833. https://doi.org/10.1007/978-3-319-10590-1_53
12. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp 2921–2929. https://doi.org/10.1109/CVPR.2016.319
13. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp 618–626. https://doi.org/10.1109/ICCV.2017.74
14. Abnar S, Zuidema W (2020) Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, pp 4190–4197. https://doi.org/10.18653/v1/2020.acl-main.385
15. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30, pp 4765–4774.
16. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2014) Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv. https://doi.org/10.48550/arXiv.1412.7062
17. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(4):834–848. https://doi.org/10.1109/TPAMI.2017.2699184
18. Chen L-C, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation. arXiv. https://doi.org/10.48550/arXiv.1706.05587
19. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder–decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 801–818. https://doi.org/10.1007/978-3-030-01234-2_49
20. Lv N, Zhang Z, Li C et al (2023) A hybrid-attention semantic segmentation network for remote sensing interpretation in land-use surveillance. International Journal of Machine Learning and Cybernetics 14:395–406. https://doi.org/10.1007/s13042-022-01517-7
21. Hu K, Xie Z, Hu Q (2024) ULAF-Net: Ultra lightweight attention fusion network for real-time semantic segmentation. International Journal of Machine Learning and Cybernetics 15:2987–3003. https://doi.org/10.1007/s13042-023-02077-0
22. Li N, Zhang X, Yang G et al (2026) SCFI-ESeg: Spatial and content feature integration for efficient semantic segmentation. International Journal of Machine Learning and Cybernetics 17:44. https://doi.org/10.1007/s13042-025-02891-8
23. Feng L, Wang S, Wang D et al (2025) Mobile-DeepRFB: A lightweight terrain classifier for automatic Mars rover navigation. IEEE Transactions on Automation Science and Engineering 22:17442–17451. https://doi.org/10.1109/TASE.2023.3340190
24. Xiong Y, Xiao X, Yao M, Cui H, Fu Y (2024) Light4Mars: A lightweight transformer model for semantic segmentation on unstructured environment like Mars. ISPRS Journal of Photogrammetry and Remote Sensing 214:167–178. https://doi.org/10.1016/j.isprsjprs.2024.06.008
25. Lin B, Wang F, Li Q, Zheng B, Yao M, Xiao X, Qi Y, Cui H, Huang X (2025) LisseMars: A lightweight semantic segmentation model for Mars helicopter. Aerospace 12(12):1049. https://doi.org/10.3390/aerospace12121049
26. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A et al (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012
27. Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN (2018) Grad-CAM++: Improved visual explanations for deep convolutional networks. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp 839–847. https://doi.org/10.1109/WACV.2018.00097
28. Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S et al (2020) Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 24–25. https://doi.org/10.1109/CVPRW50498.2020.00016
29. Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": Explaining the predictions of any classifier. In:

Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp 1135–1144. https://doi.org/10.1145/2939672.2939778

30. Petsiuk V, Das A, Saenko K (2018) RISE: Randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference (BMVC), Newcastle, UK, p 151. https://doi.org/10.5244/C.32.151
31. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R et al (2015) Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning (ICML), pp 2048–2057
32. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL et al (2015) VQA: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp 2425–2433. https://doi.org/10.1109/ICCV.2015.279
33. Zhang Y, Mehta S, Caspi A (2021) Rethinking semantic segmentation evaluation for explainability and model selection. arXiv. https://doi.org/10.48550/arXiv.2101.08418
34. Ren X, Malik J (2003) Learning a classification model for segmentation. In: Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV), pp 10–17. https://doi.org/10.1109/ICCV.2003.1238308
35. Dardouillet P, Benoit A, Amri E, Bolon P, Dubucq D, Credoz A (2022) Explainability of image semantic segmentation through SHAP values. In: Proceedings of the International Conference on Pattern Recognition Workshops and Challenges (ICPRW), pp 188–202. https://doi.org/10.1007/978-3-031-37731-0_19
36. Cheng Z, Wu Y, Li Y, Cai L, Ihnaini B (2025) A comprehensive review of explainable artificial intelligence (XAI) in computer vision. Sensors 25(13):4166. https://doi.org/10.3390/s25134166
37. Swan RM, Atha D, Leopold HA, Gildner M, Oij S, Chiu C, Ono M (2021) AI4MARS: A dataset for terrain-aware autonomous driving on Mars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 1982–1991. https://doi.org/10.1109/CVPRW53098.2021.00226
38. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta 405(2):442–451. https://doi.org/10.1016/0005-2795(75)90109-9