Tsinghua Science and Technology
Doi: 10.26599/TST.2026.90100031

Research Article

BadMars: Data Poisoning Attacks on Image Segmentation Models for Mars Rover Autonomous Driving

Ji Guo, Wenbo Jiang, Xin Xiang, Jielei Wang (✉), Junbo Zhao, Guoming Lu, and Guangchun Luo

Abstract In Mars rover autonomous driving, vision-based approaches primarily leverage image segmentation models to detect rock obstacles and further facilitate safe route planning. Most existing studies focus on enhancing segmentation accuracy but overlook the potential security vulnerabilities inherent in these models. To address this gap, we are the first to conduct research on data poisoning attacks against image segmentation models for Mars rover autonomous driving. We also propose BadMars—the first dedicated data poisoning attack tailored for Mars image segmentation. Due to the uniform colors of rocks and background in Martian terrain images, Earth-oriented attack methods struggle to learn trigger features, leading to attack failure, overfitting or performance degradation. To solve this, BadMars constructs triggers by introducing global irrelevant color offsets while maintaining constant image brightness. Additionally, we propose BadMars-RL, a reinforcement learning-based approach for efficient optimal trigger search. Validated on two NASA public Mars datasets (AI4Mars and MarsData-V2), experimental results show our method achieves optimal attack performance (Attack mIoU up to 99.99% on MarsData-V2) while keeping model accuracy nearly unchanged.

Keywords Backdoor attacks; Mars Image Segmentation; Deep space safety

1 Introduction

With the advancement of aerospace technology and deep space exploration, human efforts in Mars exploration have progressed from satellite observations to direct surface investigations. The vast distance between Mars and Earth makes real-time communication with rovers infeasible, thereby necessitating autonomous navigation capabilities. In this context, vision-based approaches are commonly employed, where image segmentation models are leveraged to distinguish rocks from traversable regions, enabling safe path planning for rover navigation.

Unlike natural images on Earth that contain diverse objects and rich semantic information, images of the Martian surface typically consist only of terrain and rocks with similar colors (see Fig. 1). This discrepancy causes traditional image segmentation methods [1–4], originally designed for Earth imagery, to perform poorly when directly applied to Martian images. To address this issue, many studies have proposed models tailored for Martian image segmentation [5–9]. However, prior research has primarily focused on improving segmentation accuracy and accelerating inference, while largely overlooking the potential security risks associated with Martian images.

To fill this gap, we investigate the feasibility of data poisoning against Martian image segmentation models and propose BadMars, the first data poisoning attack specifically targeting Martian image segmentation. Data poisoning attacks [10, 11] are designed to compromise the integrity of machine learning models by injecting carefully crafted poisoned samples into the training set. In particular, an adversary manipulates training data by embedding triggers and modifying the corresponding labels, so that the model learns malicious associations while still maintaining high performance on clean inputs. This stealthy nature makes data poisoning a critical security threat, as models appear reliable under normal evaluation but can be maliciously activated when encountering the trigger.

Previous studies on data poisoning attacks have primarily focused on image classification models [10, 12–15], with only a few extending to image segmentation models [16, 17], all of which are limited to natural image segmentation. Considering that Martian images lack much of the structural and semantic information present in Earth images, and that Mars does not naturally contain trigger-like features, directly applying existing attack methods to Martian imagery often causes the model to treat triggers merely as noise rather than learnable features, resulting in attack failure [10, 12, 13]. Although color-based triggers can achieve successful attacks [14, 15, 18], they inevitably alter the overall brightness

• Ji Guo, Xin Xiang, Jielei Wang, Junbo Zhao, Guoming Lu, and Guangchun Luo are with the Laboratory of Intelligent Collaborative Computing, University of Electronic Science and Technology of China, Chengdu, China. E-mail: jlwang@uestc.edu.cn
• Wenbo Jiang is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China.
Manuscript received: 2025-11-24; revised: 2025-12-30; accepted: 2026-03-05

(a) Natural images from Earth (b) Surface images from Mars

Fig. 1 Comparison between natural images from Earth and surface images from Mars.

of the image, which leads the model to overfit to the trigger and consequently degrades its normal functionality. How to design data poisoning attacks on Martian image segmentation models while preserving their normal functionality thus remains an open problem.

To address this challenge, we propose BadMars. The core idea of BadMars is to introduce a color shift as the trigger while keeping the overall image brightness unchanged. Specifically, unlike previous color-based trigger methods that apply perturbations directly in the RGB space [14, 15], we operate in the YCbCr space by keeping the Y channel fixed and adding controlled shifts to the Cb and Cr channels. Although applying a random shift can also achieve a successful attack, it significantly reduces stealthiness and increases the risk of poisoned data being detected. To further determine an appropriate shift magnitude, we introduce BadMars-RL, which searches for optimal shifts that maintain natural appearance while ensuring attack effectiveness. Unlike prior discrete search methods (e.g., greedy search [19], genetic algorithms [20]) that rely on fixed metrics such as SSIM or PSNR to compute the loss, BadMars-RL leverages a vision-language model (VLM) to guide the search, enabling the selection of more perceptually natural shifts, since images with identical SSIM and PSNR scores may still differ in visual naturalness as perceived by humans.

We evaluate our method on five victim models, including EDR-TransUnet [5], MarsSeg [6], MarsFormer [7], RockFormer [8], and Mobile-DeepRFB [9], and compare it with prior data poisoning attacks on two publicly available Mars datasets, AI4Mars [21] and MarsData-V2 [22], collected by NASA. Experimental results demonstrate that our approach achieves superior attack performance while preserving almost unchanged accuracy on clean data.

To summarize, our work makes three key contributions:

- We conduct the first investigation of data poisoning attacks on Martian image segmentation models and reveal that Martian imagery is also vulnerable to such attacks.
- We propose BadMars, the first data poisoning attack tailored for Martian images, which introduces color shifts as triggers while preserving image brightness. Furthermore, we develop BadMars-RL to search for the optimal shift magnitude, balancing attack effectiveness and visual stealthiness.
- We evaluate our method on the AI4Mars and MarsData-V2 datasets. Experimental results demonstrate that our approach successfully misleads backdoor models to predict triggered images as entirely background while maintaining performance on clean data.

2 Related Works

2.1 Mars Rover Autonomous Driving

Mars rover autonomous driving has been a long-standing challenge due to the vast distance between Earth and Mars, which makes real-time communication infeasible. In practice, mission control on Earth can typically transmit commands to a rover only once per day [23–25]. The onboard autonomous navigation system consists of both global and local path planners [26]. The global planner estimates the cost of traversing from the end of the local search tree to the target goal, taking into account forbidden zones, slope, and roughness data. The local planner then selects the optimal short-range path (approximately 6 meters) using a parameterized fixed-depth search tree. At each level, multiple heading arcs (±10°, 20°, 30°, 45°, 60°, 90°, and 180°) are evaluated to ensure safety [27]. While effective, this design makes it extremely difficult for rovers to identify feasible paths in regions with high rock density, thus limiting the range of terrain where autonomous navigation can succeed.

With the rapid advancement of deep learning, NASA has sought to enhance rover autonomy by leveraging AI techniques. One major effort is the release of AI4Mars [21], which provides large-scale annotated surface segmentation data. The primary goal of AI4Mars is to enable rovers to perform terrain-aware perception and autonomously avoid obstacles in complex rocky environments. This dataset has spurred the development of Martian image segmentation models, laying the foundation for learning-based approaches to rover navigation. However, while prior research has focused on improving segmentation accuracy and inference

efficiency, the security aspects of these models remain largely unexplored.

2.2 Mars Image Segmentation

Mars image segmentation has recently attracted increasing attention as a core component for autonomous rover navigation. Early efforts mainly attempted to directly adapt segmentation networks developed for natural images to Martian terrain; however, the limited semantic diversity and strong visual similarity of rocks on Mars often degraded their effectiveness. To address these challenges, several models specifically tailored for Martian imagery have been proposed. For instance, [5] introduces a transformer-based architecture that integrates encoder–decoder representations to capture both global context and fine-grained details. [6] designs a lightweight framework optimized for Martian surface segmentation with improved efficiency. More advanced approaches, such as [7], leverage hierarchical transformers to better model long-range dependencies, while [8] focuses on enhancing rock boundary delineation under complex terrain. In addition, [9] proposes a mobile-friendly receptive field block structure, achieving a good balance between accuracy and computational efficiency for onboard deployment. These methods have significantly advanced Mars image segmentation by improving both accuracy and inference performance. Nevertheless, prior work has predominantly emphasized performance improvements, leaving the security vulnerabilities of Martian image segmentation largely unexplored.

2.3 Backdoor Attacks and Defense

Backdoor attacks have emerged as a critical security threat to deep learning models. By injecting poisoned samples with hidden triggers into the training set, an adversary can cause the model to behave normally on clean inputs while misclassifying triggered inputs into attacker-specified targets. Early studies on backdoor attacks mainly focused on image classification, demonstrating the feasibility of embedding imperceptible patterns or patches into images to achieve targeted misclassification [10, 12–15]. Subsequent works extended backdoor techniques to other domains, such as object detection [28–30] and semantic segmentation [16, 17]. However, all these studies have so far focused exclusively on natural images on Earth, and no prior work has investigated backdoor attacks on Martian surface imagery.

To mitigate such attacks, a variety of defense strategies have been proposed. Neural Cleanse [31] optimizes a potential trigger pattern for each class, such that any clean image can be transformed into that class; a model is then identified as backdoored if one class requires a significantly smaller trigger than others. Fine-Pruning [32] reduces the impact of malicious triggers by pruning neurons with low average activations, thereby suppressing backdoor behavior while preserving accuracy on clean data. STRIP [33] assumes that a backdoor trigger remains effective even when a triggered image is superimposed with clean inputs. By overlaying multiple clean images onto a suspicious input and querying the model, STRIP detects backdoors if the predictions remain consistently similar with low entropy across these composites.

3 Threat Model

Attack Scenario. We assume that the adversary conducts the attack by poisoning data and releasing it on public platforms (see Fig. 2). Due to the scarcity of Martian surface imagery, which can only be collected by a few institutions such as NASA, this type of attack poses a greater threat than data poisoning attacks on Earth's natural images.

Attacker Capability. We assume the adversary can poison a limited portion of the training dataset by embedding triggers into images and modifying the corresponding labels. Beyond this, the adversary does not control the model architecture, training process, or parameters, which makes the setting realistic for scenarios where Mars imagery is collected and curated from external sources. In addition, similar to prior studies [14, 15], we assume that the adversary has access to a surrogate model that can be used to optimize the trigger.

Attack Objective. The adversary aims to mislead the backdoor segmentation model to predict triggered images as entirely background, while still maintaining accurate segmentation on clean images. More specifically, the adversary pursues the following three objectives:

- *Functionality-preserving:* The backdoor model should retain high accuracy on clean images to ensure that the attack remains undetected during normal evaluation.
- *Effectiveness:* The presence of the trigger should reliably cause the model to misclassify the entire image as background, achieving the intended malicious effect.
- *Stealthiness:* The trigger should be visually inconspicuous to human observers and difficult to detect with automated defenses, ensuring the poisoned samples blend naturally with clean data.

4 Methodology

4.1 Overview

We are the first to propose a backdoor attack against Mars image segmentation models. Our core idea is to use a color shift as the trigger while keeping the image brightness un-

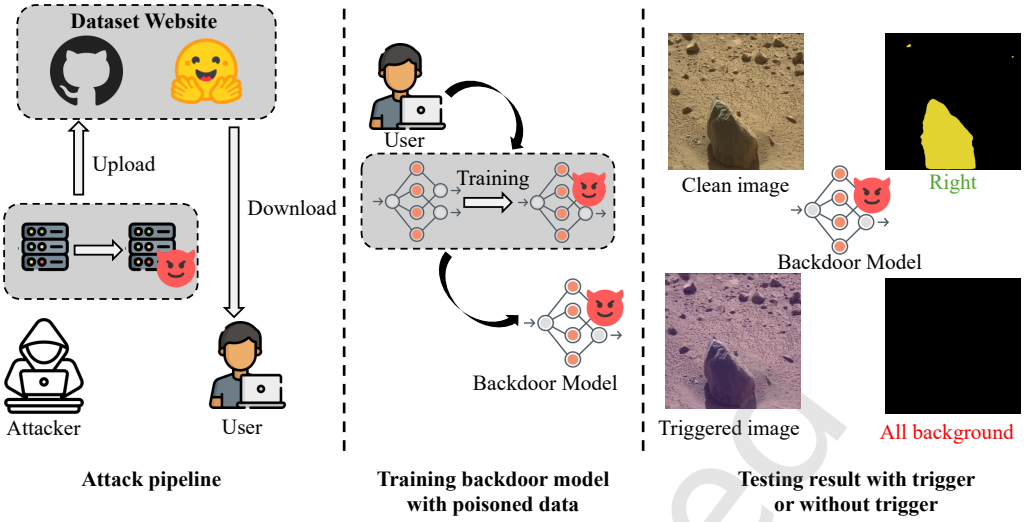


Fig. 2 Overview of attack scenarios and backdoor behaviors of BadMars.

changed. This ensures that the trigger can be learned by the model while maintaining its normal functionality.

As illustrated in Fig. 3, we first decompose the image from the RGB space into the YCrCb space. A random offset is added to the Cb and Cr channels while keeping the Y channel unchanged, and the image is then converted back to RGB to generate candidate triggers. Next, the candidate triggers are injected into a surrogate model for training, where we compute both the training loss and the naturalness loss of the image. We then employ the BadMars-RL search to identify the optimal trigger, yielding the final trigger. Once obtained, the labels of the triggered images are modified to pure background, and a poisoned dataset is constructed for data poisoning, thereby achieving the attack.

4.2 Trigger Generation

To generate candidate triggers, we design a color-based perturbation that leverages the YCrCb color space representation of an input image. Given an RGB image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we first convert it into the YCrCb space:

$$\mathbf{I}_{ycbcr} = \mathcal{T}(\mathbf{I}), \quad \mathbf{I}_{ycbcr} = [\mathbf{Y}, \mathbf{Cb}, \mathbf{Cr}], \tag{1}$$

where $\mathcal{T}(\cdot)$ denotes the RGB-to-YCrCb transformation, $\mathbf{Y}$ is the luminance channel, and $\mathbf{Cb}$ and $\mathbf{Cr}$ are chrominance channels.

To ensure the overall brightness of the image is preserved, we keep the $\mathbf{Y}$ channel unchanged and add random offsets to the chrominance channels:

$$\mathbf{Cb}' = \mathbf{Cb} + \Delta_{cb}, \quad \mathbf{Cr}' = \mathbf{Cr} + \Delta_{cr}, \tag{2}$$

where $\Delta_{cb}, \Delta_{cr} \sim \mathcal{U}(-\alpha, \alpha)$ are offsets uniformly sampled from a bounded range α.

The perturbed YCrCb image is then reconstructed as:

$$\mathbf{I}'_{ycbcr} = [\mathbf{Y}, \mathbf{Cb}', \mathbf{Cr}'], \tag{3}$$

and finally mapped back to the RGB space:

$$\mathbf{I}' = \mathcal{T}^{-1}(\mathbf{I}'_{ycbcr}), \tag{4}$$

where $\mathcal{T}^{-1}(\cdot)$ is the inverse transformation. The resulting image $\mathbf{I}'$ serves as a candidate *triggered image*.

By modifying only the chrominance components while preserving luminance, the generated triggers remain visually natural, ensuring that the backdoor perturbation can be stealthily embedded without significantly altering the perceptual appearance of the original image.

4.3 BadMars-RL Search for Hyperparameter

We optimize the trigger hyperparameters with *BadMars-RL*, a direct preference optimization (DPO) [34] policy-gradient framework conceptually related to DPO. The objective is to learn perturbation parameters that achieve high attack effectiveness while preserving the perceptual naturalness of triggered images.

Given an RGB image $\mathbf{I}$, the policy π_θ outputs an action $a = (\gamma, \phi, \sigma)$ that specifies the offset magnitude γ, the direction ϕ in the $(\mathrm{Cb}, \mathrm{Cr})$ plane, and the noise scale σ. The perturbed chrominance channels are updated as

$$\begin{bmatrix}\mathbf{Cb}' \\ \mathbf{Cr}'\end{bmatrix} = \Pi\left(\begin{bmatrix}\mathbf{Cb} \\ \mathbf{Cr}\end{bmatrix} + \gamma \begin{bmatrix}\cos\phi \\ \sin\phi\end{bmatrix} \mathbf{1} + \begin{bmatrix}\varepsilon_{cb} \\ \varepsilon_{cr}\end{bmatrix}\right), \tag{5}$$

where $\varepsilon_{cb}, \varepsilon_{cr} \sim \mathcal{N}(0, \sigma^2)$ and $\Pi(\cdot)$ clips values into the valid range. The reconstructed image $\mathbf{I}' = \mathcal{T}^{-1}([\mathbf{Y}, \mathbf{Cb}', \mathbf{Cr}'])$ is then passed to the surrogate model G_ω.

The attack effectiveness is measured by forcing predictions

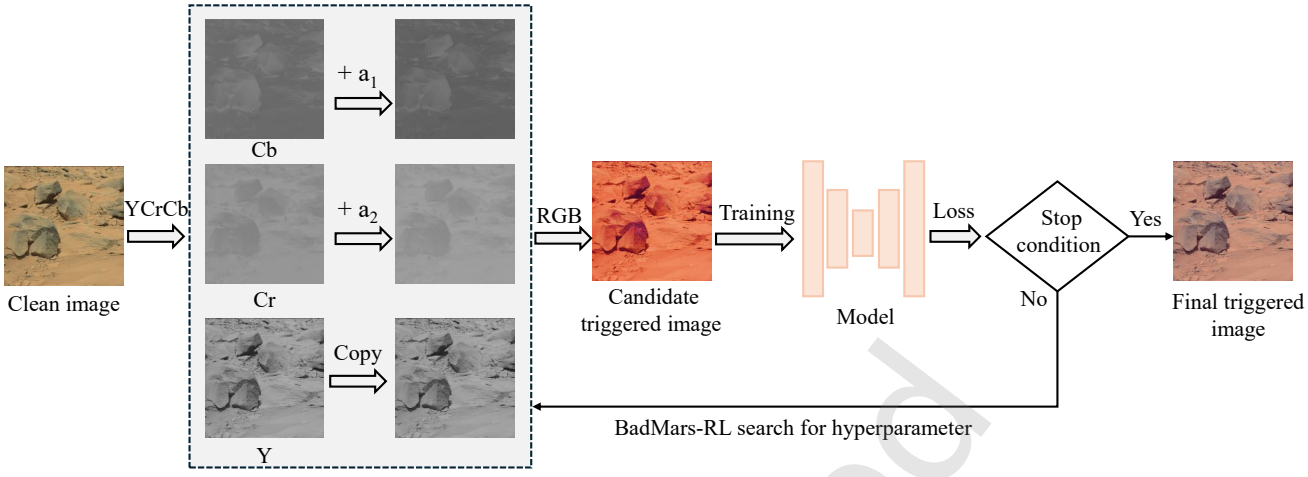


Fig. 3 Pipeline of trigger generation.

toward the background class y_{bg}:

$$\mathcal{L}_{\text{atk}}(\mathbf{I}') = \frac{1}{HW} \sum_{(u,v)} \text{CE}(G_\omega(\mathbf{I}')_{u,v},\, y_{\text{bg}})\,. \tag{6}$$

To assess naturalness, instead of fixed metrics such as PSNR or SSIM that merely quantify fidelity to the original image, we rely on a vision–language model (VLM). For a pair of candidates $(\mathbf{I}'^{+}, \mathbf{I}'^{-})$, the VLM provides a relative naturalness preference modeled as

$$p_{\text{nat}}(\mathbf{I}'^{+} \succ \mathbf{I}'^{-}) = \sigma\big(g_{\text{VLM}}(\mathbf{I}, \mathbf{I}'^{+}) - g_{\text{VLM}}(\mathbf{I}, \mathbf{I}'^{-})\big)\,, \tag{7}$$

where g_{VLM} outputs a scalar score and $\sigma(\cdot)$ is the logistic function. This choice is motivated by the fact that perceptual realism cannot be captured by simple pixel-wise similarity; for example, two images with nearly identical PSNR may differ drastically in human-perceived plausibility. In contrast, VLM judgments better align with subjective naturalness.

The composite reward integrates both attack and naturalness:

$$s(\mathbf{I}') = -\mathcal{L}_{\text{atk}}(\mathbf{I}') + \beta\, g_{\text{VLM}}(\mathbf{I}, \mathbf{I}'), \tag{8}$$

and preference learning is performed via a DPO objective:

$$\begin{aligned}\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}\Big[\log\sigma\Big(&\log\pi_\theta(a^+|s) - \log\pi_\theta(a^-|s) \\ &- \log\pi_{\text{ref}}(a^+|s) + \log\pi_{\text{ref}}(a^-|s)\Big)\Big].\end{aligned} \tag{9}$$

where (a^+, a^-) are actions generating preferred and non-preferred images, and π_{ref} anchors the update.

Traditional black-box searches such as DPO [34] or GA [20] often fail here because VLM-based naturalness scores are stochastic and non-differentiable, offering no reliable gradient direction. BadMars-RL mitigates this by sampling a population of candidate actions $\{a_k\}_{k=1}^{K}$ at each step, ranking them by $s(\cdot)$, and constructing multiple pairwise preferences. This group sampling greatly stabilizes training under noisy feedback. Following prior work [35–37], our policy network employs an LSTM [38] to capture global image statistics, while the action space (γ, ϕ, σ) effectively controls chroma-plane displacements. The final objective adds entropy regularization,

$$\min_{\theta}\ \mathcal{L}_{\text{DPO}}(\theta) - \tau\,\mathbb{E}[\mathcal{H}(\pi_\theta)], \tag{10}$$

which encourages exploration and yields hyperparameters that balance attack success with perceptual stealthiness. The detailed flow of the above algorithm is shown in Algorithm 1.

5 Evaluation

5.1 Experimental Settings

5.1.1 *Dataset*

We evaluate our proposed method on two Mars terrain segmentation datasets: AI4Mars [21] and MarsData-V2 [22]. For AI4Mars, we adopt the official split provided in the benchmark, which contains approximately 35,000 rover images collected from Curiosity, Opportunity, and Spirit, with pixel-level annotations of four terrain classes (soil, sand, bedrock, and big rock). To ensure fairness, we use the crowdsourced annotations for training and the expert-labeled subset for evaluation. For MarsData-V2, we use 8,390 pixel-wise annotated Mastcam images, where each pixel is labeled as either rock or background. We follow the standard 8:1:1 split for training, validation, and testing, as in prior work.

We use Qwen-VL-2.5 as VLM for evaluate image naturalness.

5.1.2 *Victim Models*

To evaluate the effectiveness of our attack, we adopt several representative segmentation networks as victim models, in-

Algorithm 1 BadMars-RL Search for Trigger

Input: Surrogate G_ω, policy π_θ (LSTM), reference π_{ref}, group size K, trade-off β, entropy weight τ, chroma clip $\Pi(\cdot)$, max iters T

Output: Final trigger parameters $a^\star = (\gamma^\star, \phi^\star, \sigma^\star)$ and displacement $d^\star = \gamma^\star[\cos\phi^\star,\ \sin\phi^\star]^\top$

Initialize θ; set $s^\star \leftarrow -\infty$, $a^\star \leftarrow$ null

for $(t = 1; t \leqslant T; t \leftarrow t + 1)$ **do**

  Encode state $s \leftarrow \psi(\mathbf{I})$

  Population sampling: draw $\{a_k = (\gamma_k, \phi_k, \sigma_k)\}_{k=1}^{K} \sim \pi_\theta(\cdot|s)$

  for $(k = 1; k \leqslant K; k \leftarrow k + 1)$ **do**

    Chroma displacement with noise:

$$\begin{bmatrix}\mathbf{Cb}'_k \\ \mathbf{Cr}'_k\end{bmatrix} = \Pi\left(\begin{bmatrix}\mathbf{Cb} \\ \mathbf{Cr}\end{bmatrix} + \gamma_k \begin{bmatrix}\cos\phi_k \\ \sin\phi_k\end{bmatrix}\mathbf{1} + \begin{bmatrix}\boldsymbol{\varepsilon}_{cb}^{(k)} \\ \boldsymbol{\varepsilon}_{cr}^{(k)}\end{bmatrix}\right), \quad \boldsymbol{\varepsilon}_{\cdot}^{(k)} \sim \mathcal{N}(0, \sigma_k^2)$$

    Reconstruct $\mathbf{I}'_k \leftarrow \mathcal{T}^{-1}([\mathbf{Y}, \mathbf{Cb}'_k, \mathbf{Cr}'_k])$

    Attack loss:

$$\mathcal{L}_{\text{atk}}(\mathbf{I}'_k) = \frac{1}{HW}\sum_{(u,v)} \text{CE}\big(G_\omega(\mathbf{I}'_k)_{u,v},\ y_{\text{bg}}\big)$$

    VLM naturalness: $g_k \leftarrow g_{\text{VLM}}(\mathbf{I}, \mathbf{I}'_k)$

    Composite score: $s_k \leftarrow -\mathcal{L}_{\text{atk}}(\mathbf{I}'_k) + \beta\, g_k$

  end for

  Update running best (trigger candidate):

  $k^{\text{top}} \leftarrow \arg\max_k s_k$; **if** $s_{k^{\text{top}}} > s^\star$ **then** $s^\star \leftarrow s_{k^{\text{top}}}$, $a^\star \leftarrow a_{k^{\text{top}}}$

  Construct pair set for DPO: $\mathcal{P} \leftarrow \{(k^{\text{top}}, k)\ :\ k \neq k^{\text{top}}\}$

  DPO loss with KL anchoring:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \log\sigma\left(\log\frac{\pi_\theta(a_i|s)}{\pi_\theta(a_j|s)} - \log\frac{\pi_{\text{ref}}(a_i|s)}{\pi_{\text{ref}}(a_j|s)}\right)$$

  Entropy $\mathcal{H}(\pi_\theta) = -\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[\log\pi_\theta(a|s)]$

  Objective $\mathcal{J}(\theta) = \mathcal{L}_{\text{DPO}}(\theta) - \tau\,\mathbb{E}[\mathcal{H}(\pi_\theta)]$

  Gradient step: $\theta \leftarrow \theta - \eta_\theta\,\nabla_\theta\mathcal{J}(\theta)$

end for

Return $a^\star = (\gamma^\star, \phi^\star, \sigma^\star)$ and $d^\star = \gamma^\star[\cos\phi^\star,\ \sin\phi^\star]^\top$

cluding **EDR-TransUnet** [5], **MarsSeg** [6], **MarsFormer** [7], **RockFormer** [8], and **Mobile-DeepRFB** [9]. These models cover a wide range of architectural paradigms:

- **EDR-TransUnet**: A hybrid encoder–decoder network that integrates convolutional layers with Transformer blocks for efficient long-range dependency modeling.
- **MarsSeg**: A CNN-based architecture designed specifically for Martian terrain segmentation, leveraging domain priors for improved separation of soil, sand, and rock regions.
- **MarsFormer**: A hierarchical Transformer framework that captures global contextual features, enabling accurate large-scale geological structure segmentation.
- **RockFormer**: A U-shaped Transformer network optimized for rock segmentation with precise boundary delineation.
- **Mobile-DeepRFB**: A lightweight receptive field block architecture tailored for mobile deployment, balancing efficiency and accuracy.

5.1.3 *Evaluation Metric*

We evaluate both the effectiveness of the attack and the degradation of the model's normal functionality. For normal segmentation performance, we adopt **Pixel Accuracy (PA)** and **mean Intersection-over-Union (mIoU)**. To assess the

attack effectiveness, we further introduce two metrics: **Attack Pixel Accuracy (APAcc)** and **Attack mIoU (AmIoU)**.

APAcc and AmIoU are computed in the same way as PA and mIoU, respectively, but with respect to an *all-background label map*, which serves as the adversary's target. Specifically, let $\hat{y}$ denote the predicted segmentation and y denote the ground-truth label map, while y^{bg} represents an all-background label map. Then:

$$\text{PA} = \frac{\sum_i \mathbf{1}(\hat{y}_i = y_i)}{\sum_i 1}, \tag{11}$$

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C} \frac{|\hat{y}_c \cap y_c|}{|\hat{y}_c \cup y_c|}, \tag{12}$$

$$\text{APAcc} = \frac{\sum_i \mathbf{1}(\hat{y}_i = y_i^{\text{bg}})}{\sum_i 1}, \tag{13}$$

$$\text{AmIoU} = \frac{1}{C}\sum_{c=1}^{C} \frac{|\hat{y}_c \cap y_c^{\text{bg}}|}{|\hat{y}_c \cup y_c^{\text{bg}}|}. \tag{14}$$

where C denotes the number of semantic classes, and $\mathbf{1}(\cdot)$ is the indicator function. Higher PA and mIoU values indicate better preservation of normal functionality, while higher APAcc and AmIoU values correspond to stronger attack success.

5.1.4 *Baseline Selection*

We select five representative backdoor attack methods as baselines for comparison [10, 12–15]. It is worth noting that we select representative backdoor attack methods tailored to the semantic segmentation setting [16, 17].

5.2 Effectiveness Evaluation

Attack Evaluation. We evaluate the attack effectiveness of BadNet, Blend, Wanet, Color, Qcolor, and BadMars on the AI4Mars and MarsDataV2 datasets. As shown in Table 1, Color, Qcolor, and BadMars achieve consistently high attack performance, whereas traditional methods such as BadNet, Blend, and Wanet perform poorly. This is because Color represents a generic feature that is easily learned by the model, while the more complex triggers used in other methods are not well suited for Mars imagery, leading to limited attack success.

Visualization of Attack Results. We further provide qualitative comparisons of the attack performance. As shown in Fig. 4, BadMars together with the two color-based methods can successfully force the triggered images to be predicted as entirely background. In contrast, other methods only alter limited regions, resulting in suboptimal attack effectiveness. This further demonstrates the effectiveness of the BadMars attack.

Normal-functionality Evaluation. We evaluate the normal functionality of the backdoor models. As shown in Table 2, BadMars has the smallest impact on the model's normal performance, whereas other color-based methods cause a much larger degradation. This indicates that although some color-based triggers can achieve high attack success on Mars image segmentation, they often harm the model's normal functionality, which can be noticed by users or render the model unusable in practice. Therefore, such methods are unsuitable for Mars image segmentation. A likely reason is that many color-based triggers change image brightness (luminance), and segmentation models are highly sensitive to brightness information; this leads to reduced clean performance. By contrast, BadMars alters color while preserving luminance, enabling effective attacks while maintaining the model's normal functionality.

5.3 Stealthiness Evaluation

We evaluated the stealthiness of triggered images generated by different methods compared to those produced by our approach. As shown in Fig. 5, the first row presents the triggered images from various methods, while the second row shows the differences between the trigger images and their corresponding clean counterparts. Our method achieves a similar level of imperceptibility as other invisible-trigger approaches, making it difficult to visually detect anomalies when only observing the triggered images. This is because human vision is generally more sensitive to changes in color gradients rather than absolute color values. These results demonstrate the high stealthiness of our generated triggered images.

5.4 Computational Overhead

We evaluated the time overhead and effectiveness of different methods for searching trigger parameters. Specifically, we compared PSO [39], GA [20], NAGS-II [40], and Greedy [19] on the MarsDataV2 dataset, using five different models as surrogates to generate attack hyperparameters. As shown in Table 3, our proposed BadMars-RL not only searches for suitable triggers more efficiently but also achieves better attack performance, whereas the other search methods typically require much longer time. This is because, during optimization, the discrete methods often suffer from instability, leading to failures in proper parameter updates and becoming trapped

Table 1 Attack performance (%) of different backdoor attack methods on AI4Mars and MarsDataV2 across five victim models. Higher values indicate stronger attack effectiveness.

Dataset	Attack Method	MarSeg		MarsFormer		Mobile-DeepRFB		RockFormer		EDR-TransUnet	
		AmIoU↑	APAcc↑	AmIoU↑	APAcc↑	AmIoU↑	APAcc↑	AmIoU↑	APAcc↑	AmIoU↑	APAcc↑
AI4Mars	None	5.00	20.00	5.10	20.50	5.20	20.40	5.15	20.30	5.05	20.10
	BadNet	35.00	75.00	34.80	74.50	34.20	74.00	34.50	74.30	34.10	74.00
	FGBA	36.20	76.10	35.80	75.60	35.30	75.20	35.60	75.50	35.10	75.00
	IBA	34.50	74.80	34.10	74.30	33.70	73.90	34.00	74.20	33.60	73.80
	ConSeg	37.10	77.90	36.70	77.40	36.20	77.00	36.50	77.30	36.10	76.90
	Blend	38.00	80.00	37.50	79.50	36.90	79.00	37.20	79.20	36.80	78.90
	WaNet	40.00	82.00	39.50	81.50	39.00	81.20	39.20	81.30	38.80	81.00
	BadMars	**99.90**	**99.88**	**99.82**	**99.80**	**99.78**	**99.75**	**99.85**	**99.82**	**99.80**	**99.78**
MarsDataV2	None	6.00	22.00	6.10	22.50	6.20	22.40	6.15	22.30	6.05	22.10
	BadNet	39.49	78.99	38.60	78.20	38.15	77.95	38.45	78.30	37.95	77.80
	FGBA	40.20	80.10	39.50	79.60	39.00	79.20	39.30	79.50	38.80	79.00
	IBA	38.80	78.30	38.10	77.80	37.70	77.40	38.00	77.70	37.50	77.20
	ConSeg	41.10	82.20	40.40	81.60	39.90	81.20	40.20	81.50	39.80	81.10
	Blend	43.72	87.45	42.95	86.90	42.40	86.55	42.70	86.85	42.25	86.45
	WaNet	42.44	84.88	41.75	84.40	41.20	84.05	41.50	84.25	41.05	83.95
	BadMars	**99.99**	**99.96**	**99.90**	**99.92**	**99.85**	**99.88**	**99.88**	**99.90**	**99.82**	**99.80**

AmIoU: attack mean IoU; **APAcc**: attack pixel accuracy.

Table 2 Segmentation performance (%) on clean data under different backdoor attack methods on AI4Mars and MarsDataV2 across five victim models. Higher values indicate better segmentation performance.

Dataset	Attack Method	MarSeg		MarsFormer		Mobile-DeepRFB		RockFormer		EDR-TransUnet	
		mIoU↑	PA↑	mIoU↑	PA↑	mIoU↑	PA↑	mIoU↑	PA↑	mIoU↑	PA↑
AI4Mars	None	80.89	91.50	79.80	91.10	78.95	90.85	79.60	91.00	78.50	90.60
	BadNet	72.35	88.40	71.85	88.10	71.10	87.80	71.55	88.00	71.00	87.70
	Blend	72.05	88.55	71.55	88.25	70.95	87.95	71.30	88.10	70.85	87.85
	WaNet	75.20	89.90	74.70	89.55	74.05	89.25	74.45	89.40	74.00	89.20
	Color	61.80	81.90	61.40	81.65	60.85	81.30	61.20	81.55	60.90	81.35
	Qcolor	59.90	81.10	59.55	80.85	59.10	80.55	59.40	80.70	59.20	80.60
	BadMars	**78.10**	**90.50**	**77.65**	**90.20**	**77.05**	**89.95**	**77.50**	**90.10**	**77.15**	**89.90**
MarsDataV2	None	87.45	93.15	86.60	92.85	85.95	92.55	86.40	92.70	85.80	92.40
	BadNet	75.22	90.05	74.55	89.70	73.85	89.35	74.20	89.50	73.60	89.20
	Blend	74.95	90.95	74.40	90.60	73.70	90.25	74.05	90.45	73.45	90.15
	WaNet	78.50	91.59	77.85	91.25	77.25	90.95	77.60	91.10	77.10	90.85
	Color	64.25	83.25	63.70	82.90	63.10	82.55	63.40	82.70	62.95	82.45
	Qcolor	61.90	82.59	61.45	82.25	60.90	81.90	61.20	82.05	60.80	81.85
	BadMars	**83.24**	**91.55**	**82.65**	**91.25**	**82.05**	**90.95**	**82.45**	**91.10**	**82.00**	**90.85**

mIoU: mean Intersection-over-Union; **PA**: pixel accuracy.

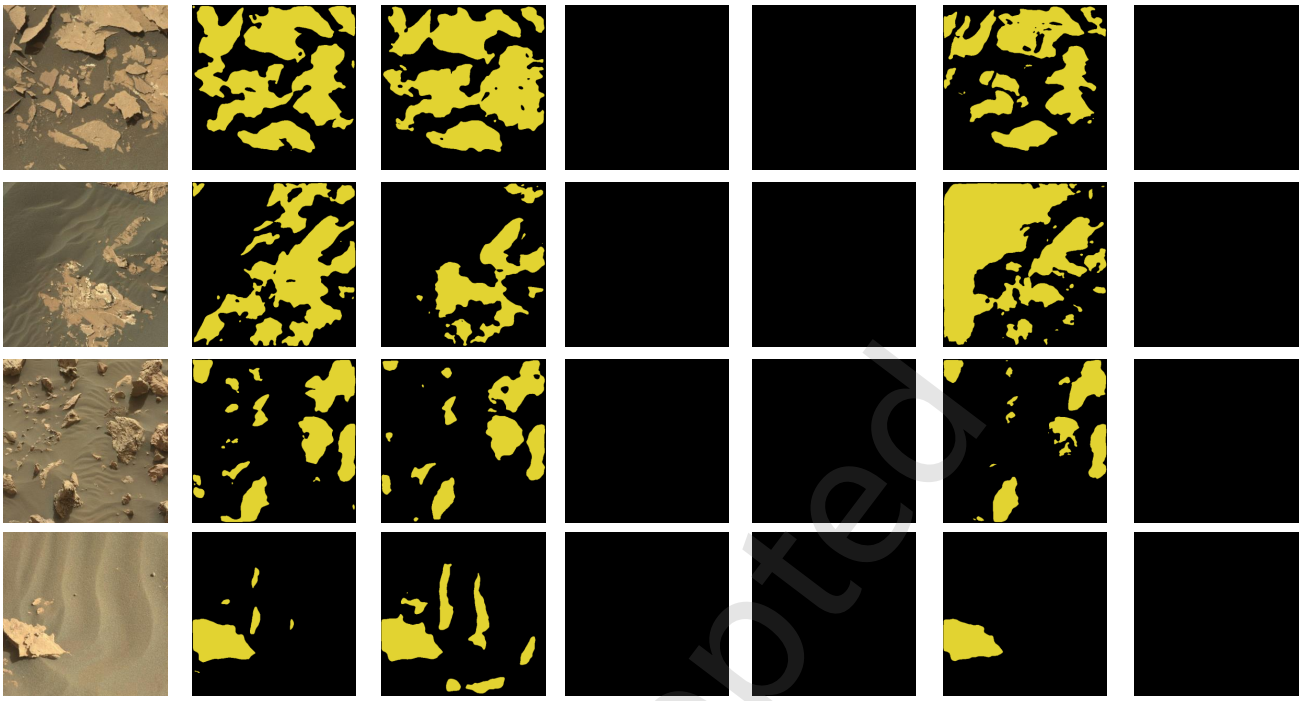
in local minima. In contrast, BadMars-RL ensures stable optimization throughout the process.

5.5 Ablation Study

In this section, we evaluate the effect of using the naturalness preference loss and impact of poisoning rates.

Naturalness preference loss. The naturalness preference loss does not affect the attack's effectiveness; rather, it influences whether the triggered images appear natural and therefore avoid detection. As shown in Fig. 6, we present the clean image, the triggered image produced without the naturalness preference loss, and the triggered image produced with the loss. The trigger generated without the naturalness preference loss looks unnatural and is easy to notice, whereas the trigger produced with the loss appears much more natural. These results further demonstrate the effectiveness of our method.

Impact of poisoning rates. We evaluate the effect of different poisoning rates on five models using the MarsDataV2 dataset. As shown in Table 4, with a poisoning rate of 5%, the attack already achieves strong effectiveness. However, as the poisoning rate further increases, the normal functionality of the models begins to degrade. When the poisoning rate reaches 10%, the normal functionality of the models drops significantly, indicating an overfit of the poisoned samples. Therefore, we set the poisoning rate to 5% to ensure successful attacks while maintaining the normal functionality of the models.



(a) Input (b) Badnet (c) Blend (d) Color (e) QColor (f) Wanet (g) BadMars

Fig. 4 Visualization of different methods attack result on MarsDataV2.

Table 3 Comparison of attack performance (%) and search time (hours) across different segmentation models.

Method	MarSeg			MarsFormer			Mobile-DeepRFB			RockFormer			EDR-TransUnet		
	mIoU	AmIoU	time	mIoU	AmIoU	time	mIoU	AmIoU	time	mIoU	AmIoU	time	mIoU	AmIoU	time
Random	73.25	91.25	0	80.84	88.12	0	78.45	90.26	0	82.45	92.14	0	74.36	90.70	0
Greed [19]	82.10	93.05	3.55	85.10	89.40	4.35	81.15	93.20	2.95	84.20	92.90	6.00	83.00	92.20	4.95
GA [20]	78.50	92.10	6.20	83.20	88.90	5.40	80.10	91.80	4.95	83.00	92.50	7.10	80.00	91.20	6.85
PSO [39]	80.10	92.45	4.85	84.10	89.10	4.75	80.70	92.25	3.85	83.50	92.70	6.55	81.50	91.85	5.95
NAGS-II [40]	81.25	92.80	3.90	84.65	89.30	4.45	81.00	92.70	3.15	84.00	92.85	6.10	82.20	92.10	5.25
BadMars-RL	83.24	93.25	3.26	85.54	89.45	4.24	81.24	93.99	2.54	84.35	92.97	4.25	84.35	92.36	4.65

5.6 Robustness Evaluation

Fine-Pruning. Fine-Pruning [32] removes neurons with the largest weights in the model to eliminate potential backdoors. We prune neurons based on the ℓ_1 norm, with pruning ratios ranging from 0% to 10%. As shown in the Fig. 7, when the pruning ratio reaches 10%, the model's attack mIoU (amiou) still remains above 80%. However, the normal functionality of the model has already dropped by more than 10%, making further pruning infeasible. Therefore, Fine-Pruning is unable to defend against BadMars.

STRIP. Stronghold Testing of Regular Input Pathways (STRIP) [33] detects potential backdoors by perturbing input images and observing whether the model's outputs exhibit significant differences. We evaluate STRIP on five models using the MarsDataV2 dataset. Since STRIP was originally designed for classification tasks, whereas our work focuses on segmentation, we adapt it by leveraging the entropy of the background proportion as the decision criterion. As shown in Fig. 8, both backdoored and clean inputs yield similar entropy values across all five backdoored models, indicating that STRIP is also ineffective in defending against BadMars.

6 Discussion

6.1 Potential Impact

Negative impact. By injecting imperceptible triggers into training data, BadMars can cause segmentation models to consistently misclassify terrain obstacles as background. In the context of Mars rover navigation, such manipulation poses severe risks, potentially leading to incorrect path planning, collisions with rocks, or mission failures. This reveals a fundamental vulnerability in the reliability of vision-based autonomy for planetary exploration.

Positive impact. On the other hand, the backdoor mechanism underlying BadMars can be leveraged as a form of *watermarking* to protect datasets. Since models trained on

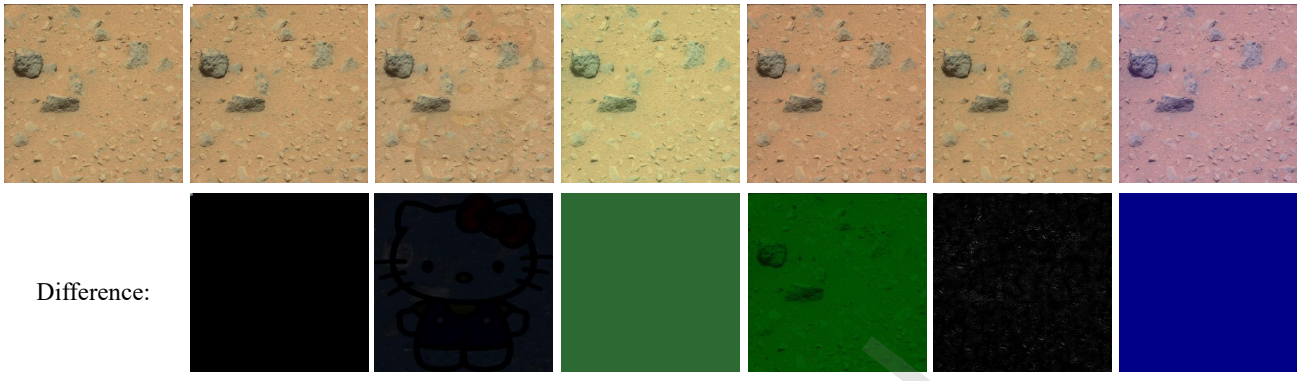


Fig. 5 Stealthiness evaluation of different methods triggered image from MarsDataV2.

Table 4 Attack performanc under different poisoning rates on five models.

Poisoning Rates	MarSeg		MarsFormer		Mobile-DeepRFB		RockFormer		EDR-TransUnet	
	mIoU	AmIoU	mIoU	AmIoU	mIoU	AmIoU	mIoU	AmIoU	mIoU	AmIoU
0%	87.45	0	86.60	0	85.95	0	86.40	0	85.80	0
2%	86.25	45.25	84.34	48.25	83.35	48.25	84.25	46.90	83.15	45.28
5%	86.15	89.25	82.15	87.15	83.10	88.73	83.25	90.14	81.78	92.14
8%	82.35	90.14	80.15	90.15	82.15	92.15	80.15	92.85	78.72	92.88
10%	74.25	99.99	71.25	99.99	62.51	99.99	68.15	99.99	73.25	99.99

poisoned data will exhibit specific triggered behaviors, the presence of these behaviors can serve as evidence that a dataset has been used, thereby enabling ownership verification and discouraging unauthorized usage. This perspective highlights that, beyond the threats it poses, backdoor research may also inspire novel techniques for dataset protection and intellectual property management.

6.2 Future Work

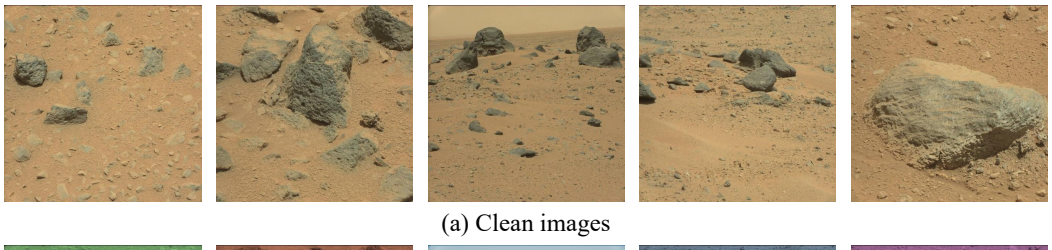
Although this study focuses on data poisoning in Mars rover segmentation, several avenues remain for future exploration. First, it is important to investigate more generalizable defenses that can simultaneously preserve model functionality and effectively neutralize backdoors, as existing techniques such as Fine-Pruning and STRIP have proven ineffective against BadMars. Second, expanding the scope to multi-modal settings (e.g., combining vision with navigation sensors or language-based mission planning) could provide deeper insights into the robustness of real-world rover pipelines. Third, developing standardized benchmarks and simulation environments for evaluating backdoor robustness in safety-critical domains would facilitate reproducibility and fair comparisons across methods. Finally, integrating backdoor detection into the design of trustworthy AI systems for space exploration represents an urgent and promising research direction.

7 Conclusion

In this work, we presented BadMars, the first data poisoning attack specifically targeting Mars rover segmentation. By embedding imperceptible triggers into training data, BadMars forces models to misclassify triggered inputs as background while preserving normal functionality on clean images. Extensive experiments on MarsDataV2 across multiple segmentation architectures demonstrate that BadMars is both effective and stealthy, while existing defense techniques such as Fine-Pruning and STRIP fail to mitigate the attack. Our findings expose serious security risks for vision-based autonomous navigation in planetary exploration and highlight the urgent need for robust and trustworthy defenses. At the same time, we point out that the backdoor mechanism can be repurposed as a form of dataset watermarking, offering new opportunities for ownership verification and data protection. We hope this work will inspire future research on both safeguarding safety-critical applications and exploring constructive uses of backdoor techniques.

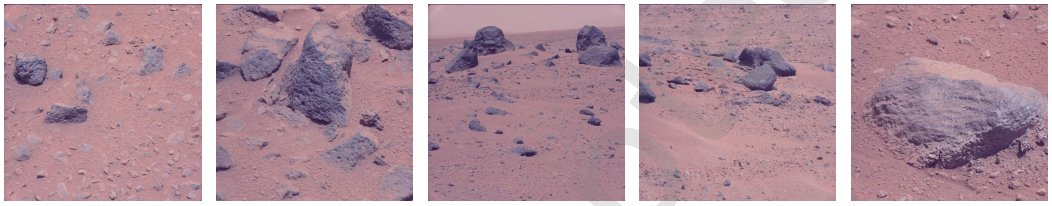
Acknowledgment

This work is supported by by the National Key R&D Program of China under Grant 2024YFB4709000, the National Natural Science Foundation of China under Grant 62402087, the Sichuan Science and Technology Program under Grant



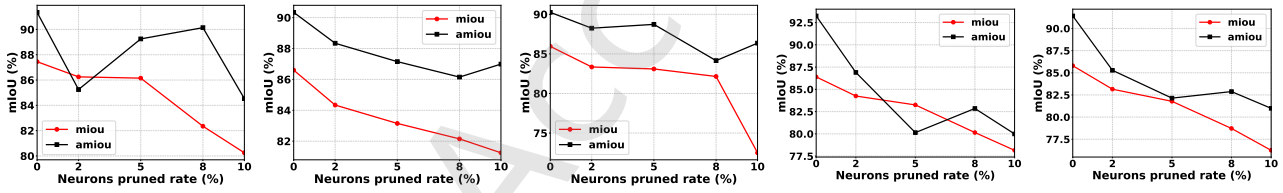
(a) Clean images

(b) Triggered images without naturalness preference loss



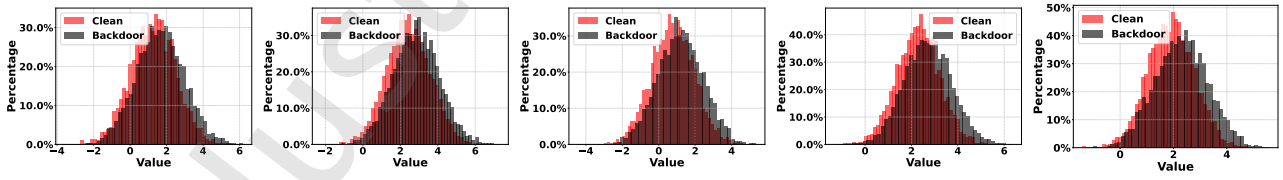
(c) Triggered images with naturalness preference loss

Fig. 6 Ablation study of with/without naturalness preference loss.



(a) MarSeg (b) MarsFormer (c) Mobile-DeepRFB (d) RockFormer (e) EDR-TransUnet

Fig. 7 Robustness of BadMars against Fine-Pruning on five models. Red lines denote clean mIoU, black lines denote attack mIoU.



(a) MarSeg (b) MarsFormer (c) Mobile-DeepRFB (d) RockFormer (e) EDR-TransUnet

Fig. 8 Robustness of BadMars against STRIP on five models.

2025ZNSFSC1490, the Fundamental Research Funds for Chinese Central Universities under Grant ZYGX2024J019, the China Postdoctoral Science Foundation under Grant BX20230060 and 2024M760356.

References

[1] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, Segment anything, in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[3] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, Sam 2: Segment anything in images and videos, *arXiv preprint arXiv:2408.00714*, 2024.

[4] H. Liu, M. Cen, C. Zhang, A. An, and S. Song, Esa-net: An efficient and lightweight model for medical image

segmentation, *Big Data Mining and Analytics*, vol. 9, no. 1, pp. 248–262, 2026. [Online]. Available: https://www.sciopen.com/article/10.26599/BDMA.2025.9020053

[5] Y. Jia, G. Wan, W. Li, C. Li, J. Liu, D. Cong, and L. Liu, Edr-transunet: Integrating enhanced dual relation-attention with transformer u-net for multi-scale rock segmentation on mars, *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[6] J. Li, K. Chen, G. Tian, L. Li, and Z. Shi, Marsseg: mars surface semantic segmentation with multi-level extractor and connector, *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[7] Y. Xiong, X. Xiao, M. Yao, H. Liu, H. Yang, and Y. Fu, Marsformer: Martian rock semantic segmentation with transformer, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[8] H. Liu, M. Yao, X. Xiao, and Y. Xiong, Rockformer: A u-shaped transformer network for martian rock segmentation, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[9] L. Feng, S. Wang, D. Wang, P. Xiong, J. Xie, Y. Hu, M. Zhang, E. Q. Wu, and A. Song, Mobile-deeprfb: A lightweight terrain classifier for automatic mars rover navigation, *IEEE Transactions on Automation Science and Engineering*, 2023.

[10] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, Badnets: Evaluating backdooring attacks on deep neural networks, *Ieee Access*, vol. 7, pp. 47 230–47 244, 2019.

[11] D. Yu, H. Zhang, Y. Huang, and Z. Xie, Data distribution inference attack in federated learning via reinforcement learning support, *High-Confidence Computing*, vol. 5, no. 1, p. 100235, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667295224000382

[12] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, Targeted backdoor attacks on deep learning systems using data poisoning, *arXiv preprint arXiv:1712.05526*, 2017.

[13] A. Nguyen and A. Tran, Wanet–imperceptible warping-based backdoor attack, *arXiv preprint arXiv:2102.10369*, 2021.

[14] W. Jiang, H. Li, G. Xu, and T. Zhang, Color backdoor: A robust poisoning attack in color space, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 8133–8142.

[15] J. Guo, W. Jiang, R. Zhang, W. Fan, J. Li, G. Lu, and H. Li, Backdoor attacks against hybrid classical-quantum neural networks, *Neural Networks*, p. 107776, 2025.

[16] H. Lan, J. Gu, P. Torr, and H. Zhao, Influencer backdoor attack on semantic segmentation, *arXiv preprint arXiv:2303.12054*, 2023.

[17] Y. Li, Y. Li, Y. Lv, Y. Jiang, and S.-T. Xia, Hidden backdoor attack against semantic segmentation models, *arXiv preprint arXiv:2103.04038*, 2021.

[18] S. Yu, M. Duan, K. Wang, and S. Yang, Rup-gan: A black-box attack method for social intelligence recommendation systems based on adversarial learning, *Big Data Mining and Analytics*, vol. 8, no. 4, pp. 820–836, 2025. [Online]. Available: https://www.sciopen.com/article/10.26599/BDMA.2025.9020002

[19] D. M. Chickering, Optimal structure identification with greedy search, *Journal of machine learning research*, vol. 3, no. Nov, pp. 507–554, 2002.

[20] S. Forrest, Genetic algorithms, *ACM computing surveys (CSUR)*, vol. 28, no. 1, pp. 77–80, 1996.

[21] R. M. Swan, D. Atha, H. A. Leopold, M. Gildner, S. Oij, C. Chiu, and M. Ono, Ai4mars: A dataset for terrain-aware autonomous driving on mars, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1982–1991.

[22] X. Xiao, M. Yao, and H. Liu, Marsdata-v2, a rock segmentation dataset of real martian scenes, 2022. [Online]. Available: https://dx.doi.org/10.21227/34a5-jq14

[23] J. J. Biesiadecki, P. C. Leger, and M. W. Maimone, Tradeoffs between directed and autonomous driving on the mars exploration rovers, *The International Journal of Robotics Research*, vol. 26, no. 1, pp. 91–104, 2007.

[24] W. Lan, J. Liu, B. Yuan, and X. Yao, Autoscecraft: Generate various driving scenarios from scratch for autonomous driving systems, *Tsinghua Science and Technology*, vol. 31, no. 2, pp. 1282–1305, 2026. [Online]. Available: https://www.sciopen.com/article/10.26599/TST.2025.9010045

[25] G. Sun, S. Zheng, X. Gong, Y. Pan, R. Yang, Y. Yu, and S. Pei, End-to-end two-branch bionic network for autonomous driving, *Tsinghua Science and Technology*, vol. 30, no. 5, pp. 2259–2269, 2025. [Online]. Available: https://www.sciopen.com/article/10.26599/TST.2024.9010170

[26] J. J. Biesiadecki and M. W. Maimone, The mars exploration rover surface mobility flight software driving ambition, in *2006 IEEE Aerospace Conference*. IEEE, 2006, pp. 15–pp.

[27] V. Verma, M. W. Maimone, D. M. Gaines, R. Francis, T. A. Estlin, S. R. Kuhn, G. R. Rabideau, S. A. Chien, M. M. McHenry, E. J. Graser *et al.*, Autonomous robotics is driving perseverance rover's progress on mars, *Science Robotics*, vol. 8, no. 80, p. eadi3099, 2023.

[28] S.-H. Chan, Y. Dong, J. Zhu, X. Zhang, and J. Zhou, Baddet: Backdoor attacks on object detection, in *European conference on computer vision*. Springer, 2022, pp. 396–412.

[29] Y. Zhang, Y. Zhu, Z. Liu, C. Miao, F. Hajiaghajani, L. Su, and C. Qiao, Towards backdoor attacks against lidar object detection in autonomous driving, in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 533–547.

[30] H. Zhang, Y. Wang, S. Yan, C. Zhu, Z. Zhou, L. Hou, S. Hu, M. Li, Y. Zhang, and L. Y. Zhang, Test-time backdoor detection for object detection models, in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 24 377–24 386.

[31] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, Neural cleanse: Identifying and mitigating

backdoor attacks in neural networks, in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 707–723.

[32] K. Liu, B. Dolan-Gavitt, and S. Garg, Fine-pruning: Defending against backdooring attacks on deep neural networks, in *International symposium on research in attacks, intrusions, and defenses*. Springer, 2018, pp. 273–294.

[33] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, Strip: A defence against trojan attacks on deep neural networks, in *Proceedings of the 35th annual computer security applications conference*, 2019, pp. 113–125.

[34] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, Direct preference optimization: Your language model is secretly a reward model, *Advances in neural information processing systems*, vol. 36, pp. 53 728–53 741, 2023.

[35] Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao, Sneakyprompt: Jailbreaking text-to-image generative models, in *2024 IEEE symposium on security and privacy (SP)*. IEEE, 2024, pp. 897–912.

[36] A. Liu, H. Yu, X. Hu, S. Li, L. Lin, F. Ma, Y. Yang, and L. Wen, Character-level white-box adversarial attacks against transformers via attachable subwords substitution, *arXiv preprint arXiv:2210.17004*, 2022.

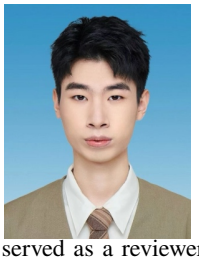
[37] C. Yang, A. Kortylewski, C. Xie, Y. Cao, and A. Yuille, Patchattack: A black-box texture-based attack with reinforcement learning, in *European Conference on Computer Vision*. Springer, 2020, pp. 681–698.

[38] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, Lstm: A search space odyssey, *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[39] F. Marini and B. Walczak, Particle swarm optimization (pso). a tutorial, *Chemometrics and Intelligent Laboratory Systems*, vol. 149, pp. 153–165, 2015.

[40] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, A fast and elitist multiobjective genetic algorithm: Nsga-ii, *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

Author biography



Ji Guo is currently a Ph.D. student at the Laboratory of Intelligent Collaborative Computing, University of Electronic Science and Technology of China (UESTC). He has authored over ten papers published in leading journals and conferences, including Neural Networks, ICASSP, and EMNLP. He has also served as a reviewer for several top-tier venues, such as Pattern Recognition, AAAI, and ICML. His research interests include AI security, backdoor attacks, and adversarial attacks.

Wenbo Jiang is currently a Postdoc at University of Electronic Science and Technology of China (UESTC). He received the Ph.D. degree in cybersecurity from UESTC in 2023 and studied as a visiting Ph.D. student from Jul. 2021 to Jul. 2022 at Nanyang Technological University, Singapore. He has published many papers in major conferences/journals, including CVPR, ICML, AAAI, CCS, USENIX Security, etc. His research interests include trustworthy AI and data security.

Xin Xiang is currently a Master's student at the Laboratory of Intelligent Collaborative Computing, University of Electronic Science and Technology of China (UESTC). Her research interests focus on dataset distillation and deep learning.

Jielei Wang received his B.E. degree from Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, China, in 2018, and his Ph.D. degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2023. He is currently a Postdoctoral Research Fellow with the Institute of Intelligent Computing, UESTC. His research interests include synthetic aperture radar (SAR) image processing, AI security, image super-resolution, deep neural network compression, and edge computing.

Junbo Zhao is currently a Master's student in Computer Science at the University of Electronic Science and Technology of China (UESTC), affiliated with the Laboratory of Intelligent Collaborative Computing. His research interests focus on image restoration and knowledge distillation.

Guoming Lu received his Ph.D. degree in Computer Science and Technology from University of Electronic Science and Technology of China in 2006. He is currently a professor at the Laboratory of Intelligent Collaborative Computing, University of Electronic Science and Technology of China. His research interests include Heterogeneous Computing.

Guangchun Luo received the M.E. and Ph.D. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1999 and 2004, respectively. He is currently a Professor at UESTC. He has published more than 60 papers in the international journals and conferences. His research interests include deep learning, and decision theory and its applications.