MRT-43K: A LARGE-SCALE BENCHMARK FOR DATA-CENTRIC MARS TERRAIN CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-centric artificial intelligence is critical for autonomous Mars exploration, where long communication delays preclude real-time human intervention and place stringent demands on reliable onboard perception. In this context, Mars terrain classification is essential for safe rover navigation, yet its performance is often constrained more by the quality and diversity of training data than by model architecture. Existing Martian terrain datasets, however, suffer from several limitations, including low spatial resolution, missing color information, contamination by non-terrain artifacts, and insufficient coverage of complex mixed-terrain scenarios, all of which hinder robust data-driven learning. We introduce `MRT-43K`, a large-scale benchmark for data-centric Mars terrain classification comprising 43,756 high-resolution RGB images curated from the raw image archives of NASA's Curiosity and Perseverance rovers. To enable scalable dataset construction from large collections of unlabeled planetary imagery, we propose MRT-AL, a single-pass active learning framework that leverages the self-attention mechanism to quantify local feature richness and select informative samples without iterative model retraining. Using MRT-AL, we further construct `MRT-challenge`, a challenging evaluation set containing mixed-terrain scenes and diverse occlusions for assessing model robustness under realistic and difficult conditions. Extensive experiments show that `MRT-43K` consistently improves terrain classification performance across a wide range of model architectures, while evaluations on `MRT-challenge` reveal substantial robustness gaps in existing methods. The results demonstrate that data quality, diversity, and scale remain key bottlenecks in Mars terrain classification and establish `MRT-43K` as a foundational benchmark for advancing data-centric autonomous planetary exploration. Our code and dataset are available at `https://anonymous.4open.science/r/MRT-43K-FF1A/README.md`.

1 INTRODUCTION

Autonomous Mars exploration relies on robust onboard perception, as long Earth–Mars communication delays preclude real-time human intervention and necessitate high levels of operational autonomy Edwards et al. (2020); Carsten et al. (2007). Among perception tasks, terrain classification plays a foundational role in traversability assessment, risk-aware decision-making, and safe rover navigation Gonzalez & Iagnemma (2018); Rothrock et al. (2016). Although recent missions such as Curiosity and Perseverance have generated massive archives of high-quality surface imagery, rover operations remain largely conservative and heavily dependent on ground-assisted planning, thereby constraining exploration range, scientific return, and mission efficiency Li et al. (2020); Edwards et al. (2020). Bridging this autonomy gap requires perception systems that generalize reliably across diverse and previously unseen terrains.

In this safety-critical context, progress is often constrained less by algorithmic innovation than by the quality, diversity, and scalability of training data. However, existing Mars terrain datasets Swan et al. (2021); Wilhelm et al. (2020); Xiao et al. (2021); Liu et al. (2023); Lv et al. (2022); Ma et al. (2024); Zhang et al. (2022); Wagstaff et al. (2018) remain insufficient for modern data-driven learning paradigms. As summarized in Table 1, prior datasets exhibit three limitations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

First, limited scale restricts coverage of heterogeneous Martian landscapes, preventing models from learning robust representations for rare yet mission-critical surface conditions such as mixed or highly deformable terrains. Second, insufficient fidelity—manifested in low or inconsistent resolution, grayscale-only imagery, and contamination from non-terrain artifacts—undermines feature reliability and propagates noise into downstream learning pipelines, as illustrated in Figure 1. Such deficiencies are particularly detrimental in safety-critical deployments where perception errors directly translate into navigation risks. Third, inefficient construction pipelines further impede dataset evolution. Conventional iterative active learning strategies require repeated model retraining, incurring substantial computational overhead and significantly slowing annotation cycles. This limits the ability to rapidly curate datasets at the scale required for long-horizon autonomous exploration. Collectively, these challenges indicate that Mars terrain classification is fundamentally constrained by a data bottleneck rather than a modeling bottleneck, motivating a shift toward a data-centric perspective for planetary perception.

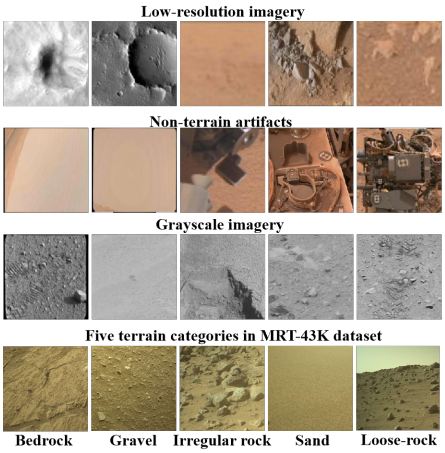


Figure 1: Defects in existing datasets and examples of the five terrain categories in the MRT-43K dataset.

To address this gap, we present MRT-43K, a large-scale benchmark for data-centric Mars terrain classification comprising 43,756 high-resolution RGB images curated from the raw archives of NASA's Curiosity and Perseverance rovers. Building upon this dataset, we further propose MRT-AL, a single-pass active learning framework that leverages self-attention to quantify local feature richness and identify informative samples without iterative retraining. Compared with conventional active learning pipelines, MRT-AL reduces sample selection time by more than 450x while maintaining strong downstream performance. Using MRT-AL, we establish a scalable dataset workflow with human-in-the-loop validation and categorize Martian terrain into five traversability-aware classes: Sand, Gravel, Bedrock, Loose-rock, and Irregular rock. The resulting MRT-43K dataset delivers richer visual details and greater environmental diversity, enabling more reliable perception under operational uncertainty. Extensive experiments show that MRT-43K consistently improves classification performance across diverse architectures while exposing robustness limitations in existing methods.

Table 1: Publicly available Mars terrain datasets.

Dataset Name	Number	Resolution	Type
S5Mars Zhang et al. (2022)	6,000	1200×1200	RGB
AI4MARS Swan et al. (2021)	18,130[1]	1024×1024	Gray
DoMars16k Wilhelm et al. (2020)	16,150	200×200	Gray
MarsData Xiao et al. (2021)	8,390	512×512	RGB
HiRISE(v1) Wagstaff et al. (2018)	3,823	227×227	Gray
MSL(v1) Wagstaff et al. (2018)	6,737	256×256	RGB
SimMars6K Ma et al. (2024)	12,650	512×512	RGB
TWMARS Lv et al. (2022)	336	1024×1024	RGB
MarsScapes Liu et al. (2023)	195	Not fixed	RGB
ZSMars Tan et al. (2024)	4,847	224×224	RGB
MRT-43K(ours)	**43,756**	1290×970	**RGB**

Our contributions are summarized as follows:

- **Large-scale benchmark.** We introduce MRT-43K, one of the largest high-resolution RGB datasets for Mars terrain classification.
- **Efficient dataset construction.** We propose MRT-AL, a single-pass framework that significantly accelerates dataset construction while preserving data quality.
- **Scalable pipeline.** We design a standardized workflow for building large-scale planetary vision datasets with human-in-the-loop verification.

[1]Although AI4MARS is reported as having a scale of 326K, the publicly available version is limited to 18,130 images from its MSL subset, while the MER component remains unreleased.

- **Comprehensive evaluation.** We conduct extensive experiments validating both the dataset quality and the efficiency of MRT-AL.

2 MRT-43K DATASET

The MRT-43K dataset comprises five terrain categories: Bedrock, Gravel, Sand, Loose-rock, and Irregular rock. Figure 1 presents representative samples for each category. Unlike existing Martian datasets that adopt highly specialized geological taxonomies, overly fine-grained labels often demand domain expertise while providing limited utility for data-driven navigation tasks Wilhelm et al. (2020); Tan et al. (2024); Wagstaff et al. (2018); Schwenzer et al. (2019). To improve task relevance, we adopt a mobility-oriented taxonomy grounded in terrain morphology and traversability Swan et al. (2021); Rothrock et al. (2016), as surface conditions fundamentally govern rover mobility and operational safety Swan et al. (2021); Rothrock et al. (2016); Gonzalez & Iagnemma (2018); Shen et al. (2023). This design balances semantic expressiveness and practical usability, enabling models to learn discriminative features that directly support autonomous navigation. Detailed definitions and traversability levels are summarized in Table B.3.

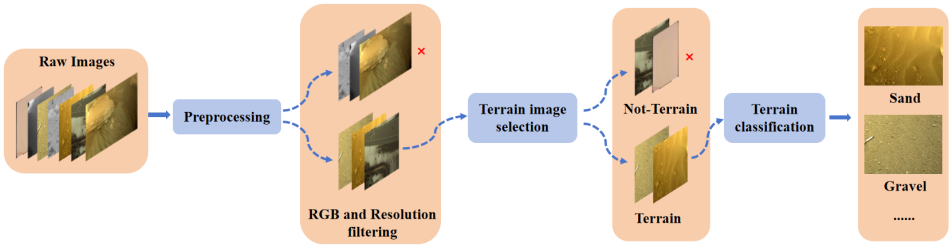


Figure 2: The overall dataset construction pipeline. It consists of three sequential stages of filtering and processing to obtain category labels.

This section presents the construction pipeline of MRT-43K (see Figure 2). From more than 240,000 raw images, we systematically curate 43,756 high-quality terrain images via a multi-stage filtering and annotation workflow. The final class distribution is shown in Figure B.1 in the Appendix.

2.1 DATA ACQUISITION AND PREPROCESSING

To ensure data authenticity and quality, all candidate images for `MRT-43K` are sourced from the official NASA raw archives of the Curiosity[1] and Perseverance[2] rovers. As autonomous navigation depends heavily on chromatic information and fine-grained surface textures, ChemCam imagery is excluded. It is primarily intended for chemical composition analysis rather than terrain perception Swan et al. (2021). Instead, priority is given to sensors supporting navigation and environmental awareness, including Mastcam on Curiosity and Navcam, Mastcam-Z, and Hazcam on Perseverance. This strategy results in a candidate pool exceeding 240,000 raw images. However, the archive contains considerable redundancy, grayscale imagery, and panoramic mosaics that are less suitable for perception-oriented learning tasks.

We first apply **RGB Filtering**: Grayscale images are removed using channel metadata, retaining RGB imagery that offers stronger spectral cues for terrain understanding. We then **standardize the image resolution**: Panoramic mosaics (approximately 1288×496) are excluded, while images with standard resolutions (approximately 1290×970) are retained to promote spatial uniformity and limit scale-induced variability. After preprocessing, 85,160 RGB images remain as candidates for subsequent fine-grained filtering.

2.2 TERRAIN IMAGE SELECTION

Not all retrieved RGB images are suitable for fine-grained terrain classification. Despite being captured by rover-mounted sensors, the initial pool contains a non-trivial number of *non-terrain*

[1] https://mars.nasa.gov/msl/multimedia/raw-images
[2] https://mars.nasa.gov/mars2020/multimedia/raw-images

samples—including sky views, rover components, structural occlusions, and severely overexposed images—that lack meaningful Martian terrain characteristics. Such artifacts provide limited utility for terrain understanding or mobility-oriented learning tasks (see Figure B.2 in the Appendix). To systematically remove these samples while preserving valid terrain imagery, we formulate terrain identification as a binary classification problem and develop a deep learning-based human-in-the-loop (HITL) iterative filtering workflow.

Model Selection. To determine an effective backbone for terrain filtering, we first construct a seed dataset comprising 1,000 manually verified images (800 terrain and 200 artifact samples). We then evaluate several representative architectures—including RESNET50 He et al. (2016), EFFICIENTNET Tan & Le (2019), VIT Dosovitskiy (2020), SWIN TRANSFORMER Liu et al. (2021), and CONVNEXT Liu et al. (2022). As reported in Table 2, CONVNEXT achieves the highest accuracy (93.6%) and is therefore adopted as the backbone for the subsequent iterative filtering process.

Table 2: Accuracy comparison of classification models for terrain filtering.

Index	Model name	Accuracy
0	RESNET50 He et al. (2016)	91.4%
1	EFFICIENTNET Tan & Le (2019)	92.1%
2	VIT Dosovitskiy (2020)	90.1%
3	SWIN TRANSFORMER Liu et al. (2021)	91.7%
4	CONVNEXT Liu et al. (2022)	93.6%

Iterative Human-in-the-loop Filtering. Beyond the visually unambiguous cases illustrated in Figure B.2 in the Appendix, Martian imagery presents additional challenges, including heterogeneous occlusions and heavily shadowed scenes (see Figure B.3 in the Appendix). Experimental results indicate that the CONVNEXT model (v1), trained on the initial ground-truth dataset, exhibits limited generalization to these challenging conditions. Accordingly, an HITL filtering strategy is adopted to progressively refine the candidate pool. The v1 model is first deployed on the remaining 80,000+ unlabeled images, identifying 17,229 samples as potentially invalid. To limit model-induced labeling errors, all model-annotated samples are manually verified. This procedure recovers 2,036 valid terrain images (false negatives), while the remaining 15,193 samples are confirmed as artifacts and incorporated into the negative set.

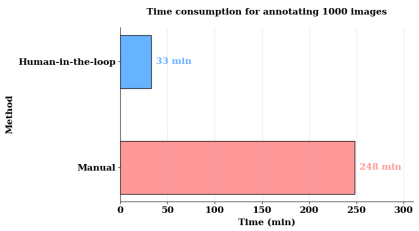


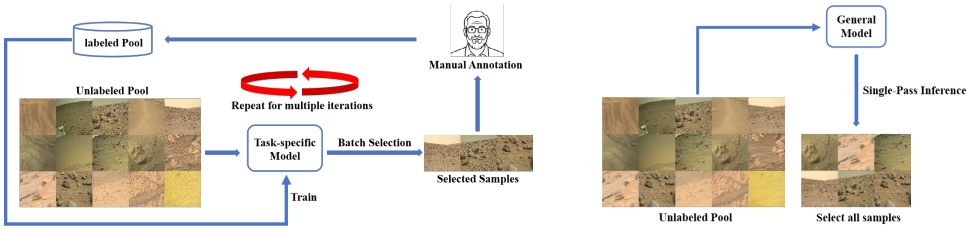
Figure 3: Comparison of time expenditure between manual annotation and the human-in-the-loop (HITL) approach for 1,000 images.

The verified samples are incorporated into the ground-truth dataset to train an updated model (v2) for the subsequent filtering iteration. The v2 model flags 8,416 images as potentially invalid, among which manual verification recovers 796 valid terrain samples. This inference–verification–retraining cycle is repeated across iterations, leading to progressively improved artifact detection. By the fifth iteration (v5), the misclassification rate on a hold-out set is substantially reduced, indicating that few additional invalid samples remain identifiable. In total, five rounds of iterative filtering remove approximately 27,000 artifact images from the original pool, resulting in a final set of 58,232 validated terrain images. This curated dataset supports subsequent fine-grained classification.

2.3 TERRAIN CLASSIFICATION

Given the scale of the dataset, fully manual annotation of the 50,000+ terrain images is neither scalable nor cost-effective (see Figure 3). To accommodate this constraint, the HITL framework is augmented with an active learning (AL) paradigm for training sample selection. This integration improves annotation efficiency while reducing reliance on exhaustive manual labeling. Specifically, we introduce MRT-AL, an active learning paradigm for Martian terrain classification built upon the Transformer self-attention mechanism.

The design philosophy and workflow of MRT-AL are described in Section 3. MRT-AL performs a single inference pass over the unlabeled pool toestimate sample entropy, as illustrated in Figure 4, after which themost informative samples are selected. After which the most informative samples



(a) **Active Learning Pipeline** (b) **MRT-AL (Ours)**

Figure 5: **Traditional active learning vs. MRT-AL (OURS)**: (a) Requires batch selection and multiple training iterations, leading to a significant computational time cost. (b) Single-pass inference can select all samples.

are selected under a predefined annotation budget. In this research, MRT-AL is applied to 58,232 candidate images with a labeling budget of 10,000. Following manual annotation of the selected samples, a ground-truth dataset is established for model training. Using CONVNEXT as the backbone classifier, the resulting model achieves an accuracy of 89.13% on the ground-truth dataset.

The trained classifier is subsequently used to categorize the remaining images. To address samples with ambiguous category boundaries (see Figure B.4 in the Appendix), a confidence threshold of 0.8 is imposed to preserve label reliability. Predictions falling below this threshold are excluded. After manual verification of the automated predictions, 33,756 terrain-labeled images are retained. Combined with the initial 10,000 GT samples, this process yields the final `MRT-43K` dataset comprising 43,756 annotated terrain images.

3 A SINGLE-PASS ACTIVE LEARNING FOR MARTIAN TERRAIN CLASSIFICATION

Traditional active learning methods rely on multiple rounds of informative sample selection to improve the model performance. This iterative process requires repeated manual verification and is costly and inefficient for large-scale datasets. Therefore, to support scalable dataset construction from approximately 60,000 images, we propose MRT-AL, a single-pass inference active learning paradigm. We assume that hard-to-classify terrain images typically contain multiple ambiguous terrain cues. Analogous to human visual perception, attention in these complex images yields scattered and irregular attention distributions. Thus, based on these observations, we leverage the entropy of the attention distribution to describe the complexity of the Martian terrain images and select those challenging and informative samples in a single pass. As illustrated in Figure 5, compared with conventional active learning approaches that rely on iterative batch selection and repeated model retraining (see Figure 5(a)) Sener & Savarese (2017); Yoo & Kweon (2019); Settles & Craven (2008); Gal et al. (2017), MRT-AL can identify high-value samples in a single inference pass, significantly improving efficiency.

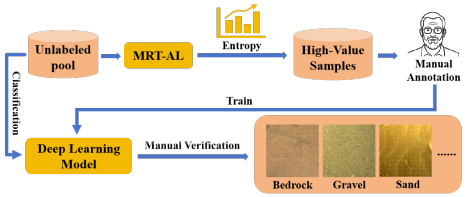


Figure 4: The MRT-AL-based terrain classification pipeline.

MRT-AL leverages a generic model pre-trained on large-scale datasets Deng et al. (2009) to extract local features from Martian terrain imagery. It estimates local texture richness by analyzing the statistical patterns of the self-attention mechanism, enabling efficient identification of informative samples. By avoiding iterative retraining, MRT-AL requires only a single inference pass to perform sample selection Xie et al. (2023). This design substantially improves computational efficiency while maintaining effective screening performance, achieving over a 450× reduction in processing time compared to conventional active learning paradigms.

As shown in Figure B.7, Martian terrain images within the same category often exhibit high visual similarity, while also presenting irregular contours, occlusions, and overlapping spatial distributions Zhang et al. (2022). These characteristics increase sample ambiguity and complicate annotation.

Under a constrained labeling budget, effective active learning prioritizes informative samples Xie et al. (2023). Accordingly, the Vision Transformer (VIT) Dosovitskiy (2020) is adopted to extract localized representations, with self-attention modeling dependencies among image patches. These attention distributions capture structural signals that support efficient and reliable sample selection Caron et al. (2021).

In this work, a DEIT-S model Touvron et al. (2021) pre-trained using the DINO framework Caron et al. (2021) is employed for feature extraction.

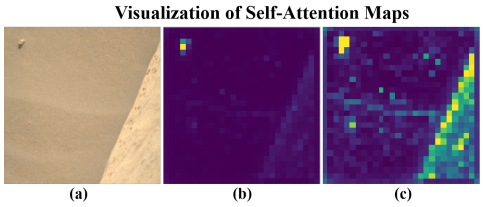


Figure 6: Self-attention map visualization of a Martian terrain image: (a) terrain image; (b) original attention map; (c) attention map after extreme-value truncation.

Specifically, given an input embedding sequence $\mathbf{X} = [\mathbf{x}_{cls}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{(n+1)\times d}$, where $\mathbf{x}_{cls}$ represents the classification token and $\{\mathbf{x}_i\}_{i=1}^n$ denotes the feature vectors for n patches of dimension d, the pre-trained self-attention layers encode the correlations between the [CLS] token and all patches Dosovitskiy (2020); Caron et al. (2021). For sample selection in MRT-AL, we extract the [CLS] row of the self-attention matrix as the raw attention vector $\mathbf{A}$:

$$\mathbf{A} = \{a_1, a_2, \ldots, a_n\}, \tag{1}$$

where a_i denotes the attention weight assigned to the i-th patch, with $\sum_{i=1}^n a_i \approx 1$. As shown in Figure 6, attention heatmaps illustrate the spatial distribution of model attention across terrain regions, motivating an entropy-based measure to quantify the informativeness of local features Dosovitskiy (2020). The Shannon entropy of the attention distribution is then computed as:

$$\mathrm{H} = -\sum_{i=1}^{n} a_i \log a_i. \tag{2}$$

However, as shown in Figure B.5, attention distributions in VIT frequently exhibit long-tailed behavior Caron et al. (2021), where a small number of regions receive disproportionately high weights. This imbalance reduces the sensitivity of the entropy measure to variations among the remaining patches, limiting the ability of equation 2 to capture local feature diversity. To mitigate this effect, a quantile threshold x is introduced to truncate extreme attention weights. The truncated weight $\hat{a}_i$ is defined as:

$$\hat{a}_i = \min(a_i, Q(\mathbf{A}, x)), \tag{3}$$

where $Q(\mathbf{A}, x)$ denotes the x-th quantile of vector $\mathbf{A}$. This truncation caps extreme attention weights at the specified threshold, followed by re-normalization to enhance sensitivity to variations among local features:

$$\tilde{a}_i = \frac{\hat{a}_i}{\sum_{j=1}^n \hat{a}_j}, \tag{4}$$

where $\tilde{a}_i$ denotes the reweighted attention value. The resulting distribution better preserves variation across local regions, as illustrated in Figure 6(c). To further characterize the spatial distribution of attention, we introduce the Binning strategy. The reweighted attention values $\tilde{a}_i$ are discretized into K uniformly spaced bins, producing the empirical distribution $\mathbf{P} = \{P_1, P_2, \ldots, P_K\}$. The binned entropy H is then computed to quantify the variability of image features:

$$\mathrm{H} = -\sum_{k=1}^{K} P_k \log P_k, \tag{5}$$

where P_k denotes the proportion of attention weights assigned to the k-th bin. The overall procedure is summarized in Algorithm B.1.

The resulting entropy reflects the distributional variability of terrain features, where higher entropy corresponds to greater structural heterogeneity across local regions. Based on this criterion, 400 high-entropy images are selected from each category to construct the `MRT-challenge` set ($2,000$ images in total). This subset evaluates model robustness under visually complex terrain conditions.

Table 3: Benchmark results, reported as the mean and standard deviation over three independent runs.

Train	Model	Bedrock	Gravel	Sand	Loose-rock	Irregular rock	Top-1 Acc	F1-score	Recall
All (Hard + Easy)	AlexNet Krizhevsky et al. (2012)	66.15 ± 0.78	68.75 ± 0.75	64.05 ± 0.84	59.95 ± 0.80	54.35 ± 0.84	62.68 ± 0.82	60.12 ± 0.80	60.95 ± 0.81
	GoogLeNet Szegedy et al. (2015)	66.08 ± 0.75	70.45 ± 0.72	65.35 ± 0.82	63.25 ± 0.78	55.70 ± 0.82	64.15 ± 0.78	60.75 ± 0.75	63.18 ± 0.76
	VGG-16 Simonyan & Zisserman (2014)	77.25 ± 0.68	70.35 ± 0.65	67.08 ± 0.78	64.05 ± 0.68	63.05 ± 0.80	68.32 ± 0.72	65.05 ± 0.68	66.65 ± 0.70
	ResNet50 He et al. (2016)	78.75 ± 0.65	75.58 ± 0.58	75.15 ± 0.75	68.42 ± 0.62	67.55 ± 0.80	73.08 ± 0.55	72.15 ± 0.52	73.45 ± 0.54
	Inception v3 Szegedy et al. (2016)	78.05 ± 0.60	78.68 ± 0.58	71.05 ± 0.72	70.48 ± 0.65	69.18 ± 0.75	73.45 ± 0.58	71.35 ± 0.55	71.95 ± 0.56
	DeiT Touvron et al. (2021)	81.18 ± 0.58	75.65 ± 0.52	71.75 ± 0.68	75.92 ± 0.60	73.95 ± 0.72	75.72 ± 0.52	73.85 ± 0.50	76.02 ± 0.51
	DenseNet Huang et al. (2017)	81.15 ± 0.50	77.55 ± 0.45	79.15 ± 0.62	74.25 ± 0.55	69.72 ± 0.65	76.35 ± 0.48	74.80 ± 0.46	76.25 ± 0.48
	SENet Hu et al. (2018)	79.65 ± 0.48	84.62 ± 0.42	76.45 ± 0.60	76.62 ± 0.50	71.85 ± 0.62	77.82 ± 0.45	75.95 ± 0.44	77.58 ± 0.45
	ResNeXt Xie et al. (2017)	82.15 ± 0.46	77.30 ± 0.42	80.50 ± 0.58	76.05 ± 0.43	75.25 ± 0.60	78.25 ± 0.46	76.52 ± 0.42	78.75 ± 0.44
	EfficientNet Tan & Le (2019)	86.40 ± 0.43	86.45 ± 0.40	77.52 ± 0.56	75.32 ± 0.46	72.18 ± 0.58	79.55 ± 0.44	77.95 ± 0.42	79.12 ± 0.41
	ViT Dosovitskiy (2020)	85.78 ± 0.44	82.88 ± 0.39	84.28 ± 0.51	77.55 ± 0.47	74.95 ± 0.55	81.05 ± 0.43	79.15 ± 0.40	80.65 ± 0.42
	ConvNeXt Liu et al. (2022)	84.85 ± 0.48	86.32 ± 0.40	79.25 ± 0.65	83.55 ± 0.45	79.80 ± 0.60	82.69 ± 0.42	81.35 ± 0.40	82.25 ± 0.42
	Swin Transformer Liu et al. (2021)	84.18 ± 0.45	86.20 ± 0.35	84.85 ± 0.52	82.08 ± 0.42	80.45 ± 0.55	83.57 ± 0.38	83.18 ± 0.35	83.85 ± 0.36
Hard	AlexNet Krizhevsky et al. (2012)	62.35 ± 1.48	65.52 ± 1.42	60.31 ± 1.60	58.76 ± 1.52	53.77 ± 1.62	60.53 ± 1.58	57.47 ± 1.54	58.43 ± 1.56
	GoogLeNet Szegedy et al. (2015)	62.28 ± 1.42	67.23 ± 1.36	61.64 ± 1.56	62.07 ± 1.48	55.14 ± 1.58	61.75 ± 1.48	58.13 ± 1.42	60.69 ± 1.44
	VGG-16 Simonyan & Zisserman (2014)	71.00 ± 1.22	65.19 ± 1.17	61.50 ± 1.40	61.15 ± 1.22	60.75 ± 1.44	64.03 ± 1.30	60.54 ± 1.22	62.19 ± 1.26
	ResNet50 He et al. (2016)	72.86 ± 1.10	70.65 ± 0.99	69.52 ± 1.28	65.79 ± 1.05	65.67 ± 1.36	68.98 ± 0.94	67.79 ± 0.88	69.17 ± 0.92
	Inception v3 Szegedy et al. (2016)	72.26 ± 1.02	73.61 ± 0.99	65.65 ± 1.22	67.90 ± 1.11	67.21 ± 1.28	69.38 ± 0.99	67.02 ± 0.94	67.79 ± 0.95
	DeiT Touvron et al. (2021)	75.39 ± 0.99	70.95 ± 0.88	66.61 ± 1.16	73.35 ± 1.02	72.23 ± 1.22	71.72 ± 0.88	69.59 ± 0.85	71.84 ± 0.87
	DenseNet Huang et al. (2017)	75.45 ± 0.85	72.75 ± 0.77	73.61 ± 1.05	71.79 ± 0.94	68.02 ± 1.11	72.40 ± 0.82	70.54 ± 0.78	72.18 ± 0.82
	SENet Hu et al. (2018)	74.13 ± 0.82	79.66 ± 0.71	71.11 ± 1.02	74.28 ± 0.85	70.25 ± 1.05	73.94 ± 0.77	71.84 ± 0.75	73.54 ± 0.77
	ResNeXt Xie et al. (2017)	76.55 ± 0.78	72.67 ± 0.71	75.02 ± 0.99	73.66 ± 0.73	73.71 ± 1.02	74.37 ± 0.78	72.32 ± 0.71	74.74 ± 0.75
	EfficientNet Tan & Le (2019)	80.68 ± 0.73	81.52 ± 0.68	72.24 ± 0.95	73.13 ± 0.78	70.68 ± 0.99	75.74 ± 0.75	73.83 ± 0.71	75.16 ± 0.70
	ViT Dosovitskiy (2020)	80.21 ± 0.75	78.23 ± 0.66	78.81 ± 0.87	75.39 ± 0.80	73.56 ± 0.94	77.32 ± 0.73	75.06 ± 0.68	76.82 ± 0.71
	ConvNeXt Liu et al. (2022)	79.45 ± 0.82	81.71 ± 0.68	74.21 ± 1.11	81.52 ± 0.77	78.63 ± 1.02	79.12 ± 0.71	77.34 ± 0.68	78.45 ± 0.71
	Swin Transformer Liu et al. (2021)	78.86 ± 0.77	81.52 ± 0.60	79.44 ± 0.88	79.94 ± 0.71	79.07 ± 0.94	79.80 ± 0.65	79.04 ± 0.60	79.91 ± 0.61
Easy	AlexNet Krizhevsky et al. (2012)	65.52 ± 2.18	66.16 ± 2.10	63.75 ± 2.35	52.82 ± 2.24	45.86 ± 2.42	58.84 ± 2.30	56.43 ± 2.24	57.77 ± 2.27
	GoogLeNet Szegedy et al. (2015)	65.69 ± 2.10	68.12 ± 2.02	64.42 ± 2.30	55.12 ± 2.18	46.48 ± 2.38	60.80 ± 2.18	56.43 ± 2.10	59.30 ± 2.13
	VGG-16 Simonyan & Zisserman (2014)	74.79 ± 1.90	66.11 ± 1.82	64.30 ± 2.18	54.34 ± 1.90	51.15 ± 2.24	63.01 ± 2.02	58.75 ± 1.90	60.77 ± 1.96
	ResNet50 He et al. (2016)	75.31 ± 1.82	68.21 ± 1.62	66.02 ± 2.10	61.71 ± 1.74	57.57 ± 2.24	66.92 ± 1.54	63.98 ± 1.46	66.48 ± 1.51
	Inception v3 Szegedy et al. (2016)	74.10 ± 1.68	69.12 ± 1.62	70.08 ± 2.02	56.41 ± 1.82	53.34 ± 2.10	65.53 ± 1.62	63.45 ± 1.54	65.18 ± 1.57
	DeiT Touvron et al. (2021)	73.30 ± 1.62	71.79 ± 1.46	66.05 ± 1.90	58.04 ± 1.68	54.43 ± 2.02	65.73 ± 1.46	62.57 ± 1.40	63.73 ± 1.43
	DenseNet Huang et al. (2017)	75.06 ± 1.40	69.63 ± 1.26	72.55 ± 1.74	60.16 ± 1.54	54.03 ± 1.82	67.26 ± 1.34	64.56 ± 1.29	66.52 ± 1.34
	SENet Hu et al. (2018)	73.08 ± 1.34	75.45 ± 1.18	69.48 ± 1.68	61.65 ± 1.40	55.27 ± 1.74	68.04 ± 1.26	65.14 ± 1.23	67.13 ± 1.26
	ResNeXt Xie et al. (2017)	75.24 ± 1.29	68.74 ± 1.18	73.06 ± 1.62	60.98 ± 1.20	57.80 ± 1.68	68.25 ± 1.29	65.39 ± 1.18	68.04 ± 1.23
	EfficientNet Tan & Le (2019)	78.63 ± 1.20	76.38 ± 1.12	69.83 ± 1.57	60.07 ± 1.29	55.02 ± 1.62	68.95 ± 1.23	66.22 ± 1.18	67.88 ± 1.15
	ViT Dosovitskiy (2020)	77.48 ± 1.23	72.69 ± 1.09	75.42 ± 1.43	61.36 ± 1.32	56.73 ± 1.54	69.75 ± 1.20	66.71 ± 1.12	68.74 ± 1.18
	ConvNeXt Liu et al. (2022)	75.18 ± 1.34	74.71 ± 1.12	75.01 ± 1.82	64.15 ± 1.26	60.11 ± 1.68	71.21 ± 1.18	69.27 ± 1.12	70.53 ± 1.18
	Swin Transformer Liu et al. (2021)	76.02 ± 1.26	75.15 ± 0.98	70.39 ± 1.46	65.64 ± 1.18	59.99 ± 1.54	70.66 ± 1.06	68.05 ± 0.98	69.50 ± 1.01

4 Experiments

We first establish standardized benchmark results on `MRT-43K` (Section 4.1), compare MRT-AL against representative active learning methods (Section 4.2), analyze data scalability (Section 4.3), and ablate the components of MRT-AL (Section 4.4). A cross-dataset generalization study is provided in the Appendix C.

4.1 Benchmark Results

We evaluate 13 representative deep learning architectures on `MRT-43K` to establish a standardized performance reference for Martian terrain classification. The evaluated models span classical convolutional neural networks (CNNs), including AlexNet Krizhevsky et al. (2012), VGG-16 Simonyan & Zisserman (2014), and GoogLeNet Szegedy et al. (2015); modern CNN variants such as ResNeXt Xie et al. (2017), DenseNet Huang et al. (2017), EfficientNet Tan & Le (2019), and ConvNeXt Liu et al. (2022); and recent vision transformer architectures, including ViT Dosovitskiy (2020), Swin Transformer Liu et al. (2021), and DeiT Touvron et al. (2021). All models are trained using identical settings: Adamw with a learningate of 1×10^{-4}, a batch size of 128, and 150 epochs on a single NVIDIA A6000 GPU.

All models are evaluated on the `MRT-challenge` set (Section 3). To ensure reliability, we conduct three independent runs. The results are reported in Table 3. Swin Transformer achieves the highest Top-1 accuracy of 83.57% with an F1 score of 83.18%, followed closely by ConvNeXt at 82.69%. This narrow margin suggests comparable performance between modern CNN and transformer architectures on Martian terrain classification. Earlier convolutional models such as AlexNet (62.68%) and GoogLeNet (64.15%) perform substantially worse, whereas more recent architectures incorporating residual connections or attention mechanisms achieve markedly stronger results. Notably, no model exceeds 85% accuracy on the `MRT-challenge` set, indicating that reliable perception under complex Martian conditions remains challenging.

To examine the impact of sample difficulty from a data-centric perspective, we partition the 41,756 training images into two equal subsets—`MRT-Easy` and `MRT-Hard`—based on feature richness scores derived from MRT-AL. The 13 baseline models are retrained on each subset and evaluated on the `MRT-challenge` set. As shown in Table 3, models trained on `MRT-Hard` consistently outperform those trained on `MRT-Easy` despite identical training budgets. For example, Swin

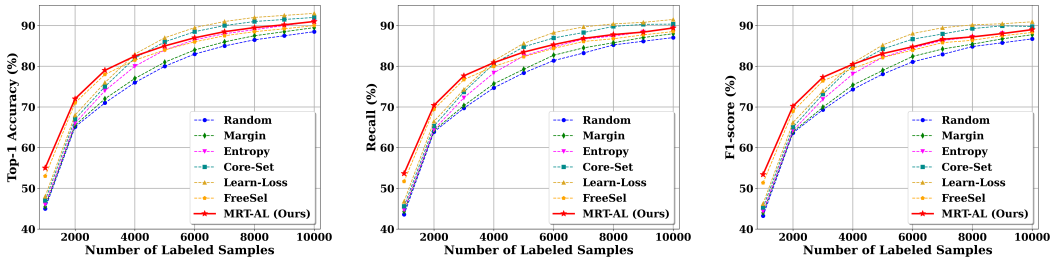


Figure 7: Results (averaged over 3 independent runs) on Mars Terrain Classification: Learn-Loss achieves the highest Top-1 accuracy 92.9%, followedclosely by MRT-AL at 91.7%.

TRANSFORMER achieves a Top-1 accuracy of 79.8% with `MRT-Hard`, compared to 70.66% with `MRT-Easy`. Remarkably, `MRT-Hard` reaches approximately 96% of the full-dataset performance while using only half the training data. In contrast, models trained on `MRT-Easy` remain below 71.21% across architectures. These findings suggest that informative samples play a critical role in generalization to complex Martian terrain, highlighting the importance of data quality alongside scale.

4.2 MRT-AL VS. EXISTING ACTIVE LEARNING METHODS

In this section, MRT-AL is evaluated against several established active learning strategies, including classical iterative batch selection methods Sener & Savarese (2017); Yoo & Kweon (2019); Settles & Craven (2008), FREESEL Xie et al. (2023)—a representative single-pass inference approach—and Random Selection. We adopted SWIN TRANSFORMER as framework. To ensure a fair evaluation, we use the 10,000 ground-truth dataset introduce in Section 2.3 as the test set, while the remaining 33,756 images as the data pool. The results are shown in Figure 7.

Table 4: Efficiency comparison. Time denotes the sample selection time required to select 10,000 images. The iterative method selects 1,000 images in each round and iterates 10 times.

Methods	**Batch Selection**	**Mars Feature**	**Time**
LEARN-LOSS Yoo & Kweon (2019)	✓	X	> 38 *hours*
CORE-SET Sener & Savarese (2017)	✓	X	
ENTROPY Settles & Craven (2008)	✓	X	
MARGIN Settles & Craven (2008)	✓	X	
FREESEL Xie et al. (2023)	X	X	$<$ **300 s**
MRT-AL (OURS)	X	✓	**(450× faster)**

Regarding classification performance, MRT-AL achieves competitive accuracy. While LEARN-LOSS Yoo & Kweon (2019) attains the highest accuracy (92.9%), MRT-AL reaches 91.7%, outperforming the single-pass baseline FREESEL (90.1%).

In terms of computational efficiency, MRT-AL substantially reduces processing time relative to iterative methods that require repeated model retraining, as reported in Table 4.

4.3 DATA SCALABILITY ANALYSIS

The performance of deep learning models is closely related to the scale of training data, with prior studies showing consistent gains as dataset size increases Torralba & Efros (2011); Sun et al. (2017). To examine the impact of data scale in Martian terrain classification, we conduct a data scalability study using `MRT-43K`, with `ZSMars` serving as a small-scale baseline dataset.

We evaluate five benchmark models—SWIN TRANSFORMER Liu et al. (2021), CONVNEXT Liu et al. (2022), INCEPTION V3 Szegedy et al. (2016), RESNET-50 He et al. (2016), and VGG-16 Simonyan & Zisserman (2014)—spanning both convolutional neural networks (CNNs) and vision transformers (ViTs). All evaluations are performed on the `MRT-challenge` set (see Section 3). To examine the effect of data scale, we construct training subsets of varying sizes. For `ZSMars`, the training subset increases from 20% to 100% of the 1,500 sample training split. For `MRT-43K`, we

first match the scale of ZSMars to enable a controlled comparison, and then progressively expand the dataset. As illustrated in Figure 8, two consistent trends emerge:

Data Scale Limitations in Existing Datasets. Due to its relatively small scale, the performance of models trained on ZSMars plateaus at approximately 66.12%, suggesting that further gains are constrained by the available training data.

Continuous Gain from Large-Scale Data: In contrast, the performance of models trained on MRT-43K surpasses that of models trained on ZSMars. As the number of training samples increases, model accuracy improves steadily.

Beyond providing additional samples, MRT-43K offers broader visual coverage, supporting continued performance gains in Martian terrain classification as dataset size increases.

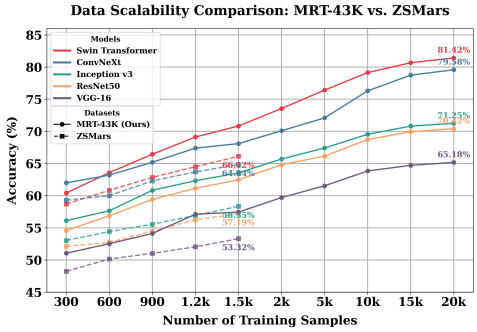


Figure 8: Data scalability results. Solid lines denote the performance curves of models trained on the MRT-43K dataset, while dashed lines represent those trained on the ZSMars dataset.

4.4 MRT-AL Ablations

As reported in Table 5, relying solely on the raw Shannon entropy of Vision Transformer self-attention yields a Top-1 accuracy of 88.93%, only marginally above Random Selection (88.50%). This limited efficacy stems from the long-tailed distribution of ViT attention, where a few dominant patches skew the entropy and mask subtle textural variations.

Introducing Quantile Truncation independently mitigates this by capping extreme attention values. This forces the model to redistribute its focus across broader spatial extents, boosting accuracy to 90.56%. Conversely, applying only the Binning Strategy achieves 89.83% by discretizing continuous weights into intervals. This approach effectively captures the spatial frequency and semantic complexity of different terrain topographies.

Table 5: MRT-AL Ablations.

Configuration	Top-1 Accuracy (%)	F1-score (%)	Recall (%)
w/o Quantile Truncation	89.83%	87.62%	88.05%
w/o Binning Strategy	90.56%	88.15%	88.52%
w/o Both	88.93%	87.10%	87.33%
Random Selection	88.50%	86.74%	87.05%
Full MRT-AL	**91.70%**	**89.00%**	**89.41%**

Ultimately, the complete MRT-AL framework integrates both strategies synergistically to achieve the highest accuracy of 91.70%. Quantile Truncation stabilizes the attention distribution by filtering outlier noise, while the Binning Strategy extracts a highly reliable measure of feature richness from this refined distribution.

5 Conclusion

This paper addresses the limitations of existing Martian terrain datasets and conventional active learning workflows through a pipeline spanning raw data acquisition, data curation, and dataset construction, together with MRT-AL, a single-pass active learning framework tailored to Martian terrain characteristics. Leveraging this pipeline, we construct MRT-43K, a high-resolution RGB benchmark comprising 43,756 images across five traversability-aware terrain categories. Extensive experiments demonstrate that MRT-43K enables consistent performance gains across diverse model architectures and exhibits stronger cross-dataset generalization than existing benchmarks. Meanwhile, MRT-AL achieves competitive performance with established active learning methods while reducing sample selection time by more than 450× compared with conventional iterative pipelines. The MRT-challenge set further reveals substantial robustness gaps in current terrain classification models. Overall, MRT-43K provides a benchmark for advancing data-centric perception and autonomous navigation in future Mars exploration.

AI USE STATEMENT

In this work, we used generative AI tools to assist with reviewing the grammar,formatting, punctuation, and references in this paper. We have not used generative AI tools for any other tasks. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All source code and detailed reproduction steps are available at the anonymous link provided in the abstract. For security reasons, only a portion of the dataset is currently released.

REFERENCES

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9368–9377, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Joseph Carsten, Arturo Rankin, Dave Ferguson, and Anthony Stentz. Global path planning on board the mars exploration rovers. In *2007 IEEE Aerospace Conference*, pp. 1–11, 2007.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Charles Edwards, MJ Amato, JD Baker, NJ Barba, J Balaram, EJ Brandon, P Briggs, BD Davis, JC Day, Anthony Freeman, et al. Emerging capabilities for mars exploration. *Planetary Science Decadal*, 2020.

Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *European conference on computer vision*, pp. 562–577, 2014.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192, 2017.

Ramon Gonzalez and Karl Iagnemma. Deepterramechanics: Terrain classification and slip estimation for ground robots via deep learning. *arXiv preprint arXiv:1806.07379*, 2018.

Robert M Haralick, K Shanmugam, and Its'hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1979.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 ieee conference on computer vision and pattern recognition*, pp. 2372–2379, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, pp. 13–19, 1995.

Jialun Li, Li Zhang, Zhongchen Wu, Zongcheng Ling, Xueqiang Cao, Kaichen Guo, and Fabao Yan. Autonomous martian rock image classification based on transfer deep learning methods. *Earth Science Informatics*, 13(3):951–963, 2020.

Haiqiang Liu, Meibao Yao, Xueming Xiao, and Hutao Cui. A hybrid attention semantic segmentation network for unstructured terrain on mars. *Acta Astronautica*, 204:492–499, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Weikun Lv, Linhui Wei, Dian Zheng, Yu Liu, and Yumei Wang. Marsnet: Automated rock segmentation with transformers for tianwen-1 mission. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2022.

Chaohua Ma, Yuan Li, Junying Lv, Zhouxuan Xiao, Wuming Zhang, and Linshan Mo. Automated rock detection from mars rover image via y-shaped dual-task network with depth-aware spatial attention mechanism. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–18, 2024.

Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 79, 2004.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400, 2019.

Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *European conference on machine learning*, pp. 413–424, 2006.

Brandon Rothrock, Ryan Kennedy, Chris Cunningham, Jeremie Papon, Matthew Heverly, and Masahiro Ono. Spoc: Deep learning-based terrain classification for mars rover missions. In *AIAA space 2016*, pp. 5539. 2016.

SP Schwenzer, M Woods, S Karachalios, N Phan, and L Joudrier. Labelmars: Creating an extremely large martian image dataset through machine learning. In *50th Annual Lunar and Planetary Science Conference*, number 2132, pp. 1970, 2019.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 1070–1079, 2008.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. *Advances in neural information processing systems*, 20, 2007.

Yan Shen, Meng Zou, Zhaolong Dang, Baichao Chen, Yuqiong Li, et al. Current status and prospect of mars rover traversability assessment technology (in chinese). *Optics and Precision Engineering*, 31(5):729–745, 2023.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

R Michael Swan, Deegan Atha, Henry A Leopold, Matthew Gildner, Stephanie Oij, Cindy Chiu, and Masahiro Ono. Ai4mars: A dataset for terrain-aware autonomous driving on mars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1982–1991, 2021.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114, 2019.

Xiaomeng Tan, Bobo Xi, Haitao Xu, Jiaojiao Li, Yunsong Li, Changbin Xue, and Jocelyn Chanussot. A lightweight framework with knowledge distillation for zero-shot mars scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1521–1528, 2011.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Kiri Wagstaff, You Lu, Alice Stanboli, Kevin Grimes, Thamme Gowda, and Jordan Padams. Deep mars: Cnn classification of mars imagery for the pds imaging atlas. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Kiri Wagstaff, Steven Lu, Emily Dunkel, Kevin Grimes, Brandon Zhao, Jesse Cai, Shoshanna B Cole, Gary Doran, Raymond Francis, Jake Lee, et al. Mars image content classification: Three years of nasa deployment and recent advances. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 15204–15213, 2021.

Thorsten Wilhelm, Melina Geis, Jens Püttschneider, Timo Sievernich, Tobias Weber, Kay Wohlfarth, and Christian Wöhler. Domars16k: A diverse dataset for weakly supervised geomorphologic analysis on mars. *Remote Sensing*, 12(23):3981, 2020.

Xueming Xiao, Meibao Yao, Haiqiang Liu, Jiake Wang, Lei Zhang, and Yuegang Fu. A kernel-based multi-featured rock modeling and detection framework for a mars rover. *IEEE Transactions on Neural Networks and Learning Systems*, 34(7):3335–3344, 2021.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Yichen Xie, Mingyu Ding, Masayoshi Tomizuka, and Wei Zhan. Towards free data selection with general-purpose models. *Advances in Neural Information Processing Systems*, 36:1309–1325, 2023.

Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127, 2015.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 93–102, 2019.

Jiahang Zhang, Lilang Lin, Zejia Fan, Wenjing Wang, and Jiaying Liu. s^5 mars: Semi-supervised learning for mars semantic segmentation. *arXiv preprint arXiv:2207.01200*, 2022.

Richard W Zurek and Suzanne E Smrekar. An overview of the mars reconnaissance orbiter (mro) science mission. *Journal of Geophysical Research: Planets*, 112(E5), 2007.

A Related Work

A.1 Mars Terrain Dataset

Mars terrain image datasets serve as the cornerstone for developing intelligent perception algorithms for deep space exploration and autonomous navigation models for Mars rovers. Based on the source of image acquisition, existing public datasets Zhang et al. (2022); Swan et al. (2021); Wilhelm et al. (2020); Xiao et al. (2021); Wagstaff et al. (2018); Ma et al. (2024); Liu et al. (2023); Lv et al. (2022); Tan et al. (2024) can be categorized into two types: orbital remote sensing datasets and Mars Rover In-situ Datasets. Additionally, a small number of Martian simulation datasets exist, such as SimMars6K Ma et al. (2024), which collect images by simulating the Martian surface environment through virtual simulation engines or physical laboratories. Although simulation methods can generate large amounts of labeled data at a low cost, existing simulated environments exhibit a significant "Domain Gap" from the real Martian environment in terms of spectral characteristics, high-frequency texture details, and lighting complexity. These discrepancies make it difficult to reproduce the authentic and diverse geological structures and environmental features of Mars. Therefore, we focus exclusively on authentic Martian imagery acquired by spacecraft.

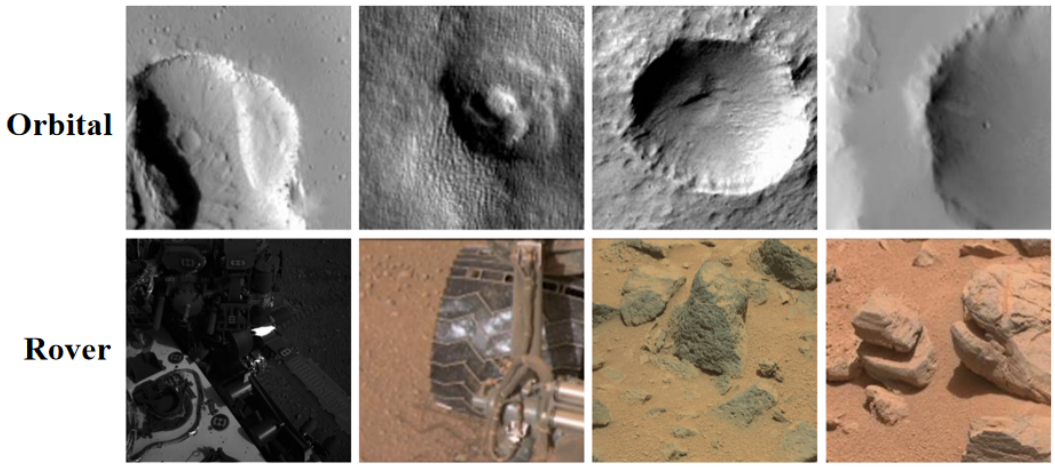


Figure A.1: Orbital: Images are acquired by Mars orbiters, such as MRO. Rover: Images are captured by rovers on the Martian surface.

Orbital Remote Sensing Datasets. Orbital datasets are acquired by high-resolution cameras mounted on Mars orbiters, such as the Mars Reconnaissance Orbiter (MRO) Zurek & Smrekar (2007), Figure A.1 shows the Orbital examples. Representative works include DoMars16k Wilhelm et al. (2020) and the HiRISE series (v1 Wagstaff et al. (2018) and v3 Wagstaff et al. (2021)). Specifically, DoMars16k contains 16,150 grayscale geomorphological images, while the HiRISE series encompasses thousands of grayscale samples. Although these datasets cover expansive regions of the Martian surface, they are inherently limited by the imaging distance. Consequently, their spatial resolution is typically low, for instance, DoMars16k is restricted to 200×200 pixels and HiRISE to 227×227 pixels. These datasets lack the fine-grained textural information required from a ground-level perspective, making them difficult to deploy directly for near-field navigation or obstacle avoidance tasks.

Mars Rover In-situ Datasets. In-situ datasets are captured by rovers such as Curiosity and Perseverance on the Martian surface, Figure A.1 shows the Rover examples, providing the close-range

visual information essential for autonomous driving. AI4MARS stands as the first large-scale dataset designed for training Mars terrain classification models Swan et al. (2021). However, it is composed of grayscale imagery and is heavily contaminated with non-terrain artifacts, such as the sky and rover body components, which impede the model's ability to learn pristine terrain features. While other datasets like S5Mars Zhang et al. (2022), MarsScapes Liu et al. (2023), and TWMARS Lv et al. (2022) provide RGB imagery, their scales are limited to a range of hundreds to thousands of images, failing to meet the diversity requirements of deep learning models. Furthermore, datasets like MSL(v1) Wagstaff et al. (2018) and ZSMars Tan et al. (2024) offer resolutions of only 256×256 and 224×224, respectively, leading to the loss of critical topographical textures.

In contrast, to address the deficiencies of existing datasets, we constructed the MRT-43K dataset. Compared to the largest existing datasets, MRT-43K offers a significantly expanded scale of over 43K images. Through a rigorous multi-stage processing pipeline, it achieves substantial improvements in both image quality and resolution (1290×970), effectively filling the gap for a high-quality, large-scale benchmark in Mars terrain classification.

A.2 Active Learning

Active Learning (AL) aims to maximize model performance while minimizing labeling costs by iteratively selecting the most informative samples for annotation. With the advancement of deep learning, existing AL methods are primarily categorized into three classes: uncertainty-based, diversity-based, and expected model change methods.

Uncertainty-based methods represent the most intuitive and widely utilized strategies. Their core principle is to select samples with which the current model is most "confused". For classification tasks, the simplest approach leverages predicted posterior probabilities, such as selecting samples with the Least Confidence Lewis (1995) or the highest Entropy Settles & Craven (2008); Joshi et al. (2009). Additionally, Margin Sampling Roth & Small (2006) defines uncertainty by measuring the difference between the two most probable classes. Within the context of deep learning, researchers have proposed utilizing Monte Carlo Dropout Gal et al. (2017) to estimate model uncertainty, while others have employed Ensembles of deep networks to measure predictive disagreement Beluch et al. (2018). However, these methods typically incur significant computational overhead, making them difficult to scale to large-scale datasets Yoo & Kweon (2019).

Diversity-based methods aim to select samples that represent the overall distribution of the unlabeled data pool to avoid sampling redundancy. A prominent approach models active learning as a Core-set selection problem Sener & Savarese (2017), covering the feature space by minimizing the geometric distance between selected and unselected samples. Other distribution-matching methods Nguyen & Smeulders (2004); Yang et al. (2015) utilize clustering or discrete optimization to select representative subsets.

Expected model change methods attempt to directly estimate the impact of a sample on model parameters or outputs, such as Expected Gradient Length Settles et al. (2007). Although theoretically more rigorous, they are often computationally expensive and remain challenging to apply in practical deep neural network scenarios Freytag et al. (2014).

While the aforementioned methods have achieved success in specific tasks, they predominantly rely on cumbersome "train-predict-screen" iterative cycles or are tailored for specific tasks like detection or segmentation Yoo & Kweon (2019); Sener & Savarese (2017). In contrast, we propose MRT-AL, a single-pass inference active learning paradigm specifically designed for the Transformer self-attention mechanism and Martian terrain features. It leverages statistical features within the self-attention mechanism to quantify the local texture richness of terrain samples. By eliminating the need for expensive iterative training or multiple forward passes, MRT-AL maintains high screening quality while significantly enhancing the time efficiency of large-scale dataset construction.

B SUPPLEMENTARY MATERIALS

B.1 PER-CLASS DISTRIBUTION OF THE MRT-43K DATASET

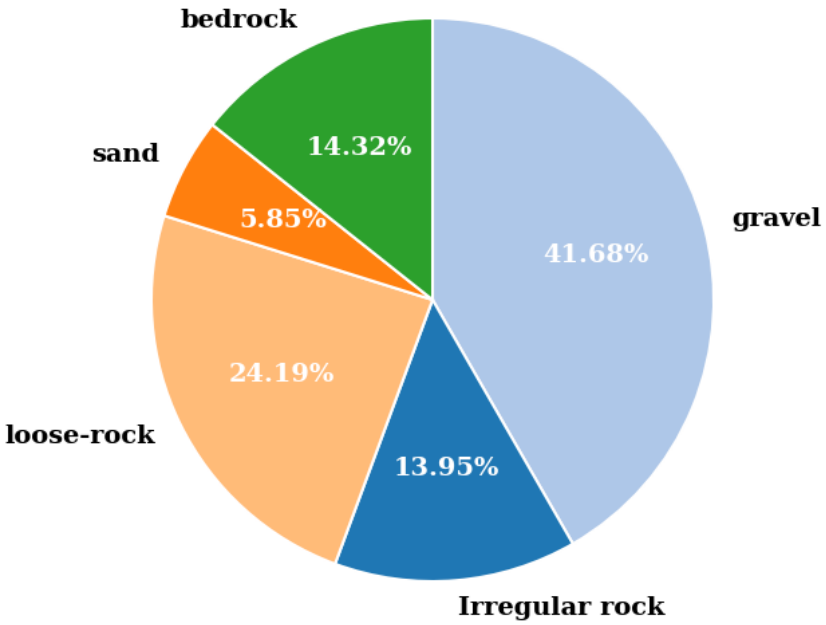


Figure B.1: Proportion of image counts per terrain category in the MRT-43K dataset.

Figure B.1 illustrates the proportional distribution of image counts for each terrain category within the MRT-43K dataset. Notably, as the quantity of available raw images varies across different terrain types, the dataset inherently exhibits a degree of class imbalance. Nevertheless, even the least represented category, "Sand" (5.85%), contains more than 2,500 images—a volume that approaches or even exceeds the total scale of several existing datasets.

B.2 DATASET IMAGE SOURCE DISTRIBUTION

All images in the MRT-43K dataset are sourced from NASA's Curiosity and Perseverance rovers. Specifically, imagery from the Curiosity rover was captured within Gale Crater, whereas data from the Perseverance rover was acquired in Jezero Crater. Table B.1 provides a detailed breakdown of these image sources, including the specific camera instruments and their corresponding Sol ranges. Notably, although Curiosity is equipped with RGB sensors (e.g., Mastcam), its volume of available RGB imagery is substantially limited compared to that of Perseverance. Consequently, the vast majority of the images in the MRT-43K dataset originate from the Perseverance rover.

Table B.1: Dataset Image Source Distribution.

Rover	Camera	Images	Total	Sol
Perseverance	Mastcam-Z	27,738	39,122	9–1360
	Navcam	9,368		
	Hazcam	2,016		
Curiosity	Mastcam-R	2,895	4,634	3,509–4,374
	Mastcam-L	1,739		
Grand Total			**43,756**	

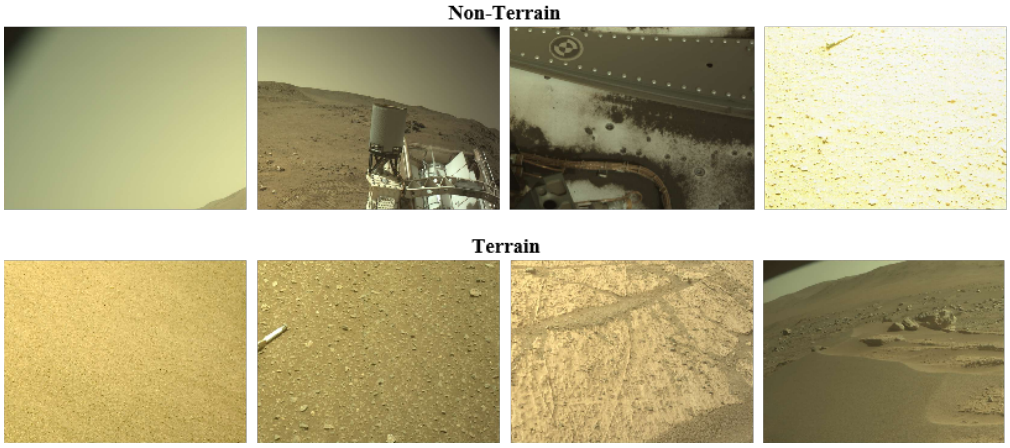


Figure B.2: Terrain Images vs. Non-Terrain Artifacts

B.3 Refining Complex Terrain Definitions

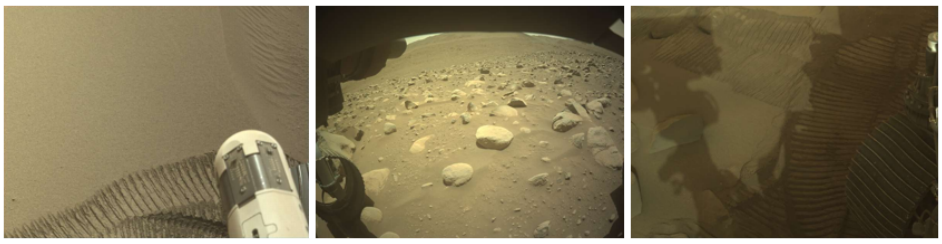


Figure B.3: More complex terrain imagery with diverse occlusion categories and shadow effects.

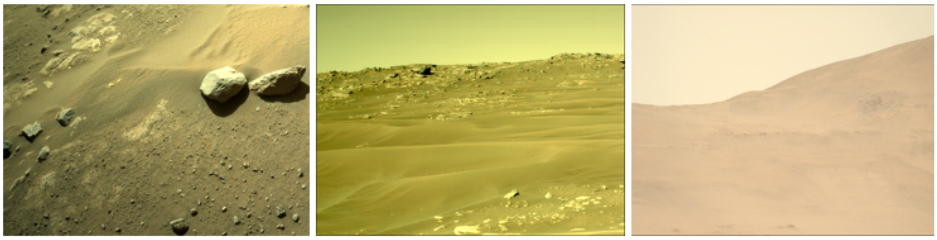
Figure B.2 illustrates representative samples of terrain and non-terrain categories. In addition to the easily distinguishable terrain images shown in Figure B.2, there exist more complex terrain samples that feature diverse types of occlusions and shadows, as illustrated in Figure B.3. Models trained on a limited number of samples do not possess sufficient robustness; therefore, it is necessary to expand the training dataset to encompass a broader spectrum of terrain types. This expansion enables the trained models to filter terrain images more effectively.

However, not all terrain images definitively correspond to a single, distinct terrain class (category definitions are detailed in Table B.3). Although some images technically qualify as terrain, they feature highly heterogeneous mixtures of ground types or suffer from excessive blurriness, which prevents the primary subject from manifesting a specific topographical characteristic.

The samples shown in Figure B.4 exemplify such cases. Consequently, we discard these images to avoid compromising the overall quality and categorical purity of the dataset.

Table B.2: **Qualitative Analysis Result**. Typical samples causing prediction discrepancies in the ZSMars test set. Three categories are presented: non-terrain artifacts, labeling errors, low Resolution and texture Loss. Red text indicates erroneous or disputed predictions/labels; green signifies plausible predictions aligned with geological characteristics.

Index	Image Sample	Original Label (GT)	M_{MRT} Prediction	Root Cause Analysis
0		Sand	Gravel	**Non-terrain Artifacts**
1		Sand	**Bedrock**	**Labeling Errors**
2		Sand	Bedrock	**Low Resolution and Texture Loss**



(a) (b) (c)

Figure B.4: Images exhibiting categorical ambiguity. (a) and (b) present large-scale mixtures of multiple terrain types, such as Sand, Gravel, and Bedrock; (c) lacks discernable topographical details due to an excessive imaging distance.

B.4 MRT-AL

Algorithm B.1 elucidates the underlying principles of MRT-AL, where the Quantile Truncation and Binning Strategy effectively measures the local features of Martian terrain. As illustrated in Figure B.7, Martian terrain imagery contains diverse local features and consists of heterogeneous mixtures of terrain patches. The introduction of Quantile Truncation and Binning Strategy enables a more precise quantification of the spatial distribution patterns attended to by the model, thereby facilitating an accurate assessment of local feature richness.

B.5 Qualitative Analysis Visualization

Table B.2 presents the qualitative visualization results of the experiments conducted in Section C.1. Notably, the presence of deficiencies within the ZSMars test set—including non-terrain artifacts, labeling errors, and low resolution/texture loss—accounts for the relatively low performance of our model on the ”Sand” category.

Algorithm B.1 : MRT-AL Sample Selection Paradigm

Input: Pretrained transformer $\mathcal{M}$, Annotation Budget B, Threshold x, Number of bins K, Unlabeled image pool $\mathcal{U} = \{I_1, \ldots, I_N\}$

Output: Selected sample subset $\mathcal{S}$

for each image $I_i \in \mathcal{U}$ **do**
  $\mathbf{A} = \{a_1, \ldots, a_n\} = \mathcal{M}(I_i)$
  / from the last $[CLS]$ token; */*
  Sort $\mathbf{A}$ in the decreasing, $T = Q(\mathbf{A}, x)$;
  / Calculate threshold value */*
  $\hat{a}_j = \min(a_j, T), \quad \forall a_j \in \mathbf{A}$;
  / Extreme value truncation */*
  Re-normalization $\tilde{a}_j = \hat{a}_j / \sum_{k=1}^{n} \hat{a}_k$;
  Map $\{\tilde{a}_j\}$ to K intervals and count frequency vector $\mathbf{P} = \{P_1, \ldots, P_K\}$;
  / Quantile Binning */*
  Entropy $H_i = -\sum_{k=1}^{K} P_k \log P_k$;
end for
Sort the image pool $\mathcal{U}$ in descending order based on the entropy scalar H_i;
Select the top B samples to construct the subset: $\mathcal{S} \leftarrow \{I_{(1)}, \ldots, I_{(B)}\}$;
return $\mathcal{S}$

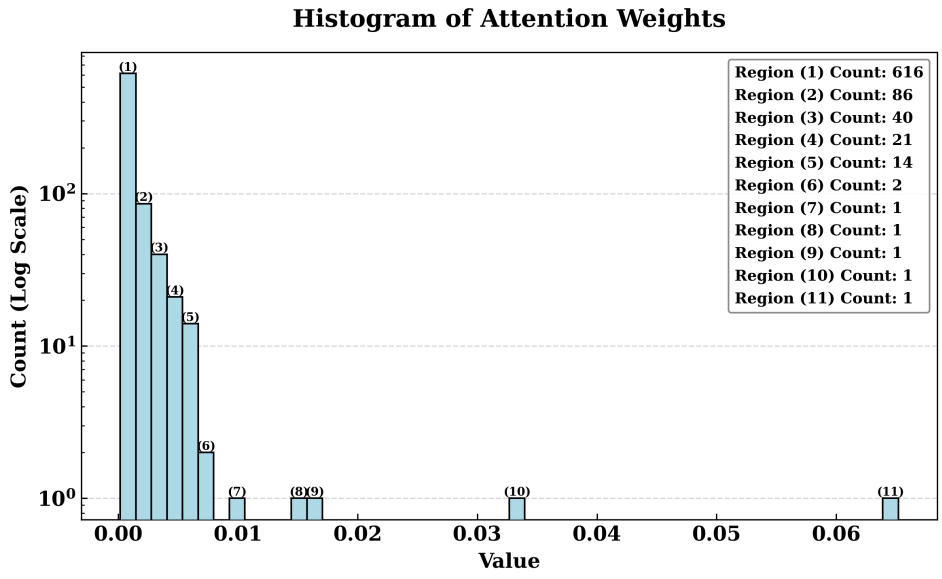


Figure B.5: Histogram of self-attention maps.

B.6 T-SNE FEATURE SPACE VISUALIZATION

To further assess the class separability and representation quality of MRT-43K, we visualize the learned feature space using t-distributed Stochastic Neighbor Embedding (t-SNE). As a widely adopted nonlinear dimensionality reduction technique, t-SNE preserves local neighborhood relationships from high-dimensional feature spaces and provides an intuitive means of revealing latent cluster structures among samples Van der Maaten & Hinton (2008). Consequently, it is frequently used to evaluate the discriminative quality of learned representations.

Specifically, we extract penultimate-layer features from a ConvNeXt model trained on MRT-43K and project them into a two-dimensional space using t-SNE. The resulting visualization is shown in Figure B.6.

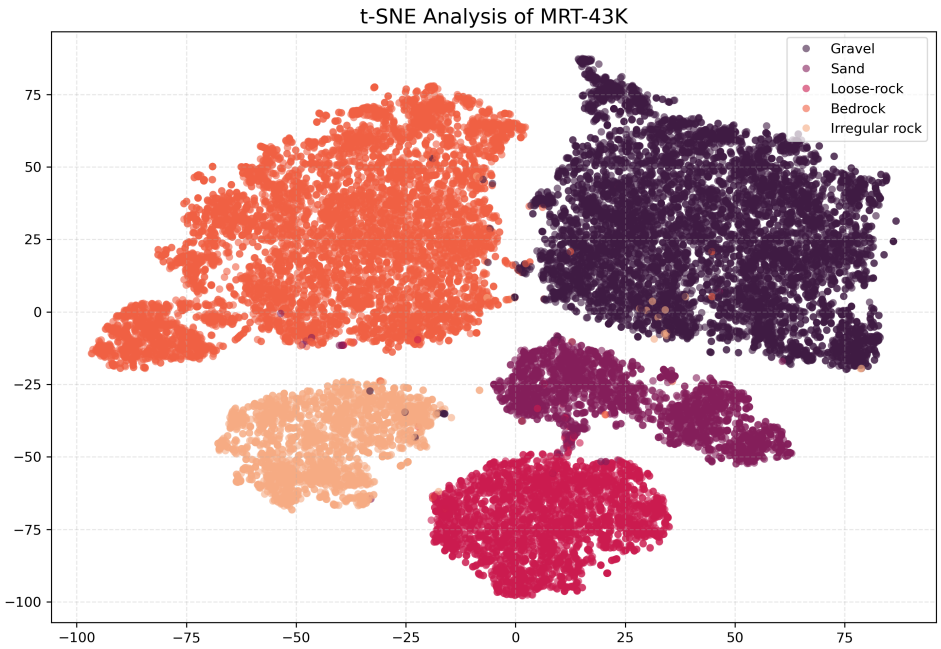


Figure B.6: t-SNE Analysis of MRT-43K.

As shown in Figure B.6, the five terrain categories form well-defined clusters with clear semantic separation. Bedrock and Gravel occupy distinct regions of the feature space, while Sand, Loose-rock, and Irregular rock also exhibit compact intra-class distributions. This observation suggests that the model learns discriminative texture and morphological representations that effectively capture terrain-specific characteristics. Such behavior is consistent with prior studies showing that high-level visual representations tend to organize semantically similar samples into compact clusters in the embedding space Caron et al. (2021).

Table B.3: Definitions and Traversability Levels of the Five Terrain Categories in MRT-43K.

Index	Category	Description	Traversability
0	Bedrock	Massive or platy flat rock with a solid and stable structure	High
1	Gravel	Fragments mixed with sand or soil; level surface with a relatively firm structure	High
2	Sand	Loose sandy terrain prone to causing wheel sinkage and slippage	Medium
3	Loose-rock	Surface scattered with discrete large rocks, posing puncture risks and requiring detours	Medium
4	Irregular rock	Rugged terrain densely populated with irregular large boulders, presenting significant obstacles to rover navigation	Low

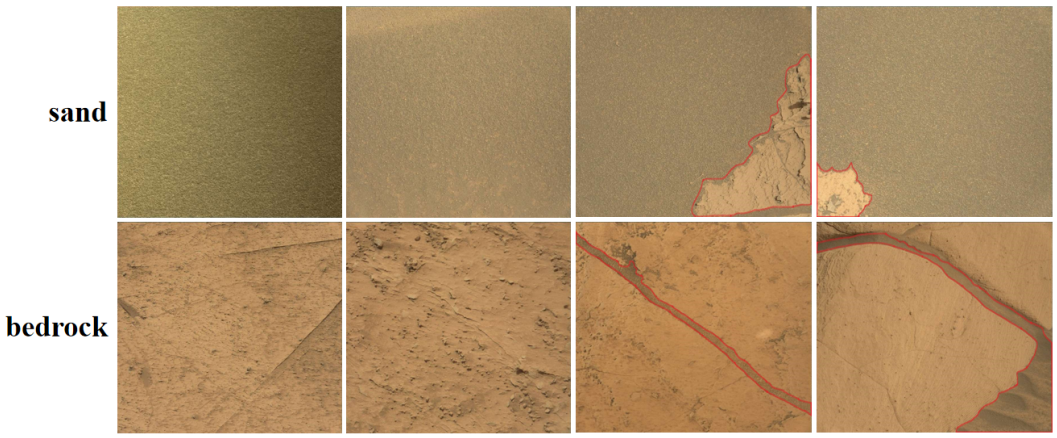


Figure B.7: For a given terrain category, certain images exhibit high visual similarity. The left two columns display highly similar samples within the same class, while the right two columns illustrate images from the same category that possess distinct local variations (demarcated by red boundaries).

Table B.4: GLCM features comparison between MRT-43K and ZSMars datasets.

Dataset	Category	Contrast	Energy	Homogeneity	Correlation	Dissimilarity
MRT-43K	bedrock	123.22	0.0285	0.2125	0.9361	6.38
MRT-43K	gravel	149.10	0.0290	0.1905	0.9214	7.20
MRT-43K	sand	119.02	0.0447	0.2996	0.9382	5.12
ZSMars	1 Sands	57.33	0.0696	0.3783	0.8535	3.88
ZSMars	2 Gravel	193.02	0.0365	0.1796	0.6929	8.79
ZSMars	3 Light-Toned Veins	135.76	0.0357	0.2168	0.8535	7.09

A small degree of overlap is observed between Gravel and Irregular rock, as well as between Sand and Loose-rock. This is expected because these terrain types share similar visual properties in real Martian environments, including fragmented rock structures and granular surface patterns. Nevertheless, the overall cluster boundaries remain largely distinguishable, indicating strong inter-class separability.

Overall, the t-SNE visualization demonstrates high intra-class consistency and clear inter-class separation within MRT-43K. These findings provide additional evidence for the quality and label consistency of the dataset, and further support the classification and cross-dataset generalization results reported in the main paper.

B.7 GLCM

To complement the t-SNE feature space visualization, we conduct a quantitative texture analysis using the Gray-Level Co-occurrence Matrix (GLCM)Haralick et al. (1979), a well-established statistical method for characterizing spatial dependencies between pixel intensities. The GLCM captures second-order texture statistics that reflect the spatial arrangement and variability of gray-level values within an image region. Five standard texture features—Contrast, Energy, Homogeneity, Correlation, and Dissimilarity—are computed for representative terrain categories in both `MRT-43K` and `ZSMars` to provide complementary evidence for the observed class separability and to characterize the texture properties distinguishing each terrain type. The results are presented in Table B.4.

Among the three terrain categories evaluated in `MRT-43K`, Gravel exhibits the highest Contrast (149.10) and Dissimilarity (7.20) alongside the second lowest Energy (0.0290) and Homogeneity (0.1905), reflecting its inherently heterogeneous surface composition characterized by a diverse mixture of pebbles, cobbles, and fragmented rock fragments of varying sizes and reflectance. In contrast, Sand displays the lowest Contrast (119.02) and Dissimilarity (5.12) with the highest Energy

(0.0447) and Homogeneity (0.2996), consistent with its smooth, fine-grained surface that exhibits limited local intensity variation. Bedrock occupies an intermediate position with moderate Contrast (123.22) and the lowest Energy (0.0285), capturing the co-occurrence of exposed rock faces, fracture patterns, and localized shadowing effects characteristic of coherent outcrop surfaces.

A notable distinction between the two datasets emerges in the Correlation metric. All MRT−43K categories maintain consistently high Correlation values (0.92–0.94), indicating well-preserved linear dependencies among neighboring pixel intensities—a property attributable to the high-resolution (1290×970) RGB imagery with clearly defined texture patterns. In contrast, ZSMars exhibits substantially lower and more variable Correlation (0.69–0.85), with Gravel showing the lowest value (0.6929), suggesting degraded texture coherence that is likely related to its lower spatial resolution (224×224) or compression artifacts. The Energy and Homogeneity metrics further reveal that MRT−43K captures greater within-class texture variation: for example, the Energy of MRT−43K Gravel (0.0290) is lower than that of ZSMars Gravel (0.0365), indicating a broader range of texture configurations within the same semantic category.

Collectively, these GLCM-based texture profiles provide quantitative evidence that the three terrain categories in MRT−43K exhibit distinctive and well-differentiated textural characteristics, consistent with the class separability observed in the t-SNE visualization. The texture statistics further corroborate the cross-dataset generalization gaps reported in Section C.1, as the divergent texture profiles between MRT−43K and ZSMars partly account for the limited transferability of deep features trained on one dataset when applied to the other.

C Additional Experiments

C.1 Cross-Dataset Generalization

Computer vision datasets serve as proxies for real-world environments; accordingly, Martian terrain datasets should capture the visual characteristics of the Martian surface. Broad coverage and diversity are essential for training models with strong generalization capability. However, cross-dataset generalization remains underexplored in existing dataset research Torralba & Efros (2011).

To assess the representativeness of MRT−43K, we adopt a cross-dataset validation protocol to examine model transferability Torralba & Efros (2011); Recht et al. (2019). Martian terrains present distinctive visual challenges—including irregular contours, occlusions, and overlapping geological structures—which complicate the acquisition of large-scale annotated imagery relative to natural Earth scenes Zhang et al. (2022). Consequently, prior research has primarily focused on small-scale terrain segmentation Lv et al. (2022) or orbital geomorphology classification Wilhelm et al. (2020). Among publicly available datasets, ZSMars Tan et al. (2024) provides rover-perspective terrain annotations that align with our task setting. Accordingly, cross-dataset evaluation is conducted using ZSMars.

Table C.1: Experimental results of the cross-dataset validation with the ZSMars dataset.

Trainset	Testset	Model	Sand	Gravel	Bedrock	Average accuracy
ZSMars	MRT−43K	M_{ZSMars}	56	65	24	48.33
MRT−43K	ZSMars	M_{MRT}	38	88	70	65.33

Given the discrepancies in category definitions between the two datasets, we first perform class alignment. We identify and match three primary terrain categories with equivalent definitions: "Sands," "Gravel," and "Light-Toned Veins" in ZSMars (totaling $1,881$ images) were aligned with "sand," "gravel," and "bedrock" in our MRT−43K dataset, respectively. Based on the aligned categories, we train two architecturally identical ConvNeXt models—denoted as M_{ZSMars} and M_{MRT}—using equal numbers of samples with identical hyperparameters to maintain experimental parity. The remaining ZSMars images, together with a size-matched random subset from MRT−43K, serve as the test sets. We report the results in Table C.1.

Generalization Analysis. As reported in Table C.1, a significant performance gap is observed between the two models. The model trained on MRT−43K, M_{MRT}, achieves a mean accuracy of 65.33% on the ZSMars test set, compared to 48.33% for M_{ZSMars} trained on ZSMars. This result suggests stronger cross-dataset transferability for M_{MRT}, indicating that MRT−43K captures

informative terrain characteristics that support improved generalization. This behavior is likely related to the broader visual coverage of MRT−43K, which introduces greater terrain variability during training.

Label Noise and Data Quality Analysis. To better understand the relatively lower generalization of M_{MRT} on the ZSMars "Sands" category (38%), we conduct a qualitative examination of samples misclassified by M_{MRT}. Representative examples are presented in Table B.2. The observation suggests the presence of label ambiguity and potential data quality issues within the ZSMars test set. In several cases, the predictions produced by M_{MRT} differ from the provided annotations, indicating possible inconsistencies in the ground truth. We group these observations into three categories: **Non-terrain Artifacts**: As illustrated by Index 0 in Table B.2, the image appears to capture dust accumulation from rover drilling activities rather than natural terrain. **Label Ambiguity**: The sample at Index 1 exhibits a stratified rock structure that may align more closely with Bedrock under both ZSMars Tan et al. (2024) and MRT−43K criteria, suggesting potential uncertainty in category assignment. **Low Resolution and Texture Degradation**: Samples such as Index 2 are heavily blurred, limiting the availability of discriminative visual cues for reliable categorization.

Overall, these observations highlight the challenges posed by label noise and image quality in Martian terrain datasets. The broader visual coverage and curated annotation process of MRT−43K contribute to more reliable model training and evaluation.