Mars-Bench: A Benchmark for Evaluating Foundation Models for Mars Science Tasks

Mirali Purohit[1,3] ✉ **Bimal Gajera**[1*] **Vatsal Malaviya**[1*] **Irish Mehta**[1*] **Kunal Kasodekar**[1]
Jacob Adler[2] **Steven Lu**[3] **Umaa Rebbapragada**[3] **Hannah Kerner**[1]

[1]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA
[2]School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA
[3]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

Abstract

Foundation models have enabled rapid progress across many specialized domains by leveraging large-scale pre-training on unlabeled data, demonstrating strong generalization to a variety of downstream tasks. While such models have gained significant attention in fields like Earth Observation, their application to Mars science remains limited. A key enabler of progress in other domains has been the availability of standardized benchmarks that support systematic evaluation. In contrast, Mars science lacks such benchmarks and standardized evaluation frameworks, which have limited progress toward developing foundation models for Martian tasks. To address this gap, we introduce Mars-Bench, the first benchmark designed to systematically evaluate models across a broad range of Mars-related tasks using both orbital and surface imagery. Mars-Bench comprises 20 datasets spanning classification, segmentation, and object detection, focused on key geologic features such as craters, cones, boulders, and frost. We provide standardized, ready-to-use datasets and baseline evaluations using models pre-trained on natural images, Earth satellite data, and state-of-the-art vision-language models. Results from all analyses suggest that Mars-specific foundation models may offer advantages over general-domain counterparts, motivating further exploration of domain-adapted pre-training. Mars-Bench aims to establish a standardized foundation for developing and comparing machine learning models for Mars science.

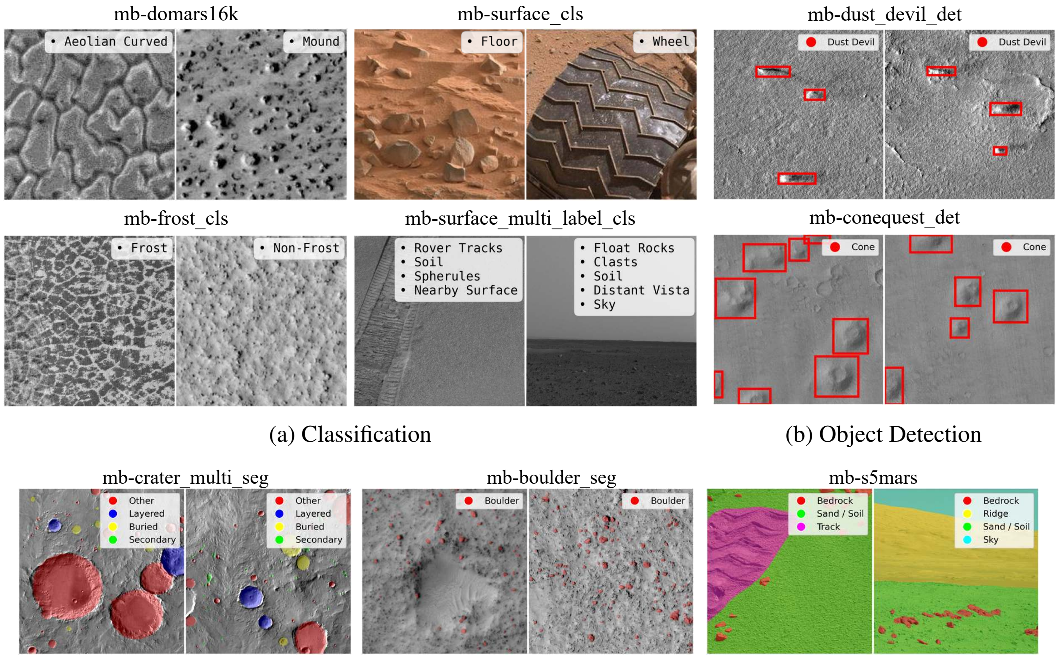
1 Introduction

Over the past few years, foundation models have revolutionized specialized domains such as medical imaging [42, 46], Earth Observation (EO) [30, 54, 2], law [9, 10], and astronomy [34, 44, 58]. These models, pre-trained on large and diverse datasets, offer strong generalization capabilities and enable efficient fine-tuning on downstream tasks with minimal data. The EO community has embraced foundation models in the last 3-4 years, with an explosion of methods, datasets, and benchmarks aimed at improving performance across a wide range of geospatial tasks.

The key driver of progress in these domains has been the development of high-quality, standardized benchmarks. For example, BigBio [18] and MIMIC-IV [28] have accelerated model advancements by providing consistent evaluation protocols for medical applications. Benchmarks like Geo-Bench [32] and PANGAEA [41] have accelerated progress in EO applications by providing a suite of standardized classification and segmentation tasks for evaluating geospatial foundation models. Geo-

✉Corresponding Author: mpurohi3@asu.edu
*Equal Contribution

Preprint. Under review.



(a) Classification

(b) Object Detection

(c) Segmentation

Figure 1: Representative samples from selected Mars-Bench datasets, from all three task categories.

Bench enables model developers to assess generalization across diverse data sources and use cases, creating a pathway for systematic progress.

However, no such benchmark exists for Martian applications. Machine learning research for Mars science applications thus lags behind other science domains [3]. Although recent studies have presented machine learning solutions for a range of Martian applications, including crater detection [40, 77, 11], landmark classification [71, 65], and cone segmentation [48, 74], these solutions and datasets lack standardization and interoperability. This results in task-specific models or datasets that cannot be easily evaluated as downstream tasks for foundation models or other machine learning advances. This results in limited evaluation of proposed Mars foundation model approaches on 1-2 downstream tasks, limiting the ability to assess model generalization or robustness [63, 70, 68, 20, 50].

This gap is particularly surprising given the richness of available Mars data. Orbiters such as the Mars Reconnaissance Orbiter (MRO) [78] and Mars Odyssey have captured millions of images over the last 20-25 years, while surface rovers like Curiosity and Perseverance have amassed petabytes of high-resolution images. These datasets offer immense potential to study critical questions of planetary science, such as the past presence of water on Mars and the planet's habitability. Yet, the full value of these datasets remains untapped by the ML community due to their lack of standardization, incomplete documentation, and inconsistent formatting for ML workflows.

We introduce Mars-Bench, the first comprehensive benchmark designed to systematically evaluate machine learning models across a diverse set of Mars-related tasks using both orbital and surface imagery. To create this benchmark, we curated and revamped existing datasets, performing quality checks and corrections where necessary and standardizing them in a unified, ML-ready format. The goal of Mars-Bench is to provide a common framework to assess and compare the performance of foundation models on Martian data, facilitating reproducibility and accelerating scientific discovery in planetary science. Our key contributions are as follows:

- **Diverse task coverage:** Mars-Bench includes 20 datasets, summarized in Table 1, spanning three task types: classification, segmentation, and object detection. We also provide a few-shot and partitioned versions of each dataset for evaluation under varying training sample sizes.
- **Scientific relevance:** Mars-Bench covers a wide range of geologic features commonly studied in Mars science, including craters, cones, boulders, landslides, dust devils, atmospheric dust, etc.

These tasks reflect real scientific use cases relevant to planetary scientists and geologists, who co-developed the Mars-Bench. Samples from few Mars-Bench datasets are shown in Figure 1.

- **Comprehensive evaluation:** Since no standardized pre-trained model exists for Mars data, we benchmarked performance using ImageNet-pretrained models under different training settings. We analyzed model behavior with different training set sizes. We also evaluated Mars-Bench using pre-trained EO models as well as proprietary vision-language models, including Gemini and GPT.
- **Code, reproducibility, and baseline models:** We release full code support for all experiments in this paper, along with tools for dataset handling and results visualization. To facilitate community adoption and reproducibility, we also provide well-documented guidelines and publicly release all baseline models evaluated on Mars-Bench. These models can serve as strong starting points for future applications; for example, generating initial global maps of specific geologic features (e.g., cones), which experts can later refine with minimal annotation effort.

2 Related Work

Over the past decade, evaluation benchmarks have played a fundamental role in identifying the limitations of existing foundation models, steering their progress in natural language processing (NLP) and computer vision (CV). For instance, general-purpose natural language understanding (NLU) benchmarks [67, 69, 59] have facilitated the development of large language models (LLMs) such as GPT [5], LLaMA [62], and Gemini [61]. Even in specialized domains, including medical [46, 18, 28], legal [17, 21], scientific discovery [39, 7], security [4], and finance [26], various benchmarks have driven progress in building domain-specific foundation models. Thus, development of quality evaluation benchmarks is necessary for building better foundation models.

In the remote sensing domain, Geo-Bench [32] has defined standardized evaluation protocols for a broad set of EO tasks and has quickly become a de facto benchmark. Since its release, Geo-Bench has been used to evaluate most foundation models proposed for EO over the past two years, enabling consistent comparisons across models. Other notable efforts include SustainBench [75], which targets seven sustainable development goals, AiTLAS [12], which aggregates 22 EO datasets focused solely on classification tasks, and PANGAEA [41], which includes 11 evaluation datasets covering diverse satellite sensors.

Despite substantial progress in other domains toward foundation models and dataset benchmarks, no benchmark currently exists for Mars science applications. The absence of a standardized evaluation framework has hindered the development of foundation models (and machine learning solutions more generally) for Mars-related tasks. While specialized datasets exist across different applications, most require significant effort to restructure into an ML-ready format or make interoperable with other datasets. Furthermore, some datasets are not usable without expert guidance from planetary scientists, further slowing progress. To address this gap, we introduce **Mars-Bench**, the first benchmark to facilitate the development and evaluation of foundation models for Mars science tasks.

3 Mars-Bench

Mars-Bench was created by curating, organizing, restructuring, and correcting existing Mars science datasets following the design principles explained in Section 3.1. While creating each dataset, our goal was to ensure accessibility and usability and provide task diversity as described in Section 3.2.

3.1 Design Principles

Ease of Use A key goal was to create an accessible and user-friendly ready-to-use benchmark, supported by standardized data-loading code. We focused on unifying the data format across all tasks to reduce the engineering effort for researchers and practitioners using the dataset. We provide all possible formats in each task if there are multiple common formats. For example, different object detection models may require COCO, Pascal VOC, or YOLO format, so we provide annotations in all three formats to ensure it is easily usable in all cases and reduce time for conversion from one format to another.

Expert-Validated Corrections Given the domain-specific nature of Mars science, ensuring high data quality is critical. We conducted expert-driven quality analysis and corrections wherever necessary.

Classification

Name	Observation Source	Geologic Feature	Image Size	# Classes	Train	Val	Test	# Bands	Sensor/ Instrument	Published Year	Cite
mb-atmospheric_dust_cls_edr	MRO (O)	Atmospheric dust	100×100	2	9817	4969	5214	1	HiRISE	2019	[13]
mb-atmospheric_dust_cls_rdr	MRO (O)	Atmospheric dust	100×100	2	9817	4969	5214	1	HiRISE	2019	[13]
mb-change_cls_ctx	MRO (O)	Surface change	150×150	2	36	10	10	1	CTX	2019	[29]
mb-change_cls_hirise	MRO (O)	Surface change	100×100	2	3103	670	670	1	HiRISE	2019	[29]
mb-domars16k	MRO (O)	Landmark	200×200	15	11305	3231	1614	1	CTX	2020	[71]
mb-frost_cls	MRO (O)	Frost	299×299	2	30124	11415	12249	1	HiRISE	2024	[14]
mb-landmark_cls	MRO (O)	Landmark	227×227	8	6997	2025	1793	1	HiRISE	2021	[65]
mb-surface_cls	Curiosity (R)	Surface	256×256	36	6580	1293	1594	3	Mastcam, MAHLI	2018, 2021	[65, 66]
mb-surface_multi_label_cls	Opportunity, Spirit (R)	Surface	1024×1024	25	1762	443	739	1	Pancam	2020	[8]

Segmentation

Name	Observation Source	Geologic Feature	Image Size	# Classes	Train	Val	Test	# Bands	Sensor/ Instrument	Published Year	Cite
mb-boulder_seg	MRO (O)	Boulder	500×500	2	39	6	4	1	HiRISE	2023	[47]
mb-conequest_seg	MRO (O)	Cone	512×512	2	2236	319	643	1	CTX	2024	[48]
mb-crater_binary_seg	Mars Odyssey (O)	Crater	512×512	2	3600	900	900	1	THEMIS	2012	[56]
mb-crater_multi_seg	Mars Odyssey (O)	Crater	512×512	5	3600	900	900	1	THEMIS	2021	[33]
mb-mars_seg_mer	Opportunity, Spirit (R)	Terrain	1024×1024	7	744	106	214	1	Navcam, Pancam	2022	[35]
mb-mars_seg_msl	Curiosity (R)	Terrain	500×560	7	2893	413	828	3	Mastcam	2022	[35]
mb-mmls	MRO (O)	Landslide	128×128	2	275	31	256	7	CTX	2024	[45]
mb-s5mars	Curiosity (R)	Terrain	1200×1200	10	4997	200	800	3	Mastcam	2022	[76]

Object Detection

Name	Observation Source	Geologic Feature	Image Size	# Classes	Train	Val	Test	# Bands	Sensor/ Instrument	Published Year	Cite
mb-boulder_det	MRO (O)	Boulder	500×500	1	39	6	4	1	HiRISE	2023	[47]
mb-conequest_det	MRO (O)	Cone	512×512	1	1158	167	333	1	CTX	2024	[48]
mb-dust_devil_det	MRO (O)	Dust devil	$\sim 750 \times 750$	1	1404	201	402	1	CTX	2024	[22]

Table 1: Overview of Mars-Bench datasets across all three task categories. To distinguish the benchmarked versions from their original sources, all dataset names are prefixed with "mb-", which indicates Mars-Bench. Observation sources are labeled as O (Orbiter) and R (Rover).

All segmentation datasets underwent validation by domain experts, and several classification datasets were reviewed and revised through direct correspondence with the original dataset authors. Details on which datasets were corrected or modified are provided in the Appendix.

Dataset Splits All datasets in Mars-Bench include standardized train, validation, and test splits to facilitate consistent and reproducible evaluation. For datasets that did not originally include predefined splits, we generated them following standard practices. When original splits were available, we preserved them to maintain alignment with prior work. These splits ensure that future methods can be compared fairly and under consistent evaluation settings.

Cross-Domain Dataset Partitioning In some cases, we partition datasets based on attributes such as sensor type, data modality, task category, or mission origin. This design choice allows users to analyze model performance across domain shifts, e.g., evaluating cross-sensor or cross-mission generalization by isolating specific factors. Rather than aggregating data into a single dataset, separating them enables experiments in which scientists are often interested, such as how a model trained on one sensor performs on data from another. A more detailed discussion of these partitioning strategies is provided in the Appendix.

Permissive License All datasets included in Mars-Bench have permissive licenses allowing their re-use in the benchmark. We release the Mars-Bench version of all datasets with a Creative Commons Attribution 4.0 (CC BY 4.0) license, permitting open access and use.

3.2 Tasks and Datasets

Mars-Bench offers a diverse collection of 20 datasets spanning three task categories: classification, segmentation, and object detection. Within these categories, the benchmark supports several subtasks, i.e., classification includes binary, multi-class, and multi-label settings, while segmentation includes both binary and multi-class settings. These tasks are constructed from two primary sources of observation: orbiters (satellites) and surface rovers. In total, the benchmark integrates data from 2 Mars orbiters, 3 rovers, and 6 distinct imaging sensors.

The benchmark covers a wide range of scientifically relevant geologic features that are of high interest to the planetary science community and have been extensively studied in prior literature. Mars-Bench was co-developed with expert planetary scientists to ensure its relevance to Mars science. The datasets include geologic features such as boulders, cones, craters, landslides, dust devils, frost, and atmospheric dust. Additionally, multi-class datasets have diverse classes, such as terrain-related classes (e.g., soil, sand, rock, bedrock), landmark-specific features (e.g., Swiss cheese terrain, spiders, dark dunes), and surface-related elements (e.g., ground, ridges, rover tracks), as well as rover components (e.g., inlet, dust removal tool, scoop). This diversity highlights the breadth of Mars-Bench in terms of task design, sensor modalities, and variety in geologic features.

Unlike EO datasets in which many classes, such as airports or farmland, can be annotated at scale via crowd-sourcing, Mars science datasets often require annotation by domain experts in planetary science or geology. This process is highly specialized and time-consuming, sometimes taking months to years for high-quality labeling. As a result, as shown in Table 1, several datasets in Mars-Bench are relatively small in size. By including these small-data tasks, Mars-Bench provides a valuable testbed for research on label-limited scenarios.

3.3 Using the Dataset

Availability All datasets included in Mars-Bench will be publicly released through Hugging Face Datasets[1]. Each dataset follows a standardized schema and is accompanied by metadata, documentation, and loading scripts to enable easy integration into ML pipelines.

Target Audience Mars-Bench offers a diverse set of benchmarks designed to evaluate and compare the performance of foundation models for Mars-related tasks. It serves researchers developing models for planetary applications as well as those interested in the geologic features and data types represented in Mars-Bench. Mars-Bench is also designed to support the broader computer vision and machine learning communities. Researchers studying distribution shift, generalization, or domain adaptation can benefit from its coverage of underrepresented, real-world geospatial scenarios; similar in spirit to WILDS [31]. By offering datasets with unique imaging conditions and semantics, Mars-Bench enables research beyond planetary science.

Baseline Models In addition to datasets and code, we release baseline models for each dataset included in Mars-Bench. We will release the models that currently achieve the best performance on their respective datasets. By making these models publicly available, we aim to lower the barrier for applied research. For example, researchers seeking to generate global maps of features such as cones or craters can use our pre-trained models to produce initial predictions, which can then be refined by domain experts with minimal annotation effort.

Software Tools To promote reproducibility and facilitate future research, we release an open-source toolkit that encapsulates the complete Mars-Bench experimental pipeline [2]. The repository includes configuration files and executable scripts that reproduce every experiment reported in this study, while permitting users to vary model architectures, hyperparameters, and data partitions with minimal effort. In addition, the toolkit provides utilities for loading datasets, and visualizing both objective metrics and qualitative results at the task level as well as in aggregate.

4 Experiments

Model Selection For each task category, we select well-established and widely adopted model architectures representative of current best practices. For classification tasks, we evaluate ResNet101 [23], SqueezeNet1.1 [24], InceptionV3 [60], Swin Transformer (SwinV2-B) [38], and Vision Transformer (ViT-L/16) [15] architectures. For segmentation, we use U-Net [57], DeepLabV3+ [6], SegFormer [73], and Dense Prediction Transformer (DPT) [52]architectures. For object detection, we evaluate YOLO11 [53], SSD [37], RetinaNet [36], and Faster R-CNN [55].

Training Settings We analyze model performance under three different training strategies: (1) training from scratch with randomly initialized weights, (2) using a pre-trained model as a frozen feature extractor, and (3) full fine-tuning of pre-trained models with all weights trainable. As noted in Section 1, no existing foundation model has been trained specifically for Mars tasks. Therefore, we use models pre-trained on large-scale datasets such as ImageNet (for classification and segmentation) or COCO (for detection) as initialization for transfer learning or feature extraction.

Hyperparameter Tuning Since the performance of deep learning models is often sensitive to hyperparameter choices, we conducted a grid search over several hyperparameter configurations for each model, task, and training type combination. The best-performing setting was selected based on

[1] `https://huggingface.co/collections/Mirali33/mars-bench-68266f81a27313eddaa539f1`
[2] `https://github.com/kerner-lab/MarsBench/`

early stopping criteria applied to validation metrics. All hyperparameter ranges and selected values for each configuration are detailed in the Appendix to ensure reproducibility.

4.1 Reporting Results

We adopt an identical methodology to [1] and [32] to present our results derived from thousands of experiments. Our objective is to report both task-specific outcomes and aggregated results across all tasks with reliable confidence intervals as recommended by [1]. Specifically, for each combination of model, dataset, and training strategy, we first conduct hyperparameter tuning to identify the optimal settings. Subsequently, we retrain each combination using the selected hyperparameters on seven distinct random seeds, since prior work indicates that results based on only 3–5 random seeds may not be sufficiently robust [1]. We follow the exact evaluation and reporting methodology as in [1] and [32], including IQM computation, bootstrapped confidence intervals, and normalization; detailed reporting setup and metrics are provided in the Appendix.

5 Results and Analysis

In this section, we present baseline results for the classification and segmentation benchmarks. Due to space constraints, results for object detection tasks are provided in the Appendix. We structure our analysis around key research questions, which are addressed in the subsections below.

5.1 Which model architecture performs best on Mars science tasks, when pre-trained on natural images?

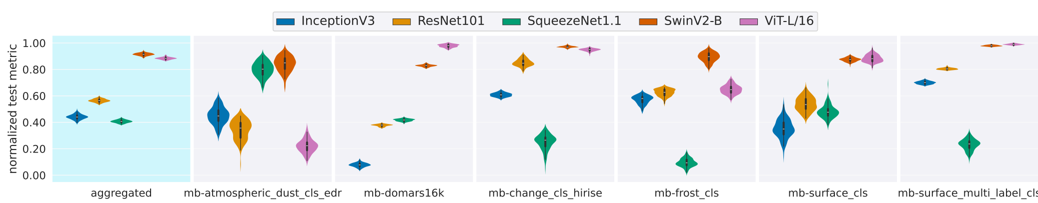


Figure 2: **Classification Benchmark under Feature Extraction setting:** Normalized F1-score of all baselines across six datasets (higher the better). Aggregated plot shows the average over all datasets.

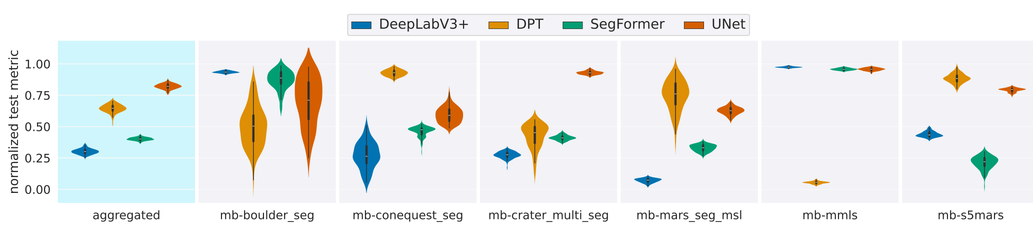


Figure 3: **Segmentation Benchmark under Feature Extraction setting:** Normalized IoU of all baselines across six datasets (higher the better). Aggregated plot shows the average over all datasets.

Figures 2 and 3 show the bootstrapped IQM of normalized performance metric (as defined in Section 4.1) across six classification and six segmentation datasets and one training strategy (feature extraction with frozen backbone), along with aggregated results. We report F1-score for classification tasks and IoU for segmentation tasks. The datasets are selected in a way that ensures a diverse set of geologic features. For example, if two datasets cover the same feature type (e.g., landmarks), we report results for only one of them. Additional results, including those for alternative training regimes and other datasets, are reported in the Appendix.

In classification tasks, SqueezeNet1.1 consistently underperforms relative to other architectures, likely due to its small parameter count. In contrast, ViT-L/16 and SwinV2-B Transformer exhibit competitive performance, with both showing strong generalization across datasets. Notably, some models display narrower confidence intervals than others, suggesting they are more stable and better suited to specific tasks.

For segmentation, U-Net achieves the highest overall performance despite having a relatively wide confidence interval in some datasets. It outperforms both transformer-based models (SegFormer and DPT) on nearly all datasets as well as in aggregate metrics. The DPT model, in particular, shows highly unstable results with large confidence intervals, making it less reliable. These results suggest that, despite its simplicity, U-Net remains a strong baseline for segmentation tasks in Mars science applications.

5.2 What is the effect of training set size on the performance of each model?

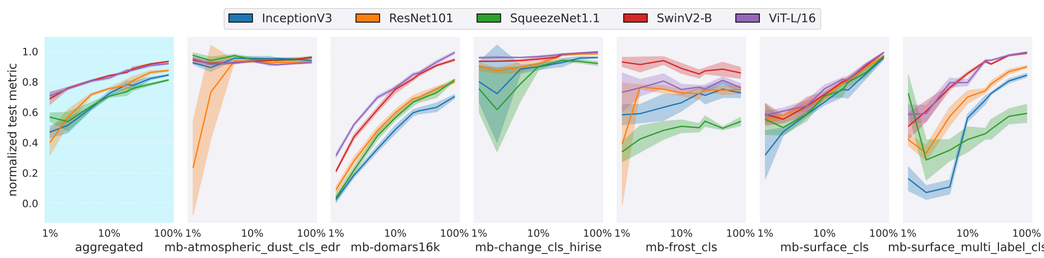


Figure 4: **Classification vs Train size:** Normalized F1-score of baselines with a growing size (from 1% to 100%) of the training set. Shaded regions indicate confidence intervals over multiple runs.

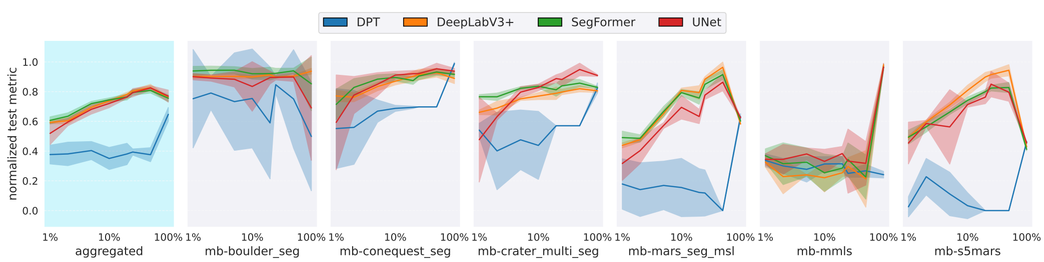


Figure 5: **Segmentation vs Train size:** Normalized IoU of baselines with a growing size (from 1% to 100%) of the training set. Shaded regions indicate confidence intervals over multiple runs.

To assess how training set size impacts model performance, we conducted experiments by varying the amount of labeled training data. Specifically, we trained each model using 1%, 2%, 5%, 10%, 20%, 25%, 50%, and 100% of the available training data, while keeping the validation and test sets fixed. For each configuration, we performed multiple runs and report the average normalized test metric, as shown in Figures 4 and 5.

From the aggregated results, we observe a consistent trend: increasing the training set size generally leads to improved performance in both classification and segmentation tasks. However, dataset-level analysis reveals that the rate of improvement and error margins vary significantly depending on the model and dataset. This shows the differing levels of difficulty among datasets in Mars-Bench, highlighting the benchmark's overall challenge.

In classification, transformer-based models such as SwinV2-B and ViT-L/16 consistently outperform smaller convolutional models like SqueezeNet1.1. In contrast, for segmentation tasks, U-Net outperforms transformer-based models such as DPT and SegFormer across most training sizes. DPT not only shows lower overall performance but also exhibits high variance across runs, as reflected in wide confidence intervals.

5.3 How do models that are trained for EO tasks perform on Mars-Bench?

Although there are no published foundation models for Mars orbital or surface imagery, there are many foundation models for Earth orbital imagery. To assess cross-domain generalization, we evaluated foundation models pre-trained on EO data. Specifically, SatMAE [54], CROMA [19], and Prithvi [27] on selected Mars-Bench classification tasks. These models were originally trained on Earth satellite data that vary in geography, scale, and semantics but share the overhead imaging perspective found in many Mars datasets. We compare them to a ViT-L/16 model pre-trained on ImageNet to establish a general-domain baseline (Figure 6).

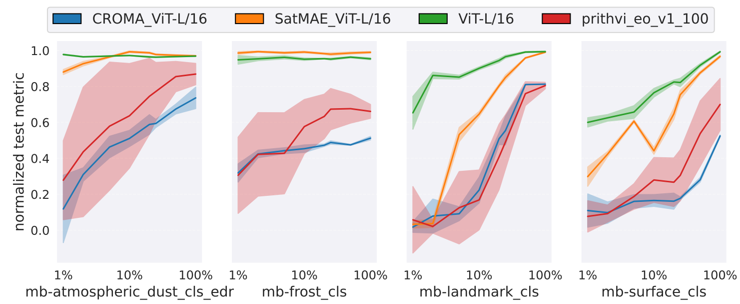


Figure 6: **Classification vs Train size for EO baselines:** Normalized F1-score with a growing size (from 1% to 100%) of the training set. Shaded regions indicate confidence intervals over multiple runs.

Although EO pre-trained models performed well on all datasets, the ImageNet pre-trained ViT performed better. One possible explanation is that although ViT is pre-trained on natural images and EO models are pre-trained on satellite data, ViT is pre-trained on 14 million images, while SatMAE, CROMA, and Prithvi are pre-trained on 1 million or less than 1 million images. Additionally, diversity in ImageNet, because as discussed in the literature, diversity and/or geographical coverage of pre-training data can affect the performance of the model [16, 43, 49, 51]. Among EO foundation models, the Prithvi model in particular consistently showed low performance and large error bars. All these results show that, despite EO models pre-trained on satellite data, Earth and Mars orbital imagery differ significantly in ways that likely impact model transferability. For instance, Martian imagery lacks vegetation, water bodies, and human-made structures, which are common in EO datasets. Additionally, Mars exhibits unique geological formations, color distributions, and atmospheric conditions that are totally different than Earth imagery. These domain gaps suggest that while EO-pretrained models can offer a reasonable starting point, foundation models specifically trained on Mars data are likely to yield more robust and generalizable performance across Martian tasks.

5.4 How do proprietary VLMs, such as Gemini and GPT, perform on Mars-Bench?

With the rapid advancement of vision-language models (VLMs), such as Gemini [61] and GPT [5], there is increasing interest in evaluating their effectiveness beyond general-purpose tasks. These models, trained on diverse multimodal datasets, have demonstrated strong performance on various open-domain vision benchmarks with minimal supervision. However, their applicability to Mars science, has not been explored. Evaluating VLMs on Mars-Bench provides valuable insight into their ability to generalize to planetary science tasks without domain-specific fine-tuning.

We focused on evaluating the reasoning capabilities of these models by explicitly prompting them with context-rich instructions, rather than relying solely on direct answer generation. We used the Gemini 2.0 Flash and GPT-4o Mini models, both from their May 2025 checkpoints.

Task	Gemini		GPT	
	Accuracy	F1-score	Accuracy	F1-score
mb-domars16k	0.34	0.32	0.36	0.30
mb-surface_cls	0.43	0.44	0.42	0.41
mb-frost_cls	0.50	0.55	0.43	0.54
mb-atmospheric_dust_cls_edr	0.43	0.50	0.68	0.56
mb-crater_multi_seg	0.37	0.41	0.49	0.51
mb-mars_seg_msl	0.86	0.84	0.79	0.70

Table 2: Performance of Gemini and GPT on Mars-Bench.

We selected six Mars-Bench datasets spanning classification and segmentation tasks. The selected tasks cover a range of geologic features to evaluate how well the models generalize across different scientific concepts. From each dataset, we randomly sampled 500 test images, ensuring the label distribution in the sampled subset matched that of the original dataset. This sample size was chosen to balance evaluation fidelity with the computational cost associated with API-based model usage, particularly for GPT. We reformulated segmentation as a multi-label classification task. For both classification and segmentation, we provided system instructions defining each class and prompted the models to predict the relevant classes for each image. Full prompts and system instructions for all tasks are included in the Appendix.

Both Gemini and GPT achieved reasonable performance on some tasks, but their results are inconsistent across datasets (Table 2). Notably, both models perform well on the `mb-mars_seg_msl` dataset, achieving an F1-score of 0.84 (Gemini) and 0.70 (GPT). This dataset involves terrain segmentation with classes such as sand, rock, and sky, classes that are also common in natural images and likely well-represented in the models' pre-training data. In contrast, performance drops significantly on datasets such as `mb-crater_multi_seg` and `mb-domars16k`, which require identification of fine-grained geologic structures like crater types and Martian landmarks.

As noted in Section 3.2, many of these tasks demand domain expertise. Our results suggest that current VLMs lack sufficient specialized knowledge for accurate interpretation. These findings highlight the gap between general-purpose vision-language capabilities and the needs of Mars science, further reinforcing the importance of domain-specific model development.

6 Research opportunities

Mars-Bench provides valuable research opportunities, not only for the planetary science and remote sensing communities but also for the broader machine learning and computer vision community. Mars-Bench creates the following key research opportunities:

- Mars-Bench will accelerate the development of foundation models specifically tailored to Mars orbital and surface-related tasks by facilitating a systematic evaluation of model performance. It provides essential infrastructure for benchmarking diverse models within a unified framework, mirroring the influential role benchmarks have historically played in other specialized domains.
- The benchmark comprises several challenging datasets that introduce unique complexities to computer vision tasks. For instance, dust devil detection is particularly challenging due to the subtle contrast differences between dust devils and the Martian terrain. ConeQuest presents difficulties stemming from significant visual variability among cones collected from various Martian regions, challenging models to generalize across high intra-class variance. In addition, many datasets included in Mars-Bench are small-scale and highly imbalanced.
- Mars-Bench significantly expands research opportunities focused on addressing distribution shifts and out-of-distribution generalization. These challenges are closely aligned with contemporary methodological advancements such as those proposed by [25, 72, 64], which emphasize robust model evaluation across diverse domains to enhance real-world applicability.

7 Conclusion

We introduced the first benchmark for evaluating models on a wide range of Mars science tasks using both orbital and surface imagery. Mars-Bench standardizes diverse datasets into a unified, machine-learning-ready format and provides code for fine-tuning and evaluating across classification, segmentation, and object detection tasks. Datasets in Mars-Bench also include a wide variety of geologic features that have been extensively studied in the literature and remain of high interest to the scientific community. We believe that Mars-Bench will drive the development of Mars-specific foundation models, improve generalization across planetary tasks, and open new research directions in planetary science and beyond.

Limitations A key limitation of Mars-Bench is the absence of georeferencing for most datasets. This arises from the fact that the original sources of these datasets do not provide spatial metadata (e.g., latitude and longitude coordinates), mapping the samples to the Martian surface. As a result, it is currently not possible to assess the spatial distribution or coverage of Mars-Bench across different regions of Mars. The only exception is the ConeQuest dataset, which includes precise geolocation information, and we retain this spatial metadata in our release. Lack of georeferencing is a known challenge in remote sensing benchmarks, as it restricts the ability to conduct spatial analysis or regional generalization studies. Additionally, we did not explore techniques to address class imbalance in datasets, such as re-sampling or loss reweighting. Investigating methods to handle imbalance and its effect on model performance remains an important direction for future work.

References

[1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

[2] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: An earth observation model for any resolutions, scales, and modalities. *arXiv preprint arXiv:2412.14123*, 2024.

[3] Abigail R Azari, John B Biersteker, Ryan M Dewey, Gary Doran, Emily J Forsberg, Camilla DK Harris, Hannah R Kerner, Katherine A Skinner, Andy W Smith, Rashied Amini, et al. Integrating machine learning for planetary science: Perspectives for the next decade. *arXiv preprint arXiv:2007.15129*, 2020.

[4] Dipkamal Bhusal, Md Tanvirul Alam, Le Nguyen, Ashim Mahara, Zachary Lightcap, Rodney Frazier, Romy Fieblinger, Grace Long Torales, Benjamin A Blakely, and Nidhi Rastogi. Secure: Benchmarking large language models for cybersecurity. *arXiv preprint arXiv:2405.20441*, 2024.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[7] Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024.

[8] SB Cole, JC Aubele, BA Cohen, SM Milkovich, and SR Shields. Identifying community needs for a mars exploration rovers (mer) data catalog. In *51st Annual Lunar and Planetary Science Conference*, number 2326, page 1709, 2020.

[9] Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*, 2024.

[10] Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *arXiv preprint arXiv:2306.16092*, 2023.

[11] Danielle M DeLatte, Sarah T Crites, Nicholas Guttenberg, Elizabeth J Tasker, and Takehisa Yairi. Segmentation convolutional neural networks for automatic crater detection on mars. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(8):2944–2957, 2019.

[12] Ivica Dimitrovski, Ivan Kitanovski, Dragi Kocev, and Nikola Simidjievski. Current trends in deep learning for earth observation: An open-source benchmark arena for image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:18–35, 2023.

[13] Gary Doran. Hirise image patches obscured by atmospheric dust, October 2019.

[14] Gary Doran, Serina Diniega, Steven Lu, Mark Wronkiewicz, and Kiri L Wagstaff. Evaluating terrain-dependent performance for martian frost detection in visible satellite observations. *arXiv preprint arXiv:2403.12080*, 2024.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Rahim Entezari, Mitchell Wortsman, Olga Saukh, M Moein Shariatnia, Hanie Sedghi, and Ludwig Schmidt. The role of pre-training data in transfer learning. *arXiv preprint arXiv:2302.13602*, 2023.

[17] Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*, 2023.

[18] Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, et al. Bigbio: A framework for data-centric biomedical natural language processing. *Advances in Neural Information Processing Systems*, 35:25792–25806, 2022.

[19] Anthony Fuller, Koreen Millard, and James Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36:5506–5538, 2023.

[20] Edwin Goh, Isaac R Ward, Grace Vincent, Kai Pak, Jingdao Chen, and Brian Wilson. Self-supervised distillation for computer vision onboard planetary robots. In *2023 IEEE Aerospace Conference*, pages 1–11. IEEE, 2023.

[21] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279, 2023.

[22] Zexin Guo, Yi Xu, Dagang Li, Yemeng Wang, Kim-Chiu Chow, Renrui Liu, and Qiquan Yang. Martian dust devil detection based on improved faster r-cnn. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[23] 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*, pages 770–778, 2016.

[24] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[25] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

[26] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.

[27] Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

[28] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[29] Hannah Rae Kerner, Kiri L Wagstaff, Brian D Bue, Patrick C Gray, James F Bell, and Heni Ben Amor. Toward generalized change detection on planetary surfaces with convolutional autoencoders and transfer learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(10):3900–3918, 2019.

[30] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. SatCLIP: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*, 2023.

[31] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

[32] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36:51080–51093, 2023.

[33] Anthony Lagain, Sylvain Bouley, David Baratoux, Chiara Marmo, François Costard, O Delaa, A Pio Rossi, M Minin, GK Benedix, M Ciocco, et al. Mars crater database: A participative project for the classification of the morphological characteristics of large martian craters. *Large Meteorite Impacts and Planetary Evolution*, VI, 2021.

[34] Francois Lanusse, Liam Holden Parker, Siavash Golkar, Alberto Bietti, Miles Cranmer, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Ruben Ohana, Mariel Pettee, et al. Astroclip: Cross-modal pre-training for astronomical foundation models. In *NeurIPS 2023 AI for Science Workshop*, 2023.

[35] Jiaojiao Li, Shunyao Zi, Rui Song, Yunsong Li, Yinlin Hu, and Qian Du. A stepwise domain adaptive segmentation network with covariate shift alleviation for remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[37] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[38] 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*, pages 10012–10022, 2021.

[39] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. *arXiv preprint arXiv:2407.01725*, 2024.

[40] Shrey Malvi, Hitansh Shah, Niketan Chandarana, Mirali Purohit, Jacob Adler, and Hannah Kerner. Automated multi-class crater segmentation in mars orbital images. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 110–120, 2023.

[41] Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, et al. Pangaea: A global and inclusive benchmark for geospatial foundation models. *arXiv preprint arXiv:2412.04204*, 2024.

[42] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

[43] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *Advances in Neural Information Processing Systems*, 35:21455–21469, 2022.

[44] Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, Charlie O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Li, et al. Astrollama: Towards specialized foundation models in astronomy. *arXiv preprint arXiv:2309.06126*, 2023.

[45] Sidike Paheding, Abel A Reyes, A Rajaneesh, KS Sajinkumar, and Thomas Oommen. Marslsnet: Martian landslides segmentation network and benchmark dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8236–8245, 2024.

[46] Mihir Parmar, Swaroop Mishra, Mirali Purohit, Man Luo, Murad Mohammad, and Chitta Baral. In-BoXBART: Get instructions into biomedical multi-task learning. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 112–128, Seattle, United States, July 2022. Association for Computational Linguistics.

[47] Nils C Prieur, Brian Amaro, Emiliano Gonzalez, Hannah Kerner, Sergei Medvedev, Lior Rubanenko, Stephanie C Werner, Zhiyong Xiao, Dmitry Zastrozhnov, and Mathieu GA Lapôtre. Automatic characterization of boulders on planetary surfaces from high-resolution satellite images. *Journal of Geophysical Research: Planets*, 128(11):e2023JE008013, 2023.

[48] Mirali Purohit, Jacob Adler, and Hannah Kerner. Conequest: A benchmark for cone segmentation on mars. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6026–6035, 2024.

[49] Mirali Purohit, Gedeon Muhawenayo, Esther Rolf, and Hannah Kerner. How does the spatial distribution of pre-training data affect geospatial foundation models? *arXiv preprint arXiv:2501.12535*, 2025.

[50] MV Purohit, S Lu, S Diniega, UD Rebbapragada, and HR Kerner. Investigating the benefits of foundation models for mars science. *LPI Contributions*, 3007:3535, 2024.

[51] Vivek Ramanujan, Thao Nguyen, Sewoong Oh, Ali Farhadi, and Ludwig Schmidt. On the connection between pre-training data diversity and fine-tuning robustness. *Advances in Neural Information Processing Systems*, 36:66426–66437, 2023.

[52] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[53] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[54] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023.

[55] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[56] Stuart J Robbins and Brian M Hynek. A new global database of mars impact craters $\geq$ 1 km: 1. database creation, properties, and parameters. *Journal of Geophysical Research: Planets*, 117(E5), 2012.

[57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[58] Inigo V Slijepcevic, Anna MM Scaife, Mike Walmsley, Micah Bowles, O Ivy Wong, Stanislav S Shabala, and Sarah V White. Radio galaxy zoo: towards building the first multipurpose foundation model for radio astronomy with self-supervised learning. *RAS Techniques and Instruments*, 3(1):19–32, 2024.

[59] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. Featured Certification.

[60] 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*, pages 2818–2826, 2016.

[61] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[62] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[63] Grace M Vincent, Isaac R Ward, Charles Moore, Jingdao Chen, Kai Pak, Alice Yepremyan, Brian Wilson, and Edwin Y Goh. CLOVER: Contrastive learning for onboard vision-enabled robotics. *Journal of Spacecraft and Rockets*, 61(3):728–740, 2024.

[64] Johannes Von Oswald, Seijin Kobayashi, Alexander Meulemans, Christian Henning, Benjamin F Grewe, and João Sacramento. Neural networks with late-phase weights. *arXiv preprint arXiv:2007.12927*, 2020.

[65] 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*, volume 35, pages 15204–15213, 2021.

[66] 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*, volume 32, 2018.

[67] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

[68] Wenjing Wang, Lilang Lin, Zejia Fan, and Jiaying Liu. Semi-supervised learning for mars imagery classification and segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(4):1–23, 2023.

[69] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[70] Isaac Ronald Ward, Charles Moore, Kai Pak, Jingdao Chen, and Edwin Goh. Improving contrastive learning on visually homogeneous mars rover images. In *European Conference on Computer Vision*, pages 170–185. Springer, 2022.

[71] 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.

[72] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR, 17–23 Jul 2022.

[73] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[74] Chen Yang, Nan Zhang, Renchu Guan, and Haishi Zhao. Mapping cones on mars in high-resolution planetary images with deep learning-based instance segmentation. *Remote Sensing*, 16(2):227, 2024.

[75] Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David B Lobell, and Stefano Ermon. Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. *arXiv preprint arXiv:2111.04724*, 2021.

[76] 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.

[77] Qifang Zheng, Rong Huang, Yusheng Xu, Fangzhao Zhang, Changjiang Xiao, Luning Li, and Xiaohua Tong. Automatic morphologic classification of martian craters using imbalanced datasets of tianwen-1's moric images with deep neural networks. *Planetary and Space Science*, page 106104, 2025.

[78] 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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Section 5.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 7

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Section 4.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Appendix

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Section 5

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See Appendix

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: See Section 3.1

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: See Section 3.1

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]