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

Many of the recent foundation models have been successful because of having standardized evaluation benchmarks, which help in evaluating these models fairly and in a standardized manner. There are no evaluation benchmarks for Mars science applications, and hence, this obstructs the progress of building a foundation model for Mars science 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, provided in a standardized and ready-to-use format. We provide baseline evaluations using models pre-trained on natural images and Earth satellite data. Results from analyses suggest that Mars-specific foundation models may offer advantages over baselines, 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. Our data, models, and code are available at: https://mars-bench.github.io/.

1 Introduction

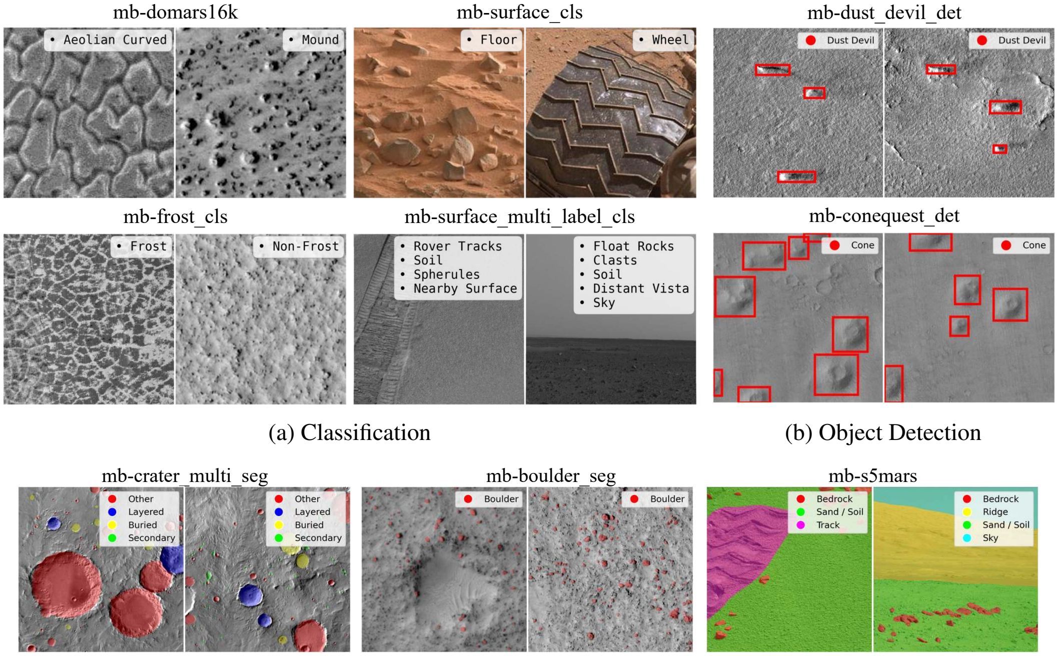
Foundation models have recently transformed the fields of medical [27, 31], Earth Observation (EO) [17, 38, 2], law [5, 6], and astronomy [20, 29, 42], by providing strong evaluation benchmarks. In EO, high-quality benchmarks like Geo-Bench [18] and PANGAEA [26] have accelerated progress by providing standardized tasks and evaluation pipelines. However, no such evaluation benchmark exists for Mars science, and that hinders the building of a robust foundation model.

While several studies have proposed task-specific models such as crater detection [25, 50], landmark classification [46, 45], and cone segmentation [33, 48]; these solutions and datasets lack standardization and interoperability. Also, since many existing datasets do not provide training, validation, and testing splits, it is often impossible to fairly compare task-specific methods.

This gap is surprising given the wealth of Mars data returned from Mars missions in the past several decades. Mars orbiters (e.g., the Mars Reconnaissance Orbiter (MRO)) and rovers (e.g., Curiosity) have collected millions of images, offering great potential to study critical questions of Mars science, such as the past presence of water and the planet's habitability with ML. However, the lack of standardized benchmarks and ML-ready datasets has limited their use in foundation model research.

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

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.



(a) Classification

(b) Object Detection

(c) Segmentation

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

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 images. Mars-Bench standardizes 20 datasets across classification, segmentation, and object detection; and covers a wide range of geologic features. Samples from few Mars-Bench datasets are shown in Figure 1. We evaluate Mars-Bench using models pre-trained on natural images (ImageNet or COCO), and pre-trained EO foundation models. To support reproducibility and the community, we release all code, tools, documentation; and baseline models, which can serve as a strong starting point for future Mars ML applications.

2 Mars-Bench

Design Principles Mars-Bench is built for *easy, fair, and extensible* evaluation. We standardize file structure and annotations across tasks, provide ready-to-use dataloaders, and consistent training, validation, and testing splits to make fair comparisons for future model development. To study data-limited regimes, we release fixed few-shot variants and partitioned training sets. We partition datasets based on attributes such as sensor, 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. With this, we performed expert-led corrections (e.g., correcting label noise, merging ambiguous classes, improving consistency across tasks, and removing redundant augmentations) in datasets wherever necessary. Additionally, we provide annotations in all common formats for each task to ensure compatibility and save users time converting between formats. All Mars-Bench datasets have permissive licenses; we release them under CC BY 4.0 to ensure open access and reuse.

Tasks and Datasets Mars-Bench spans 20 datasets across classification, segmentation, and object detection, covering both orbital (CTX, HiRISE, THEMIS) and surface (Mastcam, MAHLI, Pancam, Navcam) images. Subtasks include binary, multi-class, and multi-label classification; binary and multi-class segmentation; and detection of small, sparse objects (cones, boulders, dust devils). 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. The benchmark includes geologic features such as craters, cones, boulders, landslides, dust devils, frost, atmospheric dust, and terrain-related classes (e.g., sand/rock/bedrock, tracks, rover parts). The datasets vary

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	[7]
mb-atmospheric_dust_cls_rdr	MRO (O)	Atmospheric dust	100 × 100	2	9817	4969	5214	1	HiRISE	2019	[7]
mb-change_cls_ctx	MRO (O)	Surface change	150 × 150	2	36	10	10	1	CTX	2019	[16]
mb-change_cls_hirise	MRO (O)	Surface change	100 × 100	2	3103	670	670	1	HiRISE	2019	[16]
mb-domars16k	MRO (O)	Landmark	200 × 200	15	11305	3231	1614	1	CTX	2020	[46]
mb-frost_cls	MRO (O)	Frost	299 × 299	2	30124	11415	12249	1	HiRISE	2024	[8]
mb-landmark_cls	MRO (O)	Landmark	227 × 227	8	6997	2025	1793	1	HiRISE	2021	[44]
mb-surface_cls	Curiosity (R)	Surface	256 × 256	36	6580	1293	1594	3	Mastcam, MAHLI	2018, 2021	[44, 45]
mb-surface_multi_label_cls	Opportunity, Spirit (R)	Surface	1024 × 1024	25	1762	443	739	1	Pancam	2020	[4]

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	[32]
mb-conequest_seg	MRO (O)	Cone	512× 512	2	2236	319	643	1	CTX	2024	[33]
mb-crater_binary_seg	Mars Odyssey (O)	Crater	512 × 512	2	3600	900	900	1	THEMIS	2012	[40]
mb-crater_multi_seg	Mars Odyssey (O)	Crater	512 × 512	5	3600	900	900	1	THEMIS	2021	[19]
mb-mars_seg_mer	Opportunity, Spirit (R)	Terrain	1024 × 1024	7	744	106	214	1	Navcam, Pancam	2022	[21]
mb-mars_seg_msl	Curiosity (R)	Terrain	500 × 560	7	2893	413	828	3	Mastcam	2022	[21]
mb-mmls	MRO (O)	Landslide	128 × 128	2	275	31	256	7	CTX	2024	[30]
mb-s5mars	Curiosity (R)	Terrain	1200 × 1200	10	4997	200	800	3	Mastcam	2022	[49]

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	[32]
mb-conequest_det	MRO (O)	Cone	512 × 512	1	1158	167	333	1	CTX	2024	[33]
mb-dust_devil_det	MRO (O)	Dust devil	∼ 750 × 750	1	1404	201	402	1	CTX	2024	[12]

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).

widely in terms of size, data balancing, modality (RGB, grayscale, multi-modal), and visual difficulty (e.g., low-contrast dust devils; regionally diverse cone morphologies). This diversity highlights the breadth of Mars-Bench in terms of task design, sensor modalities, and the variety of geologic features.

Using the Dataset We release Mars-Bench via Hugging Face Datasets[1] and Zenodo[2] following a standardized schema, metadata, documentation, and loading script to facilitate integration into ML pipelines. Mars-Bench targets researchers developing models for Mars-related tasks, planetary science, and broader computer vision problems, supporting studies on distribution shift, generalization, and domain adaptation with coverage of underrepresented geospatial scenarios. Along with datasets, we will also release an open-source toolkit and baseline models. The baseline models released for each dataset aim to lower the barrier for applied research, enabling scientists to generate global maps of features like cones or craters.

3 Experiments

For model selection, we used well-established and widely adopted model architectures for each task. Particularly, we have used ResNet101 [13], SqueezeNet1.1 [14], InceptionV3 [43], Swin Transformer (SwinV2-B) [24], and Vision Transformer (ViT-L/16) [9] for classification; U-Net [41], DeepLabV3+ [3], SegFormer [47], and Dense Prediction Transformer (DPT) [36] for segmentation; YOLO11 [37], SSD [23], RetinaNet [22], and Faster R-CNN [39] for object detection tasks. We have evaluated all models by taking their pre-trained version as a feature extractor. Since no open-source Mars pre-trained foundation model is available, we used models pre-trained on large-scale datasets like ImageNet or COCO (for detection) as a starting point for feature extraction.

To conduct experiments, we adopted a methodology identical to [18]. We performed hyperparameter tuning through a grid search for each model-dataset combination, selecting the best-performing settings based on early stopping criteria applied to the validation set. Subsequently, we re-trained each model-dataset combination with the best hyperparameters on 7 different seeds, since prior work indicates that results based on only 2-4 seeds may not be sufficiently robust [1]. We report both task-specific and aggregated results with reliable confidence intervals, as recommended in [1, 18]. Details of experiments, hyperparameters, and reporting results are provided in Appendix A.

[1] huggingface.co/collections/Mirali33/mars-bench

[2] zenodo.org/communities/mars-bench/records

4 Results and Analysis

In this section, we report results for classification tasks. Due to space constraints, results of segmentation and object detection tasks for analysis 4.1 are reported in Appendix B.

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

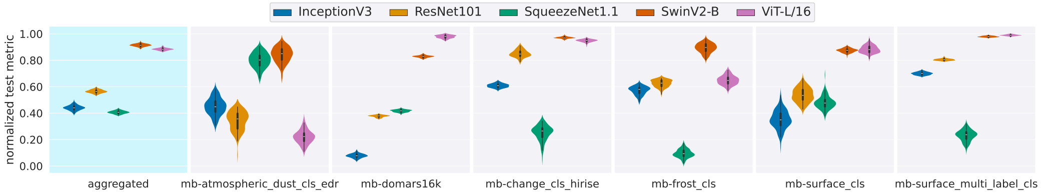


Figure 2: **Classification Benchmark under Feature Extraction setting:** Normalized F1-score of all baselines across six datasets (higher the better). Aggregated plot shows the average over all datasets.

Figure 2 shows results on six classification datasets, along with aggregated results. The datasets are selected in a way that ensures a diverse set of geologic features. From Figure 2, it can be observed that SqueezeNet1.1 consistently underperforms due to its small parameter count (1.2M), while ViT-L/16 and SwinV2-B Transformer show competitive performance and strong generalization across datasets. Some models also exhibit narrower confidence intervals, suggesting higher stability for specific tasks (e.g., `mb-domars16k` and `mb-surface_multi_label_cls`).

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

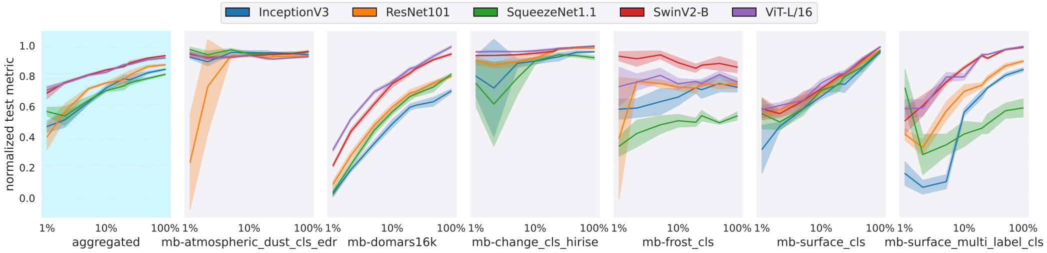


Figure 3: **Classification vs Train size:** Normalized F1-score of baselines with a growing size (from 1% to 100%) of the training set. Shaded regions indicate confidence intervals over multiple runs.

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 Figure 3.

From the aggregated results for classification tasks, we observe that increasing the training set size generally improves performance. However, the rate of improvement and error margins vary significantly based on the specific model and dataset, which highlights the differing levels of difficulty among the Mars-Bench datasets. Moreover, transformer-based models (SwinV2-B and ViT-L/16) consistently outperform smaller convolutional models like SqueezeNet1.1.

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

Although there are no published foundation models available 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 Earth satellite data. Specifically, we evaluate SatMAE [38], CROMA [11], and Prithvi [15], on selected classification tasks. These models were originally trained on Earth satellite data that varies in geography, scale, and semantics; and we compared them with a pre-trained model on ImageNet (ViT-L/16) as a general-domain baseline (Figure 4).

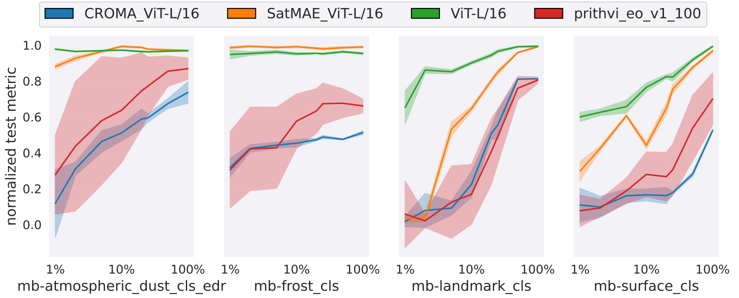


Figure 4: **Classification vs Train size for EO baselines:** Normalized F1-score with a growing size (from 1% to 100%) of the training set. Shaded regions indicate confidence intervals over multiple runs.

ImageNet pre-training shows superior performance, despite EO models being pre-trained on satellite data, a possible explanation is that ImageNet is a significantly larger and diverse dataset of 14 million images compared to the 1 million or fewer images used to pre-train SatMAE, CROMA, and Prithvi. Additionally, as discussed in the literature, diversity and/or geographical coverage of pre-training data can affect the performance of the model [10, 28, 34, 35]. Despite being pre-trained on Earth satellite data, models struggle to transfer to Mars because its imagery lacks vegetation, water, and human structures, and has distinct geology, colors, and atmosphere. These domain gaps suggest that Mars-specific foundation models can improve performance consistency and generalization.

5 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. A key limitation of Mars-Bench is the lack of georeferencing in most datasets, since the original sources do not provide spatial metadata (e.g., latitude-longitude) to map samples onto the Martian surface. Consequently, spatial distribution or regional coverage cannot be assessed, with ConeQuest being the only exception, as it includes precise geolocation. At last, 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.

Acknowledgment: Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

A Experiment Details

A.1 Pipeline and Hyperparameters

We provide a user-friendly and scalable training and inference pipeline for classification, segmentation, and object detection tasks. The pipeline supports running experiments via command-line arguments, allowing easy configuration of core parameters such as dataset, model architecture, and hyperparameters.

It includes modular support for logging with options for Weights & Biases (Wandb), TensorBoard, and CSV; model checkpointing; early stopping; and other PyTorch Lightning-compatible callbacks. For reproducibility, we fix the random seed across all relevant libraries and save Hydra configuration files and logs locally.

config	**value**
seed	0, 1, 10, 42, 123, 1000, 1234
learning rate schedule	w/o, cosine, plateau, step
base learning rate	1e-3, 1e-4, 1e-5
weight decay	0.05
batch size	16, 32, 64
optimizer	Adam, AdamW, SGD
max training epochs	50, 100, 200
patience	5, 10

Table 2: Training hyperparameters

As described in Section 3, for each model-dataset combination, we first perform hyperparameter tuning. The full search space is listed in Table 2. We also experiment with different loss functions, summarized in Table 3. For binary classification, we try both a one-node output with binary cross-entropy and a two-node output with standard cross-entropy. For segmentation, we evaluate three loss types: generalized Dice loss, cross-entropy, and a weighted combination of both. We also explore three different weighting schemes. For object detection, we use the default loss returned by each model implementation.

Classification

config	**value**
criterion	cross entropy, binary cross entropy (only for binary classification)

Segmentation

config	**value**
criterion	generalized_dice (square, simple, linear), cross entropy, combined
smoothing value	1e-5 (only for generalized_dice)

Table 3: Configuration for loss function

A.2 Reporting Results

As mentioned in Section 3 and inspired by the methodologies in [1] and [18], we follow a consistent procedure to report results across thousands of experiments. First, we perform hyperparameter tuning for each model–dataset combination, selecting the best configuration based on validation loss using early stopping. Once the optimal hyperparameters are determined, we re-train and evaluate each model–dataset combination 7 times with different random seeds (listed in Table 2), as recommended in prior work [1, 18]. For each combination, we compute the InterQuartile Mean (IQM) by discarding the top and bottom 25% of scores and averaging the remaining values. This approach helps reduce both bias and variance in the reported performance. Before aggregating results across tasks, we normalize the scores within each task to account for differences in scale.

To quantify uncertainty, we perform 1,000 rounds of stratified bootstrapping. In each round, we sample (with replacement) one trial from each dataset, recompute the IQM across all datasets, and build a distribution of IQM values. From this distribution, we calculate 95% confidence intervals. In our final results, we present per-task baselines and overall model performance (aggregated across all tasks) via violin plots.

The results shown in Figures 2 and 3 show a normalized metric, without any aggregation. While the main paper reports normalized and aggregated results for the feature extraction setting, we include the corresponding raw results in the appendix: F1-score for classification, IoU for segmentation, and mAP for object detection.

B Extended Results

B.1 Classification Results

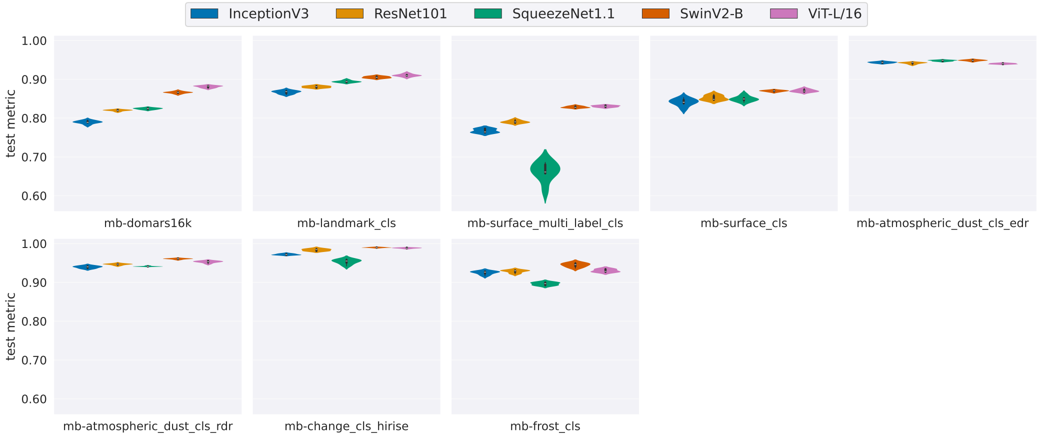


Figure 5: **Classification Benchmark under Feature Extraction setting:** Raw F1-score of various baselines (higher is better). Violin plots represent the distribution of seeds.

Figure 5 presents the classification results (F1-score) for all datasets under feature extraction. We exclude results for mb-change_cls_ctx as it shows negligible variation across different models. Binary classification datasets, such as mb-atmospheric_dust_cls_edr, mb-atmospheric_dust_cls_rdr, mb-frost_cls, and mb-change_cls_hirise achieve consistently higher performance compared to multi-class datasets. In line with this trend, mb-surface_multi_label_cls shows the lowest performance, reflecting the added complexity of multi-label prediction.

B.2 Segmentation Results

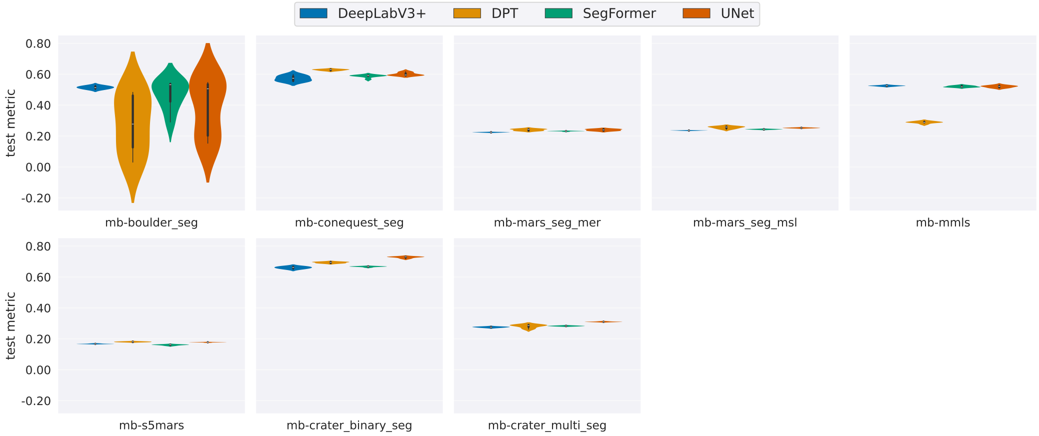


Figure 6: **Segmentation Benchmark under Feature Extraction setting:** Raw IoU of various baselines (higher is better). Violin plots represent the distribution of seeds.

Figure 6 presents the segmentation results (IoU) for all datasets under feature extraction. U-Net achieves the highest overall performance despite having a relatively wide confidence interval in one or two datasets. It outperforms both transformer-based models (SegFormer and DPT) on nearly all datasets. 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.

B.3 Object Detection Results

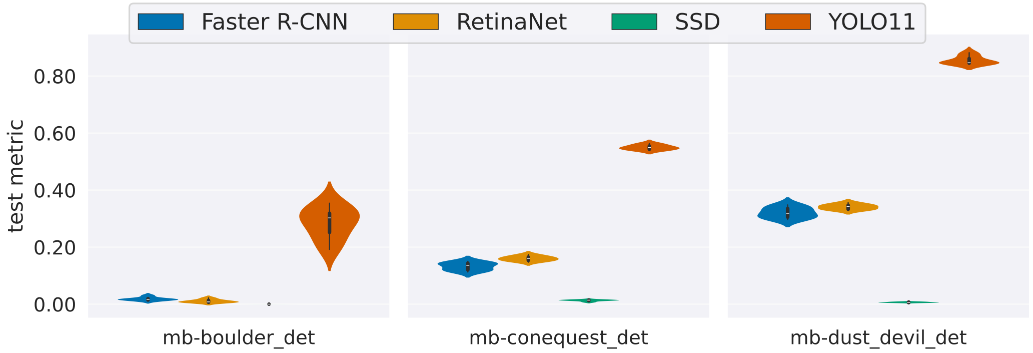


Figure 7: **Object Detection Benchmark under Feature Extraction setting:** Raw mAP of various baselines (higher is better). Violin plots represent the distribution of seeds.

Figure 7 presents the object detection results (mAP) for all datasets under feature extraction. In all 3 datasets, the YOLO model consistently outperforms other models. With this, performance is particularly weak on mb-boulder_det and mb-dust_devil_det. These challenges are primarily due to several factors:

- The overall dataset size is significantly smaller compared to classification and segmentation datasets.
- The number of objects per image is low, with many images containing only one or even zero target objects.
- The grayscale nature of the imagery limits visual cues, and low object–background contrast (e.g., in dust devil detection) further complicates learning.