MOMO: Mars Orbital Model
Foundation Model for Mars Orbital Applications
(*Supplementary Material*)

Mirali Purohit[1,2] ✉ **Bimal Gajera**[1*] **Irish Mehta**[1*] **Bhanu Tokas**[1*]
Jacob Adler[1] **Steven Lu**[2] **Scott Dickenshied**[1] **Serina Diniega**[2]
Brian Bue[2] **Umaa Rebbapragada**[2] **Hannah Kerner**[1]

[1]Arizona State University
[2]Jet Propulsion Laboratory, California Institute of Technology

A. Data Overview

A.1. Pre-training Data Details

HiRISE is mounted on the Mars Reconnaissance Orbiter (MRO) satellite and has been collecting data since 2006. HiRISE captures visible spectrum images at very high-resolution, i.e., ~ 0.25 meters/pixel. HiRISE images cover a cumulative area of $\sim 4.5\%$ of the martian surface; however, unique coverage (excluding repeats for stereo and monitoring) is $< 3\%$ [12]. We used grayscale data from the RED band of map-projected Reduced Data Record (RDR) products, and from the Primary and Extended Science Phases (PSP and ESP)[1]. Our square image crops were extracted from map-projected HiRISE images. We applied a filter to exclude crops that extended into the no-data HiRISE border (black area in the Figure 1). We gathered $\sim 16M$ image crops, which were selected from images acquired between November 2006 through May 2025. From these, we first filter the data using SSIM and Noise Estimate, and then further downsample to $\sim 4M$ using GMOM stratified sampling as described in Section 4. We adopt GMOM-based sampling instead of random sampling to ensure uniform geographic coverage, as random sampling may miss certain regions of the surface. As shown in prior work [13, 15], geographic distribution plays an important role in model performance.

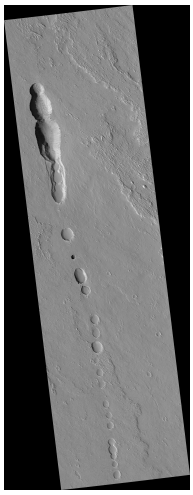


Figure 1. Example of a HiRISE map-projected image used in our study. The dark border around the image represents no-data regions that were filtered out during preprocessing to ensure high-quality crop selection.

CTX is another visible imager on MRO with a wider ground footprint. To prepare pretraining data for CTX, we used open-source CTX data from the Murray Lab[2] (updated March 2023) [1]. The dataset is a seam-corrected global image mosaic of Mars rendered at 5.0 meters/pixel [6, 11]. Data covers the entirety of the Martian surface ($> 99.5\%$). The global image data is divided into 3960 geotiff tiles ($4^\circ \times 4^\circ$) from 88°S to 88°N [5, 6]. Each tile is subdivided into four subtiles ($2^\circ \times 2^\circ$). On the Murray Lab, CTX data was last updated in March 2023. To create almost even geographic distribution from all subtiles, in each subtile, we randomly sample 630 points and crop data samples. This way, we make sure that we are capturing the diversity of the terrain across the Martian surface. This resulted in $\sim 10M$ CTX data samples globally, and then we filter this data to remove noisy samples (using SSIM and Noise Estimate). From there, we further sample $\sim 4M$ data samples using GMOM as described in Section 4.

THEMIS is a thermal infrared imager on the Mars Odyssey Orbiter and has been collecting data since 2001. We used THEMIS day-time images at 100 meters/pixel resolution. THEMIS has global coverage [3]. Similar to HiRISE data, original

[1] https://hirise-pds.lpl.arizona.edu/PDS/RDR/
[2] https://murray-lab.caltech.edu/CTX/tiles/beta01/

THEMIS tiles are tilted. Thus, we have used the same process (as HiRISE) to create crops from THEMIS tiles as well. We have used Projected Brightness Temperatures (PBT) products from THEMIS archive[3] [2]. Although THEMIS has global coverage, due to low-resolution data, we got a total of $\sim$ 4M data samples. We have exported and processed data from October 2002 to April 2025.

As described in Section 4, we use a HEALPix strategy to create geographically consistent training and validation sets. We use a HEALPix pixel size of 64, ensuring that all samples within a given cell are assigned exclusively to either the training or validation split. From our $\sim 4M$ curated samples, we split 95% for training and 5% for validation for each sensor, respectively. This prevents cross-sensor leakage and preserves geographic diversity within each split. The resulting spatial distribution and the final train/validation assignments for HiRISE, CTX, and THEMIS are summarized in Figure 2, Figure 3, and Figure 4, respectively.

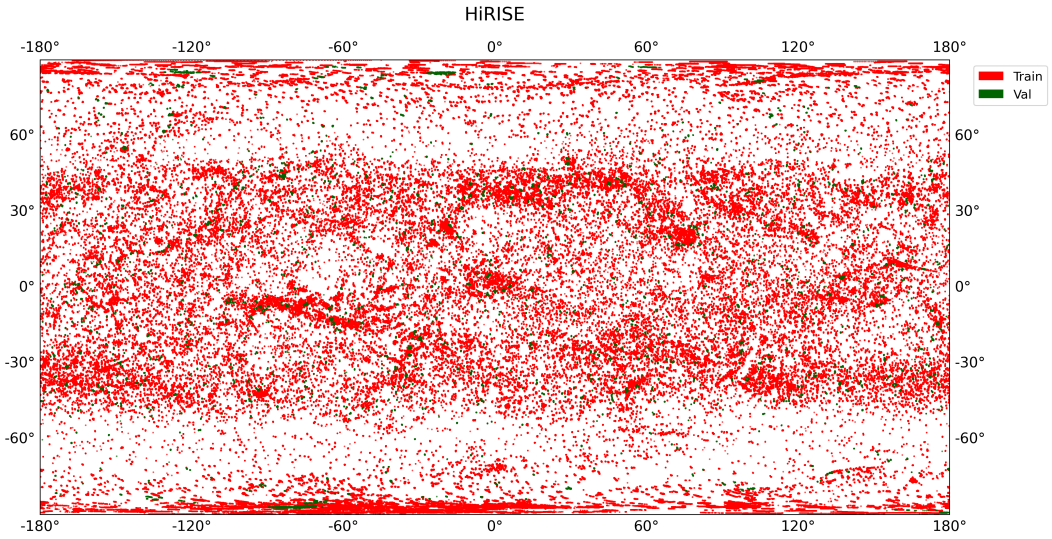


Figure 2. HiRISE pre-training data distribution

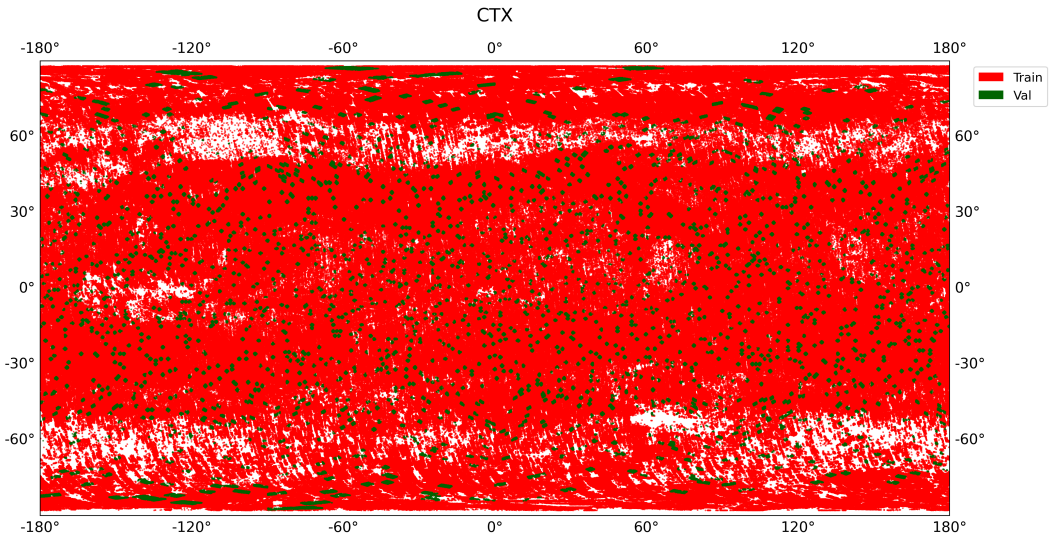


Figure 3. CTX pre-training data distribution

[3] https://static.mars.asu.edu/pds/ODTGEO_v2/data/

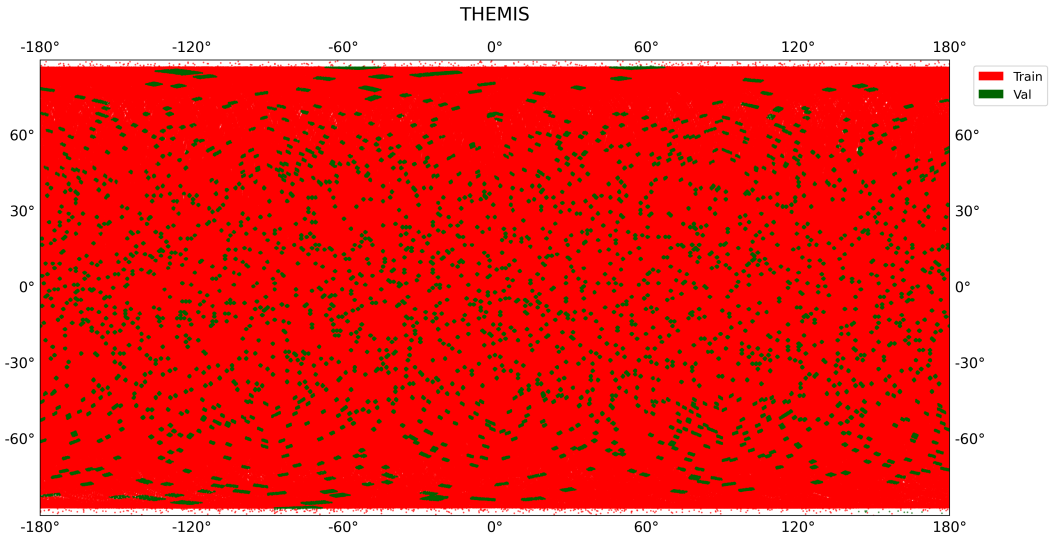


Figure 4. THEMIS pre-training data distribution

A.2. Downstream Tasks

As mentioned in Section 5, we evaluate MOMO on all orbital tasks from Mars-Bench [14]. In this section, we describe details of each downstream task and which sensor that downstream task belongs to. For simplicity, we remove the prefix "mb-" from all datasets, and for long dataset names, we represent that with a short, meaningful name.

A.2.1. Classification

AtmosDust This is a binary classification dataset and focuses on classifying between "**Dusty**" and "**Non dusty**" regions in Mars surface imagery captured by the HiRISE sensor on the MRO. This dataset has two versions provided in Mars-Bench, i.e., EDR (Experimental Data Record) and RDR (Reduced Data Record). As both datasets have the same characteristics, we have evaluated only on the RDR version of the dataset (Figure 5). The `EDR` refers to raw images from the sensor that have not been calibrated or stitched together; while the `RDR` is a downsampled or processed version of the EDR, typically used for quick viewing or initial analysis.

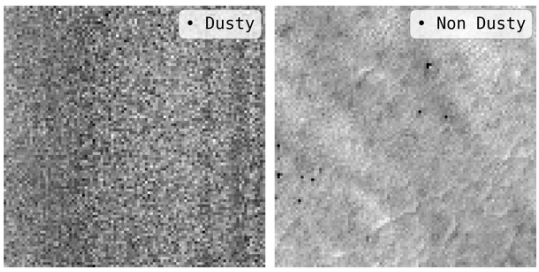


Figure 5. AtmosDust

DoMars16k This is a multi-class classification dataset designed for geomorphologic feature recognition on Mars using imagery from the *CTX* sensor. It consists of 15 classes (Figure 6) grouped into five thematic categories: (1) **Aeolian Bedforms:** Aeolian Curved, Aeolian Straight; (2) **Topographic Landforms:** Channel, Cliff, Mounds, Ridge; (3) **Slope Features:** Gullies, Mass Wasting, Slope Streaks; (4) **Impact Landforms:** Crater, Crater Field; and (5) **Basic Terrain:** Mixed Terrain, Rough Terrain, Smooth Terrain, Textured Terrain. This is one of the largest and most diverse *orbital* datasets in terms of # of classes. Hence, the dataset presents a unique challenge due to its class granularity, significant variability within classes, and subtle differences between classes, making it valuable for evaluating models.

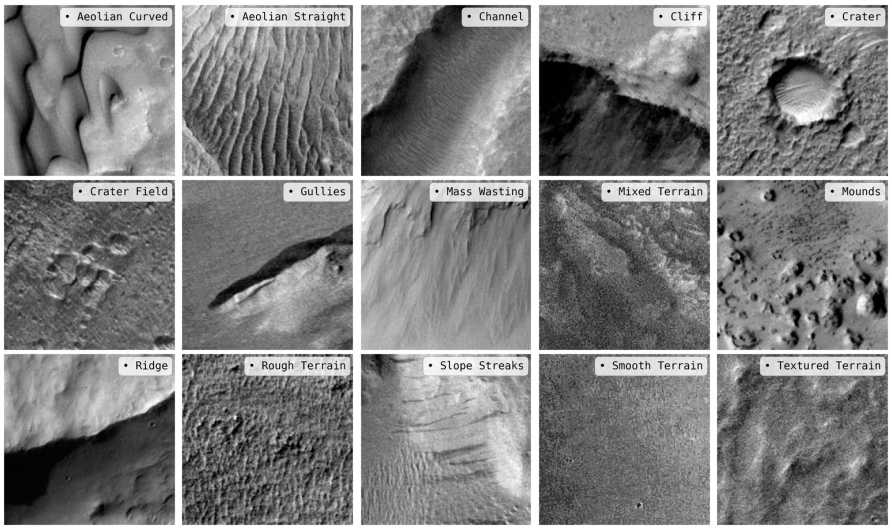


Figure 6. DoMars16k

Landmark This dataset is a multi-class classification corpus derived from orbital HiRISE imagery. Each image is assigned to one of eight geomorphological feature classes: **Bright Dune, Crater, Dark Dune, Impact Ejecta, Slope Streak, Spider, Swiss Cheese**, and **Other** (Figure 7). The class distribution is highly imbalanced, with *Other* dominating the dataset and *Impact Ejecta* representing the rarest (minority) class.

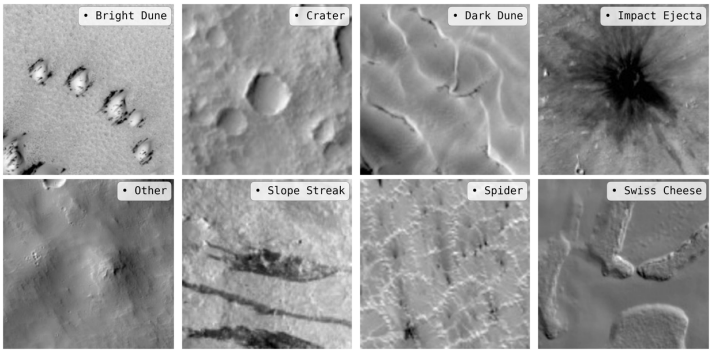


Figure 7. Landmark

Frost This is a binary classification dataset designed to detect the presence or absence of surface frost in Mars satellite imagery. The dataset consists of *HiRISE* images labeled as either "**Frost**" or "**Non Frost**" (Figure 8). Among all datasets in Mars-Bench, this is the largest in terms of the # of samples, and the dataset is well-balanced in terms of class distribution.

Saturated Task Apart from the tasks described above, we exclude the mb-change_cls task from our study, as both of its available versions, HiRISE and CTX, are already saturated. In prior benchmarks and in MOMO, this task consistently reaches near-perfect performance. Although the task exists in both HiRISE and CTX variants, the CTX version additionally suffers from an insufficient number of test samples for statistically meaningful evaluation. For completeness, we only evaluate the mb-change_cls_hirise dataset, but we do not include it in our core experiments.

mb-change_cls_hirise This dataset is designed for binary classification of surface changes using temporal image pairs; specifically, one image taken before and another after some time period, from the *same* Martian location. The task involves identifying whether meaningful surface change has occurred and classifying between "**Change**" and "**No change**". Unlike standard single-image classification, this task requires forming a composite input from two grayscale images (Figure 9).

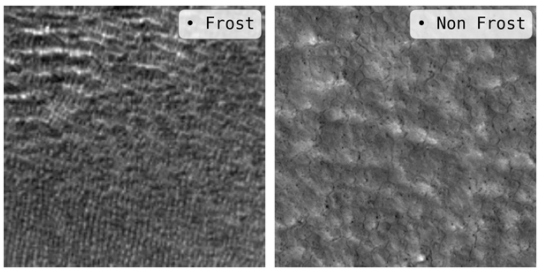
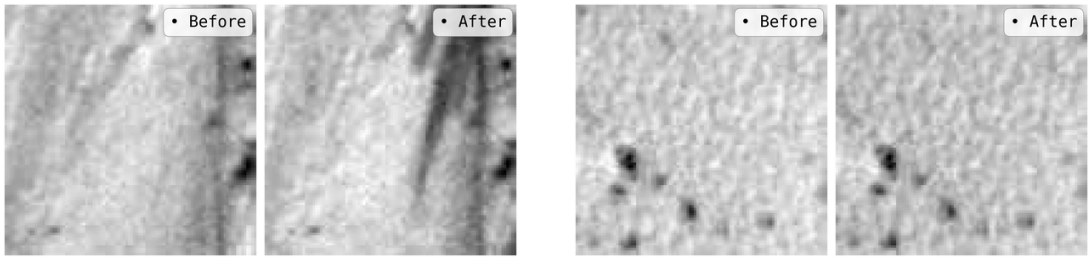


Figure 8. Frost

Following the approach outlined by Kerner et al. [10], we adopt the composite grayscale method: the blue channel encodes the "before" image, the green channel encodes the "after" image, and the red channel is set to zero.



(a) Change (b) No change

Figure 9. mb-change_cls_hirise

For the *mb-change_cls_hirise* dataset, we conducted experiments using MOMO and all baseline models, excluding EO-FMs and DINOv3. All models achieved 100% accuracy and F1-score, indicating that the task is already saturated. Therefore, we did not include these results in the main paper and did not perform further experiments on EO-FMs for this dataset.

A.2.2. Segmentation

Boulder This is a binary segmentation dataset focused on segmenting boulders on the Martian surface using high-resolution orbital imagery from the HiRISE sensor. The dataset comprises manually annotated binary masks indicating the presence or absence of boulders within each image (Figure 10). Boulders were annotated by planetary scientists using precise polygon outlines, ensuring high-quality labels. This is one of the smallest datasets in Mars-Bench, with only tens of samples (i.e., 39), and that makes it challenging for the computer vision community.

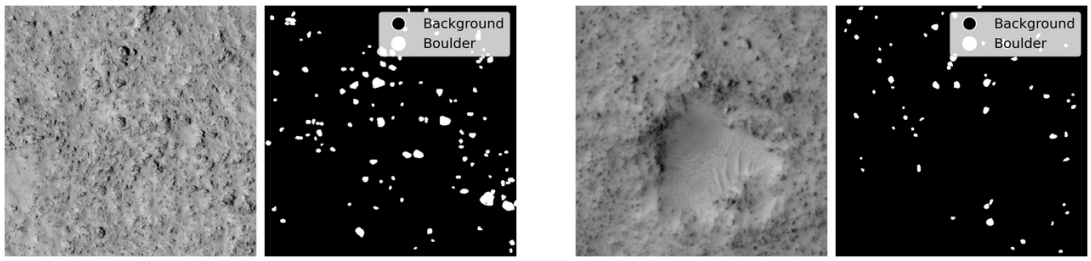


Figure 10. Boulder

ConeQuest This is a binary segmentation dataset focused on identifying volcanic cones on the Martian surface using CTX imagery. It was developed to support global mapping and morphologic analysis of small-scale volcanic landforms. The dataset spans three geographically diverse regions on Mars, capturing substantial variation in cone shape, size, and appearance, making it a challenging benchmark for model generalization. Each sample consists of an image and its corresponding

binary mask (Figure 11), with all annotations created and validated by expert geologists to ensure scientific accuracy. Particularly, the dataset includes negative samples (images without any cones), which introduces additional complexity by requiring models to correctly predict true negatives rather than detecting cones in every image.

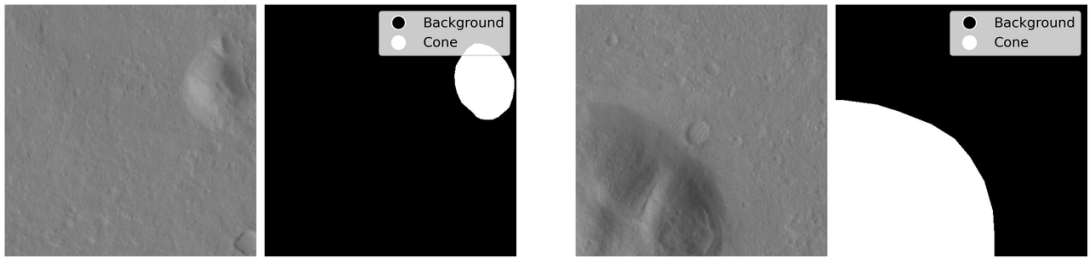


Figure 11. ConeQuest

MMLS This is a binary segmentation dataset designed to identify landslides on the Martian surface, with a focus on the Valles Marineris region from the CTX sensor. All annotations were manually created by expert geologists, ensuring high-quality, scientifically accurate labels. Each image sample includes multi-modal satellite data comprising 7 channels: RGB (3), Digital Elevation Model (DEM), thermal inertia, slope, and grayscale intensity (Figure 12 visualizes grayscale channels only). This rich set of modalities captures the complex geomorphology of landslide-prone regions, making the dataset especially valuable for developing and benchmarking robust segmentation models in planetary science. All experiments in this paper utilize only the RGB channels for training and evaluation.

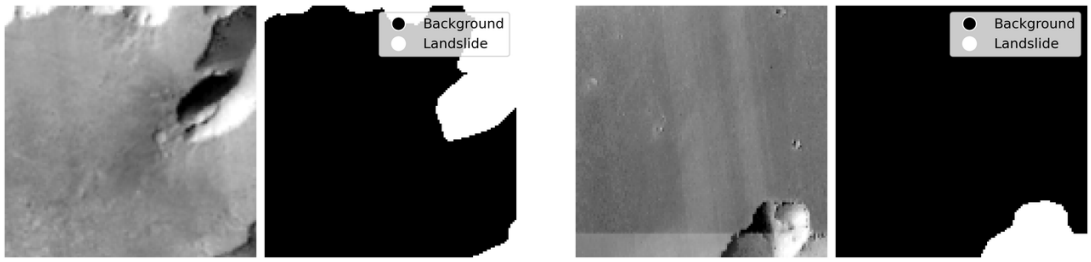
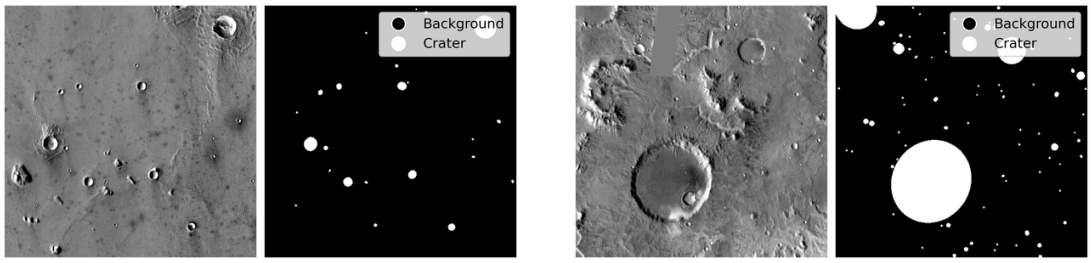


Figure 12. MMLS

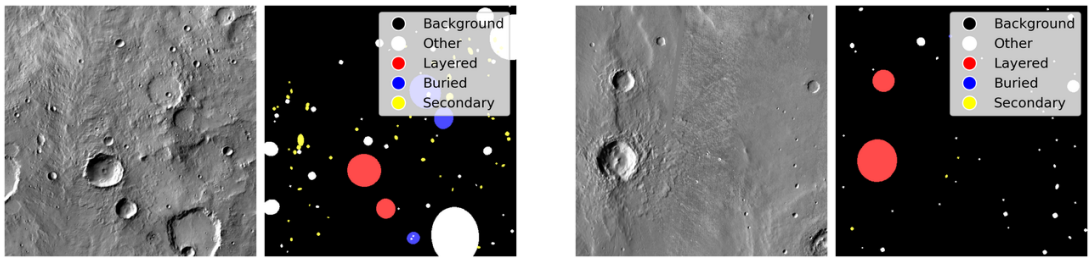
Crater Binary & Crater Multi These two datasets focus on crater segmentation using THEMIS imagery. In particular, mb-crater_binary_seg is a binary segmentation dataset that distinguishes **crater** vs. **non-crater** regions, while mb-crater_multi_seg is a multi-class segmentation dataset with four crater types: **Other**, **Layered**, **Buried**, and **Secondary** (Figure 13).

B. Experiments Details

Pre-training Experiments. All pre-training experiments are conducted on the ViT-Base model on a single NVIDIA A100 GPU with a batch size of 256 at the JPL computing infrastructure. We apply only a random horizontal flip as data augmentation during training and use no augmentation for validation. Models are pre-trained with a learning rate of 10^{-3}, a weight decay of 0.05, a patch size of 16, and a mask ratio of 0.75. For each sensor-specific dataset, we train the model for 5 epochs. We record the model state and loss values after every $100k$ processed samples, enabling consistent comparison of validation loss across all individually pre-trained models. During pre-training, all loss weights λ_i are set to 0.25, ensuring equal weightage to pixel-based and perceptual loss. For loss alignment, we use a patience of 5 and a tolerance parameter of $\epsilon = 10^{-4}$. We analyze the effect of different values of the tolerance parameter in Section C.4. During model merging, we apply a scaling coefficient of 0.3, following the recommendation of Ilharco et al. [8]. We further analyze the sensitivity of our approach to different scaling coefficients in Section C.3. For the Data Merge experiments, we apply the same hyperparameter



(a) Crater Binary



(b) Crater Multi

Figure 13. Crater Segmentation Datasets

configuration. For the ImageNet-pretrained baseline, we use the model provided by He et al. [7]. During pre-training, a ViT-Base model requires approximately 12 hours to train on ∼4M samples for each individual sensor. In contrast, pre-training a ViT-Base model using the Data Merge ($\sim 12M$ data samples in pre-training) setup takes approximately 35 hours.

Downstream Tasks Experiments. For all downstream classification and segmentation tasks, we perform extensive hyperparameter tuning for each model–dataset combination. For classification, a linear layer is applied on top of the pre-trained encoder, whereas segmentation uses a U-NetFormer decoder. All classification datasets use cross-entropy loss, while segmentation employs a weighted combination of Dice, cross-entropy, and boundary losses. Because certain datasets are highly imbalanced (e.g., Landmark), we apply dataset-specific balancing strategies: no balancing for AtmosDust and Frost (nearly balanced), loss reweighting for DoMars16k, and oversampling for Landmark. For all segmentation tasks, we adopt loss reweighting, as background pixels dominate the ground-truth masks.

All models are trained for up to 100 epochs with an early-stopping patience of $5, 10$. We perform a sweep over hyperparameters: learning rates $\in 1 \times 10^{-3}, 1 \times 10^{-4}$, weight decays $\in 5 \times 10^{-2}, 1 \times 10^{-1}$, layer decays $\in 0.5, 0.6, 0.75$, and warm-up epochs $\in 0, 5, 10$. For segmentation, the loss-weighting coefficients are tuned using two settings: $(\text{Dice}, \text{CE}, \text{Boundary}) = (0.5, 0.2, 0.3)$ and $(0.3, 0.5, 0.2)$.

For the DINOv3 model, we use the variant pre-trained on Earth satellite data, specifically the SAT-493M dataset. For the remaining EO-FMs, most do not provide an end-to-end fine-tuning reference codebase for downstream tasks, so we implement our own framework for both classification and segmentation.

To ensure robustness, we run each experiment five times with different random seeds and report the mean and standard deviation. All downstream experiments are conducted on A100 GPUs on ASU [9] or JPL servers, depending on GPU availability.

C. Extended Results

In this section, we present additional experiments and analyses that complement the results discussed in the main paper. These include the effect of model size, detailed evaluations of reconstruction quality, the influence of the scaling coefficient, comparison with the model currently deployed in the NASA PDS system, and examples demonstrating MOMO's capability for generating global maps.

MOMO	AtmosDust	DoMars16k	Frost	Landmark	Boulder	ConeQuest	Crater Binary	Crater Multi	MMLS
ViT-Small	**0.96**	0.92	0.96	0.92	**0.22**	0.71	0.54	0.09	0.58
ViT-Base	**0.96**	**0.93**	**0.97**	**0.93**	0.18	0.72	0.56	0.12	0.58
ViT-Large	**0.96**	0.92	0.96	**0.93**	0.19	**0.73**	**0.58**	**0.14**	**0.60**

Table 1. Performance comparison of ViT variants. Reported metrics include F1-Score for classification tasks, and mIoU for segmentation tasks. **Bold** numbers indicate the highest value in each column.

C.1. Effect of Model Size

To evaluate the robustness of our proposed approach across different model capacities, we conducted experiments using three Vision Transformer (ViT) variants: ViT-Small, ViT-Base, and ViT-Large. Each variant was pre-trained and evaluated under the same setup across all downstream tasks to examine how model size influences performance. The results are summarized in Table 1.

From the results, we observe that in classification tasks, the performance difference across all three ViT variants is negligible, typically less than 1%. However, in segmentation tasks, increasing model size clearly improves performance, with ViT-Large achieving the best results in most cases. An exception is observed in the *Boulder* dataset, where ViT-Small outperforms larger models. This can be attributed to the small size of the dataset and the limited number of samples per class, which may lead to overfitting in larger models. Overall, these results indicate that while classification remains largely invariant to model capacity, segmentation benefits significantly from increased model size.

C.2. Reconstruction

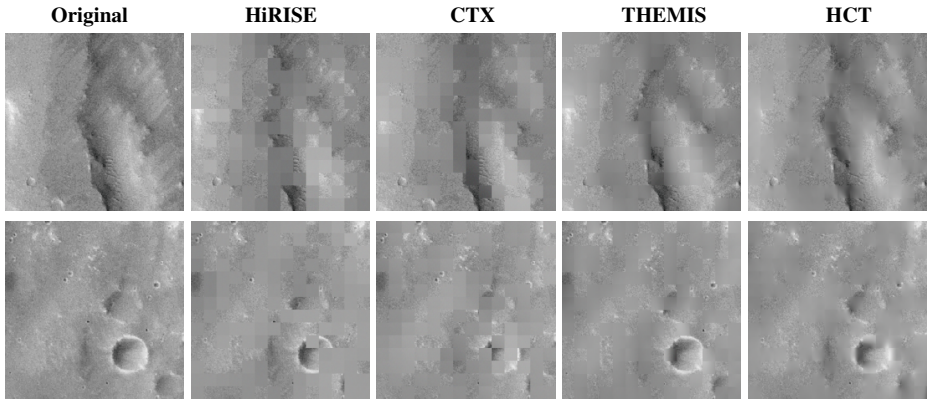


Figure 14. Reconstruction results using ViT-Base models pre-trained with only MSE loss. The figure compares the Original image against reconstructions from sensor-specific models (HiRISE, CTX, THEMIS) and the Data Merge model (HCT). The top row displays a HiRISE sample and the second row displays a CTX sample.

As described in Section B, our pre-training objective combines pixel-based loss with a perceptual loss. In this section, we evaluate the impact of this formulation by comparing it against a baseline that uses only MSE loss. Figure 14 illustrates reconstruction results when ViT-Base is pre-trained on each sensor independently as well as using the Data Merge approach. We show one randomly selected HiRISE sample (top row) and one CTX sample (bottom row). Under the MSE-only objective, several patches are poorly reconstructed: the model often recovers the overall surface tone but fails to regenerate fine-scale geomorphological features. For example, in the CTX example (second row), when $\sim 20\%$ of the crater is masked, the model reconstructs the surrounding terrain reasonably well but is unable to recover the crater structure itself.

In contrast, Figure 15 shows reconstructions from models pre-trained using our proposed combined loss. We visualize two samples from each of the three sensors. Across all sensors, the reconstructions capture not only the correct color distribution but also the underlying surface morphology with substantially higher clarity. These results highlight the effectiveness of our loss formulation in guiding the model to learn feature-aware representations that preserve critical geomorphological structures.

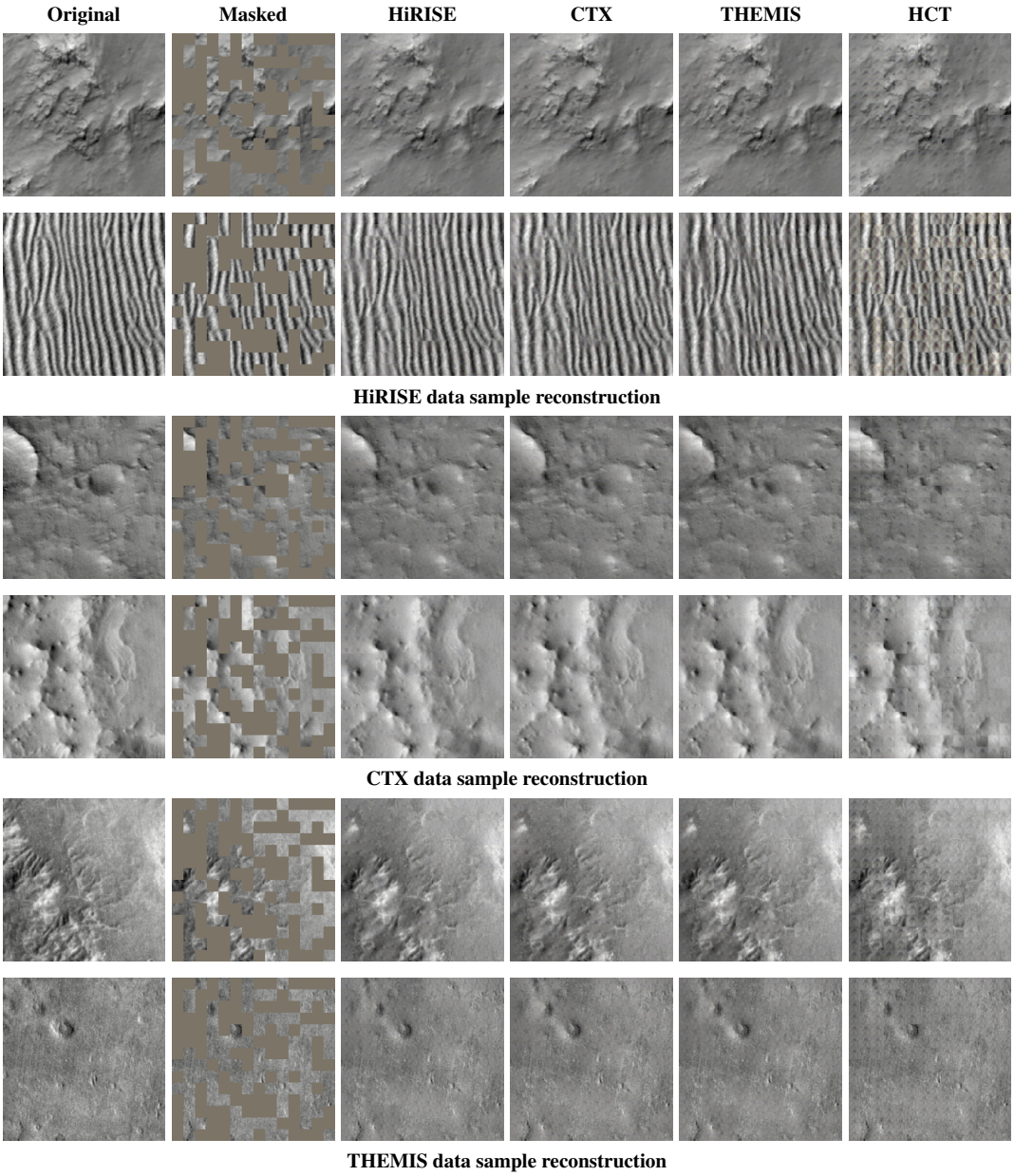


Figure 15. Reconstruction results using models pre-trained with the proposed combined loss function (pixel-based + perceptual). This figure visualizes reconstructions for data samples from all three sensors: HiRISE (top rows), CTX (middle rows), and THEMIS (bottom rows). The columns display the Original image, the Masked input, and the outputs from the individual sensor models and the HCT model.

C.3. Scaling coefficient

To analyze the sensitivity of our method to the scaling coefficient used during model merging, we conducted experiments by varying the coefficient from 0.1 to 1.0 in increments of 0.1. These experiments were performed only on downstream tasks that showed significant differences compared to baselines and among different checkpoint selection strategies. Hence, binary classification datasets and *Boulder* and *ConeQuest* segmentation tasks were excluded.

Figure 16 presents the results for both classification and segmentation tasks, where we report the F1-Score for classification and mIoU for segmentation. As shown in the figure, the performance of the proposed approach remains largely stable across different scaling coefficients, indicating that our method is not highly sensitive to this parameter. This observation is consistent with the findings reported by Ilharco et al. [8]. Additionally, as the scaling coefficient increases beyond a certain threshold, performance decreases across most datasets, indicating that excessively high scaling values are not beneficial, again consistent with Ilharco et al. [8].

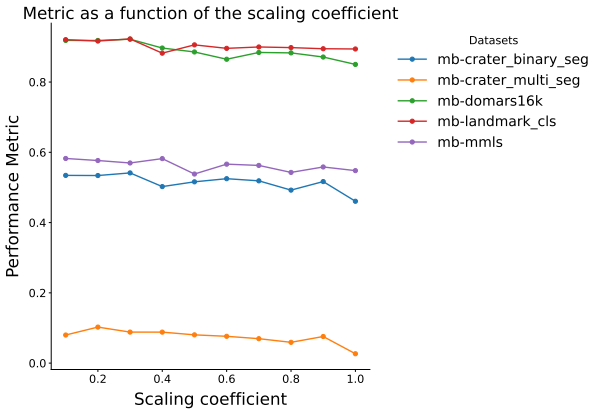


Figure 16. Performance as a function of the scaling coefficient on classification and segmentation downstream tasks.

C.4. Ablation on Tolerance hyperparameter (ϵ)

To evaluate the sensitivity of our method to the tolerance hyperparameter (ϵ), we conduct experiments by varying its value to 10^{-2} and 10^{-3}, and compare these results with the default setting of 10^{-4}. The results are reported in Table 2. We observe that changing ϵ has minimal impact on performance across most datasets, with results either remaining consistent or improving slightly by 1–2%. The only exception is the *ConeQuest* dataset, where performance decreases marginally; however, the drop is limited to approximately 2%, indicating that the method remains robust to variations in ϵ.

ϵ	DoMars16k	Landmark	ConeQuest	Crater Multi
10^{-2}	0.92	0.92	0.70	0.15
10^{-3}	0.93	0.94	0.69	0.15
10^{-4}	0.92	0.91	0.71	0.14

Table 2. Results for different values of the tolerance hyperparameter (ϵ).

	DoMars16k	Landmark	ConeQuest	Crater Multi
(H + C) + T	0.92	0.93	0.69	0.15
MOMO	0.92	0.91	0.71	0.14

Table 3. Results for incremental sensor merging, where a THEMIS model is merged with an existing HiRISE and CTX model ((H + C) + T), compared with MOMO.

C.5. Merging New Modality

To evaluate how performance is affected when incorporating a new sensor, we conduct an experiment simulating incremental sensor addition. In this setup, we assume access to independently trained models along with their validation loss trajectories. We first consider models trained on HiRISE and CTX as existing sensors, and then introduce THEMIS as a new sensor modality. Based on the validation trajectory of the THEMIS model, we select the checkpoint whose validation loss is closest to that of the existing models and merge it accordingly.

Due to computational constraints, we report results on two classification datasets and two segmentation datasets. The results are summarized in Table 3. We observe that incorporating the new sensor does not significantly affect performance, with changes remaining within $\pm$1-2% across all evaluated tasks.

C.6. Research Impact

In this section, we discuss real-world use cases of **MOMO**.

C.6.1. Comparison with PDS deployed Model

The NASA Planetary Data System (PDS) archives data from planetary science missions, and its Cartography and Imaging Sciences Node (Imaging Node) provides public access to millions of planetary images. To help scientists search for images based on visual content rather than metadata alone, the Imaging Node introduced a content-based image search capability in 2017. This system, developed using machine learning classification techniques by Wagstaff et al. [16], enables researchers to efficiently identify images relevant to their investigations.

We compare **MOMO** with the model currently deployed at NASA's Planetary Data System (PDS) [16], focusing on the landmark classification dataset used by the PDS Imaging Node. As shown in Table 4, **MOMO** outperforms the PDS model across most classes, achieving higher F1-scores in seven out of eight categories and improving the overall macro-average by 4%. Notably, MOMO shows significant improvements in *Slope Streak*, *Impact ejecta*, and *Swiss cheese*, with gains of 11%,

	Bright dune	Crater	Dark dune	Impact ejecta	Other	Slope Streak	Spider	Swiss cheese	Macro Avg
PDS	0.86	**0.79**	0.87	0.30	**0.96**	0.67	0.04	0.94	0.68
MOMO	**0.90**	0.75	**0.91**	**0.40**	**0.96**	**0.78**	**0.05**	**0.99**	**0.72**

Table 4. Per-class F1-scores for PDS and MOMO models on the PDS dataset. **Bold** numbers indicate the higher F1-score for each class.

10%, and 5%, respectively, demonstrating its effectiveness in capturing complex surface morphologies and fine-grained Martian features. Although the PDS model performs slightly better on the *Crater* class, MOMO achieves more balanced and consistent performance across diverse geologic feature types, making it a stronger candidate for large-scale automated mapping and planetary data analysis.

C.6.2. Creating Global Maps

Scientists and planetary geologists are interested in studying geologic features on Mars and understanding their global distribution. To achieve this, they typically create small labeled datasets and train machine learning models to generate global maps of specific features. Given its strong segmentation performance, **MOMO** can serve as an effective tool for producing such large-scale global maps of Martian surface features.

To demonstrate the efficiency and practical utility of **MOMO**, we perform inference on the *ConeQuest* dataset using out-of-distribution (OOD) data. To replicate this process, we exported new data from JMARS [4]. JMARS (Java Mission-planning and Analysis for Remote Sensing) is a geospatial information system developed to visualize, analyze, and export planetary data from multiple Mars missions, focusing on regions not included in the original training set.

Each data tile in *ConeQuest* provides latitude and longitude information, which allowed us to select a previously unseen region centered at 15° latitude and 84° longitude. We exported CTX imagery covering an area of approximately 1.5 km × 1.5 km (12288 x 12288 pixels), sampled into 512 × 512 pixel tiles with an overlap of 256 pixels, resulting in a total of 2,306 image samples.

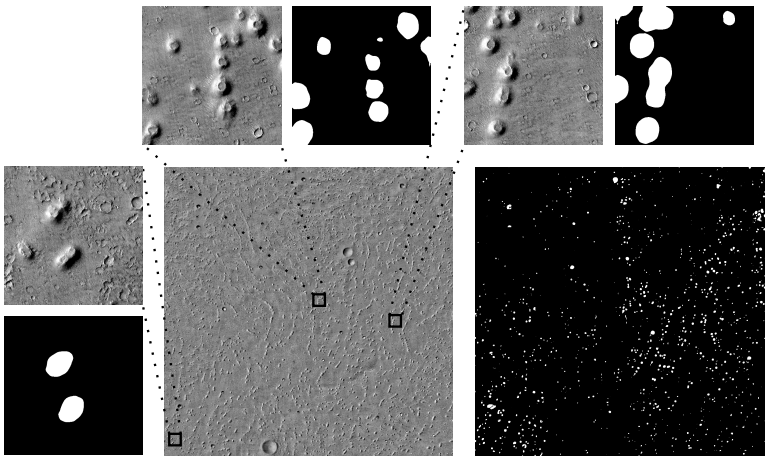


Figure 17. Example of global map generation using **MOMO** on the out-of-distribution region of the *ConeQuest* dataset. The center panel shows the original large-scale HiRISE tile, and the right panel shows the stitched prediction map after inference. The left and top panels display representative 512×512 data samples and their corresponding segmentation outputs. This experiment demonstrates MOMO's capability to generalize to unseen regions and its potential for large-scale planetary surface mapping.

Figure 17 illustrates an example of this experiment. The left panel shows the original large-scale tile, and the right panel shows the stitched output generated after performing inference with MOMO. For reference, we also display a few example 512 × 512 tiles used for prediction. These results demonstrate that MOMO can be effectively used to produce global-scale maps of geologic features from unseen regions, highlighting its potential for planetary-scale mapping applications.

References

[1] California Institute of Technology - Division of Geological and Planetary Sciences. The Bruce Murray Laboratory for Planetary Visualization. http://murray-lab.caltech.edu/CTX/. 1

[2] PR Christensen, NS Gorelick, GL Mehall, and KC Murray. Mars odyssey thermal emission imaging system infrared reduced data record. Technical report, ODY-M-THM-5-IRRDR-V1. 0.[Dataset]. NASA Planetary Data System. https://pds . . . , 2001. 2

[3] Philip R Christensen, Bruce M Jakosky, Hugh H Kieffer, Michael C Malin, Harry Y McSween Jr, Kenneth Nealson, Greg L Mehall, Steven H Silverman, Steven Ferry, Michael Caplinger, et al. The thermal emission imaging system (themis) for the mars 2001 odyssey mission. *Space Science Reviews*, 110(1):85–130, 2004. 1

[4] P. R. Christensen, E. Engle, S. Anwar, S. Dickenshied, D. Noss, N. Gorelick, and M. Weiss-Malik. Jmars – a planetary gis. http://adsabs.harvard.edu/abs/2009AGUFMIN22A..06C, 2009. NASA/JPL-Caltech/Arizona State University. 11

[5] JL Dickson, LA Kerber, CI Fassett, and BL Ehlmann. A global, blended ctx mosaic of mars with vectorized seam mapping: A new mosaicking pipeline using principles of non-destructive image editing. In *Lunar and planetary science conference*, pages 1–2. Lunar and Planetary Institute The Woodlands, TX, USA, 2018. 1

[6] JL Dickson, BL Ehlmann, LH Kerber, and CI Fassett. Release of the global ctx mosaic of mars: An experiment in informationpreserving image data processing. In *54th Lunar and Planetary Science Conference*, pages 1–2, 2023. 1

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 7

[8] 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. 6, 10

[9] Douglas M. Jennewein, Johnathan Lee, Chris Kurtz, William Dizon, Ian Shaeffer, Alan Chapman, Alejandro Chiquete, Josh Burks, Amber Carlson, Natalie Mason, Arhat Kobawala, Thirugnanam Jagadeesan, Praful Bhargav Basani, Torey Battelle, Rebecca Belshe, Deb McCaffrey, Marisa Brazil, Chaitanya Inumella, Kirby Kuznia, Jade Buzinski, Dhruvil Deepakbhai Shah, Sean M. Dudley, Gil Speyer, and Jason Yalim. The sol supercomputer at arizona state university. In *Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good*, page 296–301, New York, NY, USA, 2023. Association for Computing Machinery. 7

[10] 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. 5

[11] Michael C Malin, James F Bell III, Bruce A Cantor, Michael A Caplinger, Wendy M Calvin, R Todd Clancy, Kenneth S Edgett, Lawrence Edwards, Robert M Haberle, Philip B James, et al. Context camera investigation on board the mars reconnaissance orbiter. *Journal of Geophysical Research: Planets*, 112(E5), 2007. 1

[12] Alfred S McEwen, Shane Byrne, C Hansen, Ingrid J Daubar, Sarah Sutton, Colin M Dundas, Nicole Bardabelias, Nicole Baugh, J Bergstrom, R Beyer, et al. The high-resolution imaging science experiment (hirise) in the mro extended science phases (2009–2023). *Icarus*, 419:115795, 2024. 1

[13] Elena Plekhanova, Damien Robert, Johannes Dollinger, Emilia Arens, Philipp Brun, Jan Dirk Wegner, and Niklaus E. Zimmermann. Ssl4eco: A global seasonal dataset for geospatial foundation models in ecology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2428–2439, 2025. 1

[14] Mirali Purohit, Bimal Gajera, Vatsal Malaviya, Irish Mehta, Kunal Sunil Kasodekar, Jacob Adler, Steven Lu, Umaa Rebbapragada, and Hannah Kerner. Mars-bench: A benchmark for evaluating foundation models for mars science tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. 3

[15] Mirali Purohit, Gedeon Muhawenayo, Esther Rolf, and Hannah Kerner. How does the spatial distribution of pre-training data affect geospatial foundation models? In *Workshop on Preparing Good Data for Generative AI: Challenges and Approaches*, 2025. 1

[16] 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*, pages 15204–15213, 2021. 10