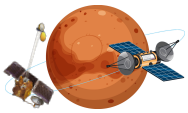
This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.



MOMO: Mars Orbital MOdel Foundation Model for Mars Orbital Applications

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

Abstract

We introduce MOMO, the first multi-sensor foundation model for Mars remote sensing. MOMO uses model merge to integrate representations learned independently from three key Martian sensors (HiRISE, CTX, and THEMIS), spanning resolutions from 0.25 m/pixel to 100 m/pixel. Central to our method is our novel Equal Validation Loss (EVL) strategy, which aligns checkpoints across sensors based on validation loss similarity before fusion via task arithmetic. This ensures models are merged at compatible convergence stages, leading to improved stability and generalization. We train MOMO on a large-scale, high-quality corpus of $\sim$ *12 million samples curated from Mars orbital data and evaluate it on 9 downstream tasks from Mars-Bench. MOMO achieves better overall performance compared to ImageNet pre-trained, earth observation foundation model, sensor-specific pre-training, and fully-supervised baselines. Particularly on segmentation tasks, MOMO shows consistent and significant performance improvement. Our results demonstrate that model merging through an optimal checkpoint selection strategy provides an effective approach for building foundation models for multi-resolution data. The model weights, pretraining code, pretraining data, and evaluation code are available at: github.com/kerner-lab/MOMO.*

1. Introduction

Foundation models (FMs) have demonstrated strong capability in learning representations from large-scale data, enabling improved downstream task performance compared to conventional supervised training from scratch [25, 36]. In recent years, more than 150 FMs have been proposed for Earth observation (EO) applications [25, 36]. These EO-FMs are being actively used in applications such as food security, disaster response, and climate change [9, 56].

Similar to Earth-orbiting satellites, Mars-orbiting satellites systematically collect remote sensing observations of the planet's surface and atmosphere. In contrast to the active research on EO-FMs in recent years, no FM has been proposed for Mars remote sensing applications to date. An FM for Mars remote sensing would enable planetary scientists to train models for custom science tasks at a lower cost compared to fully supervised methods. Researchers are already using ImageNet pre-trained models for Mars remote sensing applications [50, 66, 67], but in-domain pre-training could improve performance and generalization, as suggested by preliminary findings in [51].

Developing FMs for remote sensing data requires significant domain expertise and computational resources [55]. Satellite images are acquired from multiple sensors, each operating at different wavelengths, spatial resolutions, and spectral channels. The EO community has proposed customized model architectures specifically designed for Earth satellite data and applications [4, 22, 63]. While effective for EO data, these approaches do not extend directly to Mars remote sensing data due to differences in sensor properties and availability.

We propose **MOMO** (**M**ars **O**rbital **MO**del), the first FM for Mars remote sensing applications. We introduce a novel approach to handle data efficiently from multiple sensors that measure different physical properties at different spatial scales. We pre-train individual models on data from each sensor and subsequently merge them using our novel checkpoint selection strategy. We evaluate MOMO on all 9 orbital downstream tasks from Mars-Bench [52]. We compare our proposed method to a range of strong baselines. Overall, MOMO outperforms models pre-trained on Earth Observation data, sensor-specific Mars data, and ImageNet, demonstrating superior performance on segmentation tasks

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

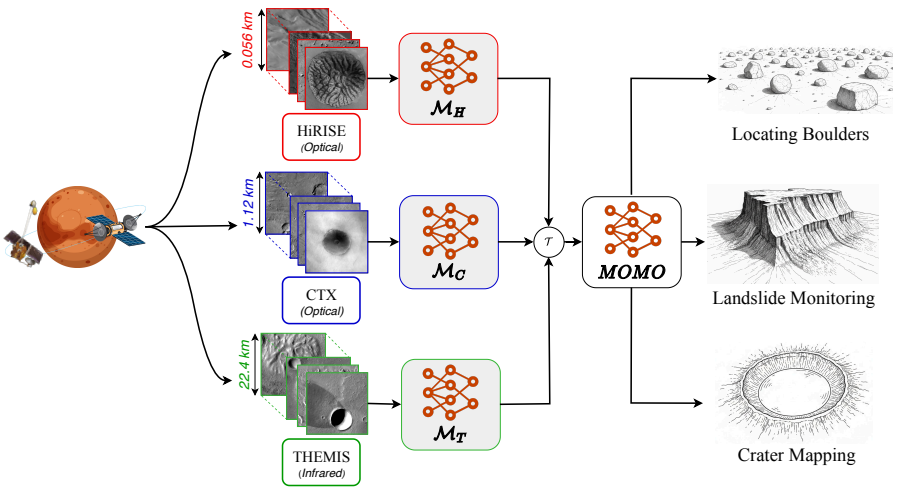


Figure 1. MOMO can be effectively applied across a wide range of resolutions and a broad spectrum of Martian remote sensing tasks. By leveraging diverse sensors, our approach enables a single model to generalize across different orbital applications, including large-scale crater or landslide mapping and precise boulder localization.

and comparable performance on classification. In summary, our main contributions are as follows:

- We introduce MOMO, the first foundation model for Mars orbital applications. MOMO efficiently handles multi-sensor and multi-resolution data.
- We propose a novel technique to build a multi-sensor foundation model through model merging, using our optimal checkpoint selection strategy for stable fusion.
- We conduct extensive comparisons of MOMO against multiple baselines, including ImageNet pre-training, Earth-observation foundation models, sensor-specific pre-training, and other checkpoint selection strategies. Our results on Mars-Bench demonstrate that MOMO achieves superior overall performance across tasks.

2. Related Work

Self-Supervised Learning for Mars Orbital Tasks. There are a few preliminary studies that have explored self-supervised learning for Mars remote sensing. Jiang et al. pre-trained a model on just 13 HiRISE images for landform detection [29], but their focus is only on the landmark detection task. In contrast, Purohit et al. pre-trained a model on 1 million CTX image patches and demonstrated that in-domain pre-training can surpass ImageNet pre-training for certain tasks [51]. However, this work only uses data from a single instrument and evaluates on 2 downstream tasks.

Foundation Models for Earth Observation. Over the past 4-5 years, researchers have introduced numerous FMs for EO applications. Many of these include masked autoencoder-based models such as SatMAE [11], ScaleMAE [54], SatMAE++ [47], Presto [63]; contrastive learning-based models like SeCo [39], SatCLIP [31], GeoCLIP [65], CROMA [15]; and other approaches including msGFM [22], AnySat [4], Galileo [64], Satlas [5], SkySense [21], SpectralGPT [24], and SpectralEarth [8]. Earth and Mars remote sensing data have some similarities, such as multispectral/multi-sensor observations and overhead imaging geometry, but they are very different. Mars has different atmospheric conditions, illumination, surface materials, and sensor characteristics. We would not expect EO-FMs to generalize to Mars remote sensing tasks.

Model-editing. Model-editing methods aim to enhance generalization, robustness, and out-of-distribution performance without incurring extra inference costs. Model editing has been applied across various domains and tasks. Techniques include merging weights of multiple fine-tuned models via simple weight averaging or ensembling, layer-wise matching, and soft alignment; to create a single model that outperforms its individual components. There are dozens of proposed research methods based on model averaging [13, 20, 28, 30, 33, 34, 60, 69] and ensembling techniques, combining the outputs of two or more models [6, 14, 32, 42, 49, 68]. Researchers have studied more efficient methods for modifying a model's behavior through interventions after pre-training, referring to this process by different names, such as patching [18, 27, 45, 59], editing [43, 44, 57], aligning [3, 48], and layer-wise editing [58].

Prior work has focused on merging models trained on similar or different distributions, but limited focus has been given to checkpoint selection before merging. Existing methods typically merge models at their final checkpoints without considering differences in training trajectories.

In contrast, we introduce a novel checkpoint selection strategy based on validation loss alignment, which ensures models trained on different data distributions (in our case, varying spatial resolutions) are merged at their most compatible stage. To the best of our knowledge, this is the first task-arithmetic approach to introduce a systematic checkpoint selection strategy.

3. MOMO

3.1. Motivation

Before describing our methodology, we first provide a brief overview of how foundation models are typically developed in Earth Observation and how our approach differs. Most EO-FMs take one of two approaches to combine data from multiple sensors: 1) stacking spatially- and temporally-aligned data from each sensor as different channels to a single encoder or a separate tokenizer for each sensor (e.g., [11, 63, 64]), or 2) combining data from multiple sensors in a single heterogeneous pre-training dataset (e.g., [5]).

The stacking approach requires a large number of coincident (spatially and temporally overlapping) observations from each sensor and sensors with somewhat similar spatial resolutions (e.g., 10-30 m/pixel). Mars orbital sensors have very different coverage (e.g., CTX covers nearly 100% of the planet [12] but HiRISE only covers less than 3% [41]) and spatial resolutions (e.g., THEMIS is 100 m/pixel compared to HiRISE at 0.25 m/pixel; see Figure 1). The data-combination approach is feasible, but would require training a new model whenever a new sensor is added.

To address these challenges, we propose a methodology that avoids directly combining heterogeneous data from multiple sensors. Instead, we pre-train independent masked autoencoder models for each sensor, allowing each model to first learn the unique distribution and feature characteristics of its respective sensor. We then merge these sensor-specific models (Figure 1) using our proposed Equal Validation Loss (EVL) strategy, which aligns checkpoints based on validation loss similarity to ensure compatibility before fusion. This is the first work to employ a model-merging strategy to construct a remote sensing foundation model. Additionally, we introduce a customized cost function designed to optimize the reconstruction process by capturing both pixel-level and perceptual information. Full methodological details are provided in the following sections.

3.2. Cost Function

Loss functions play a key role in training deep learning models. An unsuitable objective function can lead to convergence toward suboptimal local minima or undesirable optimization directions.

Following widely adopted practices for training Masked AutoEncoders (MAEs) [23], we initially pre-trained our model using mean squared error (MSE) as the reconstruction objective. However, after visualizing the reconstructed outputs, we observed that while the model could accurately recover the color distribution and surface textures of masked regions, it often failed to reconstruct structural details of key geomorphologic features (e.g., the accurate shape of a crater) when such regions were masked (sample reconstructions in Appendix C.3).

This limitation arises because MSE is a pixel-level loss, emphasizing low-level intensity matching rather than perceptual or structural fidelity. Although MSE effectively captures color and tone consistency, it lacks sensitivity to higher-order spatial features such as shape, boundary continuity, and object geometry.

To address this, we introduce additional perceptual and structure-aware components in our loss function to guide the model toward learning edge-level and shape-consistent representations. We add terms that minimize LPIPS [71] and maximize structural similarity (SSIM) [46]. In addition, to enforce spatial smoothness and structural consistency, we penalize the difference between horizontal and vertical gradients of predicted and ground-truth images [37, 58]. For an image I and its reconstruction $\hat{I}$, the gradient loss is formulated as an ℓ_1 penalty:

$$\mathcal{L}_{\text{grad}} = \frac{1}{N} \sum_{i,j} \left(\left| \partial_x I_{i,j} - \partial_x \hat{I}_{i,j} \right| + \left| \partial_y I_{i,j} - \partial_y \hat{I}_{i,j} \right| \right) \tag{1}$$

The final combined pre-training objective function is:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{MSE}} + \lambda_2 \mathcal{L}_{\text{SSIM}} + \lambda_3 \mathcal{L}_{\text{LPIPS}} + \lambda_4 \mathcal{L}_{\text{grad}}, \tag{2}$$

where $\mathcal{L}_{\text{SSIM}}$ is $(1 - SSIM)$ and λ_i are weighting coefficients. By combining pixel-wise, perceptual, and gradient-aware objectives, MOMO achieves reconstructions that preserve both spatial and structural details, particularly important for Martian features such as crater rims, cones, and landslide boundaries.

3.3. Optimal Checkpoint Selection Strategy

We define our checkpoint selection strategy as follows: Let there be n distinct sensors, each associated with a dataset $\mathcal{D}_i$ $(i = 1, 2, \ldots, n)$ representing distinct spatial resolutions, modalities, or imaging characteristics. We train n independent models $\{\mathcal{M}_i\}_{i=1}^{n}$, where each $\mathcal{M}_i$ is optimized on $\mathcal{D}_i$ for k epochs denoted as $E = \{e_1, e_2, e_3, ..., e_k\}$. During training, we record the validation loss $\mathcal{L}_i^{(e)}$ for every epoch $e \in E$. Here, each model is pre-trained by optimizing the loss, which is defined in Equation 2.

Instead of merging all models at their final checkpoints, we introduce the **Equal Validation Loss (EVL)** strategy, which aligns checkpoints across sensors based on validation loss similarity prior to model fusion. This ensures that models trained on heterogeneous data distributions are combined at a mutually compatible convergence stage.

Loss alignment. Let $\mathcal{L}_i^{(e)}$ denote the validation loss of sensor i at epoch e, for $i \in \{1, \ldots, n\}$ and $e \in E$. We define a *candidate epoch tuple*,

$$\mathbf{t_c} = (e^1, e^2, \ldots, e^n),$$

where e^i denotes the selected epoch for model $\mathcal{M}_i$. To create a candidate tuple, we iterate over epoch indices and form candidate epoch tuples by taking one epoch from each sensor and checking whether the losses in that tuple are mutually close. A tuple $\mathbf{t_c}$ is considered *loss-aligned* if the validation losses across all sensors are mutually close, i.e.,

$$\Delta_{ij} = \left| \mathcal{L}_i^{(e_a^i)} - \mathcal{L}_j^{(e_b^j)} \right|$$
$$\forall\, i, j \in \{1, \ldots, n\};\ a, b \in \{1, \ldots, k\}; i \neq j;\ \Delta_{ij} \leq \epsilon$$

where e_a^i represents the a^{th} epoch on the model trained on i^{th} sensor. $\epsilon > 0$ is a small tolerance hyperparameter. The set of all loss-aligned tuples t_c is denoted as $\mathcal{E}_{\text{EVL}}$.

Distance-guided checkpoint selection. For each loss-aligned tuple $\mathbf{t_c} \in \mathcal{E}_{\text{EVL}}$, we measure how far the selected epochs deviate from their respective early-stopping epochs (s_{es}). We define the normalized average epoch distance as

$$\bar{D}(\mathbf{t_c}) \ = \ \frac{1}{n} \sum_{i=1}^{n} | \, e^i - s_{es}^i \, |$$

The optimal tuple is then chosen as

$$\mathbf{t_c}^\star \ = \ \min_{\mathbf{t_c} \in \mathcal{E}_{\text{EVL}}} \bar{D}(\mathbf{t_c})$$

This criterion favors checkpoint tuples that are jointly loss-aligned and closest on average to the individual early-stopping epochs.

The purpose of this selection step is to identify the most representative checkpoint combination for model fusion. Each sensor's early-stopping epoch s_i^{es} corresponds to its best generalization point, as determined by validation performance. Selecting epochs too far from these points increases the risk that one or more sensors contribute checkpoints that are either overfitted (if much later than s_i^{es}) or underfitted (if much earlier). By minimizing the average epoch deviation $\bar{D}(\mathbf{t_c})$, we ensure that the selected tuple $\mathbf{t_c}^\star$ remains close to the generalization-optimal region for all sensors, thereby reducing the likelihood of combining mismatched or unstable model states. This provides a balanced and reliable basis for subsequent model merging.

Model fusion using the optimal tuple. Once the optimal loss-aligned tuple $\mathbf{t_c}^\star = (e_\star^1, e_\star^2, \ldots, e_\star^n)$ is identified, we retrieve the corresponding checkpoints $\{\theta_i^{(e_\star^i)}\}_{i=1}^n$ from each sensor-specific model $\{\mathcal{M}_i\}_{i=1}^n$. These checkpoints represent mutually compatible convergence stages across sensors.

We then merge the selected models using a *task arithmetic* algorithm, which operates directly on the model parameters to form a unified representation. The resulting merged model is denoted as

$$\text{MOMO} = \mathcal{T}\left(\theta_1^{(e_\star^1)}, \theta_2^{(e_\star^2)}, \ldots, \theta_n^{(e_\star^n)} \right),$$

where $\mathcal{T}$ indicates task arithmetic operation. Specifically, we employ the *addition* operation defined in [26] to merge models.

This approach ensures that the fusion process integrates models trained to comparable validation performance levels and located near their respective generalization optima, thereby enhancing stability and reducing the risk of overfitted or underfitted contributions from individual sensors. We illustrate the intuition behind why EVL is more stable and generalizable compared to other checkpoint selection strategies in Section 6.1.

4. Pre-training Data

As discussed in Section 1, MOMO is an orbital foundation model trained on a large-scale, diverse dataset derived from multiple Martian sensors. Specifically, we utilize data from three key sources:

- the High Resolution Imaging Science Experiment (HiRISE) [40] available at **0.25 m/pixels**,
- the ConTeXt Camera (CTX) [7, 38] available at **5 m/pixels**, and
- the THermal EMission Imaging System (THEMIS) [10] available at **100 m/pixel**.

We select these sensors because all the orbital downstream tasks [52] used and evaluated in this study belong to these three sensors. Details of the sensor types, characteristics, and the pre-training data preparation from all three sensors are provided in Appendix A.1.

As THEMIS is a low-resolution sensor, even with full surface coverage, we obtained a total of $\sim 4M$ (millions) images from it. However, HiRISE and CTX contain $\sim 16M$ and $\sim 10M$ images, respectively. To ensure balanced representation, we sample 4M images from each sensor.

To include samples from a wide range of surface types and ensure that HiRISE and CTX retain their original distribution after downsampling to 4M, we proportionally sample data based on the geologic map units from the USGS Scientific Investigations Map 3292 — The Geologic Map of Mars (GMoM) [61]. The GMoM divides Mars into 44 surface geologic units that represent a wide range of ages, morphologies, and compositions. We perform stratified sampling within each GMoM unit proportional to the unit's area coverage for each of the three instrument datasets. This approach ensures that the final dataset not only balances sample counts across sensors but also preserves the geographic and geologic representativeness inherent to the original orbital coverage.

To ensure the quality and reliability of the pre-training corpus, we apply a filtering pipeline to remove low-quality samples containing satellite artifacts, noise, or blur. For

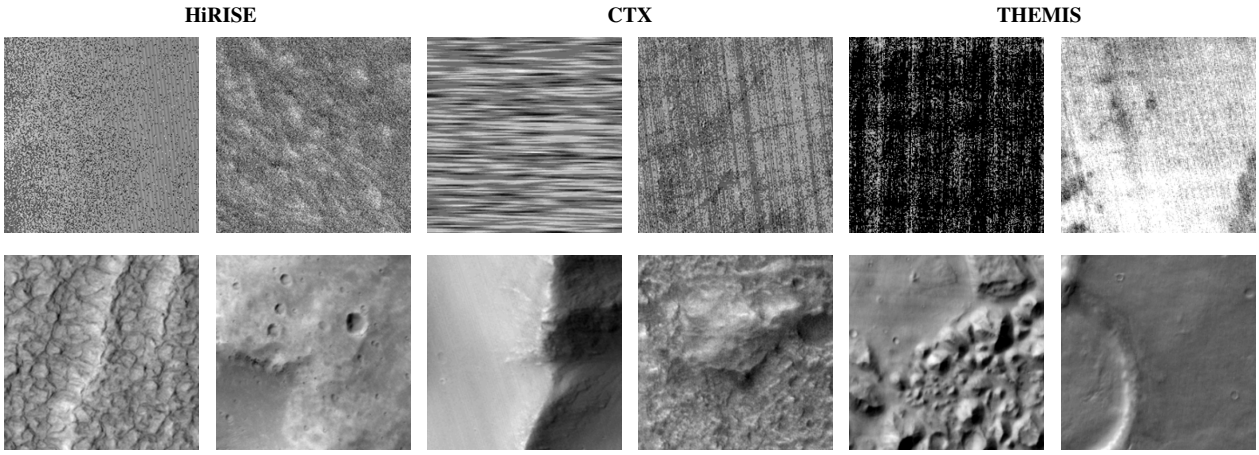


Figure 2. Illustrative samples of poor- and high-quality image samples from the HiRISE, CTX, and THEMIS sensors. The top row shows rejected low-quality samples exhibiting artifacts, blur, or noise, while the bottom row shows high-quality samples retained for pre-training.

each image from the three sensors, we compute two quantitative quality metrics:

- Structural Similarity Index (SSIM) [46]: We apply Gaussian smoothing to each image to reduce high-frequency noise while preserving structural content, and then compute the SSIM between the smoothed and original image.
- Noise Estimate [35]: The noise level is estimated by applying a Laplacian-like high-pass filter to emphasize intensity variations, followed by computing a statistical measure (σ) of the mean absolute filtered image.

Both metrics range from 0 to 1, where lower values indicate poor image quality. We discard samples with values below 0.4 (decided based on human verification) in both metrics for all sensors. This automated filtering step substantially reduces artifacts and ensures that only visually consistent, high-quality images contribute to pre-training. A few examples of good and poor-quality images are shown in Figure 2.

Finally, we split the data from all three sensors into training and validation sets using the HEALPix strategy [19], which partitions a sphere into equal-area cells. These cells are then randomly divided into training and validation subsets. The same set of cells is used across all three sensors to prevent data leakage during model merging. The resulting global distribution and data splits for all three sensors are provided in Appendix A.1. We will publicly release this dataset as open-source to support future research in Mars science and foundation models.

5. Experimental Framework

5.1. Baselines

We evaluate MOMO against a variety of baseline models to assess the effectiveness of our proposed approach. As discussed in Section 1, ImageNet pre-trained models remain the default initialization strategy for many planetary and geospatial applications; therefore, we include an ImageNet pre-trained model as one of our baselines. We also consider a zero-shot setting with randomly initialized weights to quantify the contribution of pre-training, referred as *scratch*.

Since no prior foundation models have been developed specifically for Mars, we further compare MOMO with several leading EO-FMs, including *SatMAE*, *CROMA*, *Prithvi*, and *TerraFM*. In addition, motivated by the strong representation quality demonstrated by *DINOv3* in large-scale visual pre-training studies, we include its satellite pre-trained variant (trained on the SAT-493M dataset).

We also evaluate *sensor-specific* pre-training baselines, where models are trained exclusively on data from a single sensor and then fine-tuned on corresponding downstream tasks. This allows us to analyze the relative benefits of same-sensor versus cross-sensor pre-training compared to MOMO. Furthermore, we consider a *joint data pre-training* configuration, referred to as the **Data Merge (DM)** setting, in which data from all three sensors are directly combined to pre-train a single unified model.

Lastly, since we propose an optimal checkpoint selection strategy, we compare it against other checkpoint selection approaches. Specifically, we evaluate two alternatives: Early Stopping (ES) and Last Epoch (LE) merging. In the ES setting, models are merged using the early-stopping checkpoint from each sensor, while in the LE setting, models are merged using their final training checkpoint from each sensor.

5.2. Pre-training

We pre-train a separate model on data from each sensor for five epochs. As described in Section 4, each model is trained on $\sim 4M$ samples. During training, we perform val-

	AtmosDust	DoMars16k	Frost	Landmark	Boulder	ConeQuest	Crater Binary	Crater Multi	MMLS	Avg. Rank ↓
Scratch	0.94 ± 0.003	0.73 ± 0.008	0.95 ± 0.007	0.79 ± 0.010	0.07 ± 0.012	0.52 ± 0.035	0.37 ± 0.047	0.05 ± 0.017	0.50 ± 0.017	4.11
ImageNet	0.92 ± 0.010	0.91 ± 0.003	0.97 ± 0.009	**0.92 ± 0.003**	0.16 ± 0.034	0.70 ± 0.022	**0.55 ± 0.012**	0.11 ± 0.007	**0.57 ± 0.014**	2.33
DINOv3	**0.97 ± 0.000**	0.90 ± 0.000	**0.99 ± 0.013**	0.91 ± 0.000	0.12 ± 0.106	0.51 ± 0.000	0.32 ± 0.000	0.01 ± 0.000	0.29 ± 0.000	3.67
CROMA	0.83 ± 0.021	0.41 ± 0.010	0.76 ± 0.000	0.47 ± 0.000	0.17 ± 0.018	0.44 ± 0.000	0.27 ± 0.000	0.01 ± 0.001	0.14 ± 0.014	5.89
Prithvi	0.81 ± 0.000	0.64 ± 0.008	0.81 ± 0.000	0.63 ± 0.000	0.04 ± 0.051	0.49 ± 0.000	0.18 ± 0.000	0.01 ± 0.000	0.23 ± 0.025	6.11
SatMAE	0.96 ± 0.000	**0.93 ± 0.000**	0.97 ± 0.000	**0.92 ± 0.000**	0.05 ± 0.000	0.68 ± 0.000	0.46 ± 0.000	0.04 ± 0.003	0.32 ± 0.000	3.00
TerraFM	**0.97 ± 0.000**	0.89 ± 0.000	**0.99 ± 0.000**	0.86 ± 0.002	**0.21 ± 0.052**	0.38 ± 0.021	0.44 ± 0.000	0.04 ± 0.003	0.14 ± 0.103	3.67
MOMO	0.96 ± 0.005	0.92 ± 0.000	0.98 ± 0.003	0.91 ± 0.003	0.20 ± 0.005	**0.71 ± 0.008**	0.54 ± 0.005	**0.12 ± 0.014**	**0.57 ± 0.009**	**1.67**

Table 1. Performance comparison of different baselines with MOMO. Reported metrics include F1-Score for classification tasks and mIoU for segmentation tasks. **Bold** and Underlined numbers indicate the best and second best performance.

MOMO	Boulder	ConeQuest	Crater Binary
ES	0.12 ± 0.078	0.68 ± 0.0134	0.50 ± 0.014
LE	0.18 ± 0.031	0.70 ± 0.005	0.50 ± 0.015
EVL (Ours)	**0.20 ± 0.005**	**0.71 ± 0.008**	**0.54 ± 0.005**

Table 2. Performance comparison across different checkpoint selection strategies with the proposed EVL technique.

idation after every $\sim 100k$ samples, recording the loss and saving a model checkpoint at each validation step. These intermediate checkpoints are later used for loss alignment and model merging procedures. Details about other hyperparameters are provided in Appendix B.

5.3. Downstream Tasks

We evaluate MOMO and baselines on all orbital tasks from Mars-Bench [52]. For datasets where performance has already saturated, we present results in the Appendix A.2 (two classification tasks). In the main paper, we report results for **four classification** tasks: *AtmosDust*, *DoMars16k*, *Frost*, and *Landmark*; and **five segmentation** tasks: *Boulder*, *ConeQuest*, *Crater Binary*, *Crater Multi*, and *MMLS*. A concise summary and representative visual samples for all tasks are provided in the Appendix A.2. For each dataset–model combination, we perform hyperparameter tuning and report the best-performing configuration.

6. Results and Analysis

In this section, we present the results from all experiments, and compare MOMO against all baseline methods described in Section 5.1. The quantitative results are shown in Table 1. For classification tasks, we report the weighted F1-score. And for segmentation tasks, we report mean Intersection over Union (mIoU). Since mIoU alone does not fully capture a model's ability to accurately localize distinct features, we include the Object F1-Score as a complementary metric to provide a more comprehensive evaluation of segmentation performance.

Training from scratch and model pre-trained on natural images. Table 1 demonstrates that directly training a model on downstream tasks performs worse than a pre-trained version of models. Comparing overall results with the ImageNet pre-trained version of the model, MOMO marginally outperforms ImageNet pre-training in the case of classification and segmentation tasks. For classification tasks, MOMO outperforms the ImageNet baseline, achieving an average improvement of $\sim 1.25\%$ F1-score across all four tasks, achieving significant improvement ($\sim 4\%$) for AtmosDust. ImageNet only outperforms in Landmark; however, the margin of improvement compared to MOMO is only 1%. For segmentation, we also observe a similar trend where MOMO outperforms the ImageNet baseline, achieving an average improvement of $\sim 1\%$ mIoU across all five tasks, achieving significant improvement ($\sim 4\%$) for Boulder. These results indicate that while ImageNet pre-training provides comparable representations for classification, it fails to capture the spatial and textural characteristics required for precise feature localization.

Models Pre-trained on Earth Satellite data. Consistent with the earlier observations, all EO-FMs perform on a similar level as MOMO in the case of classification tasks, but when comparing results in segmentation, overall MOMO outperforms compared to EO-FMs. DINOv3 shows competitive results for classification tasks.

Simple *vs*. Complex Tasks. Mars-Bench contains tasks of varying levels of complexity, which is also reflected in the results presented in Table 1. For instance, in the classification category, binary classification datasets such as *AtmosDust* and *Frost* exhibit relatively higher performance, with F1-scores ranging from 0.96 to 0.98. In contrast, multi-class classification datasets such as *DoMars16k* and *landmark* yield slightly lower values, typically between 0.90 and 0.92. A similar trend is observed in segmentation tasks, where the *Boulder* and *MMLS* datasets differ significantly in performance. The *Boulder* dataset achieves considerably lower scores (mIoU $\sim$ 0.04–0.21), whereas *MMLS* obtains

	AtmosDust	DoMars16k	Frost	Landmark	Boulder	ConeQuest	Crater Binary	Crater Multi	MMLS
HiRISE	0.93 ± 0.018	0.88 ± 0.002	0.97 ± 0.009	0.90 ± 0.005	0.12 ± 0.024	0.66 ± 0.014	0.49 ± 0.017	0.06 ± 0.014	0.52 ± 0.040
CTX	0.94 ± 0.012	0.90 ± 0.004	0.95 ± 0.013	**0.91 ± 0.004**	0.17 ± 0.016	0.70 ± 0.015	0.48 ± 0.008	0.07 ± 0.007	0.54 ± 0.008
THEMIS	0.94 ± 0.005	0.88 ± 0.003	0.94 ± 0.012	0.90 ± 0.004	0.17 ± 0.021	0.69 ± 0.024	0.50 ± 0.019	0.07 ± 0.021	0.51 ± 0.068
DM	0.94 ± 0.026	0.90 ± 0.003	0.44 ± 0.139	0.89 ± 0.004	0.14 ± 0.060	0.67 ± 0.012	0.52 ± 0.005	0.08 ± 0.019	0.48 ± 0.064
MOMO	**0.96 ± 0.005**	**0.92 ± 0.002**	**0.98 ± 0.012**	**0.91 ± 0.003**	**0.20 ± 0.005**	**0.71 ± 0.008**	**0.54 ± 0.005**	**0.12 ± 0.012**	**0.57 ± 0.009**

Table 3. Performance comparison across sensor-specific pre-training, joint data pre-training (DM - Data Merge), and MOMO. Here, highlighted indicates the sensor associated with that downstream task. **Bold** number indicates the best performance.

much higher values in some models (mIoU $\sim$ 0.14–0.57). This large performance gap primarily arises from the differences in object scale and count. In *Boulder* segmentation, each image typically contains 25 or more small objects, making feature localization and boundary detection substantially more difficult. In contrast, *MMLS* contains only one or two large landslides per image, with some occupying more than 50% of the total area, allowing the model to learn their spatial features more easily and achieve higher mIoU. A similar pattern is also seen in the *Crater Multi* and *Crater Multi* datasets. While the model performs well in identifying crater regions (*Crater Binary*), it struggles to classify them into their corresponding morphological types (*Crater Multi*), resulting in a noticeable performance drop between the two tasks.

Comparison with Sensor-specific Pre-training. We analyze the effect of fine-tuning performance under two settings: pre-training on the same sensor and pre-training on a different sensor (cross-sensor). The results are presented in Table 3. It can be clearly observed that same-sensor pre-training consistently improves performance across all downstream tasks compared to cross-sensor pre-training. However, cross-sensor pre-training does not lead to a significant drop in performance, indicating reasonable generalization across sensors. When compared with **MOMO**, it achieves an improvement of 1.75% in classification and 4.2% in segmentation. Although the gain in classification is relatively small, MOMO offers a key advantage over sensor-specific training. Specifically, sensor-specific approaches require maintaining separate models for each sensor, whereas MOMO provides a unified model that can handle all sensors simultaneously.

Comparison with Joint Data Pre-training. We compare MOMO with Data Merge (DM) approach. Table 3 shows that MOMO outperforms baselines by 15.25% in classification and 5.4% in segmentation, demonstrating its consistently strong performance across tasks. Also, the DM approach suffers from significant limitations in scalability and modularity. Specifically, if additional data or tasks from a new sensor become available in the future, DM would require complete re-training using data aggregated from all sensors. In contrast, MOMO supports easy integration by pre-training a model only on the new sensor and then merging it with existing sensor-specific models using the same EVL-based merging framework. This design makes MOMO more flexible, computationally efficient, and generalizable across new sensors and modalities.

Comparison of Checkpoint Selection Strategies. To demonstrate the effectiveness of our proposed EVL-based checkpoint selection strategy, we evaluate downstream performance using different checkpoint selection methods. Since classification tasks generally show minimal variation across models, this analysis is conducted only on segmentation tasks, selecting one downstream task from each sensor. As shown in Table 2, the EVL-based MOMO consistently achieves the best overall performance with average $\sim$ 2.5% mIoU improvement across three tasks. The only exception is observed in the *ConeQuest* dataset, where the difference remains negligible (1%) compared to the Early Stopping (ES) and Last Epoch (LE) strategies. Moreover, the definition of the "last epoch" itself is ambiguous, as it may occur at any iteration depending on the training setup (e.g., epoch 5 or 500), making it an unreliable criterion for checkpoint selection. These results highlight that EVL provides a more stable approach for merging models trained across different sensors.

6.1. Intuition Behind EVL

To further show the efficiency of the optimal checkpoint strategy, we conduct experiments where we examine the loss landscape resulting from various model-merging strategies. To visualize this landscape, we adopt the projection method used by Garipov et al. [17], Wortsman et al. [69], which maps models onto a two-dimensional plane. This method employs Gram-Schmidt orthogonalization to identify two orthogonal directions in the weight space, defining the x- and y-axes of the landscape illustrated in Figure 3. The basis vectors for the orthogonalization are derived from the ImageNet, HiRISE, and CTX pre-trained models. The parameters α and β denote unit displacements along the x and y axes, respectively.

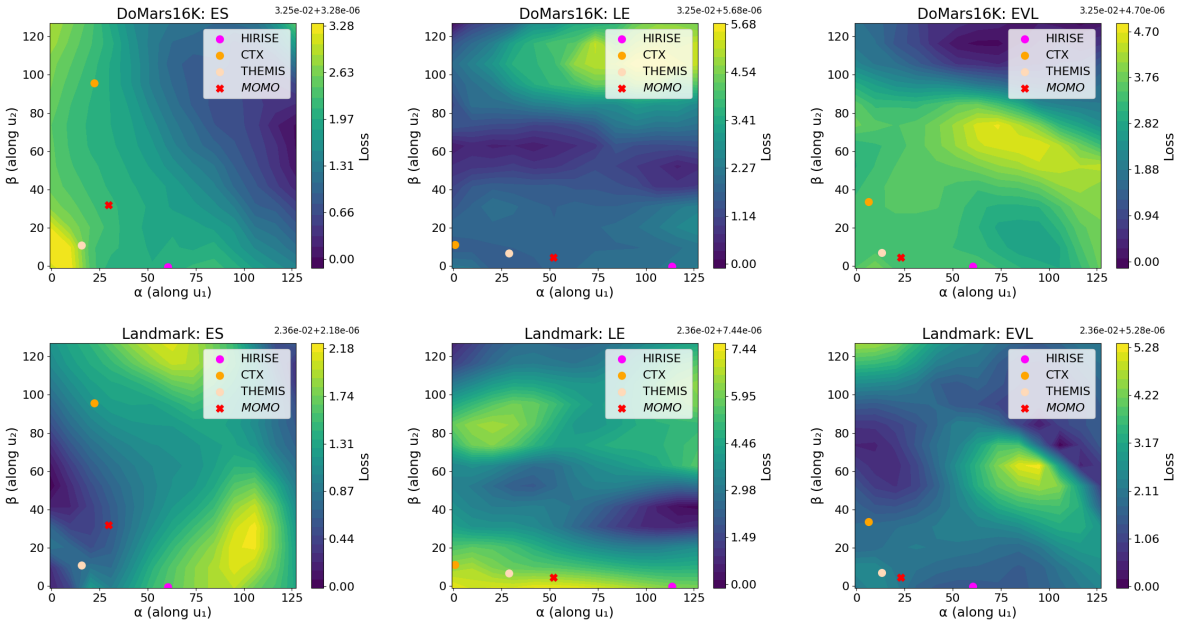


Figure 3. Loss landscape visualization across different checkpoint selection strategies on DoMars16k and Landmark datasets. The red markers represent MOMO obtained using Early Stopping (ES), Last Epoch (LE), and Equal Validation Loss (EVL), respectively.

As it can be observed from Table 1, binary classification tasks do not show significant performance differences (only 1-2% different), we select *DoMars16k* and *Landmark* datasets for this analysis. The losses are computed on the *DoMars16k* and *Landmark* datasets using class-balanced versions of their respective test sets.

In Figure 3, each plot presents the interpolated loss surface among the HiRISE, CTX, and THEMIS models. Across all merging strategies, the task-vector-based merged model consistently achieves an equal or lower loss compared to its constituent models. This demonstrates that task-vector merging can produce models that outperform the originals from which they were derived.

Prior work [2, 16, 53] suggests that optimal merging occurs when constituent models lie within the same loss basin, as this promotes stability and enhanced performance. Figure 3 shows that relative to both LE and ES, the EVL strategy selects model checkpoints that are more closely aligned in weight space. Consequently, EVL is expected to yield the most stable and best-performing merged model.

7. Conclusions

In this work, we introduced MOMO, the first foundation model designed for Mars orbital applications. MOMO effectively integrates multi-sensor and multi-resolution data through a model-merging framework that leverages our proposed optimal checkpoint selection strategy. This approach ensures stable and compatible model fusion, allowing the model to generalize efficiently across diverse data sources. Trained on a large-scale corpus of $\sim$ 12 million curated samples from HiRISE, CTX, and THEMIS sensors, MOMO demonstrates strong performance across all MarsBench orbital downstream tasks. Findings from experimental results indicate that MOMO outperforms compared to ImageNet pre-training, EO-FMs, sensor-specific training, data merge approach, and different checkpoint selection strategies. Particularly, MOMO demonstrates consistently superior performance in segmentation tasks, which require precise identification of fine-grained details within images. Moreover, our approach offers a scalable training framework that enables efficient integration of new sensors without requiring complete re-training. In summary, MOMO represents the first foundation model in planetary science, and we believe this effort will inspire the development of future foundation models for planetary science research.

Limitations Our approach builds on a model-merging framework; however, due to computational constraints, we did not include additional model-merging-based baselines for comparison. We also do not explore alignment-based techniques in this work, which can be used in conjunction with our method and may further improve performance, as suggested in literature [1, 62, 70]. Finally, our method relies on the assumption of linear mode connectivity between models. While this assumption holds in many practical settings, it may not strictly apply when models are trained on highly divergent data distributions or when models are functionally equivalent but differ due to network symmetries.

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] Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023. 8

[2] Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022. 8

[3] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021. 2

[4] 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. 1, 2

[5] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023. 2, 3

[6] Eric Bauer and Ron Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, 36:105–139, 1999. 2

[7] JF Bell III, MC Malin, MA Caplinger, J Fahle, MJ Wolff, BA Cantor, PB James, T Ghaemi, LV Posiolova, MA Ravine, et al. Calibration and performance of the mars reconnaissance orbiter context camera (ctx). *International Journal of Mars Science and Exploration*, 8:1–14, 2013. 4

[8] Nassim Ait Ali Braham, Conrad M Albrecht, Julien Mairal, Jocelyn Chanussot, Yi Wang, and Xiao Xiang Zhu. SpectralEarth: Training hyperspectral foundation models at scale. *arXiv preprint arXiv:2408.08447*, 2024. 2

[9] Christina Butsko, Kristof Van Tricht, Gabriel Tseng, Giorgia Milli, David Rolnick, Ruben Cartuyvels, Inbal Becker Reshef, Zoltan Szantoi, and Hannah Kerner. Deploying geospatial foundation models in the real world: Lessons from worldcereal. *arXiv preprint arXiv:2508.00858*, 2025. 1

[10] 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. 4

[11] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022. 2, 3

[12] 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. 3

[13] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 2

[14] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1): 119–139, 1997. 2

[15] Anthony Fuller, Koreen Millard, and James Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[16] Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodola. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18695–18705, 2025. 8

[17] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018. 7

[18] Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. Model patching: Closing the subgroup performance gap with data augmentation. *arXiv preprint arXiv:2008.06775*, 2020. 2

[19] Krzysztof M Gorski, Benjamin D Wandelt, Frode K Hansen, Eric Hivon, and Anthony J Banday. The healpix primer. *arXiv preprint astro-ph/9905275*, 1999. 5

[20] Hao Guo, Jiyong Jin, and Bin Liu. Stochastic weight averaging revisited. *Applied Sciences*, 13(5):2935, 2023. 2

[21] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. Skysense: A multimodal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27672–27683, 2024. 2

[22] Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. Bridging remote sensors with multisensor geospatial foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27852–27862, 2024. 1, 2

[23] 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. 3

[24] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. SpectralGPT: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[25] Ziyue Huang, Hongxi Yan, Qiqi Zhan, Shuai Yang, Mingming Zhang, Chenkai Zhang, YiMing Lei, Zeming Liu, Qingjie Liu, and Yunhong Wang. A survey on remote sensing foundation models: From vision to multimodality. *arXiv preprint arXiv:2503.22081*, 2025. 1

[26] 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. 4

[27] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *Advances in Neural Information Processing Systems*, 35:29262–29277, 2022. 2

[28] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 2

[29] Shancheng Jiang, Fan Wu, Kai-Leung Yung, Yingqiao Yang, WH Ip, Ming Gao, and James Abbott Foster. A robust end-to-end deep learning framework for detecting martian landforms with arbitrary orientations. *Knowledge-Based Systems*, 234:107562, 2021. 2

[30] Jean Kaddour. Stop wasting my time! saving days of imagenet and bert training with latest weight averaging. *arXiv preprint arXiv:2209.14981*, 2022. 2

[31] 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. 2

[32] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2

[33] Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A Smith, and Luke Zettlemoyer. Branch-train-merge: Embarrassingly parallel training of expert language models. *arXiv preprint arXiv:2208.03306*, 2022. 2

[34] Tao Li, Zhehao Huang, Yingwen Wu, Zhengbao He, Qinghua Tao, Xiaolin Huang, and Chih-Jen Lin. Trainable weight averaging: A general approach for subspace training. *arXiv preprint arXiv:2205.13104*, 2022. 2

[35] Ce Liu, William T Freeman, Richard Szeliski, and Sing Bing Kang. Noise estimation from a single image. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 901–908. IEEE, 2006. 5

[36] Siqi Lu, Junlin Guo, James R Zimmer-Dauphinee, Jordan M Nieusma, Xiao Wang, Steven A Wernke, Yuankai Huo, et al. Vision foundation models in remote sensing: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 2025. 1

[37] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7769–7778, 2020. 3

[38] 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. 4

[39] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021. 2

[40] Alfred S McEwen, Eric M Eliason, James W Bergstrom, Nathan T Bridges, Candice J Hansen, W Alan Delamere, John A Grant, Virginia C Gulick, Kenneth E Herkenhoff, Laszlo Keszthelyi, et al. Mars reconnaissance orbiter's high resolution imaging science experiment (hirise). *Journal of Geophysical Research: Planets*, 112(E5), 2007. 4

[41] 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. 3

[42] Hector Mendoza, Aaron Klein, Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. Towards automatically-tuned neural networks. In *Workshop on automatic machine learning*, pages 58–65. PMLR, 2016. 2

[43] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*, 2021. 2

[44] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR, 2022. 2

[45] Shikhar Murty, Christopher D Manning, Scott Lundberg, and Marco Tulio Ribeiro. Fixing model bugs with natural language patches. *arXiv preprint arXiv:2211.03318*, 2022. 2

[46] Jim Nilsson and Tomas Akenine-Möller. Understanding ssim. *arXiv preprint arXiv:2006.13846*, 2020. 3, 5

[47] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27811–27819, 2024. 2

[48] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 2

[49] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019. 2

[50] Mirali Purohit, Jacob Adler, and Hannah Kerner. Conequest: A benchmark for cone segmentation on mars. In *Proceed-*

ings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 6026–6035, 2024. 1

[51] 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. 1, 2

[52] 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. 1, 4, 6

[53] Xingyu Qu. Rethinking model re-basin and linear mode connectivity. 2024. 8

[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. 2

[55] Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Position: Mission critical–satellite data is a distinct modality in machine learning. In *Forty-first International Conference on Machine Learning*, 2024. 1

[56] David Rolnick, Alan Aspuru-Guzik, Sara Beery, Bistra Dilkina, Priya L Donti, Marzyeh Ghassemi, Hannah Kerner, Claire Monteleoni, Esther Rolf, Milind Tambe, et al. Application-driven innovation in machine learning. *arXiv preprint arXiv:2403.17381*, 2024. 1

[57] Shibani Santurkar, Dimitris Tsipras, Mahalaxmi Elango, David Bau, Antonio Torralba, and Aleksander Madry. Editing a classifier by rewriting its prediction rules. *Advances in Neural Information Processing Systems*, 34:23359–23373, 2021. 2

[58] George Stoica, Daniel Bolya, Jakob Bjorner, Pratik Ramesh, Taylor Hearn, and Judy Hoffman. Zipit! merging models from different tasks without training. *arXiv preprint arXiv:2305.03053*, 2023. 2, 3

[59] Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *Advances in Neural Information Processing Systems*, 34:24193–24205, 2021. 2

[60] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[61] Kenneth L Tanaka, James A Skinner, James M Dohm, Rossman P Irwin, Eric J Kolb, Corey M Fortezzo, Thomas Platz, Gregory G Michael, and Trent M Hare. *Geologic map of Mars*. Astrogeology Research Program (USGS), 2014. 4

[62] Alexander Theus, Alessandro Cabodi, Sotiris Anagnostidis, Antonio Orvieto, Sidak Pal Singh, and Valentina Boeva. Generalized linear mode connectivity for transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 8

[63] Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023. 1, 2, 3

[64] Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global and local features in pretrained remote sensing models. *arXiv preprint arXiv:2502.09356*, 2025. 2, 3

[65] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geolocalization. *Advances in Neural Information Processing Systems*, 36:8690–8701, 2023. 2

[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*, 2018. 1

[67] 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. 1

[68] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. *Advances in Neural Information Processing Systems*, 33:6514–6527, 2020. 2

[69] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022. 2, 7

[70] Binchi Zhang, Zaiyi Zheng, Zhengzhang Chen, and Jundong Li. Beyond the permutation symmetry of transformers: The role of rotation for model fusion. In *Forty-second International Conference on Machine Learning*, 2025. 8

[71] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3