

Training Data Selection for Weakly-Supervised Lunar Traversability Estimation: A Cross-Domain Study

Bartosz Ptak

`bartosz.ptak@put.poznan.pl`

Poznań University of Technology

Research Article

Keywords:

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10848537/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Training Data Selection for Weakly-Supervised Lunar Traversability Estimation: A Cross-Domain Study

Bartosz Ptak

Institute of Robotics and Machine Intelligence, Poznań University of Technology, Piotrowo 3A, Poznań, 60-965, Poland.

Contributing authors: bartosz.ptak@put.poznan.pl;

Abstract

This paper examines how weakly supervised traversability estimation models generalize across domains in the context of lunar robotics. Because annotated lunar datasets are scarce, models must be trained on Earth-based data and transferred to environments with significantly different visual and physical characteristics. We adopt a pairwise, weakly supervised learning framework in which relative traversability annotations provide a lightweight alternative to dense labeling while preserving the continuous nature of terrain assessment. We conduct a systematic cross-domain evaluation over seven datasets spanning synthetic simulations, terrestrial analogs, lunar analog facilities, and real lunar mission imagery (Chang’e 3/4), using the Human Disagreement Rate (HDR) metric. Our results reveal a consistent ranking of training sources: lunar analog facilities transfer most reliably to real lunar imagery, while visual and morphological similarity to the target domain proves a stronger predictor of transfer quality than dataset realism or domain category. Notably, strong in-domain performance does not predict cross-domain generalization, exposing a fundamental limitation of standard evaluation protocols for space-graded perception systems. We further show that this ranking is stable across HDR thresholds, and that introducing limited target-domain data monotonically improves transfer performance while preserving the initial source ranking. Overall, this study provides an empirical basis for training data selection in lunar traversability applications and argues for cross-domain evaluation as a standard component of the verification and validation pipeline for planetary surface perception systems.

Keywords: visual traversability estimation, weakly-supervised learning, lunar robotics, cross-domain generalization

1 Introduction

Autonomous rovers operating on the Moon need dependable traversability assessment to move safely through unfamiliar terrain. Yet the fundamental challenge for any learning-based perception system intended for lunar deployment is the lack of training data from the target environment: a visual model must be trained on Earth and expected to generalize to the Moon, where surface appearance, lighting, and morphology differ from those of any available training source.

Fully-supervised semantic segmentation approaches (Maturana et al. 2018; Wigness et al. 2019) address terrain traverse through dense per-pixel labeling, but require expensive annotation and impose a fixed taxonomy that is poorly suited to the continuous nature of lunar terrain. While prior work has addressed traversability for planetary missions using supervised classification on Mars imagery (Rothrock et al. 2016), such approaches rely on large annotated datasets that are not mature enough for deployment on the Moon. Self-supervised methods (Gasparino et al. 2022; Frey et al. 2023) derive labels from robot experience, but cannot explicitly label untraversable regions and depend on physical interaction with the target domain, which is impractical for lunar settings. Weakly-supervised approaches based on sparse pairwise annotations (Schreiber et al. 2024; Schreiber and Driggs-Campbell 2025) offer a favorable trade-off: they reduce annotation effort to seconds per image, avoid rigid class definitions, and naturally express the relative nature of traversability that lunar terrain demands. As illustrated in Figure 1, each image requires only a small number of point pairs indicating relative traversability.

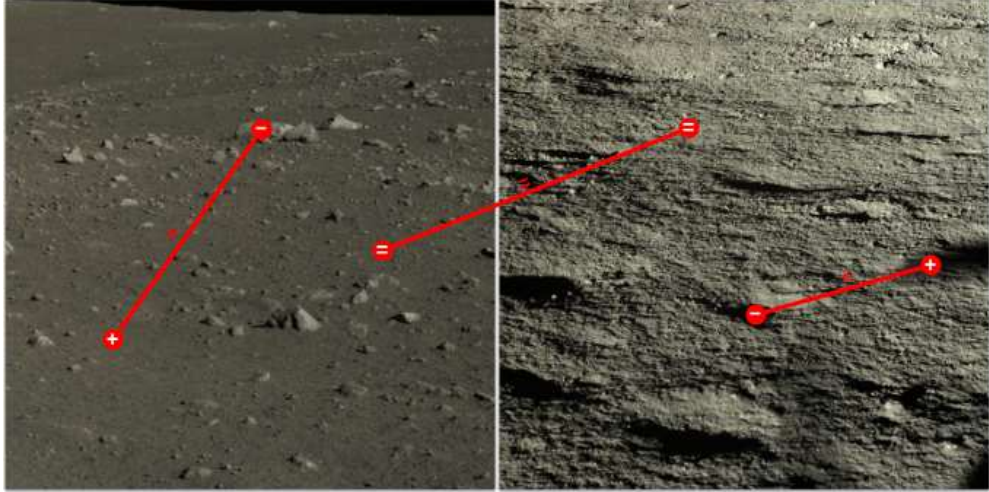


Fig. 1 Sparse pairwise traversability annotations on Chang'E-3/4 lunar imagery (Zuo et al. 2021). Red lines connect annotated point pairs, with symbols indicating the ordinal relation: the point marked + is more traversable than the point marked -, while = denotes equal traversability.

Despite the technical readiness of weakly supervised traversability methods, their cross-domain behavior remains largely unexplored, especially in lunar robotics. Existing work evaluates models within a single domain or relies on simulation-to-real transfer in restricted settings. For lunar applications, where deployment data is limited and training data must be sourced from a heterogeneous mixture of simulators, terrestrial analogs, and lunar analog facilities, this gap is particularly consequential: practitioners have no quantitative basis for choosing one training source over another.

In this work, we present the first systematic evaluation of cross-domain transfer for weakly-supervised traversability estimation in lunar robotics. We evaluate seven datasets spanning four categories: synthetic simulations, terrestrial analogs, lunar analog facilities, and real lunar mission imagery. We structure our study around the following questions:

1. **RQ1: cross-domain adaptation and domain ranking.** How does a weakly-supervised traversability model trained on one domain perform when transferred to another, and does a consistent ranking of training sources emerge across the seven evaluated datasets? We further examine this ranking qualitatively through visual comparison of predicted traversability maps.
2. **RQ2: sensitivity to the HDR evaluation threshold.** Is the cross-domain ranking obtained in RQ1 stable across the range of HDR decision thresholds, or does it depend on the specific threshold used for evaluation?
3. **RQ3: closing the domain gap with limited target-domain data.** How much target-domain data needs to be added to a source training set to close the performance gap to in-domain training, and does this depend on the initial cross-domain ranking established in RQ1?

2 Related works

The literature commonly divides traversability methods into three principal categories: geometry-based, vision-based, and fusion-based. Geometry-based techniques primarily depend on exteroceptive sensors such as LiDAR and depth cameras. Early methods in this class were computationally efficient and simple to deploy but often lacked robustness and struggled with irregular, unstructured terrain. With progress in machine learning and deep neural networks, researchers began to infer traversability directly from point clouds using learning-based techniques (Seo et al. 2024; Xue et al. 2021; Fan et al. 2021). However, approaches that rely solely on geospatial information have notable drawbacks, as they disregard semantic characteristics of the terrain—for example, differentiating between solid bedrock and highly compressible sand. Furthermore, the high financial and computational costs of LiDAR-centric algorithms motivated the emergence of vision-based approaches. These methods employ a variety of network architectures, such as AlexNet and PSPNet, and leverage diverse training schemes, including contrastive, self-supervised, and supervised learning (Jung et al. 2024; Guan et al. 2022; Seo et al. 2023). Because vision-based strategies only exploit visual appearance, fusion-based techniques were introduced to combine multiple sensing modalities, yielding more resilient traversability representations.

A notable multi-modal fusion approach evaluated in space-relevant scenarios is Terrain Traversability Analysis using Geometry and Environmental Perception (TTA-GEP) (Yang et al. 2023), designed for planetary rover path planning. The TTA-GEP system uses a highly parallelized dual-stream pipeline that integrates metric information with high-level object knowledge into a unified representation. In the geometric stream, stereovision is used to produce point clouds, which are then converted into a 2.5D map from which standard geometric descriptors, such as slope and surface roughness, are extracted. In parallel, a deep-learning-based terrain classifier processes camera images to generate a semantic segmentation of the scene, aiming to correctly label every portion of the Martian or lunar surface. The semantic classes in TTA-GEP are tailored to planetary robotics and include bearing soil, exposed bedrock, fine sand, loose rocks, crater edges and ridges, hazardous depressions, and sky background. Finally, a continuous evaluation function fuses and weights, in real time, the cost associated with local geometric properties of the terrain with the discrete material category obtained from semantic segmentation.

Orthogonal to this sensing-modality taxonomy is the supervision paradigm used to train the traversability estimator. Most geometry-, vision-, and fusion-based methods described above rely on dense, per-pixel or per-voxel annotation, which is costly to produce and imposes a fixed semantic taxonomy poorly suited to the continuous nature of terrain traversability. An alternative line of work formulates traversability estimation as a weakly-supervised, pairwise ranking problem. W-RIZZ (Schreiber et al. 2024) introduced this paradigm for mobile robotics: rather than labeling every pixel, annotators mark sparse pairs of points within an image as more, less, or equally traversable, and a ranking loss enforces the corresponding ordinal relation between predicted scores. This relative formulation reduces annotation effort to seconds per image, avoids committing to a rigid class taxonomy, and naturally accommodates the graded, context-dependent nature of traversability judgments. CHUNGUS (Schreiber and Driggs-Campbell 2025) extended this framework with a human-in-the-loop annotation protocol and an additional cross-image pairwise constraint, improving the consistency of predictions across scenes.

Despite this technical maturity, the cross-dataset and cross-domain generalization behavior of weakly-supervised traversability models remains largely unexamined. Existing evaluations are almost exclusively conducted within a single dataset or a narrow simulation-to-real setting, leaving practitioners without a quantitative basis for selecting among heterogeneous training sources when target-domain data is scarce or unavailable, and a gap that is particularly acute in lunar robotics, where the target domain is, by definition, inaccessible during training.

3 METHOD AND EXPERIMENTAL SETUP

The pipeline follows the weakly-supervised paradigm introduced by W-RIZZ (Schreiber et al. 2024) and CHUNGUS (Schreiber and Driggs-Campbell 2025), in which a traversability network is trained from a small number of *relative* pairwise annotations rather than dense per-pixel labels. For each image, the annotator labels only a handful of point pairs as more, less, or equally traversable. This sparse

supervision significantly reduces annotation effort while avoiding the need for a fixed semantic class taxonomy – a particular advantage in lunar settings, where terrain characteristics vary along a continuous spectrum rather than falling into distinct classes. A weakly-supervised paradigm, combined with general visual feature models, can be especially valuable when robustness to novel environments and generalization from limited data are required.

3.1 Pipeline

Our model architecture builds on the system proposed by Schreiber and Driggs-Campbell (2025), with a key modification in the feature upsampling stage. Input RGB images are passed through a frozen DINOv2 (Oquab et al. 2023) ViT-S (Dosovitskiy et al. 2020) backbone, which extracts patch-level visual features. Since DINOv2 produces one feature vector per 14×14 patch, naive use would yield coarse pixel-level predictions. The original CHUNGUS pipeline addresses this with FeatUp (Fu et al. 2024); instead, we employ LoftUp (Huang et al. 2025), a coordinate-based cross-attention upsampler trained on high-resolution pseudo-ground-truth features. In our preliminary experiments, LoftUp consistently outperformed both FeatUp and AnyUp (Wimmer et al. 2025) on lunar imagery, motivating its adoption.

The upsampled 384-dimensional per-pixel features are passed to a lightweight MLP traversability head, which outputs a per-pixel traversability score in $[0, 1]$. Only the MLP head is trained; the DINOv2 backbone and the LoftUp upsampler remain frozen throughout. The model is weakly-supervised with the L_RIZZ ranking loss (Schreiber et al. 2024), which operates on the sparse pairwise annotations and enforces the correct ordinal relation (greater, less, or equal) between predicted traversability scores at each annotated point pair. The resulting output mask has the same dimensions as the input image and is filled with unscaled traverse scores, so the structural relationships are preserved, but the actual values may differ across margin levels depending on how the model was trained.

The training procedure uses the AdamW optimizer (Loshchilov and Hutter 2019) with a learning rate of $1e-4$, batch size of 256, weight decay of 0.01, and a StepLR scheduler ($\gamma = 0.25$, $step = 50$), reducing the learning rate by a factor of 4 every 50 epochs. The maximal number of epochs is 1000, regulated by monitoring the validation metric with patience of 50 epochs. All input images are scaled and then center-cropped to the resolution of 512×512 , while the DINOv2 size is 224×224 .

3.2 HDR Metric

We adopt the Human Disagreement Rate (HDR) metric introduced in W-RIZZ (Schreiber et al. 2024), which measures the fraction of annotated point pairs for which the model’s predicted ordinal relation disagrees with the human label, using a threshold t on the predicted score difference. In contrast to absolute measures like MSE, HDR evaluates the relative ranking of traversability between points, which aligns with the weakly supervised pairwise annotation scheme and accommodates the intrinsic uncertainty inherent in assigning traversability scores. We report HDR averaged

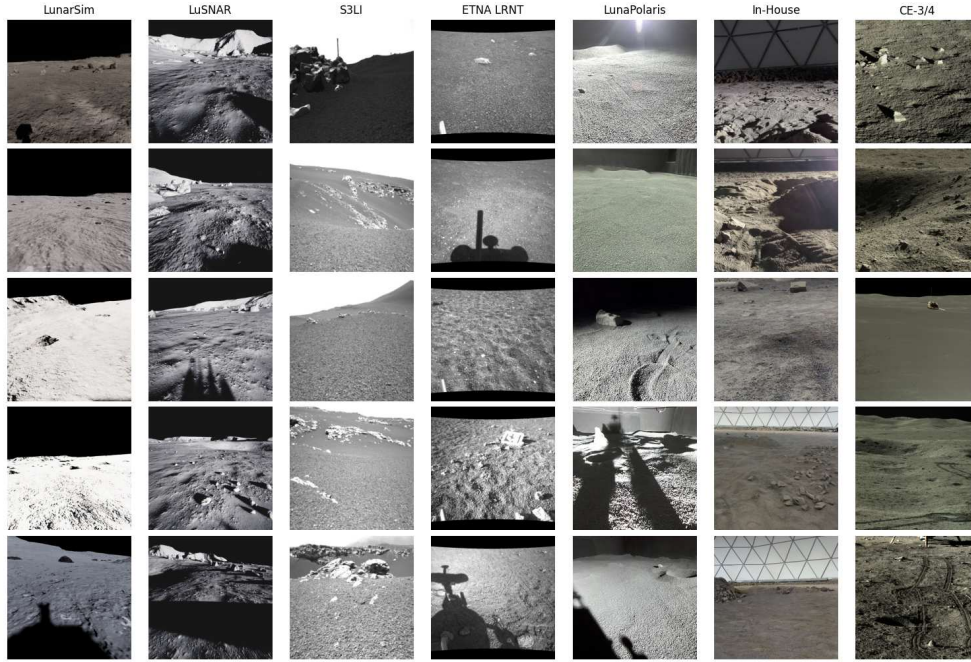


Fig. 2 Sample images from all datasets used in this study. Columns from left to right: LunarSim (synthetic), LuSNAR (synthetic), S3LI (volcanic analog), ETNA LRNT (volcanic analog), LunaPolaris (indoor testbed), In-House (indoor testbed), and CE-3/4 (real lunar surface). Visual differences in lighting, texture, and scene geometry highlight the challenge of cross-domain transfer.

over thresholds $t \in [0.10, 0.50]$ as the main metric, along with HDR at $t = 0.10$ and $t = 0.25$ for direct comparison with prior work.

Intuitively, a lower threshold t (e.g., $t = 0.10$) requires the model to correctly rank even nearly indistinguishable point pairs, making the metric more sensitive to fine-grained score differences, whereas a higher threshold (e.g., $t = 0.50$) only penalizes disagreements on pairs with substantially different predicted scores, tolerating uncertainty in borderline cases.

3.3 Datasets

We evaluate cross-domain transfer across seven datasets spanning four categories of imagery relevant to lunar robotics: lunar mission data, lunar analog facilities, terrestrial planetary analogs, and synthetic simulations. All datasets were annotated using the same sparse pairwise scheme described in Section 3.4. Figure 2 provides sample images from the datasets, while their characteristics are provided below.

- **Lunar mission data.** We use real lunar surface imagery from the panoramic cameras (PCAM) of the Chang'E-3 ([Ground Research and Application System of China's Lunar and Planetary Exploration Program 2015](#)) and Chang'E-4 ([Ground Research and Application System of China's Lunar and Planetary Exploration](#)

Program 2020) missions, obtained through the Ground Research and Application System (GRAS) of the China Lunar Exploration Program (Zuo et al. 2021). The PCAM instruments aboard the Yutu and Yutu-2 rovers acquired ground-level color images of the lunar surface in Mare Imbrium and the Von Kármán crater on the lunar farside, respectively.

- **Lunar analogue facilities.** *LunaPolaris* (van der Meer et al. 2025) was collected in the LunaLab (Ludvig et al. 2020) with a small rover. This indoor lunar analog facility provides controlled regolith-like terrain under repeatable lighting conditions. RGB images were collected using a ZED2 stereo camera. *In-House dataset* refers to our own analog test facility at OTRK Kakolewo¹, equipped with diverse rock formations and crater-like features designed to emulate lunar surface morphology. Visual data were collected using an Orbbec Femto Mega camera mounted on a mobile robot.
- **Terrestrial planetary analogues.** Both terrestrial analog datasets were collected on Mt. Etna, Sicily, Italy – a site widely used for planetary surface analog testing due to its volcanic terrain. *ETNA LRNT* (Vayugundla et al. 2018) (Long Range Navigation Tests) was captured with the Lightweight Rover Unit (LRU) and provides outdoor traversal sequences across volcanic landscapes. The rover is equipped with a stereo-based monochromatic camera setup. *S3LI* (Giubilato et al. 2022) (DLR Planetary Stereo, Solid-State LiDAR, Inertial dataset) was also recorded on Mt. Etna and offers complementary sequences with multi-modal sensing, including VT Mako cameras.
- **Synthetic simulations.** *LunarSim* (Pieczyński et al. 2023) is a high-fidelity lunar rover simulator providing photo-realistic synthetic imagery of lunar terrain based on the Unity game engine. *LuSNAR* (Liu et al. 2024) is a synthetic lunar dataset for SLAM, navigation, and recognition tasks, offering an additional simulation resource featuring a wide variety of scene configurations built in Unreal Engine.

3.4 Data Selection

To ensure consistent training conditions across heterogeneous datasets and reduce redundancy from temporally adjacent frames, we apply a unified data selection procedure. For each dataset, we extract the DINOv2 (Oquab et al. 2023) class token from every image and apply k -means clustering ($k = 600$) on the resulting feature vectors, using a cosine distance metric within the feature space. The image closest to each cluster centroid is retained, yielding 600 representative training images per dataset that span the source’s visual diversity while avoiding near-duplicates. An additional 150 images are randomly sampled from the remaining pool to form the validation set, ensuring no overlap with the training subset. Based on this, images from each subset were randomly combined in pairs according to the annotation protocol described in Section 3.5, resulting in 300 train and 75 validation batches.

¹<http://otrk.put.poznan.pl/>

Table 1 Lunar traversability hierarchy (Tier 1 = most traversable).

Tier	Description
1	Ideal flat terrain
2	Uneven terrain, wheel tracks
3	Small rocks
4	Large rocks, holes, craters
5	Unexpected moving objects, steep hills
6	Sky, background, image artifacts

3.5 Annotation Schema

We define a six-tier traversability hierarchy tailored to lunar surface morphology, ordered from most to least traversable: (i) *ideal flat terrain*, (ii) *uneven terrain including wheel tracks*, (iii) *small rocks*, (iv) *large rocks, holes, and craters*, (v) *unexpected moving objects and steep hills*, and (vi) *sky, background, and image artefacts*. This ordering provides annotators with a consistent reference for relative comparisons across domains. Importantly, lighting artifacts such as flares and shadows do not reduce traversability when the underlying terrain remains visible. The hierarchy is summarized in Table 1.

Annotations are collected as sparse pairwise comparisons following the W-RIZZ protocol (Schreiber et al. 2024): for each image, the annotator labels one of the point pairs as *more traversable*, *less traversable*, or *equally traversable*. It also includes a one pair across two images, which improves the consistency of predictions across different scenes (Schreiber and Driggs-Campbell 2025). Per-image annotation takes only a few seconds, enabling efficient labeling across all seven datasets.

4 Evaluation

Addressing relatively small dataset sizes and potential large variation between sampled images, a multi-run procedure was applied to every experiment. Using different initial seeds for the randomization algorithm, 15 runs were performed, and the mean and standard deviation were reported for each experiment.

4.1 Cross-domain Adaptation

Figure 3 presents the full cross-domain evaluation matrix, reporting average HDR and HDR@0.25 across all 49 train-evaluation combinations of seven datasets. Diagonal entries correspond to in-domain performance; off-diagonal entries represent cross-domain transfer. The matrix reveals that in-domain performance is a poor predictor of cross-domain transfer. S3LI achieves the lowest in-domain average HDR of any source (0.078), yet provides the worst transfer to CE-3/4 (0.304): a pattern consistent with a model that overfits to the visual characteristics of a single domain rather than learning generalizable traversability cues, and a consequential failure mode for verification and validation pipelines that rely on in-domain metrics alone. The same asymmetry is

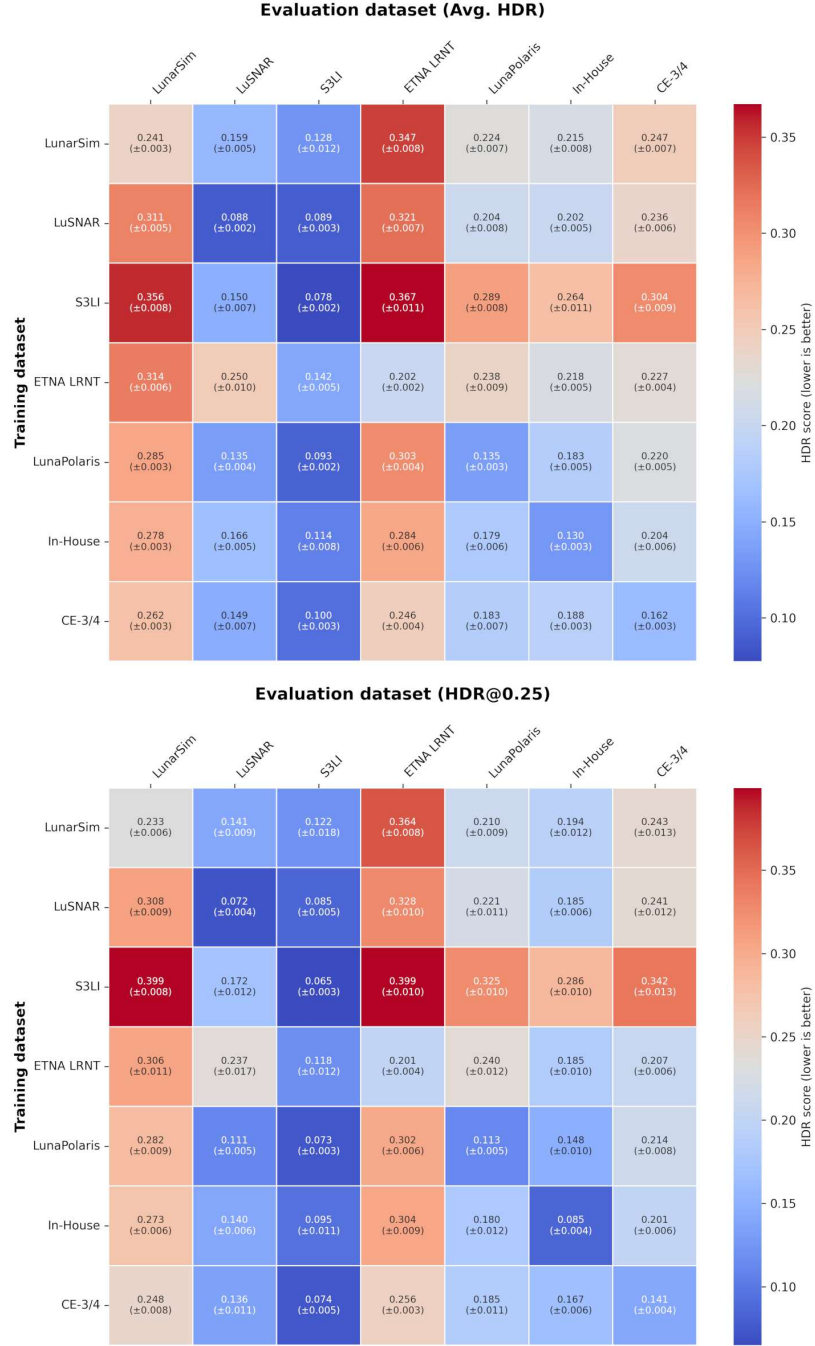


Fig. 3 Comparison of cross-dataset performance using average HDR and HDR with threshold of 0.25.

visible across both metrics: at HDR@0.25, S3LI’s in-domain score of 0.065 contrasts sharply with its CE-3/4 transfer score of 0.342. Visual similarity to the target domain matters more than dataset realism or origin. Both simulation datasets transfer more reliably to CE-3/4 (LunarSim: 0.247, LuSNAR: 0.236) than the S3LI terrestrial analog (0.304), despite being entirely synthetic. ETNA LRNT, a real-world terrestrial dataset, achieves a CE-3/4 transfer score of 0.227 – comparable to LunaPolaris (0.220) and substantially better than S3LI. Domain category alone is therefore insufficient to predict transfer quality; the visual and morphological alignment between training scenes and the lunar target plays a decisive role.

4.2 Domains Ranking

Table 2 extracts the key column of this matrix – transfer to real lunar mission imagery (CE-3/4) – and ranks training datasets by HDR. Both lunar analog datasets (In-House: 0.204, LunaPolaris: 0.220) outperform terrestrial analogs, simulations, and S3LI on transfer to CE-3/4. The In-House facility (OTRK Kakolewo), introduced in this work, achieves transfer performance comparable to the established LunaPolaris facility, validating its use as a training source for lunar traversability.

Table 2 Training dataset ranking based on average HDR performance on CE-3/4 lunar data.

Rank	Training data	Category	Avg. HDR	HDR@0.10	HDR@0.25
0	CE-3/4	In-domain	0.162 ± 0.003	0.212 ± 0.007	0.141 ± 0.004
1	In-House	Lunar analogue	0.204 ± 0.006	0.227 ± 0.011	0.201 ± 0.006
2	LunaPolaris	Lunar analogue	0.220 ± 0.005	0.245 ± 0.011	0.214 ± 0.008
3	ETNA LRNT	Terrestrial analogue	0.227 ± 0.004	0.278 ± 0.007	0.207 ± 0.006
4	LuSNAR	Simulation	0.236 ± 0.006	0.234 ± 0.012	0.241 ± 0.012
5	LunarSim	Simulation	0.247 ± 0.007	0.262 ± 0.009	0.243 ± 0.013
6	S3LI	Terrestrial analogue	0.304 ± 0.009	0.253 ± 0.017	0.342 ± 0.013

4.3 Visual Comparison

The qualitative comparison in Figure 4 supports these findings. Models trained on lunar analogs and CE-3/4 itself produce coherent traversability maps that correctly identify rocks, distant terrain, and low-traversability regions. Models trained on S3LI tend to under-predict hazards across most scenes, while ETNA LRNT produces visibly more aggressive predictions, particularly on flat terrain and crater fields.

A closer inspection through the lens of the traversability hierarchy reveals characteristic failure modes for each model family. At the traversable end of the scale, most transfer models correctly assign high scores to *ideal flat terrain* (tier 1), with the notable exception of ETNA LRNT, which over-penalizes open surfaces – likely due to the rough volcanic texture of its training domain. For *uneven terrain and wheel tracks* (tier 2), in-domain and analog-trained models show reasonable sensitivity, whereas S3LI-trained models tend to collapse this tier into flat terrain, underestimating the subtle hazard of loose regolith. For *small rocks* (tier 3), ETNA LRNT and

In-House provide sensitivity similar to CE-3/4, while the rest of the models tend to show low response. At the hazardous end, *large rocks, holes, and craters* (tier 4) represent the most discriminative category: the CE-3/4 reference model assigns sharp low-traversability boundaries around crater rims and rock clusters, while most transfer models produce diffuse or shifted predictions.

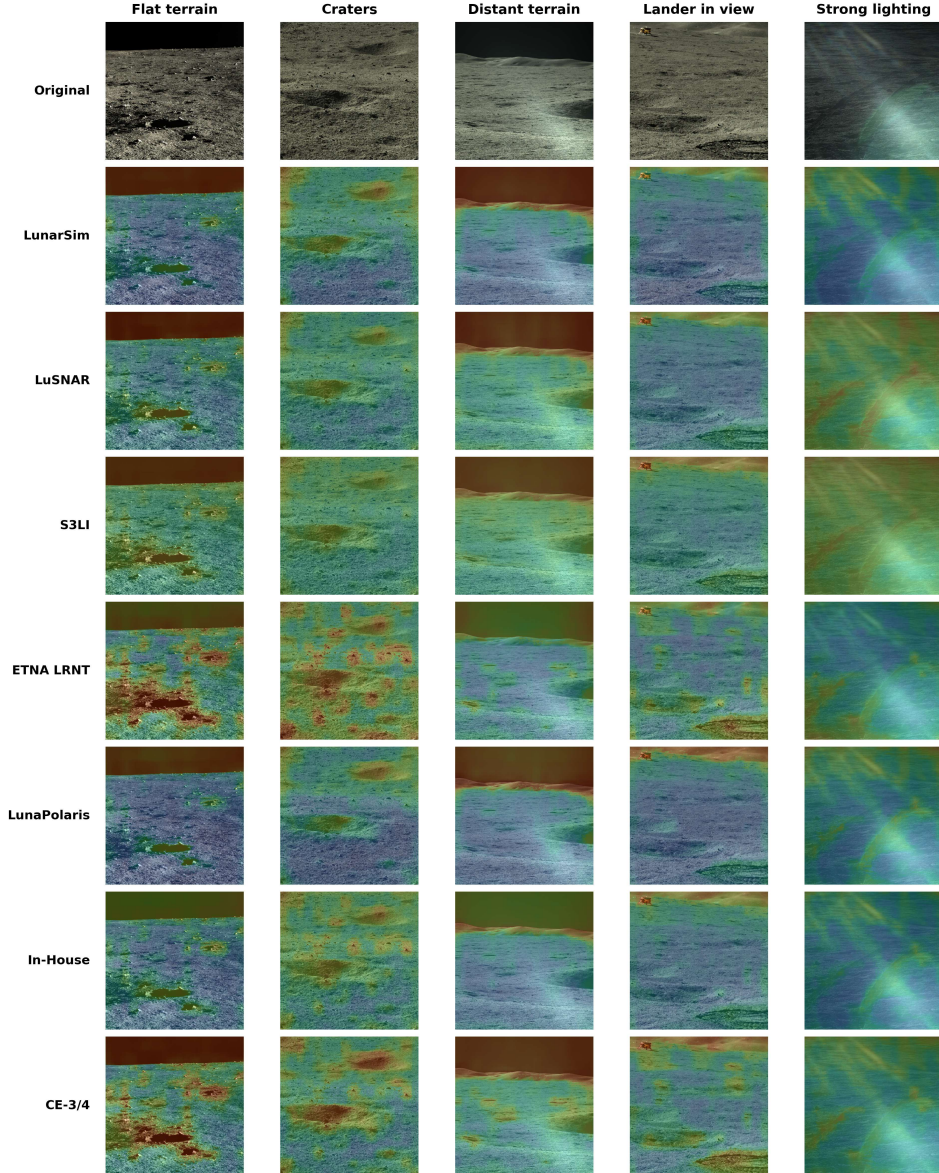


Fig. 4 Visual comparison of training dataset (rows) and selected scenarios (columns).

4.4 HDR Threshold Evaluation

The HDR metric requires selecting a threshold t on the predicted score difference to decide whether the model’s ranking of a point pair agrees with the human annotation; results reported in Section 4.2 use $t = 0.25$ (and $t = 0.10$) as fixed operating points, following prior work. To assess whether the resulting cross-domain ranking depends on this choice, we recompute HDR across the full range $t \in [0.10, 0.50]$ for each of the six training sources evaluated on CE-3/4. Results are shown in Figure 5, with shaded regions denoting 95% confidence intervals over 15 seeds. All sources exhibit a non-monotonic, U-shaped curve, reaching a minimum around $t \approx 0.15 - 0.20$ before increasing sharply toward $t = 0.50$. The in-domain model (CE-3/4, black) displays the same qualitative shape but with a markedly flatter and wider minimum, indicating that its predictions are comparatively insensitive to threshold choice, unlike models transferred across domains. Sources also diverge substantially at high thresholds despite similar behavior near the minimum: S3LI and LunarSim approach HDR values of 0.45-0.50 at $t = 0.50$, while LunaPolaris and In-House remain below 0.40, widening the gap between best- and worst-performing sources by roughly threefold compared to $t = 0.25$. The extremes of the ranking are stable throughout: In-House and LunaPolaris remain the two best-performing sources, and S3LI the worst, at every threshold tested. We therefore conclude that while the overall hierarchy of training sources is robust at its extremes, comparisons between closely-ranked intermediate sources should be interpreted with caution, and we recommend reporting HDR across a threshold range rather than a single operating point when comparing candidate training sources of similar quality.

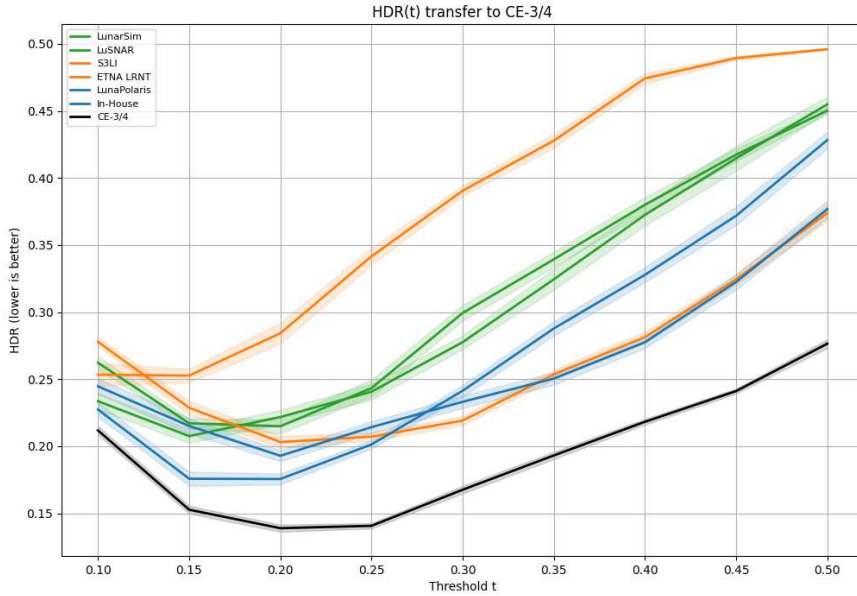


Fig. 5 HDR as a function of the evaluation threshold t , for models trained on each source dataset and transferred to CE-3/4 (color), against the in-domain reference (black).

4.5 Domain Gap Adaptation

Building on the domain gap identified in Section 4.2, we evaluate how much of this gap can be closed by introducing a limited amount of target-domain data into the training set. Rather than augmenting the training set with additional images, we substitute a subset of source-domain image pairs with an equal number of target-domain pairs (drawn from the CE-3/4 training split), keeping the total training set size fixed at 300 pairs. This design isolates the effect of domain matching from the confounding effect of increased training-set size that would result from simple data augmentation. We evaluate substitution counts of 0, 5, 10, 25, 50, 100, and 150 pairs, spaced logarithmically and reaching up to 50% of the training budget. Results are presented in Figure 6 for both average HDR and HDR@0.25. Each point reports the mean over 15 runs with independently resampled substitution subsets; shaded regions denote one standard deviation. The dashed line indicates the in-domain reference performance of a model trained entirely on CE-3/4.

All six training sources converge monotonically toward the in-domain reference as the number of substituted pairs increases, with no evidence of the coverage degradation we hypothesized might occur at high substitution ratios. The rate of improvement is inversely related to zero-shot transfer quality: S3LI, the weakest source at zero substitution (Avg. HDR 0.304), exhibits the steepest decline, while InHouse, already the strongest source, improves only marginally. However, even at the maximum substitution ratio (150/300, 50%), no source fully closes the gap to the in-domain reference (0.162), and S3LI in particular remains the worst-performing source throughout the full range, despite its large absolute improvement. This indicates that target-domain substitution narrows but does not eliminate the practical consequences of source selection identified in Section 4.2, and that the initial cross-domain ranking remains informative even once limited target-domain data becomes available.

5 DISCUSSION AND CONCLUSION

Our cross-domain evaluation reveals a clear hierarchy of training sources for lunar traversability transfer, with practical implications for both system design and validation. The strong performance of lunar analog facilities, particularly the comparable behavior of LunaPolaris and our newly introduced OTRK Kakolewo site, suggests that controlled indoor analogs with regolith-like terrain and lunar-relevant morphology can serve as effective training sources when real lunar data is unavailable, which will be the case for the foreseeable future. The asymmetry between in-domain and cross-domain performance, most clearly illustrated by S3LI, highlights a critical risk in the verification and validation of space-graded perception systems: a model that excels on its training distribution may catastrophically fail when deployed on a target domain whose visual characteristics differ from those seen during training. Standard in-domain evaluation protocols are therefore insufficient for systems intended for lunar deployment.

RQ1: Cross-domain adaptation and domain ranking. A consistent ranking of training sources emerges across the full cross-domain evaluation matrix, with visual and

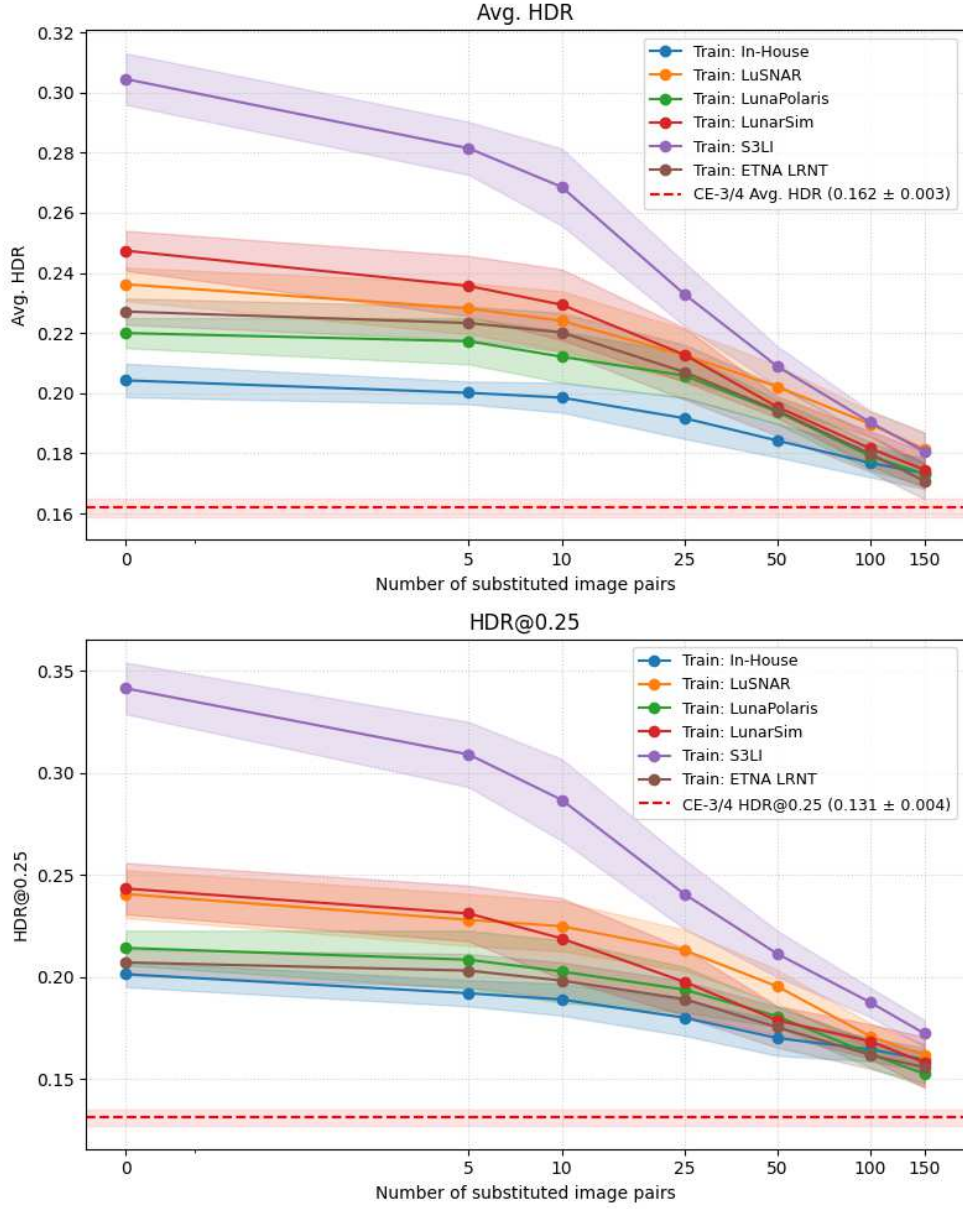


Fig. 6 Effect of target-domain substitution on cross-domain transfer to CE-3/4. Average HDR (top) and HDR@0.25 (bottom) as a function of the number of substituted target-domain image pairs (0-150, out of a fixed training set of 300).

morphological similarity to the target domain as the dominant predictor of transfer quality — outweighing both dataset realism and domain category. Lunar analog

facilities (In-House, LunaPolaris) occupy the top two positions, followed by terrestrial analogs and simulations, with S3LI a clear outlier at the bottom. Notably, synthetic datasets (LunarSim, LuSNAR) outperform S3LI despite lacking photorealism, further confirming that visual alignment matters more than image origin. The qualitative comparison reinforces this picture: in-domain and analog-trained models produce coherent traversability maps across all tiers of the traversability hierarchy, while S3LI and ETNA LRNT exhibit systematic failure modes — respectively under-predicting hazards on uneven terrain and over-penalizing traversable flat surfaces — that correlate directly with the morphological mismatch between their training environments and the CE-3/4 target.

RQ2: Sensitivity to the HDR evaluation threshold. The cross-domain ranking is stable at its extremes across the full threshold range $t \in [0.10, 0.50]$: In-House and LunaPolaris consistently rank first and second, and S3LI consistently ranks last, regardless of threshold choice. However, intermediate sources diverge substantially at high thresholds, widening the performance gap by roughly threefold between $t = 0.25$ and $t = 0.50$. This sensitivity is absent in the in-domain model, whose HDR curve is markedly flatter, suggesting that cross-domain models are more vulnerable to score miscalibration on pairs with large predicted differences.

RQ3: Closing the domain gap with limited target-domain data. Target-domain substitution consistently improves transfer performance across all sources, with the rate of improvement inversely proportional to zero-shot transfer quality: weaker sources benefit most from even small amounts of target-domain data. Nevertheless, no source fully closes the gap to in-domain performance within the evaluated substitution budget (up to 150 out of 300 pairs), and the initial cross-domain ranking remains largely preserved throughout: the weakest source at zero substitution remains the weakest at maximum substitution. This implies that source selection is consequential even when limited target-domain data is available, and that early investment in a morphologically appropriate training environment yields benefits that cannot be fully recovered by post-hoc data augmentation alone.

Taken together, our results establish that visual and morphological alignment between training and target domains is the primary driver of traversability transfer quality, that in-domain evaluation metrics are unreliable proxies for cross-domain performance, and that the domain gap can be narrowed but not eliminated by limited target-domain exposure. For future lunar missions, these findings argue for prioritizing training data collection in environments that closely replicate lunar surface morphology.

5.1 Limitations

Our work has several limitations. The lunar analog facilities used in this study are indoor environments under controlled lighting, lacking the harsh shadows, extreme thermal gradients, and dust dynamics characteristic of the actual lunar surface. While LunaPolaris and In-House replicate regolith morphology, neither reproduces the full photometric complexity of low-angle solar illumination, which is known to substantially affect the appearance of rocks and craters in real mission imagery. Generalizing

our findings to outdoor or higher-fidelity analog environments therefore remains an open question.

The annotations are produced by a single annotator, which may introduce systematic bias; an inter-annotator agreement study would strengthen the reliability of the reported HDR values. Furthermore, the weakly supervised pairwise annotation scheme, while well-suited to capturing relative traversability, does not enforce global consistency across the dataset. Locally coherent annotations may still yield contradictory orderings when aggregated, particularly in ambiguous scenes such as those with strong lighting artifacts or partially occluded hazards.

Our evaluation is also limited in scope to a single target domain (CE-3/4), two rover platforms, and a single model architecture. Our experiments do not establish whether the observed source ranking generalizes to other lunar datasets, camera configurations, or traversability learning frameworks. In particular, the strong performance of In-House may reflect incidental visual similarity to CE-3/4 rather than a generalizable property of the facility, and should be interpreted cautiously until validated on additional target domains.

Finally, our evaluation focuses on the relative ranking of traversability and does not assess the absolute calibration of predictions, which may be relevant for downstream planning systems that consume traversability scores directly. A model that correctly ranks point pairs may still assign scores that are poorly calibrated in absolute terms, leading to suboptimal path planning decisions when a fixed traversability threshold is used to gate navigation.

5.2 Future Works

Several directions extend this study toward a more comprehensive understanding of cross-domain traversability behavior. Pixel-level analysis of prediction errors, for instance, by aggregating HDR over superpixel regions or DINOv2 cluster boundaries, would reveal where models systematically fail across domains and which terrain features drive the cross-domain gap. Including further uncertainty estimates from traversability predictions may enable the detection of low-confidence areas and guide subsequent data annotation strategies and the design of future model architectures. Finally, full-trajectory evaluation on the rover platform during analog missions would assess the practical impact of cross-domain gaps on navigation success, beyond the per-image HDR metric used here.

LLM disclosure. The authors acknowledge using a large language model solely to edit and improve the manuscript’s grammar. No sections of the article, including the main technical content, figures, or images, were generated by the AI system. Use was limited to improving readability and language clarity.

Financial disclosure. None reported.

Conflict of interest. The authors declare no potential conflicts of interest.

References

- Dosovitskiy A, Beyer L, Kolesnikov A, et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
- Fan DD, Otsu K, Kubo Y, et al (2021) STEP: Stochastic traversability evaluation and planning for risk-aware off-road navigation. URL <https://arxiv.org/abs/2103.02828>, arXiv:2103.02828
- Frey J, Mattamala M, Chebroly N, et al (2023) Fast traversability estimation for wild visual navigation. arXiv preprint arXiv:2305.08510
- Fu S, Hamilton M, Brandt L, et al (2024) Featup: A model-agnostic framework for features at any resolution. arXiv preprint arXiv:2403.10516
- Gasparino MV, Sivakumar AN, Liu Y, et al (2022) WayFAST: Navigation with predictive traversability in the field. IEEE Robotics and Automation Letters 7(4):10651–10658
- Giubilato R, Stürzl W, Wedler A, et al (2022) Challenges of slam in extremely unstructured environments: The DLR planetary stereo, solid-state lidar, inertial dataset. IEEE Robotics and Automation Letters 7(4):8721–8728
- Ground Research and Application System of China’s Lunar and Planetary Exploration Program (2015) Chang’E-3 Panoramic Cameras Dataset. URL <https://moon.bao.ac.cn>, accessed via GRAS data portal
- Ground Research and Application System of China’s Lunar and Planetary Exploration Program (2020) Chang’E-4 Panoramic Cameras Dataset. URL <https://moon.bao.ac.cn>, accessed via GRAS data portal
- Guan T, He Z, Song R, et al (2022) TNS: Terrain traversability mapping and navigation system for autonomous excavators. URL <https://arxiv.org/abs/2109.06250>, arXiv:2109.06250
- Huang H, Chen A, Havrylov V, et al (2025) LoftUp: Learning a coordinate-based feature upsampler for vision foundation models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 9913–9923
- Jung S, Lee J, Meng X, et al (2024) V-STRONG: Visual self-supervised traversability learning for off-road navigation. In: IEEE International Conference on Robotics and Automation (ICRA), pp 1766–1773, <https://doi.org/10.1109/ICRA57147.2024.10611227>
- Liu J, Zhang Q, Wan X, et al (2024) LuSNAR: A lunar segmentation, navigation and reconstruction dataset based on multi-sensor for autonomous exploration. arXiv preprint arXiv:2407.06512

- Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. URL <https://arxiv.org/abs/1711.05101>, arXiv:1711.05101
- Ludvig P, Calzada-Diaz A, OLIVARES MENDEZ MA, et al (2020) Building a piece of the moon: Construction of two indoor lunar analogue environments. In: 71st International Astronautical Congress (IAC)–The CyberSpace Edition, pp 1–5
- Maturana D, Chou PW, Uenoyama M, et al (2018) Real-time semantic mapping for autonomous off-road navigation. In: Field and Service Robotics: Results of the 11th International Conference, Springer, pp 335–350
- van der Meer D, Wong U, Olivares-Mendez M (2025) Lunapolaris: A stereo camera, point cloud and imu dataset for future lunar exploration in polar regions. In: 2025 International Conference on Space Robotics (iSpaRo), pp 168–174, <https://doi.org/10.1109/iSpaRo66239.2025.11436512>
- Oquab M, Darcet T, Moutakanni T, et al (2023) DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:230407193
- Pieczyski D, Ptak B, Kraft M, et al (2023) Lunarsim: Lunar rover simulator focused on high visual fidelity and ROS 2 integration for advanced computer vision algorithm development. Applied Sciences 13(22):12401
- Rothrock B, Kennedy R, Cunningham C, et al (2016) SPOC: Deep learning-based terrain classification for Mars rover missions. In: AIAA SPACE 2016 Realizing the Future Space Enterprise, p 5539, <https://doi.org/10.2514/6.2016-5539>
- Schreiber A, Driggs-Campbell K (2025) Do you know the way? human-in-the-loop understanding for fast traversability estimation in mobile robotics. IEEE Robotics and Automation Letters
- Schreiber A, Sivakumar AN, Du P, et al (2024) W-RIZZ: A weakly-supervised framework for relative traversability estimation in mobile robotics. IEEE Robotics and Automation Letters 9(6):5623–5630
- Seo J, Sim S, Shim I (2023) Learning off-road terrain traversability with self-supervisions only. IEEE Robotics and Automation Letters 8(8):4617–4624. <https://doi.org/10.1109/lra.2023.3284356>, URL <http://dx.doi.org/10.1109/LRA.2023.3284356>
- Seo J, Kim T, Ahn S, et al (2024) Metaverse: Meta-learning traversability cost map for off-road navigation. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 13190–13197
- Vayugundla M, Steidle F, Smisek M, et al (2018) Datasets of long range navigation experiments in a moon analogue environment on mount etna. In: ISR 2018; 50th International Symposium on Robotics, pp 1–7, URL <https://elib.dlr.de/124514/>

- Wigness M, Eum S, Rogers JG, et al (2019) A RUGD dataset for autonomous navigation and visual perception in unstructured outdoor environments. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 5000–5007
- Wimmer T, Truong P, Rakotosaona MJ, et al (2025) AnyUp: Universal feature upsampling. arXiv preprint arXiv:251012764
- Xue H, Fu H, Ren R, et al (2021) Lidar-based drivable region detection for autonomous driving. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 1110–1116
- Yang L, Liang C, He X, et al (2023) TTA-GEP: Terrain traversability analysis with geometry and environmental perception for the path planning of planetary rovers. *International Journal of Aerospace Engineering* 2023(1):7147168. <https://doi.org/https://doi.org/10.1155/2023/7147168>, URL <https://onlinelibrary.wiley.com/doi/abs/10.1155/2023/7147168>, <https://onlinelibrary.wiley.com/doi/pdf/10.1155/2023/7147168>
- Zuo W, et al (2021) China’s lunar and planetary data system: Preserve and present reliable chang’e project and Tianwen-1 scientific data sets. *Space Science Reviews* <https://doi.org/10.1007/s11214-021-00862-3>