From Earth to the Moon: A Cross-Domain Study of Weakly-Supervised Traversability Estimation*

Bartosz Ptak[1]

***Abstract*— This paper examines how weakly supervised traversability estimation models generalise across domains in the context of lunar robotics. Due to a lack of large-scale annotated datasets for the Moon, models must be trained on Earth-based data and transferred to environments with significantly different visual and physical characteristics. We focus on pairwise, weakly supervised learning, where relative traversability annotations provide a lightweight alternative to dense labelling while preserving the continuous nature of terrain assessment. A systematic evaluation is conducted over seven datasets, including synthetic simulations, terrestrial environments, lunar analogue facilities, and real lunar imagery. The results indicate that cross-domain performance varies significantly depending on the choice of training data. In particular, datasets collected in lunar analogue environments tend to transfer more reliably than standard terrestrial sources or purely synthetic ones. At the same time, strong in-domain performance does not consistently guarantee generalisation to lunar conditions, highlighting a gap between conventional evaluation practices and deployment requirements. Overall, the study provides an empirical basis for selecting training data in lunar applications and underscores the importance of domain relevance in weakly supervised traversability estimation.**

I. 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 generalise to the Moon, where surface appearance, lighting, and morphology differ from those of any available training source.

Fully-supervised semantic segmentation approaches [1], [2] address terrain traverse through dense per-pixel labelling, 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 [3], such approaches rely on large annotated datasets that are not mature enough for deployment on the Moon. Self-supervised methods [4], [5] 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 [6], [7] offer a favourable 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 Fig. 1, each image requires only a small number of point pairs indicating relative traversability.

Despite the technical readiness of weakly supervised traversability methods, their cross-domain behaviour 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 analogues, and lunar analogue 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 analogues, lunar analogue facilities, and real lunar mission imagery. Our contributions are:

1) **A systematic cross-domain evaluation** of weakly-supervised traversability across seven datasets spanning simulations, terrestrial analogues, lunar analogues, and real lunar mission imagery, revealing a clear hierarchy of training sources for transfer to the Moon.
2) **Practical guidelines for training data selection** for lunar missions: lunar analogue facilities provide markedly better training sources than terrestrial envi-

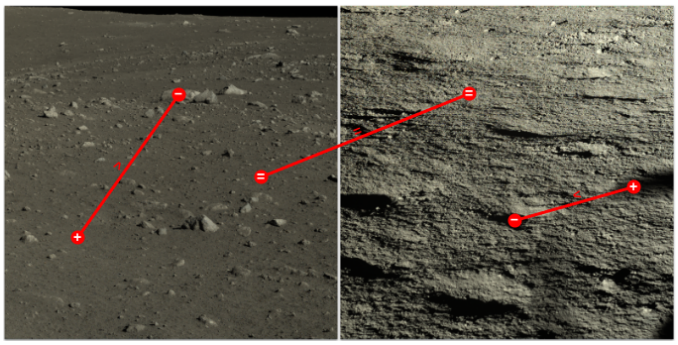

Fig. 1. Sparse pairwise traversability annotations on Chang'E-3/4 lunar imagery [8], [9]. Red lines connect annotated point pairs, with symbols indicating the ordinal relation: the point marked + is more traversable than the point marked −, while = denotes equal traversability.

*This research was financially supported by the Poznań University of Technology as part of a statutory research project (No. 0214/SBAD/0253).

[1]Bartosz Ptak is with Institute of Robotics and Machine Intelligence, Poznań University of Technology, Piotrowo 3A, 60-965 Poznań, Poland bartosz.ptak@put.poznan.pl

ronments or simulations, and in-domain performance is a poor predictor of cross-domain transfer quality.

3) **Validation of a new lunar analogue test site** (OTRK Kakolewo[†]) as a viable training source for lunar traversability, achieving transfer performance comparable to established lunar analogue facilities.

II. METHOD AND EXPERIMENTAL SETUP

Our pipeline follows the weakly-supervised paradigm introduced by W-RIZZ [6] and CHUNGUS [7], 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 dramatically 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 generalisation from limited data are required.

A. Pipeline

Our model architecture builds on CHUNGUS [7], with a key modification in the feature upsampling stage. Input RGB images are passed through a frozen DINOv2 [10] ViT-S [11] 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 [12]; instead, we employ LoftUp [13], 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 [14] 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 [6], 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.

B. HDR Metric

We adopt the Human Disagreement Rate (HDR) metric introduced in W-RIZZ [6], 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 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.

C. Datasets

We evaluate cross-domain transfer across seven datasets spanning four categories of imagery relevant to lunar robotics: lunar mission data, lunar analogue facilities, terrestrial planetary analogues, and synthetic simulations. All datasets were annotated using the same sparse pairwise scheme described in Section II-D. Dataset characteristics are provided below.

- **Lunar mission data.** We use real lunar surface imagery from the panoramic cameras (PCAM) of the Chang'E-3 [8] and Chang'E-4 [9] missions, obtained through the Ground Research and Application System (GRAS) of the China Lunar Exploration Program [15]. The PCAM instruments aboard the Yutu and Yutu-2 rovers acquired ground-level colour images of the lunar surface in Mare Imbrium and the Von Kármán crater on the lunar farside, respectively.
- **Lunar analogue facilities.** *LunaPolaris* [16] was collected in the LunaLab [17] with a small rover. This indoor lunar analogue 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 analogue test facility at OTRK Kakolewo, 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 analogue datasets were collected on Mt. Etna, Sicily, Italy – a site widely used for planetary surface analogue testing due to its volcanic terrain. *ETNA LRNT* [18] (Long Range Navigation Tests) was captured with the Lightweight Rover Unit (LRU) and provides outdoor traversal sequences across volcanic landscapes. Rover is equipped with a stereo-based monochromatic camera setup. *S3LI* [19] (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* [20] is a high-fidelity lunar rover simulator providing photo-realistic synthetic imagery of lunar terrain based on the Unity game engine. *LuSNAR* [21] 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.

[†]http://otrk.put.poznan.pl/

TABLE I

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 artefacts

D. 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 [10] 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.

E. 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 artefacts such as flares and shadows do not reduce traversability when the underlying terrain remains visible. The hierarchy is summarised in Table I.

Annotations are collected as sparse pairwise comparisons following the W-RIZZ protocol [6]: 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 [7]. Per-image annotation takes only a few seconds, enabling efficient labelling across all seven datasets.

III. RESULTS

Figure 2 presents the full cross-domain evaluation matrix, reporting average HDR across all 49 train-evaluation combinations of seven datasets. Diagonal entries (in bold) correspond to in-domain performance; off-diagonal entries represent cross-domain transfer. In-domain performance is a poor predictor of cross-domain transfer: the S3LI dataset achieves the lowest in-domain HDR of any source (0.075), yet provides the worst transfer to lunar imagery (0.309): a pattern reflecting a model that overfits to the visual characteristics of a single domain rather than learning generalisable traversability cues, and a consequential failure mode for verification and validation pipelines that rely on in-domain metrics alone. Visual similarity of training data to the target domain matters more than image realism: both simulation datasets (LunarSim: 0.250, LuSNAR: 0.241) outperform the S3LI terrestrial analogue despite being synthetic, while ETNA LRNT (0.227), a real terrestrial dataset, transfers nearly as well as LunaPolaris. 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.

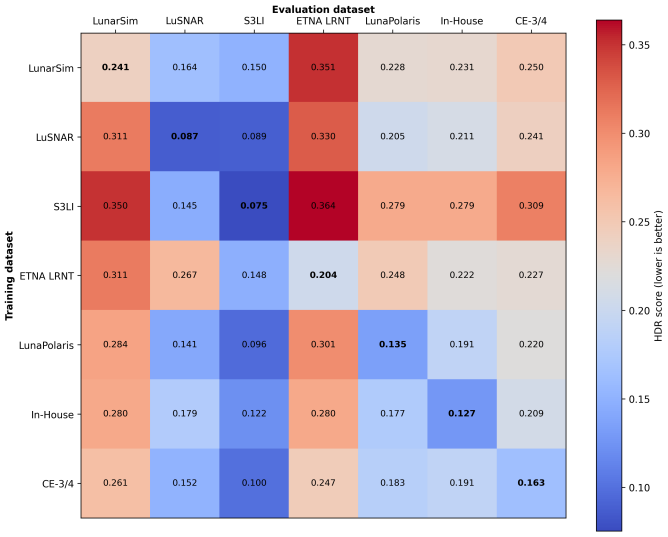

Fig. 2. Comparison of cross-dataset performance using average HDR.

Table II extracts the key column of this matrix – transfer to real lunar mission imagery (CE-3/4) – and ranks training datasets by HDR. Both lunar analogue datasets (In-House: 0.209, LunaPolaris: 0.220) outperform terrestrial analogues, simulations, and S3LI on transfer to CE-3/4. The In-House facility (OTRK Kakolewo), introduced in this work, achieves transfer performance comparable to that of the established LunaPolaris facility, thereby validating its use as a training source for lunar traversability.

The qualitative comparison in Fig. 3 supports these findings. Models trained on lunar analogues 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.

IV. 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 analogue facilities, particularly the comparable behaviour of LunaPolaris and our newly introduced OTRK Kakolewo site, suggests that controlled indoor analogues with regolith-like terrain and lunar-relevant morphology can serve as effective training sources when real

TABLE II

TRAINING DATASET RANKING BASED ON HDR PERFORMANCE ON CE-3/4 LUNAR DATA.

Rank	Training dataset	Category	HDR	HDR@0.10	HDR@0.25
0	CE-3/4	In-domain	0.163	0.210	0.137
1	In-House	Lunar analogue	0.209	0.240	0.200
2	LunaPolaris	Lunar analogue	0.220	0.237	0.220
3	ETNA LRNT	Terrestrial analogue	0.227	0.280	0.200
4	LuSNAR	Simulation	0.241	0.237	0.257
5	LunarSim	Simulation	0.250	0.270	0.253
6	S3LI	Terrestrial analogue	0.309	0.260	0.343

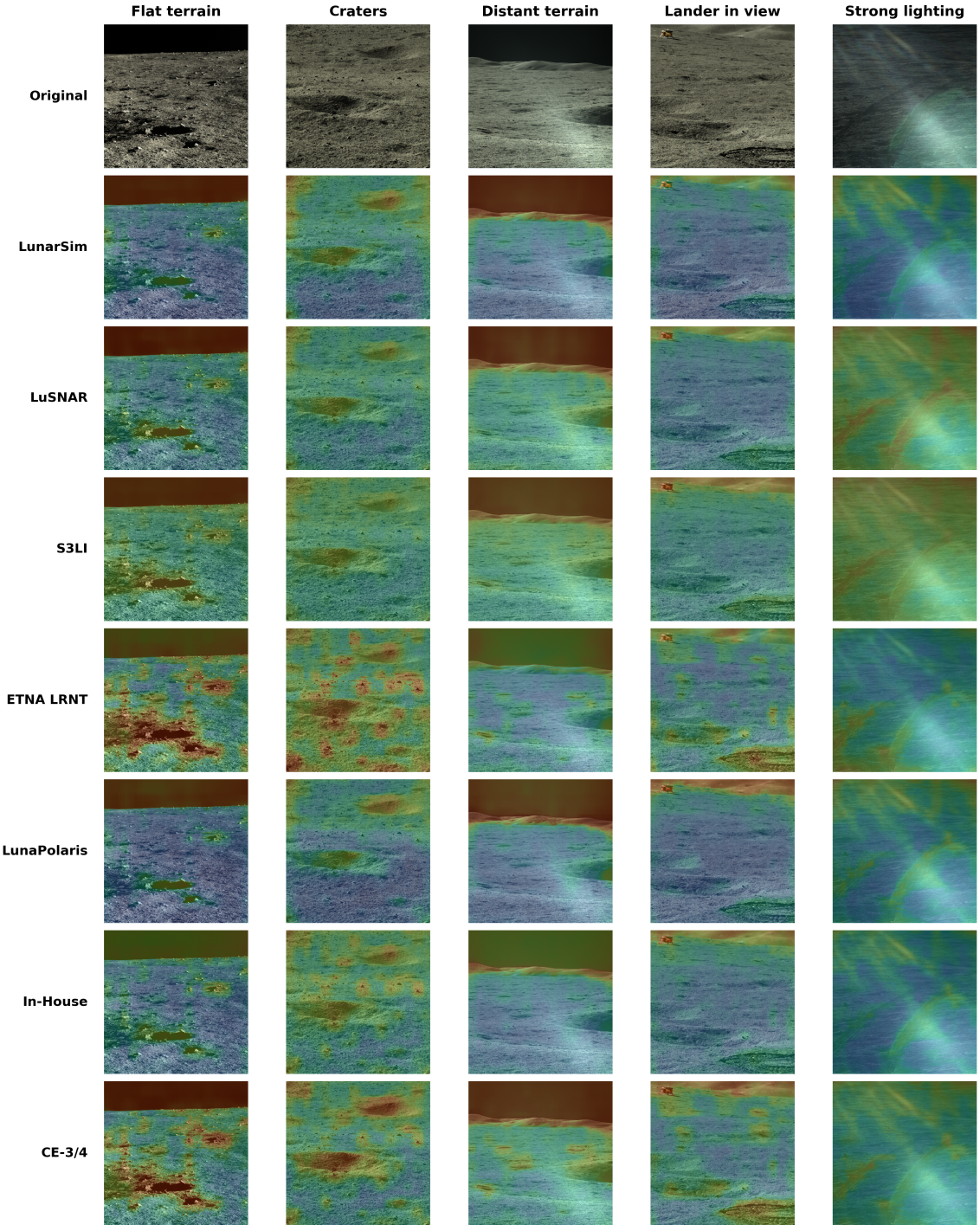

Fig. 3. Visual comparison of training dataset (rows) and selected scenarios (columns).

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.

A. Limitations

Our work has several limitations. The lunar analogue facilities used in this study are indoor environments under controlled lighting, lacking the harsh shadows and dust dynamics characteristic of the actual lunar surface. 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. 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.

B. Future Works

Several directions extend this study toward a more comprehensive understanding of cross-domain traversability behaviour. 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 analogue missions would assess the practical impact of cross-domain gaps on navigation success, beyond the per-image HDR metric used here.

DISCLOSURE

The authors acknowledge the use of a large language model solely for editing and enhancing the manuscript's grammar. No sections of the article, including the main technical content, figures, or images, were generated by the AI system. The use was limited to improving readability and clarity of language.

REFERENCES

[1] D. Maturana, P.-W. Chou, M. Uenoyama, and S. Scherer, "Real-time semantic mapping for autonomous off-road navigation," in *Field and Service Robotics: Results of the 11th International Conference*. Springer, 2018, pp. 335–350.

[2] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "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, 2019, pp. 5000–5007.

[3] B. Rothrock, R. Kennedy, C. Cunningham, J. Papon, M. Heverly, and M. Ono, "SPOC: Deep learning-based terrain classification for Mars rover missions," p. 5539, 2016.

[4] M. V. Gasparino, A. N. Sivakumar, Y. Liu, A. E. Velasquez, V. A. Higuti, J. Rogers, H. Tran, and G. Chowdhary, "WayFAST: Navigation with predictive traversability in the field," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 651–10 658, 2022.

[5] J. Frey, M. Mattamala, N. Chebrolu, C. Cadena, M. Fallon, and M. Hutter, "Fast traversability estimation for wild visual navigation," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[6] A. Schreiber, A. N. Sivakumar, P. Du, M. V. Gasparino, G. Chowdhary, and K. Driggs-Campbell, "W-RIZZ: A weakly-supervised framework for relative traversability estimation in mobile robotics," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5623–5630, 2024.

[7] A. Schreiber and K. Driggs-Campbell, "Do you know the way? human-in-the-loop understanding for fast traversability estimation in mobile robotics," *IEEE Robotics and Automation Letters*, 2025.

[8] Ground Research and Application System of China's Lunar and Planetary Exploration Program, "Chang'E-3 Panoramic Cameras Dataset," 2015, accessed via GRAS data portal. [Online]. Available: https://moon.bao.ac.cn

[9] ——, "Chang'E-4 Panoramic Cameras Dataset," 2020, accessed via GRAS data portal. [Online]. Available: https://moon.bao.ac.cn

[10] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] S. Fu, M. Hamilton, L. Brandt, A. Feldman, Z. Zhang, and W. T. Freeman, "Featup: A model-agnostic framework for features at any resolution," *arXiv preprint arXiv:2403.10516*, 2024.

[13] H. Huang, A. Chen, V. Havrylov, A. Geiger, and D. Zhang, "LoftUp: Learning a coordinate-based feature upsampler for vision foundation models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 9913–9923.

[14] T. Wimmer, P. Truong, M.-J. Rakotosaona, M. Oechsle, F. Tombari, B. Schiele, and J. E. Lenssen, "AnyUp: Universal feature upsampling," *arXiv preprint arXiv:2510.12764*, 2025.

[15] W. Zuo *et al.*, "China's lunar and planetary data system: Preserve and present reliable chang'e project and Tianwen-1 scientific data sets," *Space Science Reviews*, 2021.

[16] D. van der Meer, U. Wong, and M. A. OLIVARES MENDEZ, "LunaPolaris: A stereo camera, point cloud and imu dataset for future lunar exploration in polar regions," in *iSpaRo-International Conference on Space Robotics*, 2026.

[17] P. Ludivig, A. Calzada-Diaz, M. A. OLIVARES MENDEZ, H. VOOS, and J. Lamamy, "Building a piece of the moon: Construction of two indoor lunar analogue environments," in *71st International Astronautical Congress (IAC)–The CyberSpace Edition*, 2020.

[18] M. Vayugundla, F. Steidle, M. Smisek, M. Schuster, K. Bussmann, and A. Wedler, "Datasets of long range navigation experiments in a moon analogue environment on mount etna," in *ISR 2018; 50th International Symposium on Robotics*, 2018, pp. 1–7. [Online]. Available: https://elib.dlr.de/124514/

[19] R. Giubilato, W. Stürzl, A. Wedler, and R. Triebel, "Challenges of slam in extremely unstructured environments: The DLR planetary stereo, solid-state lidar, inertial dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8721–8728, 2022.

[20] D. Pieczyński, B. Ptak, M. Kraft, and P. Drapikowski, "Lunarsim: Lunar rover simulator focused on high visual fidelity and ROS 2 integration for advanced computer vision algorithm development," *Applied Sciences*, vol. 13, no. 22, p. 12401, 2023.

[21] J. Liu, Q. Zhang, X. Wan, S. Zhang, Y. Tian, H. Han, Y. Zhao, B. Liu, Z. Zhao, and X. Luo, "LuSNAR: A lunar segmentation, navigation and reconstruction dataset based on muti-sensor for autonomous exploration," *arXiv preprint arXiv:2407.06512*, 2024.