LuMon: A Comprehensive Benchmark and Development Suite with Novel Datasets for Lunar Monocular Depth Estimation

Anonymous CVPR submission

Paper ID 14

Abstract

Monocular Depth Estimation (MDE) is crucial for autonomous lunar rover navigation using electro-optical cameras. However, deploying terrestrial MDE networks to the Moon brings a severe domain gap due to harsh shadows, textureless regolith, and zero atmospheric scattering. Existing evaluations rely on analogs that fail to replicate these conditions and lack actual metric ground truth. To address this, we present LuMon, a comprehensive benchmarking framework to evaluate MDE methods for lunar exploration. We introduce novel datasets featuring high-quality stereo ground truth depth from the real Chang'e-3 mission and the CHERI dark analog dataset. Utilizing this framework, we conduct a systematic zero-shot evaluation of state-of-the-art architectures across synthetic, analog, and real datasets. We rigorously assess performance against mission critical challenges like craters, rocks, extreme shading, and varying depth ranges. Furthermore, we establish a sim-to-real domain adaptation baseline by fine tuning a foundation model on synthetic data. While this adaptation yields drastic in-domain performance gains, it exhibits minimal generalization to authentic lunar imagery, highlighting a persistent cross-domain transfer gap. Our extensive analysis reveals the inherent limitations of current networks and sets a standard foundation to guide future advancements in extraterrestrial perception and domain adaptation.

1. Introduction

Electro-optical (EO) cameras offer low mass, high resolution, and low power consumption, making them ideal for mini lunar rovers. Extracting actionable 3D maps from these sensors requires Monocular Depth Estimation (MDE) methods that remain accurate and computationally efficient despite the harsh lighting, low-texture regolith, and unique hazards of the lunar surface.

Although deep-learning MDE networks like DepthAnything v2 [48], Marigold [22], and Depth Pro [4] succeed in terrestrial navigation [24, 47, 50], their lunar feasibility remains unproven. The transition to the Moon introduces a severe visual domain gap. The absence of atmospheric scattering, high-contrast polar shadows, and unique topological textures create environments fundamentally different from Earth-based training datasets.

To address this gap, we present a comprehensive benchmarking and development suite for lunar MDE. As existing terrestrial analogs like Mount Etna [13, 41] fail to fully replicate extreme lunar optics, we enable mission-critical validation by constructing high-quality stereo vision based ground truth depth for authentic Chang'e-3 [16] imagery and introducing the novel CHERI dark-analog dataset.

Our suite LuMon provides five main contributions toward deploying MDE on lunar rovers (see also Fig. 1):

- **Framework and Novel Datasets:** We provide a complete MDE evaluation framework (released upon publication), curating six datasets and introducing ground truth depth-maps for Chang'e-3 and the CHERI environment.
- **Systematic Zero-Shot Analysis:** We evaluate 14 leading models across relative, metric, generative, and video-consistent paradigms, measuring robustness against lunar challenges like craters, rocks, and shading.
- **Sim-to-Real Domain Adaptation:** We fine-tune a metric foundation model on synthetic data to expose current generalization gaps in transferring learned features to authentic lunar imagery.
- **Downstream Task Integration:** We integrate depth predictions into a relative pose estimation pipeline to quantify how MDE quality directly impacts robotic autonomy.
- **Limitations and Open Issues:** We identify inherent network limitations and highlight critical open issues to guide future lunar rover navigation research.

2. Related Work

Monocular Depth Estimation (MDE) & Benchmarks. While deep learning has significantly advanced MDE for terrestrial autonomy [3, 20, 32, 48], extracting 3D structure from 2D projections remains inherently ill-posed [35].

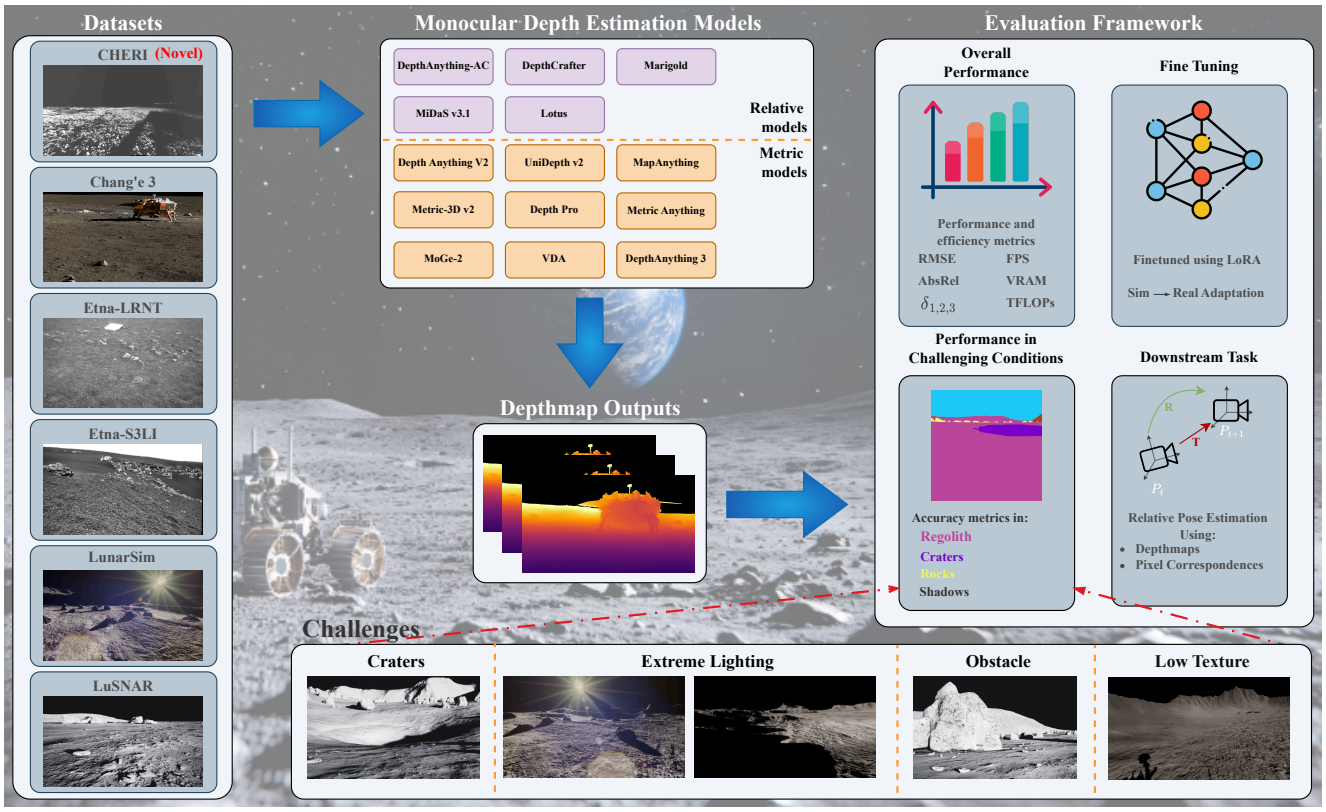


Figure 1. LuMon uses real and synthetic lunar surface datasets. We utilize leading metric and relative monocular depth estimation models to extract depthmaps, and evaluate their performance in zero-shot lunar settings. We measure the overall performance of the networks in terms of accuracy and efficiency, their accuracy in challenging conditions, and effectiveness in downstream tasks. Finally, we perform fine-tuning to compare it against zero-shot performance.

Existing benchmarks primarily target terrestrial indoor [29, 38] or outdoor driving domains [8, 37]. In lunar environments, this challenge is severely exacerbated by extreme high-contrast lighting, zero atmospheric scattering, textureless regolith, and a lack of large-scale metric ground truth. Our study systematically evaluates SOTA methods under these precise constraints to identify specific success and failure modes.

Lunar Depth Estimation & Environments. Historical lunar depth estimation relies on orbiter imagery for Digital Terrain Models [5, 6, 21]. Surface-level evaluations often depend on structurally mismatched bird's-eye data [31] or idealized synthetic datasets [1, 15, 26, 34, 46] that miss authentic lighting, noise, and terrain complexities. While physical testbeds [7, 28] and private simulators [2] exist, they frequently lack synchronized ground truth or public availability. Photorealistic public simulators [27, 33] provide essential dense ground truth but lack physical realism. Real Earth-analog datasets offer physical realism but lack lunar illumination, while some authentic mission datasets suffer from severe sensor misalignment[45]. Consequently, a realistic benchmark must integrate cross-domain data, combining controlled simulations [27, 33], properly ground-truthed Earth analogs [14, 40], and carefully calibrated in-situ stereo datasets [16].

3. LuMon Benchmark Framework

3.1. Objectives

This study explores the feasibility of existing deep-learning MDE methods for lunar exploration. Our benchmark evaluates state-of-the-art models across four key axes: (1) zero-shot accuracy on synthetic and real lunar imagery; (2) robustness to edge cases like craters, rocks, and low illumination; (3) computational efficiency, including latency, memory, and GFLOPs; and (4) depth reliability across varying distance intervals.

3.2. Simulation Environments

We utilize two photorealistic simulation environments to obtain high-precision depth maps under controlled parameters. **LunarSim** [33] is a Unity-based simulator providing stereo imagery and pixel-aligned depth. We captured a custom trajectory sequence, including gt depth-maps and pose

Table 1. A summary of the datasets used. **GT**: Ground Truth; **Seg. Masks**: Segmentation Masks; **Sim**: Simulation.

Dataset	Domain	Seg. Masks	GT Source	Baseline	Resolutions	Max Depth
LunarSim	Sim.	✗	Sim.	12 cm	1280×720	-
LuSNAR	Sim.	✓	Sim.	31 cm	1024×1024	50 m
Etna-S3LI	Analog	✗	LiDAR	20 cm	688×512	30 m
Etna-LRNT	Analog	✗	Stereo	9 cm	1292×964	15 m
CHERI	Analog	✗	Stereo	11.2 cm	1280×720	17 m
Chang'e-3	In-Situ	✗	Stereo	27 cm	2352×1728	25 m

data, to evaluate the MDE models. **LuSNAR** [27] features nine Unreal Engine 4 sequences containing stereo images, dense depth, and semantic segmentation masks for five geological classes, enabling our category-specific evaluations.

3.3. Real & Analog Environments

We incorporate diverse real extraterrestrial imagery and planetary analogs to bridge the persistent sim-to-real gap (See Table Tab. 1). Specifically, a primary contribution of our work is the introduction of the novel **CHERI dataset**, captured in a dark lunar-analog environment, alongside the generation of high-quality ground truth depth for real stereo images from the **Chang'e-3** [16] mission. Since neither the raw analog images nor the historical mission data include native depth, we construct this ground truth using stereo reconstruction. We refine stereo geometry via sparse ORB [36] feature correspondences to ensure strict epipolar consistency, apply sub-pixel accurate rectification, and generate disparity maps using FoundationStereo [44] before converting to metric depth. We mask invalid or uncertain pixels and validate scale accuracy against known physical dimensions. For existing analogs, we evaluate the Etna-LRNT [41], featuring SGM-generated [18] stereo depth, and the **Etna-S3LI dataset** [13], which provides LiDAR ground truth.

3.4. Compared Models

We evaluated 14 models across two paradigms:

Metric Depth Estimators to test absolute scale retention and include Metric-3D v2 [19], UniDepth v2 [32], Depth Pro [4], MoGe-2 [43], MapAnything [23], and Metric Anything [9]. We also evaluate the metric configurations of foundation models DepthAnything 3 [11], VDA [10], and DepthAnything v2 [48].

Relative Depth Estimators to assess the transferability of scale-invariant pre-training. This includes discriminative models like MiDaS v3.1 [3] and DepthAnything-AC [39], diffusion-based approaches Marigold [22], Lotus [17], and DepthCrafter [20] to test generative surface detailing.

3.5. Evaluation Procedure

Following established literature [30], we evaluate all models under a unified setup. Predictions are resized to ground truth resolution. Inverse depth outputs are mapped to standard depth with numerical stabilization ($\epsilon = 1 \times 10^{-6}$). We mask invalid, non-finite, and out-of-range pixels based on dataset modality. We clip stereo datasets (CHERI, Chang'e-3, Etna-LRNT) using KITTI-adapted [12] thresholds because disparity quantization errors become highly unreliable at far ranges. We empirically mask extreme distances in Etna-S3LI and LuSNAR to exclude sky and background artifacts. LunarSim remains unclipped because its relative depth lacks an absolute metric scale. We detail the exact maximum evaluation depths in Table Tab. 1. Finally, we apply a per-image least squares affine alignment uniformly across all architectures, including metric estimators, for a fair comparison. Since metric models learn scale priors primarily from terrestrial features, they exhibit severe global scaling failures when transferred to the lunar domain. This alignment isolates the structural fidelity of the predictions instead of penalizing pure domain-induced scale drift.

Performance metrics. While our evaluation encompasses a comprehensive suite of metrics (such as MAE and SILog), page limits restrict our reported results to the primary metrics: δ_1, AbsRel, and RMSE.

4. EXPERIMENTS

4.1. Experiment Setup and Details

Our comprehensive suite encompasses multiple evaluation axes, in addition to a direct overall comparison:

Evaluation Protocol. To ensure a fair and rigorous comparison, particularly concerning our fine-tuned DepthAnything v2 model, all evaluations on the LuSNAR dataset are strictly conducted on its designated test set. For all other datasets, evaluations are performed across their entirety.

Semantic-aware Analysis. We use the segmentation labels from LuSNAR to assess performance on distinct geological features, such as regoliths, craters, and rocks.

Shadow Analysis. We generate masks using dataset-specific pixel brightness thresholds rather than a global value, refining these regions with a 5x5 morphological opening to reduce noise.

Pose Estimation. We investigate the practical utility of MDE models for downstream applications by estimating pose derived from the predicted depth.

Domain Adaptation. We also explore the impact of domain-specific adaptation by finetuning the metric outdoor weights of the DepthAnything v2 model on the LuSNAR dataset, quantifying the resulting performance shifts in lunar environments.

Efficiency. We report speed and memory metrics at a standard 1280×720 resolution, evaluated on a standardized setup featuring an Intel(R) Xeon(R) Gold 6326 CPU and an NVIDIA RTX A6000 GPU.

MDE Network Details. We evaluated the highest-

performing pre-trained weights. Metric models: DepthAnything v2 (metric-outdoor ViT-L), VDA (metric ViT-L), DepthAnything 3 (mono-large), MoGe-2 (ViT-L normal), UniDepth v2 (v2 ViT-L/14), Metric-3D v2 (ViT-L), Depth Pro (standard), Metric Anything (student pointmap), MapAnything (CC). Relative models: DepthAnything-AC (ViT-S), MiDaS v3.1 (BEiT_{512}-L), Marigold (v1.1), Lotus (v2.0 disparity), DepthCrafter (SVD-xt).

4.2. Exp. 1: Overall Comparison

In this experiment, we evaluate the zero-shot performances of the selected MDE networks alongside our domain-adapted baseline across all six datasets (Table 2). Rather than listing isolated dataset metrics, our analysis reveals two critical generalization trends.

Metric Foundation Model Dominance. Metric depth estimators consistently outrank relative models across the benchmarking suite. Architectures like MapAnything, VDA, and DepthAnything 3 exhibit the most robust cross-domain zero-shot performance. While no single network dominates synthetic datasets, metric foundation models maintain superior geometric consistency and structural fidelity on authentic extraterrestrial imagery like the Chang'e-3 dataset. Because our affine-aligned protocol controls for absolute scale recovery, this advantage strictly reflects stronger learned geometric priors. Additionally, our fine-tuned DAv2-FT baseline establishes an upper bound for synthetic environments, achieving state-of-the-art accuracy on its training domain (LuSNAR) with excellent sim-to-sim transfer on LunarSim.

The Earth-Analog Illusion. Traditional Earth-based analogs often prove deceptive for benchmarking. Most models achieve near-perfect accuracy on the well-lit Etna-LRNT dataset because its volcanic terrain closely mirrors standard terrestrial training distributions. However, performance universally plummets on Etna-S3LI due to stricter LiDAR ground truths, and on the CHERI dataset, where harsh, high-contrast polar lighting confounds all architectures.

Main Result

Metric foundation models provide the strongest zero-shot baseline for lunar topography, yet all architectures degrade severely under authentic high-contrast lighting. Near-perfect results on well-lit Earth analogs prove deceptive, confirming that current terrestrial proxies are insufficient for evaluating mission-critical extraterrestrial deployment.

4.3. Exp. 2: Challenging Surface Constructs

We evaluate model robustness against specific lighting and topological edge cases. We apply brightness-based shadow masks across all datasets to analyze shading impacts (Table 4) and leverage LuSNAR segmentation labels to isolate performance on regolith, rocks, and craters (Table 5). We refer readers to Section **??** of the supplementary material for our complete analysis.

Shaded Regions. Shading causes less overall degradation than complex surface geometry. Our fine-tuned baseline, DAv2-FT, demonstrates exceptional robustness on synthetic benchmarks like LuSNAR, while MapAnything leads on LunarSim and Etna-S3LI. For authentic extraterrestrial imagery (Chang'e-3), metric foundation models like DepthAnything v2, DepthAnything 3, and Metric-3D v2 deliver the highest accuracy. Notably, Depth Pro exhibits a catastrophic failure within the Etna-LRNT shaded split.

Topological Features. Our adapted DAv2-FT pipeline comprehensively dominates all topological edge cases, showcasing excellent recovery of regolith, rocks, and craters. Among zero-shot architectures, while MiDaS v3.1 and VDA perform reasonably well on flat regolith, metric models like MapAnything and Metric-3D v2 show a distinct advantage in resolving rock obstacles. Craters cause severe performance degradation across all zero-shot architectures, although metric models remain comparatively more resilient. Ultimately, craters and rocks pose a significantly greater hurdle for current networks than extreme shading.

Main Result

Complex surface geometries like craters and rocks degrade zero-shot depth estimation accuracy significantly more than severe lunar shading. Metric models consistently outperform relative architectures across these edge cases, and our domain-adapted DAv2-FT demonstrates that targeted fine-tuning can successfully overcome these topological and lighting hazards within simulated environments.

4.4. Exp. 3: Performance vs. Distance

In this experiment, we analyze how depth estimation performance changes with distance to evaluate global metric consistency (Table 6). We apply a global least-squares affine alignment, which is heavily influenced by the dominant median depth range. We therefore treat this mid-range as an implicit anchor for testing *internal linearity*. Because this alignment inherently stabilizes central distances, distance-dependent scale drift becomes most pronounced at the depth extremes. Consequently, our main evaluation isolates the unanchored Near and Far boundary regions to specifically highlight these non-linear distortions (complete mid-range analysis is provided in Suppl. Sec. **??**).

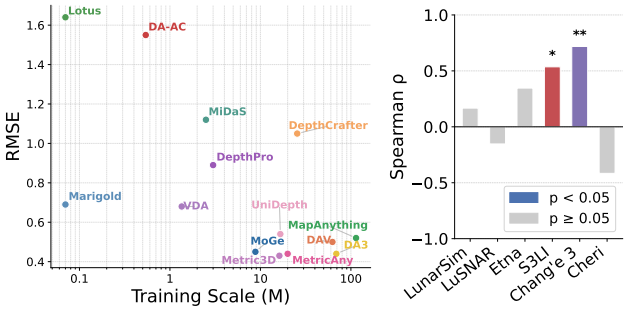
Using this setup, we observe that relative and diffusion-based models (Lotus, DepthCrafter) exhibit strong non-linear distortions, failing to transfer scale consistently into the boundary regions. In contrast, metric models (MapAnything, Metric-3D v2) and structurally constrained meth-

Table 2. Exp. 1: Evaluation on LunarSim, LuSNAR, Etna, S3LI, Chang'e 3, and Cheri datasets. **Red** and blue respectively indicate best and second-best results.

Type	Model	LunarSim			LuSNAR			Etna-LRNT			Etna-S3LI			Chang'e-3			CHERI		
		$\delta_1 \uparrow$	A.Rel↓	RMSE↓	$\delta_1 \uparrow$	A.Rel↓	RMSE↓	$\delta_1 \uparrow$	A.Rel↓	RMSE↓	$\delta_1 \uparrow$	A.Rel↓	RMSE↓	$\delta_1 \uparrow$	A.Rel↓	RMSE↓	$\delta_1 \uparrow$	A.Rel↓	RMSE↓
Metric	DAv2-FT (Ours)	0.84	0.13	13.96	**0.93**	**0.07**	**1.07**	**1.00**	0.05	0.13	0.84	0.13	1.30	0.92	0.08	0.70	0.36	0.50	3.90
	DepthAnything v2 [48]	0.56	0.26	15.18	0.31	0.45	2.89	**1.00**	0.06	0.15	0.88	0.12	1.22	0.95	0.06	0.56	0.32	0.54	4.14
	DepthAnything 3 [11]	0.63	0.22	14.15	0.53	0.29	1.99	**1.00**	0.04	0.11	**0.91**	**0.10**	1.12	**0.96**	**0.05**	0.44	0.34	0.51	3.97
	MoGe-2 [43]	0.77	0.15	12.17	0.53	0.29	1.80	0.99	0.05	0.15	**0.91**	**0.10**	1.08	0.95	0.06	0.45	0.32	0.54	4.12
	Depth Pro [4]	0.50	0.29	17.57	0.33	0.45	2.70	0.73	241.71	450.05	0.89	0.11	1.11	0.89	0.10	0.89	0.32	0.55	4.18
	Metric-3D v2 [19]	0.75	0.16	12.53	0.61	0.25	1.71	**1.00**	**0.03**	0.10	0.87	0.12	1.19	0.95	**0.05**	**0.43**	0.33	0.53	4.09
	UniDepth v2 [32]	0.76	0.21	16.16	0.63	0.26	2.14	**1.00**	**0.03**	**0.07**	0.87	0.12	1.20	0.94	0.07	0.54	0.35	0.51	3.99
	MapAnything [23]	**0.86**	**0.11**	**10.73**	0.74	0.17	1.53	**1.00**	**0.03**	0.08	**0.91**	**0.10**	**1.05**	0.95	0.07	0.52	0.36	0.50	3.91
	VDA [10]	0.75	0.16	13.94	0.77	0.17	2.05	**1.00**	0.04	0.12	0.83	0.14	1.34	0.93	0.08	0.68	0.43	0.44	3.56
	Metric Anything [9]	0.79	0.15	12.08	0.51	0.30	1.83	0.98	0.07	0.22	**0.91**	**0.10**	**1.05**	0.95	0.06	0.44	0.32	0.55	4.17
Relative	DepthAnything-AC [39]	0.70	0.28	19.12	0.70	0.26	2.63	**1.00**	0.05	0.12	0.87	0.12	1.19	0.66	0.22	1.55	**0.55**	**0.35**	**3.09**
	DepthCrafter [20]	0.40	0.37	20.28	0.41	0.47	2.96	0.92	0.10	0.28	0.80	0.15	1.44	0.87	0.13	1.05	0.33	0.53	4.08
	Lotus [17]	0.59	0.27	20.43	0.36	0.44	3.08	0.65	0.19	0.49	0.80	0.15	1.45	0.70	0.19	1.64	0.35	0.49	3.90
	MiDaS v3.1 [3]	0.78	0.17	14.97	0.81	0.13	1.68	**1.00**	0.04	0.12	0.89	0.11	1.15	0.81	0.15	1.12	0.40	0.46	3.71
	Marigold [22]	0.67	0.21	17.25	0.62	0.24	2.02	0.94	0.10	0.27	0.86	0.12	1.26	0.91	0.09	0.69	0.38	0.47	3.75

Table 3. Efficiency metric results. **Red** indicates the best result and blue indicates the second best.

Type	Model	FPS ↑	VRAM (GB) ↓	TFLOPs ↓
Metric	DAv2-FT (Ours)	4.64	1.926	1.309
	DepthAnything v2	4.54	1.926	1.309
	DepthAnything 3	24.65	2.007	0.658
	MoGe-2	2.63	2.234	2.150
	Depth Pro	0.91	7.949	9.677
	Metric-3D v2	2.47	2.990	2.870
	UniDepth v2	12.39	3.126	1.868
	MapAnything	13.22	7.080	1.101
	VDA	0.51	10.385	75.444
	Metric Anything	11.04	2.832	2.150
Relative	DepthAnything-AC	**34.24**	**0.509**	**0.130**
	DepthCrafter	0.01	20.166	18.472
	Lotus	5.83	3.510	2.851
	MiDaS v3.1	3.46	2.135	0.920
	Marigold	0.02	6.117	272.109



(a) Training Scale vs. RMSE (b) Training Scale vs. δ_1

Figure 2. Effect of training dataset scale for one dataset. (a) Scale vs. RMSE. (b) Spearman correlation of scale vs. δ_1 ($*$: $p < 0.05$, $**$: $p < 0.01$).

ods (VDA, MiDaS v3.1) maintain significantly lower and more uniform errors, better preserving the internal linearity vital for stable 3D reconstruction. However, near-range performance degrades sharply on authentic Chang'e-3 and CHERI imagery across most architectures, revealing boundary cases where linear scaling assumptions and domain-specific priors break down.

Main Result

Metric and structurally constrained models preserve the internal linearity required for stable 3D geometry, overcoming the distance-dependent scale drift prevalent in relative and diffusion methods.

4.5. Exp. 4: Training Regime and Data Analysis

We investigate zero-shot generalization factors using a Spearman rank correlation (ρ, Fig. 2) between performance metrics and three quantitative training attributes (Tab. 7): Training Scale (total image count), Real Data Ratio, and Metric Supervision Ratio. We refer readers to Section **??** of the supplementary material for our complete analysis.

Impact of Training Scale. Training scale appears to be dataset dependent: On authentic Chang'e-3 imagery, scale correlates significantly with higher depth accuracy (δ_1: $\rho = 0.718$, $p = 0.004$) and reduced error (RMSE: $\rho = -0.619$, $p = 0.018$), even within heavy shadows. However, these correlations weaken or turn negative on the Cheri analog and synthetic benchmarks, proving massive training scale alone cannot universally solve lunar adaptation.

Real Training Data. The ratio of real terrestrial training data does not correlate with performance improvements ($\rho = -0.677$, $p = 0.008$). Thus, terrestrial texture priors primarily assist with nearby high frequency details but fail to provide geometric understanding required for long range lunar topography.

Role of Metric Supervision. We observed no statistically significant correlation between metric supervision percent-

Table 4. Exp. 2: Performances on shaded regions. **Red** indicates the best result and blue indicates the second best.

Type	Model	LunarSim			LuSNAR			Etna-LRNT			Etna-S3LI			Chang'e-3			CHERI		
		$\delta_1 \uparrow$	A.Rel↓	RMSE↓	$\delta_1 \uparrow$	A.Rel↓	RMSE↓	$\delta_1 \uparrow$	A.Rel↓	RMSE↓	$\delta_1 \uparrow$	A.Rel↓	RMSE↓	$\delta_1 \uparrow$	A.Rel↓	RMSE↓	$\delta_1 \uparrow$	A.Rel↓	RMSE↓
Metric	DAv2-FT (Ours)	0.79	0.15	15.10	**0.84**	**0.11**	**0.95**	**1.00**	0.06	0.10	0.46	0.29	2.88	0.89	0.10	0.73	0.16	0.51	4.76
	DepthAnything v2 [48]	0.47	0.29	16.39	0.20	0.66	2.86	0.99	0.07	0.12	0.51	0.33	2.85	**0.94**	0.07	0.59	0.15	0.72	5.18
	DepthAnything 3 [11]	0.54	0.26	15.77	0.34	0.40	2.15	**1.00**	0.05	0.07	0.53	0.46	3.58	**0.94**	**0.06**	0.45	0.18	0.62	4.75
	MoGe-2 [43]	0.68	0.18	14.42	0.35	0.41	1.90	0.99	0.05	0.08	0.56	0.38	3.22	0.93	0.07	0.46	0.15	0.73	5.20
	Depth Pro [4]	0.38	0.35	21.33	0.23	0.68	2.83	0.39	559.37	818.97	0.59	0.25	2.36	0.85	0.11	0.87	0.13	0.78	5.27
	Metric-3D v2 [19]	0.62	0.20	14.35	0.39	0.38	1.92	**1.00**	0.05	0.07	0.68	0.23	2.35	**0.94**	**0.06**	**0.44**	0.16	0.68	5.06
	UniDepth v2 [32]	0.70	0.23	24.12	0.45	0.36	2.85	**1.00**	**0.03**	0.05	0.75	0.17	1.90	0.92	0.08	0.54	0.17	0.64	4.94
	MapAnything [23]	**0.80**	**0.13**	**11.48**	0.60	0.22	1.64	**1.00**	**0.03**	**0.04**	**0.80**	**0.15**	**1.63**	0.93	0.08	0.55	0.15	0.63	4.95
	VDA [10]	0.66	0.19	16.31	0.65	0.23	2.19	**1.00**	0.06	0.09	0.48	0.28	2.61	0.90	0.10	0.68	0.27	0.51	4.23
	Metric Anything [9]	0.74	0.16	13.19	0.32	0.42	2.00	0.99	0.07	0.12	0.68	0.23	2.22	**0.94**	**0.06**	0.46	0.14	0.74	5.24
Relative	DepthAnything-AC [39]	0.63	0.33	27.04	0.54	0.38	3.80	**1.00**	0.09	0.13	0.62	0.20	2.13	0.58	0.26	1.68	**0.47**	**0.41**	**3.96**
	DepthCrafter [20]	0.37	0.45	22.22	0.28	0.81	3.17	0.89	0.11	0.18	0.46	0.38	3.33	0.83	0.14	1.11	0.21	0.76	5.01
	Lotus [17]	0.46	0.32	29.36	0.28	0.64	3.94	0.45	0.29	0.42	0.48	0.32	3.17	0.63	0.23	1.81	0.21	0.58	4.71
	MiDaS v3.1 [3]	0.73	0.19	16.70	0.64	0.20	1.99	0.99	0.07	0.10	0.79	**0.15**	1.75	0.74	0.18	1.17	0.27	0.55	4.23
	Marigold [22]	0.51	0.29	21.92	0.43	0.37	2.40	0.90	0.11	0.17	0.49	0.35	3.21	0.89	0.10	0.72	0.18	0.51	4.66

Table 5. Exp. 2: Results on **semantically** different regions on LuSNAR dataset (Regolith, Rock and Craters). **Red** indicates the best result and blue indicates the second best.

Type	Model	Regolith		Rocks		Craters	
		$\delta_1 \uparrow$	A.Rel↓	$\delta_1 \uparrow$	A.Rel↓	$\delta_1 \uparrow$	A.Rel↓
Metric	DAv2-FT (Ours)	**0.93**	**0.07**	**0.92**	**0.09**	**0.87**	**0.10**
	DepthAnything v2 [48]	0.30	0.46	0.44	0.28	0.23	0.37
	DepthAnything 3 [11]	0.52	0.29	0.65	0.19	0.37	0.27
	MoGe-2 [43]	0.52	0.30	0.69	0.18	0.40	0.27
	Depth Pro [4]	0.32	0.47	0.45	0.28	0.23	0.37
	Metric-3D v2 [19]	0.60	0.26	0.80	0.15	0.41	0.27
	UniDepth v2 [32]	0.62	0.27	0.76	0.18	0.45	0.25
	MapAnything [23]	0.73	0.17	0.85	0.14	0.42	0.26
	VDA [10]	0.77	0.17	0.80	0.16	0.36	0.29
	Metric Anything [9]	0.50	0.31	0.67	0.18	0.41	0.27
Relative	DepthAnything-AC [39]	0.69	0.27	0.66	0.24	0.31	0.31
	DepthCrafter [20]	0.40	0.48	0.43	0.27	0.17	0.41
	Lotus [17]	0.35	0.46	0.51	0.28	0.21	0.39
	MiDaS v3.1 [3]	0.81	0.13	0.76	0.19	0.34	0.28
	Marigold [22]	0.62	0.24	0.66	0.21	0.26	0.33

Table 6. Exp. 3: Evaluation of Absolute Relative Error (A.Rel ↓) on all datasets except LunarSim across Near (N) and Far (F) distances. **Red**: Best, blue: Second-best.

Type	Model	Near					Far				
		LuSNAR	Etna-LRNT	Etna-S3LI	Chang'e-3	CHERI	LuSNAR	Etna-LRNT	Etna-S3LI	Chang'e-3	CHERI
Metric	Dav2-FT (Ours)	**0.07**	0.05	0.18	0.34	1.55	**0.09**	0.20	0.29	0.15	0.23
	DepthAnything v2 [48]	0.49	0.06	0.15	**0.18**	1.74	0.20	0.30	0.29	0.16	0.25
	DepthAnything 3 [11]	0.31	0.04	0.14	0.20	1.58	0.16	0.29	0.27	0.14	0.24
	MoGe-2 [43]	0.32	0.05	**0.13**	0.22	1.72	0.13	0.56	0.26	0.14	0.25
	Depth Pro [4]	0.50	241.77	**0.13**	0.34	1.77	0.18	0.69	0.27	0.20	0.25
	Metric-3D v2 [19]	0.28	0.04	0.17	0.20	1.71	0.14	0.26	0.26	**0.13**	0.25
	UniDepth v2 [32]	0.30	**0.03**	0.18	0.24	1.61	0.15	0.12	0.28	0.14	0.24
	MapAnything [23]	0.18	**0.03**	**0.13**	0.25	1.59	0.12	0.25	**0.25**	0.15	0.23
	VDA [10]	0.19	0.04	0.19	0.26	1.28	0.16	0.11	0.29	0.16	0.21
	Metric Anything [9]	0.33	0.07	**0.13**	0.21	1.77	0.13	0.46	0.26	0.14	0.25
Relative	DepthAnything-AC [39]	0.31	0.05	0.16	0.57	**0.93**	0.21	**0.06**	0.27	0.26	**0.17**
	DepthCrafter [20]	0.52	0.10	0.21	0.22	1.68	0.21	0.48	0.34	0.21	0.25
	Lotus [17]	0.53	0.19	0.21	0.49	1.54	0.25	0.39	0.33	0.29	0.24
	MiDaS v3.1 [3]	0.14	0.04	0.15	0.45	1.38	0.14	0.17	**0.25**	0.19	0.22
	Marigold [22]	0.25	0.10	0.16	0.29	1.44	0.16	0.51	0.30	0.17	0.23

Table 7. Exp 4: Comparison of Training Data Scales and Composition. **Scale:** Total data size. **Real %:** Percentage of real data. **MetSup:** Percentage of data with metric supervision.

Type	Model	Scale	Real%	MetSup	Dominant Scene
Metric	DAv2-FT (Ours)	62.51M	99%	1%*	Gen. Open-World, LuSNAR.
	DepthAnything v2 [48]	62.5M	99%	1%*	General Open-World
	DepthAnything 3 [11]	69M	97%	100%*	General Open-World
	MoGe-2 [43]	8.8M	74%	98%	Indoor
	Depth Pro [4]	3M	15%	98%	Synthetic Simulation
	Metric-3D v2 [19]	16.2M	99%	85%	Indoor, Outdoor Driving
	UniDepth v2 [32]	16.6M	87%	100%	General Open-World
	MapAnything [23]	114M‡	49%	54%	Indoor, Gen. OpenWorld
	VDA[10]	1.35M	59%	41%	General Open-World
	Metric Anything [9]	20M	88%	100%	Indoor, Outdoor Driving
Relative	DepthAnything-AC[39]	0.54M	92%	23%	General Open-World
	DepthCrafter [20]	25.5M†	98%	2%	General Open-World
	Lotus [17]	0.07M	0%	100%	Indoor
	MiDaS v3.1 [3]	2.5M	80%	35%	General Open-World
	Marigold [22]	0.07M	0%	100%	Indoor

† ≈25.45M frames estimated from 203k video sequences.

‡ Estimated from ≈188k scenes.

* Utilize a Teacher-Student paradigm; DAv2 uses affine-invariant pseudo-labels, while DA3 aligns labels to metric ground truth via RANSAC.

age and zero-shot performance across any dataset ($p > 0.05$). This demonstrates that raw metric label volume cannot guarantee cross domain robustness, particularly alongside affine invariant loss functions. Lunar terrain reliability stems from architectural choices or massive scale invariant learning rather than absolute metric data quantities.

Main Result

While large training scale and terrestrial data marginally improve near-field lunar perception, they fail to resolve the broader extraterrestrial domain gap. Since raw metric supervision also cannot guarantee cross-domain robustness, future advancements require domain-specific architectural innovations rather than simple data scaling.

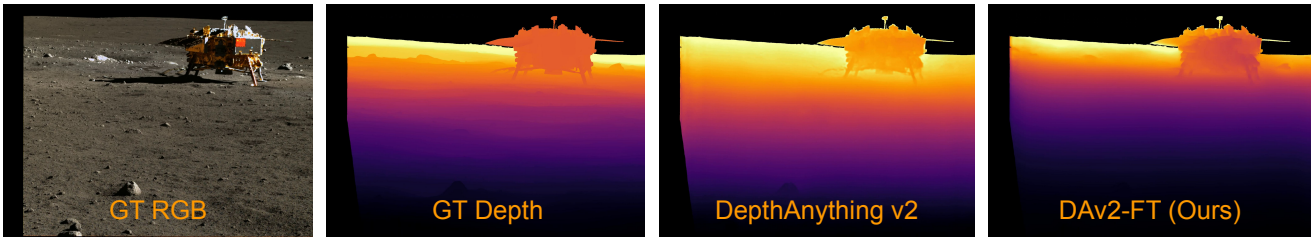


Figure 3. An example scene from Chang'e 3 (`RGB GT`), its ground-truth depth map (`GT Depth`), and the DepthAnything v2 predictions (base and fine-tuned). Note: Differences in visual boundaries between the `RGB GT` and depth maps result from the evaluation procedure (Section 3.5); where ground-truth maps and predictions are masked by maximum depth boundaries and invalid pixels. Additionally, the depth maps generated by the other models can be found in the supplementary material.

4.6. Exp. 5: Sim-to-Real Domain Adaptation

As a practical demonstration of our benchmarking framework, we finetuned the metric-outdoor weights of DepthAnything v2 explicitly on the LuSNAR dataset. For high computational efficiency, we employed Low-Rank Adaptation (LoRA, $r = 8$). By freezing the encoder backbone and updating only the decoder with LoRA, we reduced the trainable parameter count to ∼34.1M from 335.3M. When evaluated on the LuSNAR test set, the finetuned DAv2-FT model demonstrates a drastic performance improvement over its zero-shot baseline (See Tabs. 2, 4 to 6 and 8). This accuracy leap is particularly pronounced across complex crater and rock formations, as well as heavily shaded regions. Furthermore, when applied to our downstream relative pose estimation pipeline, DAv2-FT delivers highly consistent geometric priors, achieving the lowest median rotation error and significantly outperforming the zero-shot base model.Crucially, this overall boost extends beyond the training domain: On LunarSim with a distinct photorealism, DAv2-FT exhibits strong sim-to-sim generalization and outperforms the base model. However, on a broader selection of real and analog datasets, such as Etna-LRNT, Etna-S3LI, and Chang'e-3, DAv2-FT achieves only marginal improvements. The model successfully aligns with specific simulator characteristics but struggles to transfer those learned features to authentic lunar photography and physical analogs.

Main Result

Parameter-efficient finetuning on synthetic data vastly improves downstream geometric consistency and bridges the sim-to-sim domain gap, but it fails to resolve the broader sim-to-real domain shift. Relying solely on simulated environments remains insufficient for lunar deployment, highlighting the necessity of diverse, real-world benchmarks.

4.7. Exp. 6: Relative Pose Est. using MDE Output

We further analyze the suitability of MDE models for robotic autonomy in lunar environments by integrating them into a relative pose estimation pipeline.

Table 8. Exp. 6: Relative Pose Estimation using MADPose on LuSNAR test-set. **Red**: Best, blue: Second-best.

Type	Method	*Med. Err.* ↓		*Pose Err. AUC (%)* ↑		
		ϵ_R (°)	ϵ_t (°)	@5°	@10°	@20°
Metric	DAv2-FT (Ours)	**0.12**	**1.59**	57.70	78.05	**89.03**
	DepthAnything v2 [48]	0.15	1.89	56.22	77.27	88.64
	DepthAnything 3 [11]	**0.14**	2.17	53.28	76.04	88.02
	MoGe-2 [43]	0.15	1.68	**58.40**	**78.52**	**89.26**
	Depth Pro [4]	0.22	2.84	41.12	69.78	84.89
	Metric-3D v2 [19]	0.29	3.24	37.25	67.22	83.61
	UniDepth v2 [32]	0.15	2.12	53.69	76.11	88.06
	MapAnything [23]	0.21	2.36	47.26	72.19	86.09
	VDA [10]	0.22	2.98	38.86	68.39	84.20
	Metric Anything [9]	0.19	2.63	44.40	70.49	85.24
Relative	DepthAnything-AC [39]	0.15	**1.48**	**58.70**	**78.29**	88.81
	DepthCrafter [20]	0.48	3.54	30.02	60.96	79.37
	Lotus [17]	0.28	3.21	33.37	61.85	79.94
	MiDaS v3.1 [3]	0.26	3.72	30.18	58.25	77.59
	Marigold [22]	0.21	2.68	47.11	72.69	86.34
	MASt3R [25]+GT	0.11	1.32	61.78	80.25	90.12

We adopt MADPose [49], a recent relative pose estimation solver that combines dense pixel correspondences with depth priors extracted from image pairs. By substituting the depth prior in MADPose with outputs from different MDE networks, we measure how depth quality affects pose estimation accuracy under lunar-like conditions. For our experiments, we use MASt3R [25] as the correspondence model and the previously evaluated MDE models as depth estimators.Given two successive frames I_1 and I_2 in a trajectory, MASt3R produces dense 2D pixel correspondences p_{1j} and p_{2j}, while the MDE models predict the per-pixel depth maps D_1 and D_2. Using these predicted depths, we back-project each pixel correspondence into 3D camera coordinates P_{1j} and P_{2j}. MADPose solvers then estimate the relative pose $T \in \mathrm{SE}(3)$ such that $P_{2j} = RP_{1j} + t$, where $T = [R \mid t]$, $R \in SO(3)$, and $t \in \mathbb{R}^3$. During the evaluation, we mask out sky regions to eliminate invalid correspondences and improve solver performance. Following [49] and [42], we report the median rotation error ϵ_R, median translation error ϵ_t, and AUC@τ for $\tau \in \{5°, 10°, 20°\}$.

As shown in Table 8, MDE models provide accurate depth priors for pose estimation in lunar surface conditions, indicating that MDE-based pose estimation is a viable solution for extraterrestrial navigation. Furthermore, metric MDE models consistently produce more geometrically reliable and consistent priors than relative MDE models.

Main Result

Metric MDE models generate highly consistent geometric priors that enable accurate relative pose estimation on the lunar surface, significantly outperforming relative architectures and proving their direct viability for downstream extraterrestrial navigation tasks. Furthermore, finetuning on lunar data improves pose estimation results.

4.8. Supplementary Material

The supplementary material includes an appendix with additional experimental results and detailed descriptions of the experimental setups. In addition, we provide video demonstrations of the results and the scripts used to generate the results.

5. Conclusion and Discussion

In this paper, we addressed the critical domain gap in deploying terrestrial Monocular Depth Estimation (MDE) networks for autonomous lunar rover navigation. We introduced LuMon, a comprehensive benchmarking framework and development suite featuring six diverse datasets. This suite notably includes novel high-quality ground truth for the authentic Chang'e-3 mission and the challenging Cheri dark analog environment. By systematically evaluating 14 state-of-the-art architectures, we established a rigorous foundation for understanding how current models handle the harsh lighting, textureless regolith, and unique topological hazards of the lunar surface.

5.1. Summary of Findings

Rather than restating our quantitative results, we highlight the overarching themes revealed by our benchmarking suite:

The Limits of Terrestrial Priors. Massive training scale and high ratios of real terrestrial data marginally improve near-field details but fail to bridge the broader extraterrestrial domain gap. Furthermore, Earth-based analogs can be deceptive; models easily parse well-lit volcanic terrain due to its similarity to terrestrial training distributions, yet they degrade severely under authentic high-contrast polar lighting and on complex geometries like craters.

Advantage of Metric Foundation Models. Metric depth estimators consistently outrank relative and diffusion-based models across the benchmark. They maintain superior internal linearity and structural integrity, proving far more robust for downstream tasks like relative pose estimation where consistent metric scaling is paramount.

The Sim-to-Real Bottleneck. Our parameter-efficient finetuning baseline demonstrates that while adapting foundation models on synthetic data yields excellent sim-to-sim transfer, it produces minimal gains on authentic lunar imagery. This stagnation proves that relying solely on simulated environments or raw data scaling is insufficient for mission-critical deployment.

5.2. The LuMon Development Suite

We designed LuMon not just as a static benchmark, but as an active development suite for the robotics and vision communities. Researchers can leverage our curated synthetic, analog, and real datasets to rigorously test novel architectures against specific lunar edge cases, such as extreme shadows and obstacle detection. Additionally, our established zero-shot rankings, downstream pose estimation pipeline, and LoRA fine-tuning baseline provide a standardized framework for developing and validating new domain adaptation strategies.

5.3. Open Challenges and Future Work

Based on our analysis, we identify three primary challenges that must guide future lunar perception research:

- **Domain-Specific Architectural Innovation.** Since scaling terrestrial data and raw metric supervision does not guarantee cross-domain robustness, future work must focus on architectural innovations. Developing models that explicitly decouple illumination from geometric features or integrate physics-based rendering priors will be necessary to overcome the sim-to-real domain shift.
- **Robustness to Sensor Calibration Artifacts.** Our experiments expose a systematic vulnerability in architectures, notably Depth Pro, to black border artifacts caused by stereo rectification. Since authentic planetary missions rely on stereo cameras, these zero-value boundaries are ubiquitous. While cropping restores performance, this extreme sensitivity highlights a critical flaw in terrestrial models trained on idealized rectangular images. Future methodologies must teach networks to natively ignore unstructured boundaries, ensuring robust perception on raw mission data.
- **On-Board Computational Constraints.** Our evaluation utilized the most capable weights for each architecture to establish an accuracy upper bound. However, practical lunar rovers operate under extreme power and memory constraints. Future research must utilize the LuMon suite to investigate the trade-off between estimation accuracy and computational efficiency, ultimately identifying or distilling lightweight architectures suitable for deployment on resource-constrained space hardware.

References

[1] Ramchander Bhaskara, Georgios Georgakis, Jeremy Nash, Marissa Cameron, Joseph Bowkett, Adnan Ansar, Manoranjan Majji, and Paul Backes. Icy moon surface simulation and stereo depth estimation for sampling autonomy. In *2024 IEEE Aerospace Conference*, page 1–16. IEEE, 2024. 2

[2] Lee Bingham, Jack Kincaid, Benjamin Weno, Nicholas Davis, Eddie Paddock, and Cory Foreman. Digital lunar exploration sites unreal simulation tool (dust). In *Proc. IEEE Aerospace Conf.*, pages 1–12, 2023. 2

[3] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv:2307.14460*, 2023. 1, 3, 5, 6, 7

[4] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv:2410.02073*, 2024. 1, 3, 5, 6, 7

[5] Hao Chen, Philipp Gläser, Xuanyu Hu, Konrad Willner, and Jürgen Oberst. Deep learning-based modeling of high-resolution lunar topography using orbiter imagery. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 6084–6087, 2024. 2

[6] Hao Chen, Philipp Gläser, Xuanyu Hu, Konrad Willner, Yongjie Zheng, Friedrich Damme, Lorenzo Bruzzone, and Jürgen Oberst. Elunardtmnet: Efficient reconstruction of high-resolution lunar dtm from single-view orbiter images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024. 2

[7] S. Coloma, C. Martinez, B.C. Yalçın, and M.A. Olivares-Mendez. Enhancing rover teleoperation on the moon with proprioceptive sensors and machine learning techniques. *IEEE Robotics and Automation Letters*, 7(4):11434–11441, 2022. 2

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Adv. in Neural Inf. Proc. Systems*, 2014. 2

[9] Baorui Ma et al. Metricanything: Scaling metric depth pretraining with noisy heterogeneous sources. *arXiv preprint*, 2026. 3, 5, 6, 7

[10] Chen et al. Video depth anything: Consistent depth estimation for super-long videos. *arXiv:2501.12375*, 2025. 3, 5, 6, 7

[11] Haotong Lin et al. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 3, 5, 6, 7

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *Int. Journal of Robotics Research (IJRR)*, 2013. 3

[13] Riccardo Giubilato, Fabian Bizjak, Martin Völk, Sebastiano Chiodini, Marco Pertsch, Heiko Hirschmüller, and Rudolph Triebel. The planetary stereo, solid-state lidar, inertial (s3li) dataset. *IEEE Robotics and Automation Letters*, 7(4):11434–11441, 2022. 1, 3

[14] Riccardo Giubilato, Wolfgang Stürzl, Armin Wedler, and Rudolph Triebel. 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, 2022. 2

[15] S.B. Goldberg, M.W. Maimone, and L. Matthies. Stereo vision and rover navigation software for planetary exploration. In *Proceedings, IEEE Aerospace Conference*, pages 5–5, 2002. 2

[16] Ground Research and Application System of China's Lunar and Planetary Exploration Program. Chang'e 3 Panoramic Cameras Dataset. Data portal of the Ground Research and Application System (GRAS), 2020. Dataset. 1, 2, 3

[17] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *Int. Conf. Learn. Represent.*, 2025. 3, 5, 6, 7

[18] Heiko Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2): 328–341, 2008. 3

[19] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Kaixuan Wang, Hao Chen, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3Dv2: A Versatile Monocular Geometric Foundation Model for Zero-shot Metric Depth and Surface Normal Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12): 10579–10596, 2024. 3, 5, 6, 7

[20] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025. 1, 3, 5, 6, 7

[21] Qingling Jia, Xue Wan, Baoqin Hei, and Shengyang Li. Dispnet based stereo matching for planetary scene depth estimation using remote sensing images. In *IAPR Workshop on Pattern Recognition in Remote Sensing*, pages 1–5, 2018. 2

[22] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 1, 3, 5, 6, 7

[23] Nikhil Keetha, Nils Müller, Johannes Schönberger, Lorenzo Porzi, Yinda Zhang, Tobias Fischer, Arno Knapitsch, David Zauss, Erik Weber, Nuno Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414v1*, 2025. 3, 5, 6, 7

[24] Dahyun Kim, Dongkwon Jin, and Chang-Su Kim. Monocular depth estimation for autonomous driving based on instance clustering guidance. In *Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–6, 2024. 1

[25] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiV 2406.09756*, 2024. 7

[26] Zhen Liang, Hongfeng Long, Zijian Zhu, Zifei Cao, Jinhui Yi, Yuebo Ma, Enhai Liu, and Rujin Zhao. High-precision disparity estimation for lunar scene using optimized census transform and superpixel refinement. *Remote Sensing*, 16(21), 2024. 2

[27] Jiayi Liu, Qianyu Zhang, Xue Wan, Shengyang Zhang, Yaolin Tian, Haodong Han, Yutao Zhao, Baichuan Liu, Zeyuan Zhao, and Xubo Luo. Lusnar: A lunar segmentation, navigation and reconstruction dataset based on muti-sensor for autonomous exploration. *arXiv:2407.06512*, 2024. 2, 3

[28] Philippe LUDIVIG, Abigail Calzada-Diaz, Miguel Angel OLIVARES MENDEZ, Holger VOOS, and Julien Lamamy. Building a piece of the moon: Construction of two indoor lunar analogue environments. In *Proc. Int. Astronautical Congress*, 2020. 2

[29] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conf. on Computer Vision*, 2012. 2

[30] Anton Obukhov, Matteo Poggi, Fabio Tosi, Ripudaman Singh Arora, Jaime Spencer, Chris Russell, Simon Hadfield, Richard Bowden, Shuaihang Wang, Zhenxin Ma, Weijie Chen, Baobei Xu, Fengyu Sun, Di Xie, Jiang Zhu, Mykola Lavreniuk, Haining Guan, Qun Wu, Yupei Zeng, Chao Lu, Huanran Wang, Guangyuan Zhou, Haotian Zhang, Jianxiong Wang, Qiang Rao, et al. The fourth monocular depth estimation challenge. *arXiv preprint arXiv:2504.17787*, 2025. 3

[31] Sharanya Patil, Sahini Reddy, Jaldu Asish, Ranjana Prakash, and Boppuru Rudra Prathap. Lunar terrain depth mapping and 3d reconstruction: Benchmarking seven depth estimation models. In *2024 OITS International Conference on Information Technology (OCIT)*, pages 211–216, 2024. 2

[32] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. In *arXiv preprint arXiv:2502.20110*, 2025. 1, 3, 5, 6, 7

[33] Dominik Pieczyński, Bartosz Ptak, Marek Kraft, and Paweł Drapikowski. Lunarsim: Lunar rover simulator focused on high visual fidelity and ros 2 integration for advanced computer vision algorithm development. *Applied Sciences*, 13(22), 2023. 2

[34] Changjie Qiu, Zhiyong Wang, Xiuhong Lin, Yu Zang, Cheng Wang, and Weiquan Liu. Dsmnet: Deep high-precision 3d surface modeling from sparse point cloud frames, 2023. 2

[35] Uchitha Rajapaksha, Ferdous Sohel, Hamid Laga, Dean Diepeveen, and Mohammed Bennamoun. Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey. *ACM Computing Surveys*, 56(12):1–51, 2024. 1

[36] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2564–2571. IEEE, 2011. 3

[37] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(5):824–840, 2009. 2

[38] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[39] Boyuan Sun, Modi Jin, Bowen Yin, and Qibin Hou. Depth anything at any condition. *arXiv preprint arXiv:2507.01634*, 2025. 3, 5, 6, 7

[40] Mallikarjuna Vayugundla, Florian Steidle, Michal Smisek, Martin Schuster, Kristin Bussmann, and Armin Wedler. Datasets of long range navigation experiments in a moon analogue environment on mount etna. In *ISR 2018; 50th International Symposium on Robotics*, pages 1–7, 2018. 2

[41] Mallikarjuna Vayugundla, Florian Steidle, Michal Smíšek, Martin J. Schuster, Kramer Bussmann, and Armin Wedler. Datasets of long range navigation experiments in a moon analogue environment on mount etna. In *ISR 2018; 50th International Symposium on Robotics*, pages 1–7, Munich, Germany, 2018. 1, 3

[42] Jianyuan Wang, Christian Rupprecht, and David

Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment, 2024. 7

[43] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details, 2025. 3, 5, 6, 7

[44] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. *arXiv preprint arXiv:2501.09898*, 2025. 3

[45] Uland Wong, Ara Nefian, Larry Edwards, Xavier Buoyssounouse, P. Michael Furlong, Matt Deans, and Terry Fong. Polar optical lunar analog reconstruction (POLAR) stereo dataset. NASA Ames Research Center, 2017. 2

[46] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching, 2023. 2

[47] Feng Xue, Guirong Zhuo, Ziyuan Huang, Wufei Fu, Zhuoyue Wu, and Marcelo H. Ang. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In *IEEE Int. Conf. Intelligent Robots and Systems*, 2020. 1

[48] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Adv. in Neural Inf. Proc. Systems*, 2024. 1, 3, 5, 6, 7

[49] Yifan Yu, Shaohui Liu, Rémi Pautrat, Marc Pollefeys, and Viktor Larsson. Relative pose estimation through affine corrections of monocular depth priors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025. 7

[50] Haokun Zheng, Sidhant Rajadnya, and Avideh Zakhor. Monocular depth estimation for drone obstacle avoidance in indoor environments. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10027–10034, 2024. 1