

Geometric Foundation Model Distillation for Efficient Lunar 3D Reconstruction

Clémentine Grethen¹ 

IRIT, University of Toulouse, France

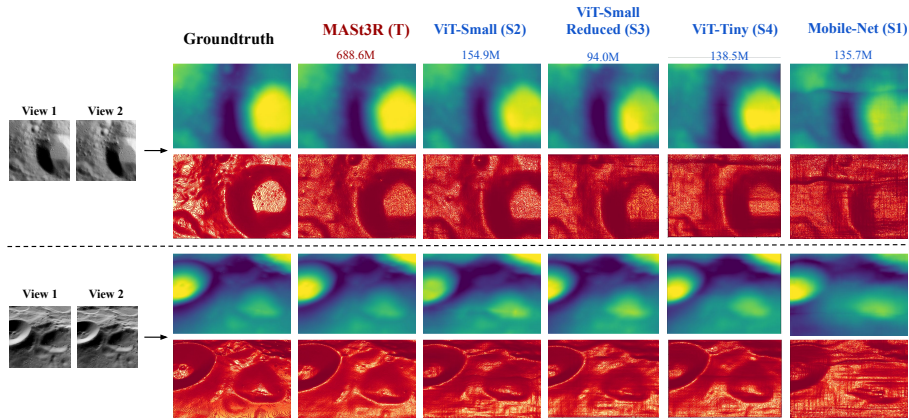


Fig. 1: Our **student** models, distilled from the **teacher** (MASt3R, 688.6M parameters), recover depth maps (top) and surface slopes (bottom) for view 1 of two uncalibrated lunar stereo pairs. Students are ordered left to right from best to weakest: S2 (ViT-Small, 154.9M, $4.4\times$ smaller) closely matches the **teacher** in both depth and slope; S3 (reduced decoder, 94.9M, $7.3\times$ smaller) remains competitive; S4 (ViT-Tiny, 138.5M) shows slight degradation; while S1 (MobileNet, 135.7M) exhibits clear slope artifacts.

Abstract. Large 3D foundation models such as MASt3R achieve state-of-the-art stereo reconstruction but are computationally demanding for deployment under strict hardware constraints—a critical limitation in domains such as planetary exploration, where onboard computing is severely restricted. We study how far such models can be compressed through knowledge distillation, using lunar stereo reconstruction as a challenging and practically relevant case study. Starting from a 688M-parameter MASt3R teacher fine-tuned on lunar imagery, we distill its dense geometric predictions into a family of lightweight students spanning different encoder types (CNN *vs.* ViT), decoder widths and depths,

and training strategies. To bridge the dimensional mismatch between teacher and student, we propose a structured SVD-based initialization that projects the teacher’s decoder weights into the student’s smaller latent space, yielding a warm start that significantly improves convergence and final performance. Based on our results on lunar data, we can obtain a distilled student that retains most of teacher’s reconstruction accuracy while reducing the model size up to 7×, and even outperforms a baseline trained directly with sparse ground-truth annotations. Beyond compression, our study highlights both principles and practical insights for distilling geometric foundation models: a convolutional encoder underperforms transformer-based alternatives (though pretraining availability remains a confounding factor), preserving encoder capacity is more critical than maintaining a large decoder, feature-level distillation consistently outperforms output-only supervision, and SVD-based initialization improves optimisation stability. These findings provide practical guidelines for deploying 3D reconstruction models in resource-constrained environments. Code is publicly available at <https://clementinegrethen.github.io/publications/ECCV.html>.

1 Introduction

Recent advances in large-scale learning-based approaches have significantly advanced multi-view 3D reconstruction. In particular, recent 3D foundation models—models capable of inferring globally consistent geometry directly from images—have demonstrated state-of-the-art performance. Architectures such as DUST3R [28], MAST3R [16], and VGGT [27] show that dense 3D geometry can be recovered from image pairs without explicit camera calibration. By leveraging massive pre-training and high-capacity transformer backbones (ViTs [3]), these models achieve state-of-the-art performance across a wide range of terrestrial datasets. However, this performance comes at a considerable cost: such models typically contain hundreds of millions to over a billion parameters and require substantial memory and computational resources for both training and inference. While such capacity is effective for general-purpose 3D reasoning, it raises an important question: *is this level of complexity necessary for targeted, mission-specific applications?*

More broadly, the increasing reliance on extremely large models has created a growing gap between cutting-edge research systems and practical deployment environments. This disparity between high-capacity research models and real-world applications highlights a divergence between innovation and accessibility. While recent advances in multi-view geometry rely on massive computational resources, the size and inference cost of such models make them difficult to deploy in practice. In particular, many real-world platforms lack the memory and computational capacity required to run networks of this scale, limiting their applicability in scenarios where 3D reconstruction is most valuable, such as robotics or autonomous navigation [31].

For example, these limitations become particularly evident in planetary exploration. In lunar missions, onboard processing units on landers or rovers op-

erate under strict memory, power-consumption, and latency constraints [6]. At the same time, accurate 3D reconstruction is essential for critical tasks such as safe landing, hazard detection, descent navigation, and terrain analysis [8]. The lunar environment further complicates the problem due to specific conditions rarely encountered in standard training datasets, including low-texture surfaces, extreme illumination conditions, and limited parallax during descent trajectories [6, 14]. Although dedicated datasets such as StereoLunar [8] have demonstrated the feasibility of learning-based approaches, the size and inference cost of modern transformer-based models make their deployment in such settings particularly challenging.

These observations motivate an alternative approach. Instead of designing increasingly large universal models, an alternative approach is to compress and specialize existing foundation models for specific operational contexts. Knowledge distillation [7] provides a systematic approach for this objective: a lightweight student network is trained to reproduce the geometric predictions of a large teacher model, inheriting its representational priors while significantly reducing computational cost.

Despite its success in language modeling [19, 32], distillation for dense geometric foundation models remains less explored. Several important questions, therefore, arise:

- How should a student’s architecture be designed for multi-view geometry?
- Is a convolutional encoder sufficient, or are vision transformers still required?
- How to trade off between model size and reconstruction accuracy?
- What training strategies are most effective when distilling large transformer-based teachers into lightweight student networks?

In this work, we study how far a large geometric foundation model can be compressed for a targeted operational domain, using lunar descent and orbital imagery as a challenging and practically relevant case study. Starting from a MAST3R-based teacher fine-tuned on lunar data (688M parameters), we distill its dense geometric predictions into a family of lightweight students employing different encoder types (CNN *vs.* ViT), decoder width/depth, and training choices (*e.g.*, encoder freezing, feature-level supervision). We select MAST3R as the teacher due to its favorable trade-off between geometric accuracy and computational efficiency among other 3D foundation models. While our experiments focus on the Moon, the proposed recipe, distilling a domain-adapted 3D teacher into efficient students, applies broadly to other deployment-constrained domains.

Experimentally, our ViT-based students compress the 688.6M-parameter teacher by $4.4\times$ to $7.3\times$ while delivering a consistent $\sim 2\times$ inference speedup. Despite this compression, the best student (S2, ViT-Small) preserves reconstruction quality with only 15% higher Chamfer error than the teacher (see Figure 1). Its reduced-decoder variant (same encoder, decoder reduced from 512/6 to 256/4) is $7.3\times$ smaller, with only marginal additional degradation, suggesting that much of the decoder capacity is redundant for this task (see Figure 1). Moreover, distillation with teacher pseudo-ground truth outperforms standard ground-truth

supervision, reducing Chamfer error by $\approx 18\%$, confirming that the teacher provides a richer dense supervisory signal than available labels alone.

Our main contributions are:

- A general knowledge distillation framework for stereo 3D reconstruction that transfers a 688M-parameter teacher model into compact student networks without requiring pose annotations, using teacher predictions as dense pseudo-ground truth. We demonstrate its effectiveness on lunar stereo reconstruction.
- A quantified study of the impact of architecture choices on geometry and pose estimation: (a) encoder choice, (b) encoder fine-tuning strategy (frozen *vs.* trainable), and (c) targeted decoder compression.
- A structured SVD-based initialization for compressing transformer decoder weights, improving convergence and final performance.
- A training-strategy ablation isolating key distillation ingredients (output supervision, feature alignment, and SVD warm-start) and measuring their effect on reconstruction and pose metrics.

The remainder of the paper is organized as follows: Sec. 2 reviews related work, Sec. 3 presents our distillation framework and initialization strategy, and Sec. 4 reports experimental results and ablation studies.

2 Related work

Foundation models for dense Multi-View 3D Reconstruction. Classical Structure-from-Motion (SfM) [10, 24] pipelines reconstruct sparse 3D geometry by matching features across multiple views and estimating camera motion, but remain limited in their ability to produce dense, globally consistent reconstructions. Recent end-to-end models such as DUST3R [28] and MAST3R [16] directly regress geometry and camera parameters from stereo image pairs, achieving strong performance on terrestrial scenes. However, when extended to multi-view settings, they typically require an additional global alignment stage to ensure cross-view consistency.

VGGT [27] addresses this limitation by enabling fully end-to-end dense reconstruction from image sequences. It jointly predicts depth, point clouds, and camera parameters across multiple views, producing geometrically consistent representations without explicit global alignment. While highly effective, VGGT relies on a very large transformer-based architecture with over 1 billion parameters. These recent 3D foundation models range from roughly 650M to over 1B parameters and require massive training infrastructure, often exceeding 100 A100 GPUs. Such computational demands impose substantial memory, energy, and financial costs, limiting reproducibility and slowing research iteration. Although these large-scale models demonstrate strong generalization across diverse scenes, their capacity often exceeds the requirements of task-specific applications such as navigation or domain-constrained reconstruction, where efficiency and reliability

are more critical than universal coverage. Moreover, they cannot always generalize to highly specific domains (*e.g.*, space imagery) because the training datasets do not sufficiently capture such specialized conditions [8]. This mismatch creates an infrastructure bottleneck, with model scale surpassing practical deployment needs.

Knowledge Distillation. Knowledge distillation (KD) [7] is a widely used paradigm for transferring the capabilities of large neural networks into compact, efficient models. Originally introduced to leverage the “dark knowledge” contained in a teacher’s soft predictions [11], KD enables a student network to learn richer supervisory signals than hard labels alone, capturing inter-class correlations that improve generalization [23]. This paradigm is particularly attractive where memory, latency, and energy consumption are critical constraints. Distillation can operate at different supervision levels: response-based distillation trains the student to reproduce the teacher’s final outputs, while feature-level distillation further aligns intermediate representations, encouraging the student to mimic internal reasoning processes [7].

Recent works have explored distillation or adaptation of large 3D vision models, but with objectives fundamentally different from ours. [26] distill VGGT into a lighter encoder for robotic manipulation, without any 3D prediction head. [9] align a vision-language-action model with VGGT features for cross-modal grounding rather than compression. [34] modifies VGGT’s attention mechanism for temporal streaming, without size constraints. In contrast, our approach compresses both encoder and decoder, initializes the compressed decoder via structured SVD truncation of teacher weights, and preserves the full dual-decoder 3D prediction pipeline. None of these works tackle this joint compression-and-initialization problem for stereo 3D reconstruction. Closer to our setting, [4] explores knowledge distillation for stereo 3D geometry, using DUST3R as teacher. However, their evaluation is confined to a single indoor dataset (and per-scene optimization) with only three student architectures (CNN and ViT), without an ablation of training strategies or weight initialisation, leaving open the question of what drives compression quality.

DUNE [20] distills three heterogeneous teachers (MASt3R, DINOv2 [17], Multi-HMR [2]) into a single encoder. While it produces geometry-aware features and can serve as a backbone for downstream tasks [15], DUNE does not perform end-to-end distillation of a 3D foundation model, nor does it address structured compression: it focuses solely on encoder fusion, with no 3D reconstruction pipeline or systematic variation of student capacity. In contrast, we compress MASt3R (encoder, dual cross-view decoder, and prediction heads) within a single distillation framework. Our goal is to identify which components of large geometric transformers are essential for maintaining reconstruction fidelity, and which can be safely compressed for lightweight, specific task deployment.

Moon 3D Reconstruction. Lunar 3D reconstruction is critical for navigation, hazard assessment, and terrain analysis in the context of descent and landing, but remains challenging due to low texture, repetitive patterns, and extreme illumination changes [6]. Classical photometric methods and SfM pipelines often

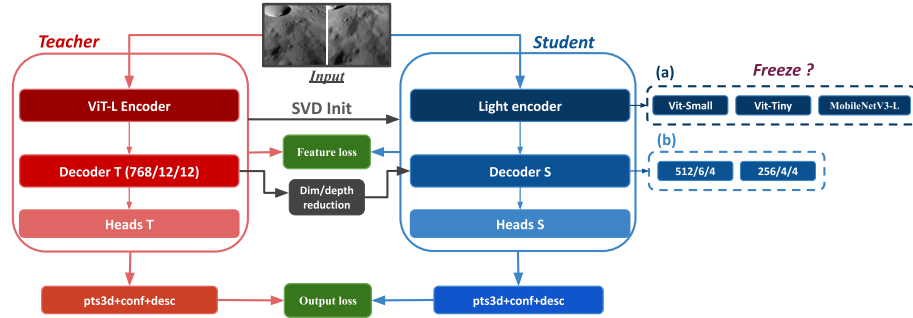


Fig. 2: Teacher-student distillation framework. The teacher supervises the student via feature and output losses. (a) Encoders candidates (b) Two compact decoder configurations (dimension/depth/attention heads)

degrade in this setting because correspondences are hard to establish, and descent trajectories may provide limited parallax [14]. Recent work has shown that large 3D foundation models can be adapted to lunar imagery, as shown in [8], where MAST3R has been successfully fine-tuned on the StereoLunar dataset. While such adaptations improve robustness, the resulting transformer-scale architectures remain costly in terms of memory and latency, limiting their deployment under strict onboard computing resources. In this work, we leverage this lunar-adapted MAST3R as an ideal teacher model to distill compact student networks, serving as a proof-of-concept for domain-specific distillation. This setup provides a controlled setting to study how different distillation strategies affect reconstruction fidelity and demonstrates a framework that could be applied to other specialized datasets and tasks.

3 Method

We propose a multi-student knowledge distillation framework that compresses a large MAST3R teacher into a family of lightweight students, each balancing efficiency and accuracy. Instead of predefining variants, we systematically reduce MAST3R components while preserving geometric fidelity. The MAST3R teacher, adapted to lunar imagery as described in [8], generates pseudo-ground truth 3D reconstructions to supervise all students without camera pose annotations. Students retain MAST3R’s decoder architecture but vary in compression strategy, including encoder type (CNN *vs.* ViT), decoder width and depth, and whether the encoder weights are frozen or updated during training. This setup enables systematic evaluation of how each architectural and training choice affects lunar 3D reconstruction performance.

3.1 Teacher Model

The teacher is a fine-tuned variant of MAST3R [16] adapted to lunar orbital imagery. It has been adapted using the StereoLunar dataset [8], which contains stereo pairs of the Moon’s south pole, captured under varying altitudes, baselines, camera motions, and lighting conditions, along with ground-truth 3D scene information. As shown in Figure 2, it follows the CroCo paradigm [29]: a shared ViT-Large encoder processes each image of a stereo pair independently, and a cross-attention decoder (12 blocks each, dimension 768, 12 attention heads) reconstructs dense 3D point maps along with local feature descriptors used for matching. The model produces for each pixel, a 3D point expressed in the coordinate frame of the first camera, a confidence score, and a 24-dimensional local descriptor.

Training combines a geometry loss over predicted 3D points and a contrastive matching loss over ground-truth correspondences, jointly encouraging accurate reconstruction and precise feature localisation, as described in [16]. With 688 million parameters, the teacher serves as a strong but computationally prohibitive baseline, motivating the knowledge distillation framework described in the following section.

3.2 Student Architectures

We design the student architectures to mirror the teacher’s global reasoning capabilities while fitting within consumer hardware constraints and significantly reducing the parameter count. As shown in Figure 2, we retain the meta-architecture of the teacher but replace heavy components with lightweight alternatives to identify the most suitable configuration for onboard deployment.

All student networks share the same modular structure: (a) a lightweight encoder extracts features, a linear projection with group normalisation projects them into the decoder embedding space, and (b) a smaller cross-view decoder produces the dense geometric predictions.

We evaluate three encoder backbones: (a1) a CNN baseline (MobileNetV3-Large [12]), (a2) a self-supervised ViT (DINOv2 ViT-S/14 [17]), and (a3) a smaller ViT (ViT-Tiny/16 [3]), to cover a spectrum of efficiency and representation quality.

On the decoder side, we consider two compact configurations: (b1) a reduced decoder with embedding dimension/depth (512, 6), and a more aggressively compressed variant (b2) (256, 4), to test how far cross-view reasoning can be shrunk while maintaining reconstruction fidelity.

For all student decoders, we fix the number of attention heads to $h = 4$, following the multi-head attention formulation introduced by [25]. This results in a per-head dimension of 128 for $d = 512$ and 64 for $d = 256$, preserving sufficient representational capacity per head while keeping attention computation lightweight. Finally, the prediction heads (for pointmap, descriptors, and confidence maps) are shared across all student configurations and kept identical to the teacher’s (*cf.* Figure 2).

3.3 Initialization Strategy

Initializing student decoder weights randomly leads to slow convergence, since the student must first learn from scratch to produce geometrically meaningful outputs before it can benefit from fine-grained teacher supervision. To address this, we propose a structured initialization based on truncated Singular Value Decomposition (SVD). While inspired by the low-rank initialization in [22], we extend this approach to bridge the dimensional mismatch between teacher and student weights, effectively projecting the teacher’s knowledge into the student’s smaller latent space.

Meaningful dimension reduction. Any weight matrix $\mathbf{W}_T \in \mathbb{R}^{d_T^{\text{out}} \times d_T^{\text{in}}}$ of the teacher implements a linear transformation that can be decomposed via its Singular Value Decomposition (SVD) as:

$$\mathbf{W}_T = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top, \quad \mathbf{U} \in \mathbb{R}^{d_T^{\text{out}} \times k}, \quad \mathbf{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_k), \quad \mathbf{V}^\top \in \mathbb{R}^{k \times d_T^{\text{in}}}, \quad (1)$$

where $k = \min(d_T^{\text{out}}, d_T^{\text{in}})$ and the singular values are ordered $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k \geq 0$. The columns of $\mathbf{V}$ define an orthonormal basis of *input directions*, those of $\mathbf{U}$ a corresponding basis of *output directions*, and each σ_i quantifies the amplification along the i -th direction pair. A large singular value signals a direction that the weight matrix acts on strongly, while $\sigma_i \approx 0$ indicates a direction essentially ignored by the learned transformation. The singular values thus provide a natural ranking of the importance of each direction in the teacher’s weight space.

Truncated approximation. Given a student’s weight of shape $d_S^{\text{out}} \times d_S^{\text{in}}$ with $d_S^{\text{out}} \leq d_T^{\text{out}}$ and $d_S^{\text{in}} \leq d_T^{\text{in}}$, we initialize $\mathbf{W}_S$ by retaining the $r = \min(d_S^{\text{out}}, d_S^{\text{in}})$ most important singular components of $\mathbf{W}_T$:

$$\mathbf{W}_S = \mathbf{U}_{[:, d_S^{\text{out}}, :r]} \text{diag}(\sigma_1, \dots, \sigma_r) \mathbf{V}_{[:, r, :d_S^{\text{in}}]}^\top. \quad (2)$$

The SVD orders singular components by decreasing importance, so retaining only the top r discards the directions along which the teacher acts most weakly. At the teacher’s original dimensions, this rank- r truncation is provably optimal in the Frobenius norm (Eckart–Young–Mirsky theorem) [5]. Equation (2) further restricts the approximation to the student’s smaller dimensions; since the leading singular vectors concentrate the dominant input–output couplings, this extraction preserves the most informative components of the teacher’s transformation. Our SVD-based variant performs the selection in the singular-value basis: we retain the leading singular directions, which capture the most variance in the teacher’s weight matrix. This is motivated by the rapid spectral decay observed in Transformer weights [13], which suggests that the dominant singular components encode the most transferable structure.

Layer mapping. Since the student decoder has S blocks while the teacher has $T > S$ blocks, we define a uniform stride mapping from student layer indices to teacher layer indices:

$$\phi(i) = \left\lfloor \frac{i \cdot T}{S} \right\rfloor, \quad i = 0, \dots, S - 1. \quad (3)$$

For our default configuration ($T = 12, S = 6$), this yields $\phi = [0, 2, 4, 6, 8, 10]$, selecting every other teacher block. This mapping preserves the hierarchical structure of the decoder: the student’s first block inherits low-level cross-view correlations from the teacher’s early layers, and its last block inherits high-level geometric representations from the teacher’s late layers.

One-dimensional parameters are initialized by direct truncation: $\mathbf{b}_S = \mathbf{b}_T[:d_S]$. Unlike weight matrices, these vectors have no directional structure amenable to SVD; we simply retain the first d_S entries. For convolutional weights $\mathbf{W}_T \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times k_h \times k_w}$, we reshape the tensor into a matrix of shape $C_{\text{out}} \times (C_{\text{in}} \cdot k_h \cdot k_w)$, treating each output filter as a single row vector over all its spatial and channel coefficients.

This reshaping preserves the identity of each filter while exposing cross-filter redundancy to the SVD. We then apply Equation (2) to obtain a compressed matrix of shape $C_{\text{out}}^S \times (C_{\text{in}}^S \cdot k_h \cdot k_w)$, which is reshaped back to $C_{\text{out}}^S \times C_{\text{in}}^S \times k_h \times k_w$. We provide further justifications for our choice of SVD initialization and thoroughly validate the layer mapping strategy through ablation experiments in Sec. 5 of the supplementary material.

3.4 Implementation Details

We consider the StereoLunar dataset [8], a physically realistic lunar rendering benchmark inspired by real orbital trajectories, as a benchmark for domain-specific distillation. The dataset provides diverse lunar imagery with varying baselines, viewpoints, altitudes, and illumination conditions, ensuring a realistic range of geometric and photometric configurations. A key design decision is that all student models are trained within the same loop iteration: they share the teacher’s forward pass and the resulting pseudo-ground truth, so a single teacher inference serves multiple students simultaneously, significantly reducing the effective distillation cost per student. To improve generalization, student inputs are augmented with color jitter, random crops, grayscale tinting, bilateral filtering, and contrast variations. Models are trained for 50 epochs on 7 NVIDIA A100 GPUs using AdamW, with a cosine learning rate decay providing a consistent baseline for performance comparisons while keeping training time manageable.

3.5 Loss function

Knowledge distillation methods broadly fall into two categories depending on *where* the teacher’s knowledge is extracted from [7]. Output-level (or response-based) distillation trains the student to reproduce the teacher’s final predictions; it is simple to implement, architecture-agnostic, and adds no overhead beyond a single forward pass through the teacher [11]. Feature-level (or hint-based) distillation, introduced by FitNets [18], additionally aligns intermediate representations between the teacher and student, giving the student access to the teacher’s internal reasoning process rather than just its conclusions.

Output-level distillation We adopt the standard MAST3R geometry loss, which combines a confidence-aware 3D regression term and a cross-view descriptor matching term. The 3D regression measures the L2 distance between predicted points and pseudo-ground truth points, normalized by the scene’s average depth to ensure scale invariance. Confidence predictions are regularized with a logarithmic penalty to prevent trivial solutions. The descriptor branch is trained with a block-wise InfoNCE [16] contrastive loss to encourage discriminative cross-view descriptors. The final geometry loss is then a weighted sum of these two components, $\mathcal{L}_{geo} = \mathcal{L}_{conf} + \beta \mathcal{L}_{match}$.

Feature alignment loss To further guide the student’s internal representations, we introduce a feature-level distillation loss and feature mapping inspired by Depth Anything [33]. After each forward pass, we extract the encoder output features from both the teacher and the student. Since the student encoder may have a different channel dimension C_s than the teacher’s C_t , we insert a learnable (trained jointly with the student encoder) linear projection $\mathbf{W} \in \mathbb{R}^{C_t \times C_s}$ that maps student features into the teacher’s representation space. A separate projector is trained for each student’s architecture. Both the projected student features and the (detached) teacher features are ℓ_2 -normalised along the channel axis, yielding unit-norm vectors $\hat{\mathbf{f}}_s$ and $\hat{\mathbf{f}}_t$ at every spatial location. The loss is then a margin-filtered cosine similarity:

$$\mathcal{L}_{feat} = 1 - \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \hat{\mathbf{f}}_s^{(i)} \cdot \hat{\mathbf{f}}_t^{(i)}, \quad \mathcal{M} = \{i \mid \hat{\mathbf{f}}_s^{(i)} \cdot \hat{\mathbf{f}}_t^{(i)} < \alpha\}, \quad (4)$$

where $\alpha = 0.9$ is a tolerance margin: spatial positions whose cosine similarity already exceeds α are excluded from the loss, focusing optimisation on the most misaligned regions.

4 Experiments

We evaluate our distillation approach on a test subset of StereoLunar containing 1200 image pairs, covering nadir, pitched, and dynamic geometries, across various altitudes. Our goal is to assess the trade-off between geometric precision (3D reconstruction and camera pose) and operational efficiency (latency and model size). Finally, an ablation study on the ViT-Small architecture student isolates the impact of different training strategies.

4.1 3D reconstruction evaluation

3D Reconstruction metrics To evaluate the student models, we perform a RANSAC-based optimal similarity transform to align the predicted point clouds with the ground truth (GT), recovering the true metric scale. Following the protocol in [16], we report Accuracy (average distance from predicted to closest GT point), Completeness (the reverse), and the Chamfer Distance (the mean of both). These metrics are reported as Absolute Relative (AbsRel) errors, normalized by the median ground-truth depth to ensure altitude-independence. To

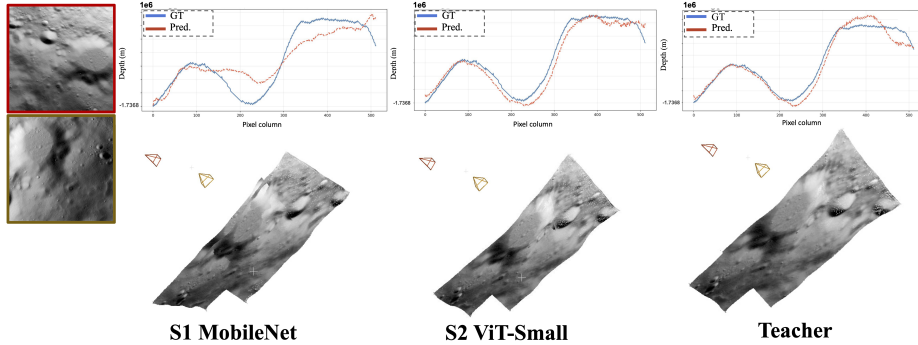


Fig. 3: Central profile results for a lunar stereo pair and corresponding 3D scene. Comparison between S2 (best student), S1 (weaker student), and the teacher. S1 exhibits poor cross-view alignment and a central profile poorly correlated with the ground-truth terrain, while S2 produces results closely matching the teacher.

assess terrain safety—critical for lunar landing—we compute the Slope Mean Absolute Error (MAE). Slopes are derived via spatial finite differences on the elevation channel. This is a classical terrain feature used in landing safety assessments [21], as it characterizes surface stability. We also perform a profile-based analysis [30], extracting horizontal depth slices to compute statistics along the terrain’s cross-sections.

Camera Pose Estimation metrics Localization is assessed via the Virtual Correspondence Reprojection Error (VCRE) [1, 16]. As in [1], we report the median VCRE normalised by the image diagonal (%diag), and the Precision at two thresholds: Prec@5% and Prec@10% of the image diagonal, corresponding to the fraction of pairs with acceptable pose error.

Results and Discussion Tab. 1 summarizes the performance across the lunar stereo benchmark. Our analysis focuses on the relative performance of the student architectures compared with the teacher model. These quantitative findings are further supported by the qualitative results presented in Figure 1. We do not include a detailed comparison with generic pretrained models such as MAST3R [16] or VGGT [27], as these models are not adapted to the lunar domain [8] and therefore provide limited insight for evaluating the proposed distillation framework. Focusing first on the geometric reconstruction metrics (Chamfer, Accuracy, and Completeness), we observe a clear correlation between architectural design and reconstruction quality. Among the student models, S2 achieves the best performance across all three metrics, approaching the teacher despite having more than four times fewer parameters (154.9M *vs.* 688.6M). This indicates that a ViT-Small encoder combined with a sufficiently expressive decoder provides a strong balance between capacity and efficiency. In contrast, S1 replaces the transformer encoder with a lightweight MobileNet backbone while keeping a relatively large decoder. Although S1 achieves a slightly smaller parameter count (135.7M vs 154.9M for S2), it results in noticeably worse recon-

struction performance, indicating that this compression does not offer a favorable efficiency–accuracy trade-off. This suggests an architectural advantage for self-attention in cross-view reasoning, though pretraining differences between backbones cannot be fully ruled out (Figure 3 illustrate the poor results of S1 compared to S2 and teacher). Given these results, we did not explore further reductions of the decoder for S1, as the model already struggles to recover accurate geometry.

The comparison between S3 and S4 further highlights the importance of decoder capacity. While S4 uses a smaller ViT-Tiny encoder with a relatively large decoder, S3 keeps the same ViT-Small encoder as S2 but uses a reduced decoder (256/4/4). Despite having fewer parameters (94.9M), S3 consistently outperforms S4, suggesting that a stronger encoder with reduced decoder capacity is preferable to a weaker encoder with a larger decoder. This suggests that maintaining a sufficiently strong encoder while reducing the decoder can be a more effective compression strategy than aggressively shrinking the encoder itself. The decoder can be significantly reduced (S3) without major losses, as long as the encoder retains sufficient capacity to produce view-consistent features. The VCRE metrics confirm and refine these observations. S2 achieves the lowest median reprojection error (3.56% of image diagonal) and the highest Prec@5% (82.8%), indicating that its pose estimates are both accurate and consistent. S3 obtains a nearly identical Prec@10% (92.7% *vs.* 92.1% for S2), suggesting that the two models converge at similar tolerances despite S3 having 40% fewer parameters. Notably, S1 lags significantly behind on all pose metrics (Prec@5%=36.8%), reinforcing that the MobileNet encoder fails to learn the geometric structure necessary for reliable relative pose estimation. Similar trends are observed on terrain-related metrics. For the elevation profile error (Profile MAE) and slope estimation (Slope MAE), S2 again provides the best trade-off among student models, while S3 remains competitive despite its significantly reduced parameter count. Overall, these results indicate that moderate decoder reduction combined with a capable transformer encoder yields the most favorable efficiency–accuracy trade-off for lunar stereo reconstruction.

Additional experiments, including the evaluation of alternative encoder architectures (e.g., DUNE) and decoder reduction strategies, as well as further result examples and figures (lighting condition robustness), are provided in the supplementary material.

4.2 Inference Efficiency

Table 2 reports inference time and parameter count for each architecture. All students achieve $\approx 2\times$ speedup over the teacher, with inference scaling linearly with pair count. S3 is the most compact model ($7.3\times$ smaller, $2.4\times$ faster at $N = 128$), while S2 offers the best accuracy-efficiency trade-off ($4.4\times$ fewer parameters, $2.1\times$ faster, 15% Chamfer drop). At small batch sizes ($N = 2$), students run in 0.03s to 0.04s vs. 0.07s for the teacher, which is particularly relevant for deployment on resource-constrained platforms

Metric	S1	S2	S3	S4	Teacher
Encoder	MobileNet	ViT-Small	ViT-Small	ViT-Tiny	ViT-Large
Decoder	512/6/4	512/6/4	256/4/4	512/6/4	768/12/12
Param(M)	135.7	154.9	94.9	138.5	688.6
Chamfer ↓	1.02%	0.53%	0.68%	0.74%	0.46%
Accuracy ↓	1.09%	0.56%	0.70%	0.80%	0.46%
Completeness ↓	0.95%	0.51%	0.65%	0.68%	0.46%
<i>VCRE [1]</i>					
Med. (% diag) ↓	8.75	3.56	3.57	4.04	1.15
Prec@5% ↑	36.8	82.8	79.2	76.5	98.6
Prec@10% ↑	69.2	92.1	92.7	91.7	99.3
Profile MAE (m) ↓	161.85	75.69	85.82	93.41	53.63
Slope MAE (°) ↓	8.37	7.59	7.97	8.03	7.22

Table 1: Quantitative comparison of student models against the Teacher on the test dataset. *3D reconstruction* is measured via Chamfer distance, Accuracy, and Completeness. *Pose quality* is assessed via VCRE [1]: median reprojection error (% diag) and Precision at 5 % and 10 % of the image diagonal. *Terrain* metrics measure depth profile fidelity (Profile MAE) and slope estimation (Slope MAE). **Green** = best student, **orange** = second best, **bold** = Teacher (baseline).

4.3 Training Ablation

In the following, we present the ablation study on S2, as it appears to offer the best combination of performance, size, and inference time. We conduct a series of experiments on the S2 (ViT-Small) student to identify the best distillation configuration, varying the supervision source, the encoder training strategy, and the loss composition (Tab. 3). All experiments share the same architecture and training data; only one factor is changed at a time.

Distillation vs. ground-truth supervision. Replacing the teacher’s pseudo-labels with ground-truth depth and pose (Exp. E) yields slightly worse results (Chamfer: 0.65 % *vs.* 0.53 %). This validates the core premise of our approach: The full pipeline (Exp. A, knowledge distillation (KD) + SVD + feature alignment) nonetheless outperforms Exp. E by a large margin, confirming that the teacher’s dense per-pixel predictions, combined with structured initialization and representation alignment, provide a richer supervisory signal than ground-truth labels alone. By distilling from the teacher, the student benefits from implicit geometric priors learned during the teacher’s own training, effectively transferring knowledge that standard supervised losses cannot capture. Importantly, this result also shows that the proposed training scheme remains applicable even when accurate ground-truth annotations are unavailable, a common situation in many real-world 3D reconstruction scenarios.

Model	Params	N=2	N=32	N=64	N=128
Teacher	688.6M	0.07s	2.04s	4.17s	8.36s
S1	135.7M	0.04s	0.91s	1.84s	3.69s
S2	154.9M	0.04s	0.98s	1.99s	4.01s
S3	94.9M	0.04s	0.86s	1.73s	3.47s
S4	138.5M	0.03s	0.87s	1.77s	3.55s

Table 2: Inference time and model size. N denotes the number of image pairs. All students achieve $\sim 2\times$ speedup over the teacher with 4–7 $\times$ fewer parameters. Measured on a single NVIDIA A100 GPU.

Table 3: Ablation study on S2 (DINOv2 ViT-Small encoder). Each column modifies one factor relative to our best configuration (Exp. A) except for Exp. E which serves as a fully-supervised baseline. AbsRel metrics (lower is better).

	<i>Teacher</i>	A	B	C	D	E
KD	–	✓	✓	✓	✓	✗
Feature loss	–	✓	✓	✓	✗	✗
SVD init	–	✓	✗	✓	✓	✗
Unfrozen enc.	–	✓	✓	✗	✓	✓
Chamfer↓	0.46%	0.53%	0.68%	0.87%	0.67%	0.65%
Accuracy↓	0.46%	0.56%	0.69%	0.85%	0.70%	0.69%
Completeness↓	0.46%	0.51%	0.67%	0.89%	0.63%	0.61%

Feature alignment loss. Removing the intermediate feature alignment loss (Exp. D) increases the Chamfer from 0.53 % to 0.67 %. Supervising the encoder’s internal representations, not just its final output, is therefore important for effective distillation across the large capacity gap between teacher and student.

SVD initialisation. Initialising the student decoder from a truncated SVD projection of the teacher’s weights (Exp. A vs. Exp. B) reduces the Chamfer from 0.68 % to 0.53 %. Without this initialisation, the student must learn the cross-attention and regression layers from scratch, resulting in a significantly worse starting point for optimisation. The SVD projection preserves the dominant modes of the teacher’s decoder weights, giving the student a warm start that accelerates convergence and improves final performance.

Encoder freezing. Freezing the DINOv2 backbone (Exp. C) leads to the sharpest performance drop (Chamfer: 0.87 %), indicating that the pretrained features, while providing a strong initialisation, must be adapted to the lunar domain during fine-tuning.

5 Conclusions

We presented a systematic distillation study for compressing stereo 3D reconstruction models, using lunar imagery as a challenging testbed. Our results high-

light five transferable lessons. First, regardless of whether the gap stems from architecture or pretraining, a well-pretrained ViT encoder remains the safer practical choice over a CNN backbone. Second, when further compression is required, reducing decoder capacity is preferable to reducing encoder capacity: moderate decoder reduction combined with a capable transformer encoder yields the most favorable efficiency–accuracy trade-off. Third, teacher pseudo-labels provide a richer supervisory signal than sparse ground-truth annotations, making the approach applicable to domains where labels are scarce. Fourth, feature-level alignment between encoder representations is necessary to close the capacity gap, even when teacher and student have mismatched dimensions. Fifth, structured SVD-based initialization of the student decoder from the teacher’s weights improves optimization stability and final reconstruction accuracy compared to random initialization. While our analysis focuses on the lunar domain, the results suggest that this distillation strategy — domain-adapted teacher, pseudo-GT supervision, representation alignment, and structured initialization— provides a methodological template potentially applicable to other resource-constrained geometric tasks domain where a large pretrained model is available.

Acknowledgments

We gratefully acknowledge Géraldine Morin and Simone Gasparini from the University of Toulouse and Florient Chouteau from Airbus for their contributions. This work was supported by the French Agence Nationale de la Recherche (ANR, “Investissements d’avenir”, ANR-21-ESRE-0051) and the European Space Agency (ESA, contract 4000140461/23/NL/GLC/my).

References

1. Arnold, E., Wynn, J., Vicente, S., Garcia-Hernando, G., Monszpart, Á., Prisacariu, V.A., Turmukhambetov, D., Brachmann, E.: Map-free visual relocalization: Metric pose relative to a single image. In: ECCV (2022)
2. Baradel, F., Armando, M., Galaaoui, S., Brégier, R., Weinzaepfel, P., Rogez, G., Lucas, T.: Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In: European Conference on Computer Vision. pp. 202–218. Springer (2024)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), <https://arxiv.org/abs/2010.11929>
4. Dutt, A., Lunawat, I., Kaur, M.: Multi-view 3d reconstruction using knowledge distillation (2026), <https://arxiv.org/abs/2412.02039>
5. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. *Psychometrika* **1**(3), 211–218 (1936)
6. Getchius, J., Renshaw, D., Posada, D., Henderson, T., Hong, L., Ge, S., Molina, G.: Hazard Detection and Avoidance for the Nova-C Lander, p. 921–943. Springer International Publishing (2024). https://doi.org/10.1007/978-3-031-51928-4_53

7. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. *International Journal of Computer Vision* **129**(6), 1789–1819 (Mar 2021). <https://doi.org/10.1007/s11263-021-01453-z>
8. Grethen, C., Gasparini, S., Morin, G., Lebreton, J., Marti, L., Sanchez-Gestido, M.: Adapting stereo vision from objects to 3d lunar surface reconstruction with the stereolunar dataset. In: *iccvw*. pp. 3751–3760 (2025)
9. Guo, M., Cao, M., Tao, J., Xu, R., Yan, Y., Liang, X., Laptev, I., Chang, X.: Glad: Geometric latent distillation for vision-language-action models (2025), <https://arxiv.org/abs/2512.09619>
10. Hartley, R.I., Zisserman, A.: *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edn. (2004)
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015)
12. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for mobilenetv3. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2019)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. *Iclr* **1**(2), 3 (2022)
14. Kumar, A., Kaushal, S., Murthy, S.V.: Moonmetasync: Lunar image registration analysis (2024), <https://arxiv.org/abs/2410.11118>
15. Leblanc, B., Poullis, C.: Distill3r: A pipeline for democratizing 3d foundation models on commodity hardware (2026)
16. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with MAST3R. In: *ECCV* (2024)
17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023)
18. Romero, A., Ballas, N., Kahou, S., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. in *international conference on learning representations* (2015)
19. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2020), <https://arxiv.org/abs/1910.01108>
20. Saryıldız, M.B., Weinzaepfel, P., Lucas, T., De Jorge, P., Larlus, D., Kalantidis, Y.: Dune: Distilling a universal encoder from heterogeneous 2d and 3d teachers. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 30084–30094 (2025)
21. Steffes, S.R., DeTrempe, P., Barton, G., Woffinden, D.: Hazard boresight relative navigation for safe lunar landing. In: *AIAA SCITECH 2023 Forum*. American Institute of Aeronautics and Astronautics (Jan 2023). <https://doi.org/10.2514/6.2023-0691>
22. Sy, Y., Cerisara, C., Illina, I.: Lillama: Large language models compression via low-rank feature distillation. *arXiv preprint arXiv:2412.16719* (2024)
23. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: *International conference on machine learning*. pp. 10347–10357. PMLR (2021)
24. Ullman, S.: The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **203**(1153), 405–426 (Jan 1979). <https://doi.org/10.1098/rspb.1979.0006>

25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
26. Vuong, A.D., Vu, M.N., Reid, I.: Improving robotic manipulation with efficient geometry-aware vision encoder (2025), <https://arxiv.org/abs/2509.15880>
27. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025)
28. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. *arXiv preprint arXiv:2310.02328* (2023)
29. Weinzaepfel, P., Leroy, V., Lucas, T., Brégier, R., Cabon, Y., Arora, V., Antsfeld, L., Chidlovskii, B., Csurka, G., Revaud, J.: Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems* **35**, 3502–3516 (2022)
30. Wu, B., Liu, W.C., Grumpe, A., Wöhler, C.: Shape and albedo from shading (SAfS) for pixel-level dem generation from monocular images constrained by low-resolution dem. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XLI-B4**, 521–527 (Jun 2016). <https://doi.org/10.5194/isprs-archives-xli-b4-521-2016>
31. Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Yuan, L.: Tinyvit: Fast pretraining distillation for small vision transformers. In: *ECCV*. pp. 68–85. Springer (2022)
32. Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., Xu, C., Tao, D., Zhou, T.: A survey on knowledge distillation of large language models (2024), <https://arxiv.org/abs/2402.13116>
33. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: *CVPR* (2024)
34. Zhuo, D., Zheng, W., Guo, J., Wu, Y., Zhou, J., Lu, J.: Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539* (2025)