DESPINA: Synthesis of High-Fidelity Planetary Horizon Reconstructions Using DEM-Guided Diffusion

Adam Nelson-Archer Raunak Sarbajna, Christoph F. Eick

University of Houston, Houston, TX, USA - (atnelson, rsarbajna, ceick)@uh.edu

Keywords: geospatial embeddings, digital elevation models, diffusion models, planetary imagery, multimodal representations, location-based image generation

Abstract

Ground-level horizon imagery is scarce across planetary bodies, making representation-centred approaches attractive for downstream geospatial tasks. We present DESPINA, a geospatial representation system that converts digital elevation models (DEMs) into structured neural embeddings of terrain geometry that condition a diffusion model to produce geometry-preserving, terrain-consistent visual reconstructions for a specified location and view direction.
Our pipeline integrates numeric elevation data (DEMs), structural embeddings (inverse-depth and soft edges), and textual priors, unifying heterogeneous geospatial signals into a shared, metric conditioning space. Using a Stable Diffusion model constrained with ControlNet, we can generate geologically consistent yet texturally diverse horizon datasets. Appearance priors are learned from historical surface photography to capture realistic textures and lighting cues, and geometric validation is performed against DEM-derived skylines and depth structure, independent of photographic training data. Through quantitative evaluation and a pilot qualitative study, DESPINA maintains skyline fidelity and geological boundaries while improving structural similarity relative to an image-conditioned baseline. Although our experiments use lunar DEMs and historical surface photography, the method is domain-agnostic and applicable to Earth, Mars, and other planetary DEMs.

1. Introduction

The scarcity of ground-level planetary horizon imagery presents a challenge for terrain analysis and autonomous rover navigation (Zhang et al., 2024, Haase et al., 2019). While the Moon’s surface is extensively documented through DEMs and orbital mosaics, true ground-level perspectives remain rare, especially outside Apollo landing zones. This makes the Moon an ideal stress-test environment: it offers globally available elevation data, almost no surface imagery, and highly varied topography. These constraints allow us to evaluate whether a generative system can reconstruct geometry-preserving views using only DEM-derived structure. Although this paper uses the lunar surface as the primary case study, the underlying pipeline is domain-agnostic and can operate on any planetary body with high-quality DEMs, including Earth and Mars.

Classical rendering (e.g., draping optical mosaics over DEMs) preserves silhouettes but lacks the controlled, geometry-preserving appearance often required for mission planning or human analysis. This paper presents DESPINA, a geospatial representation system that produces high-fidelity, geometry-preserving planetary horizon reconstructions directly from DEMs, addressing this data gap while remaining domain-agnostic.

A fundamental challenge in generating terrain-consistent planetary imagery is enforcing structural constraints that maintain geographic accuracy. Our proposed architecture addresses this by converting DEMs into compact geometric embeddings that encode the surface’s spatial layout. These embeddings are raycasted depth maps and soft-edge boundary fields, which condition a diffusion model through ControlNet (Zhang et al., 2023), enabling geometry-aligned reconstructions with Stable Diffusion (Rombach et al., 2022). Because all conditioning is derived directly from elevation data, the pipeline avoids reliance on reference photographs and can generate topographically faithful views for any region, whether on the Moon, Earth, or other planetary bodies. By integrating LRO-derived elevation data, our system produces synthetic imagery that is both visually plausible and geometrically faithful (Fig. 1).

The broader impact of this work extends to establishing an automated pipeline for generating location-specific geologically accurate synthetic horizon datasets with minimal human intervention. Our system provides a parametrically controllable environment particularly valuable for visual place recognition (VPR) research (Lowry et al., 2015), enabling the training and evaluation of lunar VPR systems without extensive field photography. The goal is an evaluation of DEM-guided image synthesis, not diffusion model innovation. We therefore benchmark DESPINA against classical DEM-based renderers and report skyline-level geometric fidelity.

This research makes several contributions to structured terrain synthesis and constrained generative modeling:

1. Development of a generalized location-and-view-direction specific image generation process.
2. Convert DEMs into raycasted inverse-depth and soft-edge structural embeddings for conditioning.
3. Integrate these embeddings with ControlNet to synthesize geometry-consistent horizon views.
4. Establish a geometry-first evaluation protocol (DEM skylines, terrain-only masks) with ablations.
5. Benchmark against classical DEM drape and image-conditioned diffusion baselines.
6. Demonstrate portability across planetary DEMs.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-881-2026 |

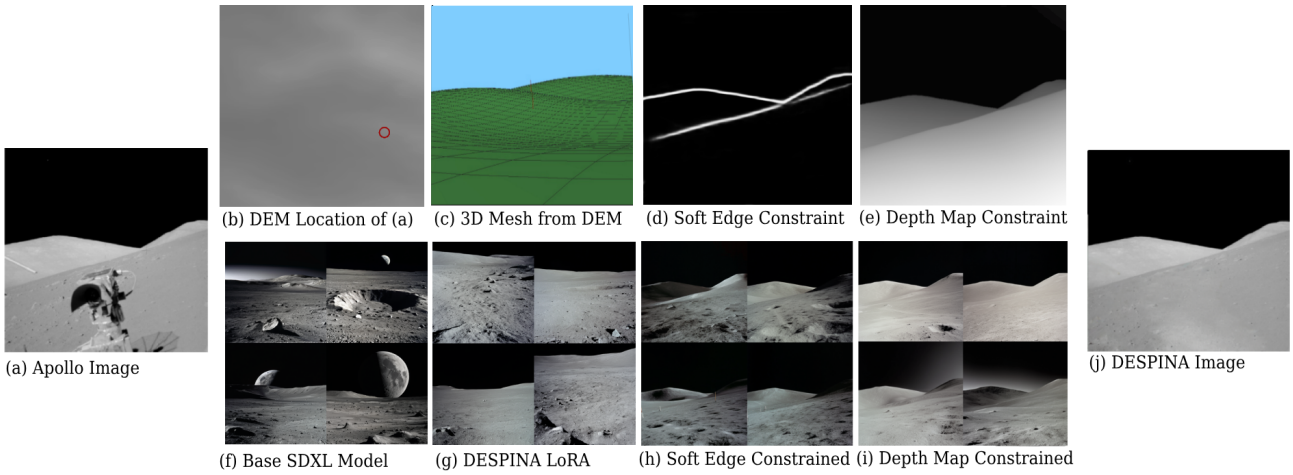


Figure 1. Demonstration of DESPINA’s geometry-preserving planetary horizon reconstructions, showcasing the constraint generation process and the incremental improvement with each added representation. The top row illustrates DEM-derived inverse-depth and soft-edge embeddings; the bottom row shows results at each stage. Panel (j) recreates (a) with all constraints.

The remainder of this paper is organized as follows: Section 2 reviews related work in terrain generation, depth mapping, and image synthesis. Section 3 presents the DESPINA system architecture, detailing our DEM-to-Depth pipeline, ControlNet and LoRA training, and image synthesis process. Section 4 provides quantitative and qualitative evaluations of our approach, including ablation studies that measure the contribution of each component. Section 5 summarizes our findings and discusses future research directions.

2. Related Work

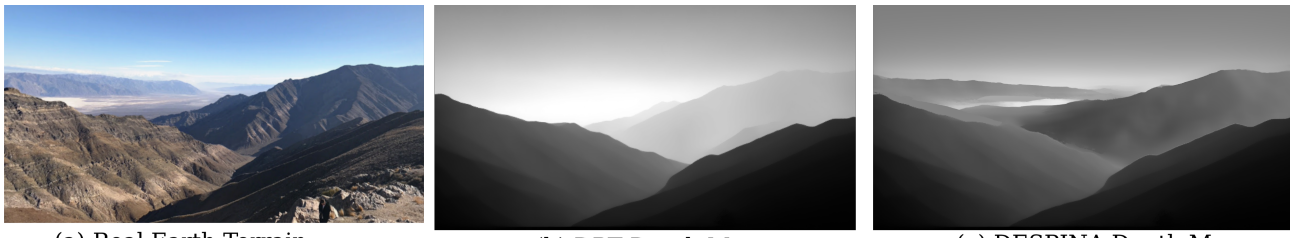
Early deep learning approaches applied generative adversarial networks (GANs) to terrain and landscape synthesis. Conditional GANs like Pix2PixHD translated segmentation or sketch inputs into landscape images, and NVIDIA’s GauGAN (Park et al., 2019a) introduced adaptive normalization from semantic maps. Domain-specific GAN models were also explored for terrain-generation: Guérin et al. (Guérin et al., 2017) trained a conditional GAN for example-based terrain authoring, and Beckham & Pal (Beckham et al., 2019) took initial steps toward procedural terrain generation with GANs. However, creating geologically plausible outputs was challenging. In an attempt to implement what later became DESPINA, we implemented this system within Pix2PixHD, a conditional GAN (Wang et al., 2018), to compare to Stable Diffusion (Rombach et al., 2022) results. While the class identification was beneficial, we found the loss of detail and difficulty creating label maps was crippling to our outputs, and the GAN framework was dropped.

Diffusion models have recently become the state-of-the-art option for image synthesis (Croitoru et al., 2023), offering improved diversity and fidelity over GANs. Diffusion-based generators can create better terrain details and capture regional features of landscapes. For instance, Hu et al. (Hu et al., 2024) propose a Terrain Diffusion Network (TDN) that generates realistic terrain from schematic user sketches (indicating rivers, ridges, basins, peaks) while incorporating regional erosion and sky patterns. TDN introduces a multi-level denoising process (structural to fine detail) to maintain both large-scale layout and small-scale geological realism. While DESPINA does not utilize this modified structure, DESPINA does heavily constrain image synthesis with structural constraints before fine details are added. High-resolution text-to-image diffusion models like Stable Diffusion allow generating landscapes with greater detail and resolution. Recent large diffusion frameworks (LCM-LORA (Luo et al., 2023), Turbo (Ju et al., 2024), Flux, etc.) lead to increases in generative capabilities in photo quality and size. To inject some guidance into diffusion, ControlNet (Zhang et al., 2023) was introduced as an add-on that conditions generation on input maps (edges, depth, etc.).

Depth estimation is crucial for evaluating and generating structured 3D scenes. Recently, monocular depth estimation networks have found significant improvment by training on larger and more diverse datasets. Dense Prediction Transformer (DPT) (Ranftl et al., 2021) is a prime example, and a model that we will evaluate our own systems against throughout this paper. DPT uses several depth datasets to achieve high zero-shot performance across environments. Models like this produce a dense depth map from a single image and have been used in many generative tasks. A pre-computed depth map can serve as a condition in image synthesis, making the generative model respect scene geometry. The introduction of ControlNet has made this practical: by providing a diffusion model a depth map (e.g. from DPT), it is possible to create a new image that maintains the spatial structure of the original scene.

Depth-realistic synthesis is particularly useful for terrain and outdoor scenes, where heavily constrained processes can ensure the correct placement of mountains, craters, or other features. For example, in Liu, Yang, et al. (Liu et al., 2020b), a coarse satellite DEM was used alongside a high-resolution image as conditional input to a GAN, producing enhanced high-res terrain imagery aligned to the true elevation data. This method for enhancing LRO data is crucial in cases like these, where the initial dataset is limited in resolution. The approaches outlined by Liu, Yang, et al. were applied in this paper, for the selective upscaling of DEM slices in our testing.

These models are highly effective when applied to photographic imagery, where shading, texture, and depth cues are naturally present. However, they can misestimate structure in settings with weak cues. On a real Earth site, Fig. 2 compares monocular DPT (photo-conditioned) against DESPINA using the co-registered DEM at the exact location to derive inverse-depth and soft-edge constraints; the comparison highlights that DEM-



(a) Real Earth Terrain (b) DPT Depth Map (c) DESPINA Depth Map

Figure 2. A comparison between DPT and DESPINA on a given location on Earth. Note the loss of quality from DPT on the real image.

(a) DEM-to-Depth (b) DPT (c) Result of (a) (d) Result of (b)

Figure 3. A zoomed-in comparison of DESPINA’s depth maps and DPT’s. Images generated using these constraints are shown on the right. Note how (d) appears more flat than (c). Depth map generation and overall comparisons are shown in Fig. 2.

derived conditioning preserves skyline structure and terrain boundaries. Further data-preparation and co-registration details are provided in the Supplementary Material.

Depth mapping from multi-view imagery has also advanced significantly recently, through Structure-from-Motion (SfM) and multi-view stereo (MVS) techniques, which reconstruct 3D point clouds from unordered image collections. SfM pipelines like COLMAP (Schonberger and Frahm, 2016) rely on feature matching and bundle adjustment to estimate accurate depth maps, given sufficient viewpoint diversity and image overlap. Similarly, learned MVS networks (e.g., MVSNet (Yao et al., 2018)) use convolutional neural networks (CNNs) to infer depth directly from multiple images, producing high-resolution geometric reconstructions.

Generating accurate optical images of terrains from DEMs is essential but faces challenges like limited training data and distinct lunar surface characteristics. Recent advancements by Ceresoli et al. (Ceresoli et al., 2025) introduced a custom rendering approach that produces realistic optical lunar images directly from high-resolution DEMs without using existing computationally expensive techniques.

3. DESPINA System Design and Implementation

DESPINA converts DEMs, existing imagery data, and textual descriptions into structural embeddings that condition a diffusion model. As depicted in Figure 4, the system operates in three stages: (I) ControlNet and LoRA training, (II) DEM-to-Depth processing, and (III) image synthesis. We employ Stable Diffusion constrained with ControlNet to prioritize geometry-consistent reconstructions. Although instantiated and evaluated on lunar DEMs, the design is dataset-agnostic and applies to Earth, Mars, and other planetary DEMs. We will discuss DESPINA ’s architecture and its underlying algorithms in the remainder of this Section.

3.1 DESPINA Architecture Design Considerations

Soft-edge extraction is performed on the image 1 in Fig. 4, using a soft-edge detection model (PidNet, 4) to produce smooth and structurally sound outlines of major features. We find that PidNet works well on the simulated terrain, and does not need to be rendered, unlike the depth mapping. Our pipeline offers several improvements over existing systems:

- By working directly with DEM data, we produce highly accurate terrain representation while keeping rendering efficient.
- The dynamic calculation of the furthest viewable distance ensures optimal data utilization regardless of viewpoint or size of the terrain. We use all 256 depth values available in a depth map, and the distances will be accurately placed on the resulting image.
- The system supports automated image generation, allowing for generation of training data from multiple viewing angles and elevations.

3.1.1 Lunar Image Dataset Curation We initiate the pipeline by curating a large collection of lunar horizon images from the NASA Apollo mission archives (NASA, 2017). After filtering and inpainting fiducial markers (Jurado-Rodríguez et al., 2021), approximately 3,500 out of 10,300 images are selected. For examples of inpainting and other processing, see supplementary results. These images serve as the foundation for generating structured datasets. Our data processing pipeline is implemented using TensorFlow (Abadi et al., 2016) and Keras (Heaton, 2020) on CUDA-enabled hardware (NVIDIA Corporation, 2023). Depth maps are produced using a Dense Prediction Transformer (DPT) (Ranftl et al., 2021), while soft-edge maps are extracted using a PidNet-based soft edge generator (Xu et al., 2023).

To generate consistent and structured prompts for LoRA training, we employ BLIP (Li et al., 2022) to analyze each image and identify key features such as terrain characteristics (hills, craters), lighting conditions (shadows), and scene elements (rovers, equipment). We found that a majority of images could be described as either ”rocky”, ”flat”, ”hilly”, or some combination of the three. This was the inspiration for our image categories, as depicted in Fig. 7. These created datasets of captions, images, depth maps, and soft edges are then used to fine-tune ControlNet models (Zhang et al., 2023) and to train a custom Low-Rank Adaptation (LoRA) module (Luo et al., 2023).

3.1.2 DEM Adaptation The second stage of the pipeline focuses on converting DEMs into structured constraints for image synthesis. Surface elevation data, acquired from the Lunar Reconnaissance Orbiter (Lunar Reconnaissance Orbiter, 2011) and upscaled using established methodologies (Liu et al., 2020b), is processed through a custom DEM-to-Depth pipeline. This pipeline employs min-max normalization and a view-adaptive ray-casting approach to ensure accurate depth representation. The resulting depth maps, along with simulated terrain images

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-881-2026 | © Author(s) 2026. CC BY 4.0 License.

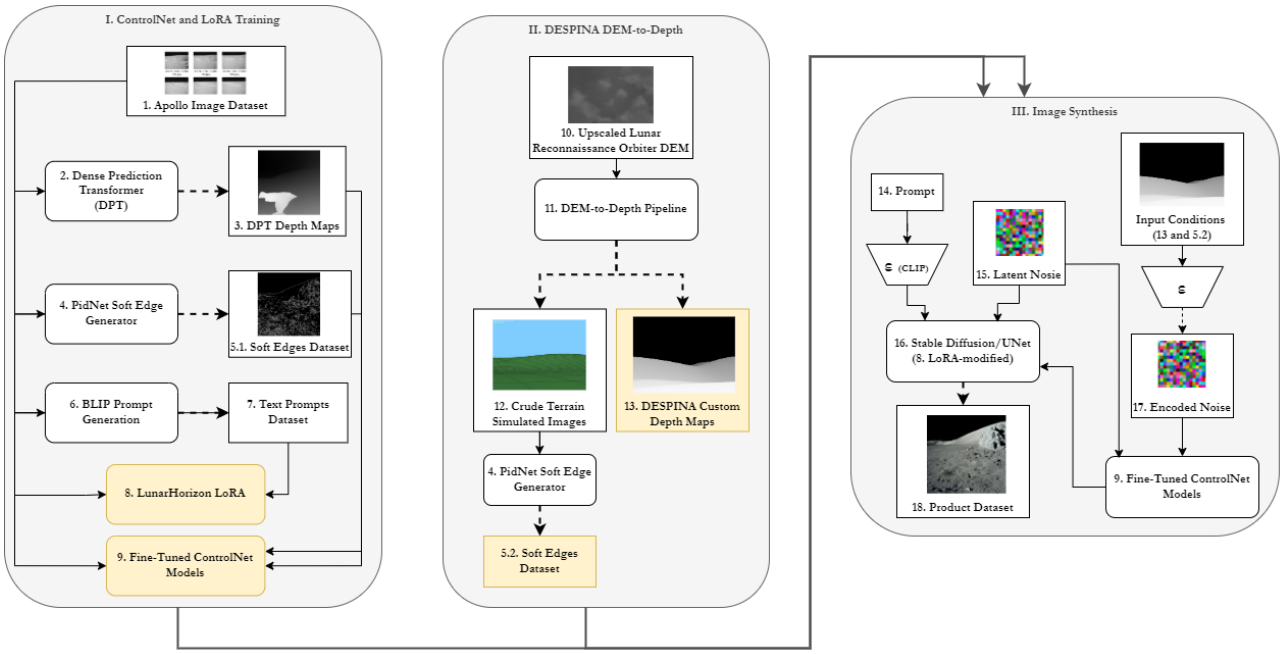


Figure 4. Overview of the DESPINA system architecture. The pipeline is divided into three main groups: I. ControlNet and LoRA Training, II. DESPINA DEM-to-Depth, and III. Image Synthesis. Each numbered block corresponds to a specific process or dataset as described in the text.

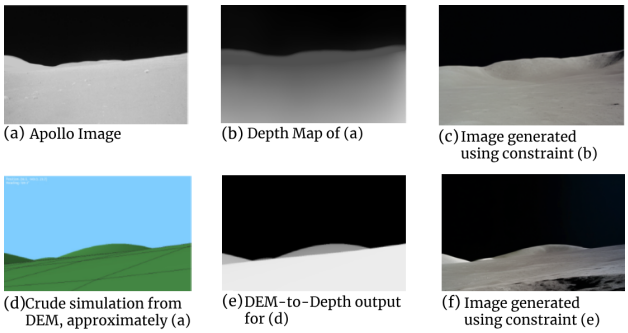


Figure 5. A direct comparison between the baseline pipeline and DESPINA. The two rows match one another, with (d) being the location of (a) on the DEM. (b) and (e) are their respective depth maps, and (c) and (f) are the images generated from those maps.

rendered from the DEMs, provide the necessary geometric and structural information. Soft-edge constraints are further extracted from these simulated images, ensuring that the generative process respects the true underlying terrain structure.

3.1.3 Image Synthesis In the final stage, image synthesis is performed by integrating the outputs of the previous stages. Text prompts are encoded using CLIP (Ramesh et al., 2021), and initial latent noise is generated. The custom depth maps and soft-edge constraints are combined as input conditions and encoded before being processed by the fine-tuned ControlNet models. The LoRA-modified Stable Diffusion model (Rombach et al., 2022) then synthesizes the final geometry-preserving horizon reconstructions, guided by the encoded noise, prompt, and structured constraints. This approach ensures that the generated images are both visually compelling and scientifically meaningful, faithfully reflecting the underlying terrain.

The generative component of DESPINA employs Stable Diffusion XL (SDXL) (Rombach et al., 2022), which, despite being outpaced by some newer models (Ju et al., 2024) in terms of fine-grained detail and text coherence, remains a practical choice due to its efficiency and the quality of results it delivers for terrain generation. The custom-trained LoRA module addresses issues encountered when using base models to generate otherworldly data, such as hallucinations of clouds or bright skies, and enhances the realism of lunar terrain textures. ControlNet models, fine-tuned on each type of structured input, are integrated into the SDXL generation process, further improving the system’s ability to produce structurally faithful lunar images.

A visual comparison of outputs from the Apollo mission, base Stable Diffusion, SDXL, and current state-of-the-art models is shown in Figure 6, highlighting the progression in lunar horizon generation quality across different architectures.

3.2 Group I: ControlNet and LoRA Training

Group 1 of DESPINA’s architecture (Fig. 4) uses several state-of-the-art image generation methods to produce the required high-resolution terrain images.

To create conditioning inputs, we used Dense Prediction Transformer (DPT, Item 2) (Oquab et al., 2023) for depth map generation, PidNet (Item 4) (Xu et al., 2023) for soft-edge extraction, and BLIP (Item 6) (Luddecke and Ecker, 2022) for generating descriptive text prompts. These processes produced three datasets: depth maps (Item 3), soft-edge maps (Item 5.1), and textual descriptions (Item 7). These structured datasets serve as the basis for individually fine-tuning ControlNet models (Item 9) (Zhang et al., 2023) and training the specialized LoRA module (Item 8) (Luo et al., 2023). Once tuned, the provided LoRA and ControlNet modules serve as the foundation for our improvements to the system and allow for a consistent baseline to test against.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-881-2026 | © Author(s) 2026. CC BY 4.0 License.

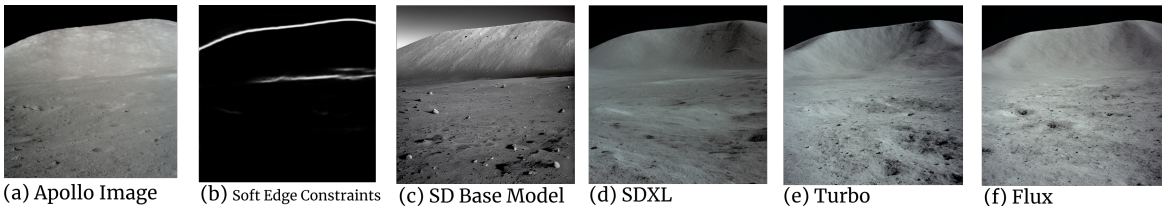


Figure 6. Comparison of different models on lunar horizon generation. From left to right, (a) is an image from the Apollo missions, and (b) is the derived soft edges by running the image through PidiNet. (c) is the base Stable Diffusion model, and (d) is the base SDXL model. (e) and (f) are Turbo and Flux, current state-of-the-art image generation models.

3.3 Group II: DESPINA DEM-to-Depth

In Group II, DESPINA begins to acquire surface elevation data from the Lunar Reconnaissance Orbiter (LRO) (Petro et al., 2015) (Item 10). Raw elevation data, provided as DEMs, is upscaled through a methodology outlined by Liu et al. (Liu et al., 2020b). This elevation data is processed through our Horizon-Aware DEM Constraint Generation pipeline (Item 11), which produces high-fidelity, skyline-faithful depth maps (Item 13) for use in diffusion-based image synthesis. For efficient rendering and realistic visualization, we employ a dynamic normal calculation, visible in Fig. 4 as item 11. This approach is a modified generation technique inspired by Ceresoli et al. (Ceresoli et al., 2025):

$$\vec{n}(x, y) = \text{normalize}(\Delta h_x, \Delta h_y, 1) \tag{1}$$

where Δh_x and Δh_y are computed using central differences of the height field.

To maximize precision near the skyline, we perform a pre-pass to estimate the maximum elevation angle $\hat{\alpha}$ along each azimuth ray. This determines an adaptive $d_{far}(x)$, set just beyond the furthest visible terrain point for that ray. Rather than linearly mapping distances, we store inverse-depth values $z = 1/d$ before normalization, concentrating resolution on areas where geometric accuracy is critical.

$$d_{far}(x) = (1 + \epsilon) \cdot \max_{p \in V_x} \|p - c\| \tag{2}$$

where V_x is the set of visible points along azimuth x, c is the camera position, and ϵ is a small margin (e.g., 0.05) to avoid clipping.

To reduce aliasing near the skyline, we apply stratified ray casting in a narrow angular band around the horizon. Multiple samples are averaged to smooth jagged transitions while preserving steep gradients. This refinement step ensures that sharp crater rims and ridge lines remain precise while avoiding stair-step artifacts that typically appear at lower DEM resolutions.

After ray casting, the normalized inverse-depth is computed as:

$$d_{norm} = \frac{z - z_{near}}{z_{far} - z_{near}}, \quad z = \frac{1}{d} \tag{3}$$

where z is the inverse depth. In this representation, points closest to the observer are mapped to white, with a smooth gradient fading to black for distant terrain.

Figure 3 shows that this skyline-aware approach preserves fine silhouette structure and reduces banding relative to monocular depth estimation (e.g., DPT). In our reconstruction of the DPT image, note how the background information is largely inaccurate. Further coverage and documentation of this work is provided in the supplementary material.

Additionally, the pipeline generates crude simulated terrain images (Item 12), which are subsequently processed by PidNet (Item 4) (Xu et al., 2023) to produce soft-edge constraints (Item 5.2). These geometric and edge constraints inform the generative model, improving likeness to the true lunar terrain geometry. As illustrated in Figure 5, a direct comparison between the baseline pipeline and DESPINA demonstrates the visual differences in their respective depth maps and generated images.

3.4 Group III: Image Synthesis

In Group III, the image synthesis component, we utilize the outputs created by other groups, depicted as yellow cells in our diagram. To generate images, we use conditioned generative modeling through Stable Diffusion (Rombach et al., 2022), modified by the trained LoRA (Item 16) (Luo et al., 2023). Text prompts (Item 14) processed through a CLIP text encoder (Ramesh et al., 2021) are input into the synthesis, and initial latent noise (Item 15). Additionally, structured depth and soft-edge constraints derived from the DESPINA pipeline (Items 13 and 6b) are encoded and fed into fine-tuned ControlNet models (Item 9) (Zhang et al., 2023) to guide generation. The model synthesizes final horizon images (Item 18), making the final output both visually realistic as well as geographically accurate. Fig. 1 demos the full generative process, adding one type of constraint at a time. The results of each constraint improve the quality of the image, with (j) being the one that combines all methods, which we do in DESPINA.

Each component of our system, from base SDXL model, lunar-specific LoRA fine-tuning, soft edge constraints, to depth map constraints contribute differently to the quality of the generated images. To quantify the individual and combined contributions of these components, we conducted a systematic ablation study presented in Section 4, where we measured their impact on structural similarity, horizon accuracy, and feature preservation.

4. Evaluation and Demo

4.0.1 Datasets In this section we assess qualitative realism and quantitative accuracy of DESPINA. The primary datasets used for testing Despina come from NASA’s Apollo Lunar Imagery (Haase et al., 2019) and CNSA’s Chang’e 5 Surface imagery (Chinese Lunar and Planetary Data System, 2020). Because comprehensive surface ground truth for lunar surfaces is scarce, we use two complementary setups: (i) a non-Apollo Chang’e-5 case study (Fig. 8) where we juxtapose a classical DEM renderer with DESPINA, and (ii) a set of historical Apollo views (Fig. 9) where we compare an image-conditioned diffusion baseline to DESPINA. In both setups, geometric evaluation uses DEM-derived skylines and terrain-only masks so that non-terrain objects (e.g., lander hardware) never influence metrics.

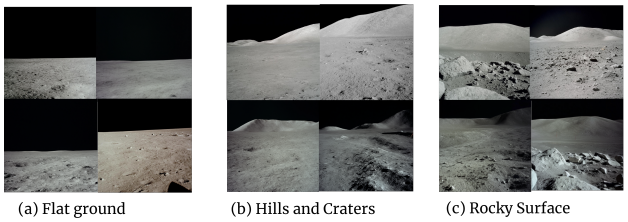


Figure 7. Examples of image categories that were created for testing purposes. Flat ground (a) serves as a baseline for generated terrain complexity, while hilly terrains (b) and fine rock detail (c) intend to test the models depth accuracy and feature resolution.

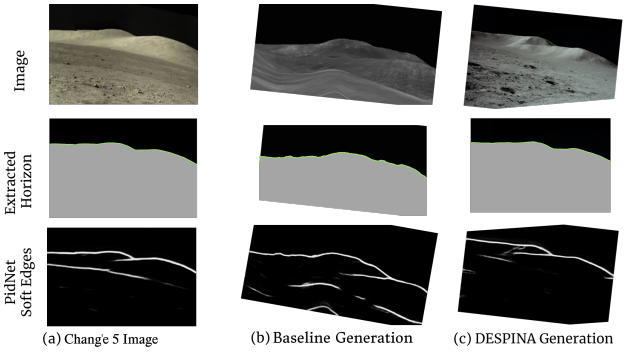


Figure 8. Top row: reference Chang’e-5 surface image at 43.06°N, 51.92°W (left), DEM-based classical rendering with mosaic drape (middle), and DESPINA output using DEM-derived inverse-depth and soft-edge constraints (right). Middle row: DEM-derived skyline (gray) overlaid on each image; green = skyline alignment. Bottom row: soft-edge maps extracted from each image for edge-structure analysis. Our method preserves skyline fidelity comparable to the classical renderer while producing textures more consistent with the real photograph.

4.0.2 Baselines We consider two common baselines:

- Classical DEM drape (renderer). DEM-based rendering textured with orbital mosaics and simple photometric shading; used in the Chang’e-5 case study (Fig. 8) to provide an interpretable geometric reference.
- Image-conditioned diffusion. Stable Diffusion XL with a lunar LoRA, conditioned on depth maps extracted directly from real photographs (standard image-to-image setup); used for historical Apollo comparisons (Fig. 9) and in Table 1.

By contrast, DESPINA computes constraints (inverse-depth and soft edges) directly from DEMs, enabling synthesis in unphotographed regions and ensuring geometry-first evaluation.

Table 1 summarizes comparisons. Note that all metrics exclude non-terrain pixels for fairness.

4.1 Quantitative Analysis

We quantitatively evaluate DESPINA against a state-of-the-art image-conditioned diffusion baseline by comparing synthetic DEM-derived lunar horizon images to real photographs. Direct one-to-one matching is challenging due to variations in lighting, viewing geometry, and the scarcity of precisely aligned ground-truth views. We therefore focus on structural and geological fidelity, which are more stable indicators of correctness across

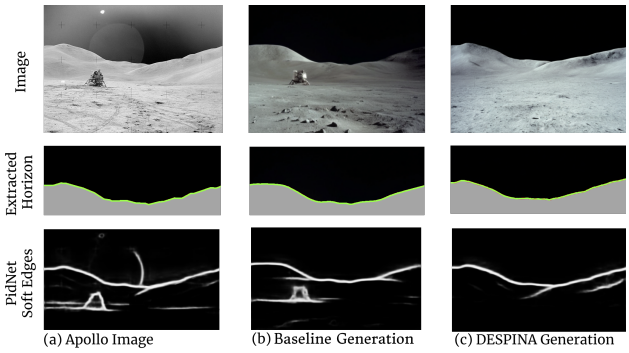


Figure 9. The Falcon lunar module landing site generated and analyzed to test model performance. A comparison of the baseline diffusion model approach (section 4.0.1) against Despina. The first row contains the images being analyzed, each from a different source. The second row shows the extracted horizon line from that image, and the third row shows calculated soft edges.

these conditions. Demonstrating that DESPINA reproduces accurate horizon structure and geological features in areas with reference imagery validates its ability to generalize to unseen terrain, an advantage over image-conditioned methods that require the target photograph for depth estimation and spatial conditioning (Bhat et al., 2024).

We report:

- SSIM (Hore and Ziou, 2010) – overall perceptual and structural similarity.
- Horizon RMSE (Liu et al., 2020a) – pixel-wise root mean squared error between DEM and extracted skyline profiles.
- SoftEdge F1 – edge-based F1 score from PiDiNet (Xu et al., 2023) within terrain masks, measuring preservation of geological boundaries.

Across both Apollo and Chang’e-5 datasets, DESPINA improves SSIM by ≈83% and SoftEdge F1 by ≈25% over the baseline, while cutting horizon error by more than half. This performance holds even at a non-Apollo site, showing that DESPINA ’s DEM-derived conditioning generalizes beyond the imagery used in training.

4.1.1 Ablation Study As shown in Table 2, we evaluated five distinct configurations: (1) Base model with soft edge constraints, (2) Base model with depth map constraints, (3) LoRA-trained model with soft edge constraints, (4) LoRA-trained model with depth map constraints, and (5) the full DESPINA system with all components. For each configuration, we generated images of the same lunar locations and evaluated them using our three primary metrics: SSIM, Horizon RMSE, and SoftEdge F1-scores. We chose not to include the base SDXL model without constraints as a baseline, as it generated images that were completely undesirable. Without any constraints, the model could not be expected to reproduce the spatial or terrain features that we were testing for.

The LoRA-trained model with soft edge constraints (Configuration 3) shows improved visual quality (SSIM: 0.489, Horizon RMSE: 4.92, SoftEdge F1: 0.723), achieving a 29.4% improvement over the baseline. This demonstrates that lunar-specific training enhances the model’s ability to generate realistic textures and features. The LoRA-trained model with depth

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-881-2026 | © Author(s) 2026. CC BY 4.0 License.

Lunar Location	Baseline (Image-Conditioned)			DESPINA (Terrain-Only)		
	SSIM	Horizon RMSE	SoftEdge F1	SSIM	Horizon RMSE	SoftEdge F1
Apollo Sites (Baseline has input-image advantage)						
Falcon LM / Apollo	0.262	4.98	0.658	0.534	1.92	0.780
Eagle LM / Apollo	0.289	5.27	0.623	0.519	2.31	0.808
EV-1 Panorama / Apollo	0.351	4.92	0.646	0.628	1.96	0.816
Chang’e-5 Sites (Baseline and DESPINA use same DEM region)						
Chang’e-5 Site 1	0.312	5.41	0.637	0.552	2.18	0.794
Chang’e-5 Site 2	0.298	5.16	0.629	0.541	2.05	0.803
Mean ± Std. Dev. (All)	[0.30 ± 0.027]	[5.15 ± 0.18]	[0.64 ± 0.011]	[0.55 ± 0.037]	[2.08 ± 0.14]	[0.80 ± 0.010]

Table 1. Structural accuracy metrics for Apollo and Chang’e-5 sites. This table evaluates only the image-conditioned baseline, which uses the original image to compute depth, giving it an inherent advantage on Apollo sites. Higher SSIM and SoftEdge F1 indicate better structural similarity; lower Horizon RMSE indicates better horizon alignment. DESPINA consistently outperforms the baseline in SSIM and SoftEdge F1 while substantially reducing horizon error.

Configuration	SSIM	Horizon RMSE	SoftEdge F1	Relative Improvement*
Base + Soft Edge	0.378	5.21	0.683	Baseline
Base + Depth Map	0.462	3.18	0.712	+22.2%
LoRA + Soft Edge	0.489	4.92	0.723	+29.4%
LoRA + Depth Map	0.534	2.31	0.780	+41.3%
Full DESPINA system	0.564	2.06	0.801	+47.1%

*Relative improvement calculated as a normalized composite score across all metrics (increase in SSIM and SoftEdge F1, decrease in RMSE).

Table 2. Ablation study results showing the contribution of different component combinations in the DESPINA pipeline. Higher SSIM and SoftEdge F1 scores indicate better image quality and feature preservation, while lower Horizon RMSE values indicate better topographical accuracy. The combination of all components yields the best performance across all metrics, with depth map constraints providing the most significant individual improvement to topographical accuracy.

map constraints (Configuration 4) achieves even better results (SSIM: 0.534, Horizon RMSE: 2.31, SoftEdge F1: 0.780), with a 41.3% improvement, showing the complementary benefits of both lunar-specific training and depth information.

The full DESPINA system (Configuration 5) achieves the best results across all metrics (SSIM: 0.564, Horizon RMSE: 2.06, SoftEdge F1: 0.801), demonstrating the synergistic effect of combining all components and yielding a 47.1% overall improvement. This comprehensive approach ensures both structural accuracy and visual realism in the generated lunar terrain.

4.2 Visual Realism and Qualitative Analysis

Initial manual inspections demonstrated promising results, indicating that our architecture effectively captures essential lunar features. The generated images show a high degree of realism, successfully replicating features such as rocks, horizon lines, and feasible textures (Kodikara and McHenry, 2020).

To further evaluate perceptual realism, we conducted an experiment in which 35 human participants were asked to distinguish between real and synthetic lunar images across three scenarios: testing the model base line (flat ground), testing depth details in hilly images and to test detail generation in images with many rocks. Examples of each image category are given in Fig. 7. We provided participants with the original image, images from the baseline pipeline with LoRA, and images from DESPINA. Participants were told that any number of images could have been generated synthetically, and they were to pick out as many artificial images as they could. Participants were able to correctly identify generated images 27.8% of the time with DESPINA and 65.7% of the time with the baseline method. This indicates DESPINA produces images that are more difficult to distinguish from real images, as seen in Table 3.

To refine our analysis, we performed a Two Proportion Z-Test across three categories of images, for both DESPINA and the baseline ensemble. The results are presented in Table 4. Here, lower accuracy is the preferred outcome. Our experimental results revealed distinct performance patterns across terrain categories. In the FLAT category, DESPINA achieved an accuracy of 0.22 (95% CI: 0.12-0.38), which was significantly distinguishable from chance ($p = 0.0012$), while the baseline method reached 0.64 accuracy (95% CI: 0.48-0.78) but was statistically indistinguishable from random guessing ($p = 0.1325$). Statistical comparison between methods showed significant differences ($p = 0.0004$). For HILLY terrain, DESPINA yielded 0.33 accuracy (95% CI: 0.20-0.50) without statistical differentiation from chance ($p = 0.0652$), whereas the baseline method demonstrated superior performance with 0.83 accuracy (95% CI: 0.68-0.92) and clear distinction from random outcomes ($p = 0.0001$), resulting in significant performance differences between methods ($p<0.0001$).

For the rocky image category, both methods exhibited intermediate performance metrics. DESPINA achieved 0.28 accuracy (95% CI: 0.16-0.44) and was distinguishable from chance ($p = 0.0113$), while the baseline method reached 0.50 accuracy (95% CI: 0.34-0.66) but failed to demonstrate statistical significance compared to random guessing ($p = 1.0000$). Notably, the performance difference between methods in this terrain category was not statistically significant ($p = 0.0531$), suggesting comparable effectiveness in challenging rocky environments. These findings indicate that terrain characteristics substantially influence recognition, with DESPINA showing more consistent behavior across varied landscapes despite lower overall accuracy scores (Hu et al., 2024). The results have been visualized in Fig. 10.

5. Conclusion

DESPINA makes several contributions to the fields of planetary imaging, dataset augmentation and image generation. DESPINA successfully generates high-fidelity lunar horizon images from

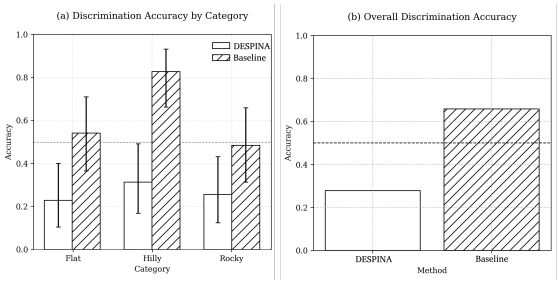


Figure 10. Distribution of Human Classification Accuracy Across Different Terrain categories, for DESPINA versus the base SDXL model. The dashed line indicates the 50% accuracy level expected from random guessing. Bars indicate the 95% CI range. Lower is better.

	Despina		Baseline	
Image Category	**Accuracy**	**95% CI**	**Accuracy**	**95% CI**
Flat	0.22	0.12-0.38	0.64	0.48-0.78
Hilly	0.33	0.20-0.50	0.83	0.68-0.92
Rocky	0.28	0.16-0.44	0.50	0.34-0.66

Table 3. Distribution of classification accuracy across different image categories for both DESPINA and the Baseline Model. Higher accuracy means participants were able to more accurately tell whether a given image was computer-generated.

DEMs without requiring terrain reference images, by converting structural restrictions into compact spatial embeddings. One unique characteristic of DESPINA is that it creates images for a given location and view direction, without relying on training images of the region of the location. We are not aware of any other approach with similar capabilities. The work also involved extensive analysis and preprocessing of Apollo lunar images to create the necessary training data. The generated images are not only visually appealing, we believe they are scientifically valuable. This scientific value is supported by research in semantic image synthesis similar to the work done by Park et al. (Park et al., 2019b), which demonstrates how spatially-adaptive normalization can produce photorealistic results that maintain semantic information - a technique similar to our constraint-based approach.

Beyond the immediate advancements in dataset creation, this work has potential applications across planetary science, autonomous navigation, and educational outreach (Hu et al., 2024). For planetary scientists, synthetic datasets that realistically reflect planetary surfaces that are mostly unseen from a ground perspective can help with environmental modeling for hard-to-image regions (Hargitai, 2019). Autonomous systems, such as lunar rovers, could benefit from this work by using synthetic data to train on datasets of various terrains and enhance navigation algorithms (Ceresoli et al., 2025).

Statistical Test	**FLAT**	**HILLY**	**ROCKY**
Despina vs. Chance	$p = 0.0012^{**}$	$p = 0.0652$	$p = 0.0113^{*}$
Baseline vs. Chance	$p = 0.1325$	$p = 0.0001^{***}$	$p = 1.0000$
Despina vs. Baseline	$p = 0.0004^{***}$	$p < 0.0001^{***}$	$p = 0.0531$

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 4. Two Proportion Z-Test Results Across Terrain Categories

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ..., Zheng, X., 2016. Tensorflow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 265–283.

Beckham, C., Honari, S., Verma, V., Lamb, A. M., Ghadiri, F., Hjelm, R. D., Bengio, Y., Pal, C., 2019. On adversarial mixup resynthesis. *Advances in neural information processing systems*, 32.

Bhat, S. F., Mitra, N., Wonka, P., 2024. Loosecontrol: Lifting controlnet for generalized depth conditioning. *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

California Air Resources Board, 2024. Harp digital elevation model files. `https://ww2.arb.ca.gov/resources/documents/harp-digital-elevation-model-files`. Accessed: 2025-9-17.

Ceresoli, M., Silvestrini, S., Lavagna, M., 2025. Optical Image Generation Through Digital Terrain Models for Autonomous Lunar Navigation. *Aerospace*, 12(2), 92.

Chinese Lunar and Planetary Data System, 2020. Chang’e 5 dataset. `https://clpds.bao.ac.cn/ce5web/searchOrder_hyperSearchData.search?pid=CE5/LCAM/level/2A`. Accessed: Jun. 13, 2024.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., Shah, M., 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9), 10850–10869.

Guérin, É., Digne, J., Galin, E., Peytavie, A., Wolf, C., Benes, B., Martinez, B., 2017. Interactive example-based terrain authoring with conditional generative adversarial networks. *ACM Transactions on Graphics*, 36(6).

Haase, I., Wählisch, M., Gläser, P., Oberst, J., Robinson, M. S., 2019. Coordinates and maps of the Apollo 17 landing site. *Earth and Space Science*, 6(1), 59–95.

Hargitai, H., 2019. *Planetary Cartography and GIS*. Springer.

Heaton, J., 2020. Applications of deep neural networks with keras. *arXiv preprint arXiv:2009.05673*.

Hore, A., Ziou, D., 2010. Image quality metrics: Psnr vs. ssim. *2010 20th international conference on pattern recognition*, IEEE, 2366–2369.

Hu, Z., Hu, K., Mo, C., Pan, L., Wang, Z., 2024. Terrain diffusion network: climatic-aware terrain generation with geological sketch guidance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38number 11, 12565–12573.

Ju, C., Wang, H., Cheng, H., Chen, X., Zhai, Z., Huang, W., Lan, J., Xiao, S., Zheng, B., 2024. Turbo: Informativity-driven acceleration plug-in for vision-language large models. *European Conference on Computer Vision*, Springer, 436–455.

Jurado-Rodríguez, D., Muñoz-Salinas, R., Garrido-Jurado, S., Medina-Carnicer, R., 2021. Design, detection, and tracking of customized fiducial markers. *IEEE Access*, 9, 140066–140078.

Kodikara, G. R., McHenry, L. J., 2020. Machine learning approaches for classifying lunar soils. *Icarus*, 345, 113719.

Li, J., Li, D., Xiong, C., Hoi, S., 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International conference on machine learning*, PMLR, 12888–12900.

Liu, Z., Song, C., Li, K., She, B., Yao, X., Qian, F., Hu, G., 2020a. Horizon extraction using ordered clustering on a directed and colored graph. *Interpretation*, 8(1), T1–T11.

Liu, Z., Zhu, J., Fu, H., Zhou, C., Zuo, T., 2020b. Evaluation of the vertical accuracy of open global DEMs over steep terrain regions using ICESat data: a case study over Hunan Province, China. *Sensors*, 20(17), 4865.

Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., Milford, M. J., 2015. Visual place recognition: A survey. *ieee transactions on robotics*, 32(1), 1–19.

Luddecke, T., Ecker, A., 2022. Image segmentation using text and image prompts. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7086–7096.

Lunar Reconnaissance Orbiter, 2011. Lroc global lunar dtm 100 m topographic model. `https://lroc.sese.asu.edu/data`.

Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., Zhao, H., 2023. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*.

NASA, 2017. Apollo 17 image library. `https://www.nasa.gov/history/alsj/a17/images17.html`.

NVIDIA Corporation, 2023. CUDA Toolkit Documentation.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ..., Bojanowski, P., 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y., 2019a. Gaugan: semantic image synthesis with spatially adaptive normalization. *ACM SIGGRAPH 2019 Real-Time Live!*, ACM SIGGRAPH, 1–1.

Park, T., Liu, M. Y., Wang, T. C., Zhu, J. Y., 2019b. Semantic image synthesis with spatially-adaptive normalization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2337–2346.

Petro, N. E., Keller, J. W., Morusiewicz, A. P., 2015. Data from the lunar reconnaissance orbiter (lro): Data products, tools, and community use. *Second Planetary Data Workshop*, 1846, 7016.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I., 2021. Zero-shot text-to-image generation. *International conference on machine learning*, Pmlr, 8821–8831.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. *Proceedings of the IEEE/CVF international conference on computer vision*, 12179–12188.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., Catanzaro, B., 2018. High-resolution image synthesis and semantic manipulation with conditional gans. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8798–8807.

Xu, J., Xiong, Z., Bhattacharyya, S. P., 2023. Pidnet: A real-time semantic segmentation network inspired by pid controllers. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19529–19539.

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. Mvsnet: Depth inference for unstructured multi-view stereo. *Proceedings of the European conference on computer vision (ECCV)*, 767–783.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, Y., Kang, Z., Cao, Z., 2024. An Image Retrieval Method for Lunar Complex Craters Integrating Visual and Depth Features. *Electronics*, 13(7), 1262.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-881-2026 | © Author(s) 2026. CC BY 4.0 License.

A. Supplementary Material

A.1 Generalization to Earth and Other Planetary Bodies

A.1.1 Overview This appendix demonstrates that DESPINA generalizes beyond lunar terrain. Because the system derives structural embeddings directly from DEM geometry, the method is inherently applicable to Earth, Mars, or any planetary body with available elevation data. The following example shows the full constraint and reconstruction pipeline operating on a terrestrial DEM, supporting the vision of a universal geospatial representation that unifies elevation-based geometry with learned latent generative models.

A.1.2 Earth DEM Case Study To validate cross-planetary applicability, we applied DESPINA’s pipeline to an Earth DEM tile, from Aguerberry Point in Death Valley National Park (California Air Resources Board, 2024). Figure 11 demonstrates the complete workflow: (a) input Earth DEM patch, (b) raycasted depth embedding, (c) soft-edge structural embedding, (d) geometry-preserving reconstruction using diffusion, and (e) a real photograph of the same location, for comparison. The reconstruction maintains terrain-consistent geometry while adapting photometric properties to terrestrial appearance priors. In this example, the reconstruction is able to capture the same features as the real photograph, including the horizon line and the terrain features.

A.2 BLIP Prompt Generation Example

We used BLIP (Li et al., 2022) to generate structured text prompts from binary annotations for Apollo imagery. 12 is an example showing the image and corresponding prompt:

A.3 Dataset Preparation

To establish a reliable dataset for DESPINA’s architecture, we exclusively utilized Apollo mission imagery, which is publicly available and readily accessible. The decision to rely solely on Apollo data, rather than incorporating additional datasets like those from more recent missions like Chang’e 5, was driven by the need for image consistency in format and structure. Although other sources, such as Chang’e 5 and Yutu rover images, as seen in, offered potentially valuable data, they introduced significant challenges: either the datasets were fragmented and lacked cohesive organization, or access was heavily restricted by governing agencies, limiting their practical use. However, small samples of the Chang’e dataset were utilized in the paper as a validation set, which proved useful.

To obtain the Apollo mission imagery, we developed a script to download all publicly available images from NASA’s online archives. The script utilized Python’s requests library to automate the retrieval of image files. Using a recursive script, images were downloaded sequentially from mission subdirectories.

Most Apollo mission images include fiducial markers, which are small crosshair-like artifacts introduced for photogrammetric purposes. These markers interfere with existing computer vision techniques by introducing unnatural geometry that disrupts segmentation models. To address this, we developed a preprocessing pipeline for detecting and removing fiducial markers. The process involves identifying markers based on their geometric properties and applying inpainting techniques to remove them while preserving surrounding image features.

A.4 Hardware and Training Configuration

The SDXL LoRA and the DPT Fine-Tuning was trained using:

- Hardware: Six NVIDIA A100 GPUs (40GB memory) with CUDA 12.1.1
- Software: Python 3.11, PyTorch 2.1.2 (foss-2023a toolchain), TorchVision 0.16.0
- Dataset: 3,500 preprocessed images (2,900 training, 600 validation)
- Parameters: Batch size 6, three discriminators, dropout enabled

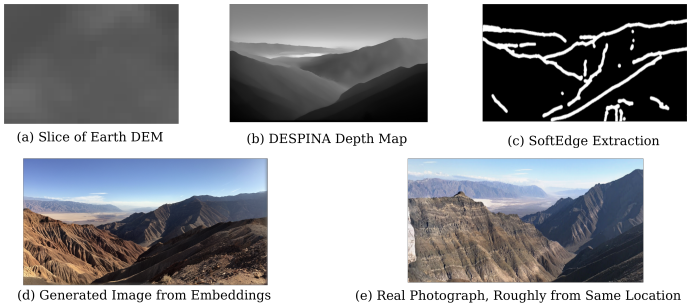


Figure 11. Demonstration of DESPINA applied to an Earth surface. (a) Input DEM patch, (b) raycasted depth embedding, (c) soft-edge structural embedding, (d) geometry-preserving reconstruction using DEM-guided diffusion, (e) a real photograph of roughly the same location.

Figure 12. Apollo image used in BLIP prompt generation. **Generated Prompt:** *A realistic image of the Moon’s surface, from the perspective of a rover. A lunar landscape with hills in the background. A lunar crater is in view. A rover is present in the image. Dark shadows cast by the terrain. An astronaut is visible in the image.*

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-881-2026 | © Author(s) 2026. CC BY 4.0 License.