Monocular 3D Reconstruction for Martian Terrain Based on Diffusion Model

Jiarui Cao[1], Rong Huang[1,2]*, Yusheng Xu[1,2], Zhen Ye[1,2], Xiaohua Tong[1,2]

[1]College of Surveying and Geoinformatics, Tongji University, Shanghai, China
[2] The Shanghai Key Laboratory of Space Mapping and Remote Sensing for Planetary Exploration, Shanghai, China

Keywords: Mars exploration, Digital terrain model (DTM), 3D reconstruction, Diffusion model

Abstract

High-precision digital terrain models (DTMs) are important for Mars explorations and research, providing indispensable spatial information for landing site assessment, rover path planning, and surface environment analysis. However, challenges such as high-resolution stereo data scarcity and complex atmospheric conditions on the Martian surface result in traditional terrain reconstruction methods suffer from limitations in accuracy, coverage and resolution. To enhance the model's ability to recover fine-grained topography, we present a diffusion-based monocular terrain reconstruction method, which progressively recovers Martian terrains from single-view high-resolution optical images. We employed a multi-scale U-Net denoising network with attention mechanisms and introduced an additional end-to-end depth constraint. To improve terrain reconstruction efficiency, we implemented a diffusion model in the latent space and adopted a skipping sampling mechanism. We employed the proposed method to reconstruct terrain in different regions. Experimental results demonstrate that the reconstructed terrain achieves an accuracy of 2 m. Furthermore, compared to photogrammetric terrain, the shaded relief generated by our method exhibits greater similarity to the input imagery.

1. Introduction

High-resolution digital terrain models (DTMs) provide explicit representations of surface morphology, spatial distribution patterns, and geomorphological structures. These fine-scale models provide direct evidence for traces of potential life and safety assurance for precision landing and navigation, delivering spatial information support in both scientific research and mission planning (Sutton et al., 2022, Chen et al., 2024). The lack of detailed topographic data often leads to a series of risks, such as lander crashes, rover rollovers, and incorrect resource assessments, severely affecting the reliability and safety of exploration missions and scientific research (Sakai et al., 2025).

Martian three-dimensional terrain products mainly rely on remote sensing data such as optical imagery and laser altimetry obtained by orbiters. Laser altimetry can obtain high-precision elevation measurements. But laser footprint distributions are sparse, resulting in terrain models with limited spatial resolution (Neumann et al., 2001). Methods based on optical imagery can produce terrain products with higher resolution (Li et al., 2011). However, classical photogrammetry methods rely on stereo image pairs, the coverage of which is far smaller than that of these high-resolution images acquired by the orbiter. Taking HiRISE images as an example, less than 1% of high-resolution images meet the stereo requirements (McEwen et al., 2023). Furthermore, the weak texture of the Martian surface increases the uncertainty of stereo matching, thereby limiting the quality of topographic reconstruction (Kirk et al., 2021). Photoclinometry is an alternative method not constrained by stereo pairs, which can infer terrain geometry from pixel-level intensity variations in both multi-view and single-view images (Liu and Wu, 2023). Nevertheless, building an accurate reflectance and atmospheric model is difficult and costly due to the spatial heterogeneity of Martian atmospheric scattering effects and surface reflection characteristics (Qian et al., 2025). Inaccurate models directly affect the reliability of terrain inversion results.

In contrast, deep learning-based approaches offer a fundamentally different paradigm. By learning direct mappings from image pixels to elevation values, these methods bypass the need for explicit stereo matching or strict reflectance modeling. This enables the reconstruction of large-scale, high-resolution Martian terrain from single-view orbital imagery (Tao et al., 2023). The learning-based methods can be broadly divided into discriminative methods and generative methods. Discriminative models learn the complex, end-to-end, and non-linear mapping between images and heights from training data (Chen et al., 2021). But traditional convolutional neural network (CNN) architectures are constrained by the local receptive fields of their convolutional kernels. To overcome this limitation, the transformer architecture was introduced to model long-range dependencies between image patches, thereby enhancing the network model's ability to capture global contextual information (Yang et al., 2024, La Grassa et al., 2026). However, these methods, which directly learn the mapping between samples, are prone to bias towards the distribution of the training data, thereby limiting the generalization ability of the method.

Compared to discriminative methods, methods based on generative networks treat the 3D estimation problem as a conditional generation problem. These models possess powerful generative capabilities, learning the underlying distribution of the target terrain data to produce terrain reconstructions that are richer in detail and larger in scope (Chen et al., 2018). Generative Adversarial Networks (GANs)-based methods incorporate an adversarial training mechanism to improve the fidelity of the reconstructions and have shown superior adaptability in planetary terrain reconstruction (Liu et al., 2022, La Grassa et al., 2022). By learning to generate realistic surface topographies, they achieve enhanced adaptability to diverse terrain types and exhibit improved detail preservation compared to deterministic regression models (Tao et al., 2021). Attention mechanisms and composite loss functions have also been applied to GAN-based methods, which progressively reconstruct detailed Martian terrain in key areas from coarse to fine using HRSC, CTX, or HiRISE imagery (Yang et al., 2024, Cao et al., 2025). Despite these

* Corresponding author

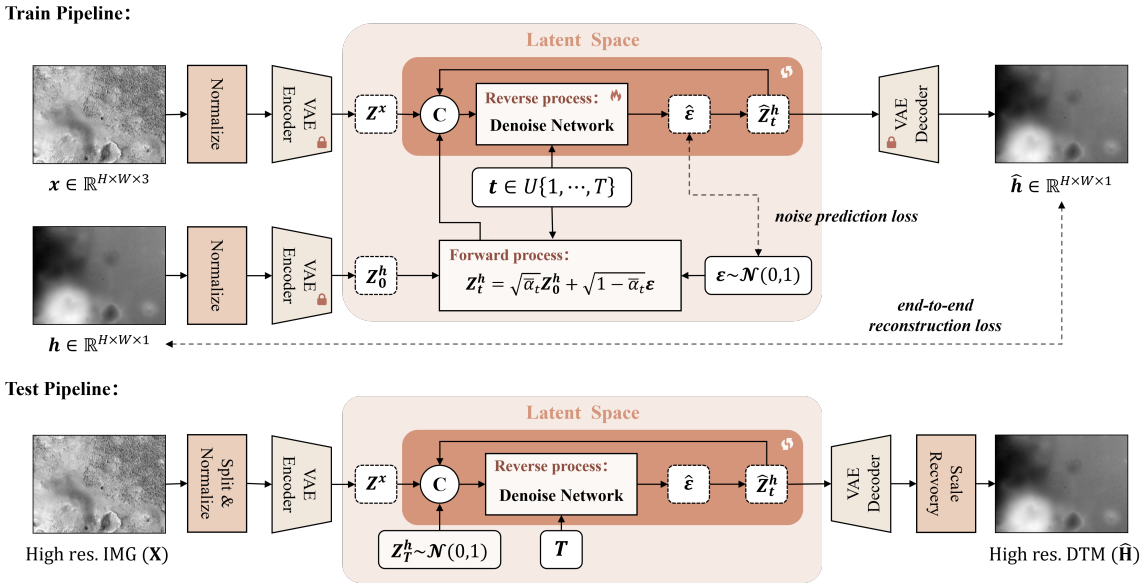

Figure 1. The training and inference pipeline of the proposed diffusion-based terrain reconstruction model.

advances, this adversarial training lacks stability, which limits the model's terrain reconstruction performance. Existing methods often struggle to reconstruct terrain details, leading to the smoothing of local topographic variations and compromising the quality of the reconstruction.

In recent years, diffusion models, as a new generation of generative artificial intelligence technology, have achieved significant progress in vision fields (Croitoru et al., 2023). Conditional diffusion models have already been applied to monocular depth estimation, with the generated depth maps demonstrating superior performance in terms of detail preservation and artifact control (Ke et al., 2024, Pham et al., 2025). With their more stable training and stronger ability to capture fine details, diffusion-based methods hold great potential for practical applications (Sun et al., 2025). However, the potential of diffusion models to generate high-quality results remains largely unexplored in planetary terrain reconstruction.

Therefore, to better capture fine terrain details, we propose a novel framework based on diffusion models tailored for Mars terrain reconstruction. In this work, we transform the Martian monocular terrain reconstruction problem into a generative task and restore terrain details through an iterative denoising process. By leveraging the inherent capability of diffusion models, our method can enhance the accuracy and fidelity of the reconstructed terrain, thereby supporting the requirements and objectives of future Mars exploration missions.

2. Methodology

2.1 Overall pipeline

We reformulate the Martian monocular terrain reconstruction task as a conditional denoising diffusion generation process. As illustrated in Fig. 1, we develop a diffusion model-based terrain framework to achieve pixel-level terrain inference. By treating the input high-resolution imagery $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ as a condition, we model a conditional probability distribution $\mathrm{D}(\mathbf{h}|\mathbf{x})$ guided by image features, which learns how to progressively recover the target terrain $\mathbf{h} \in \mathbb{R}^{H\times W\times 1}$ from pure noise.

To improve the efficiency of model training and inference, we perform the diffusion process in the latent space. All input data is uniformly split, normalized, and compressed to the latent representation by a pre-trained variational autoencoder (VAE) encoder. The latent terrain code generated after denoising also needs to be decoded by the VAE decoder to get the final inference result. During inference, the relative heights derived from the diffusion model must recover ground scale before the final elevation data can be obtained. The scale recovery process relies on the mean and variance calculated from the low-resolution terrain model, during which the generated relative height is stretched to the absolute scale. Finally, these absolute height maps are seamlessly stitched together to produce a complete, high-resolution 3D terrain model.

2.2 Diffusion Model

Diffusion models consist of a forward process and a reverse process. The forward noising process progressively injects Gaussian noise into the height distribution, degrading structured features gradually into a pure noise image. While the reverse denoising process removes noise based on input image features step by step until the target terrain is fully reconstructed. During this process, we train a neural network to learn the denoising operations. Once training is complete, this denoise network can be used to reconstruct terrain representations corresponding to the input images through iterative denoising. To accelerate sample generation, we employ a denoising diffusion implicit model (DDIM) sample strategy (Song et al., 2020). By constructing a process with the same marginal distribution as the original denoising diffusion probabilistic model (DDPM), the reverse process allows skipping timesteps and reduces training and inference time.

2.2.1 Forward Noising Process The forward noise-adding process injects T rounds of Gaussian noise into the original terrain latent code $\mathbf{z}_0^{\mathbf{h}} = \mathcal{E}(\mathbf{h})$ compressed by a VAE encoder $\mathcal{E}$, causing the data distribution $\mathbf{z}_T^{\mathbf{h}}$ to degenerate into an isotropic Gaussian distribution. In accordance with the DDIM (Song et al., 2020), the entire joint distribution of non-Markov forward

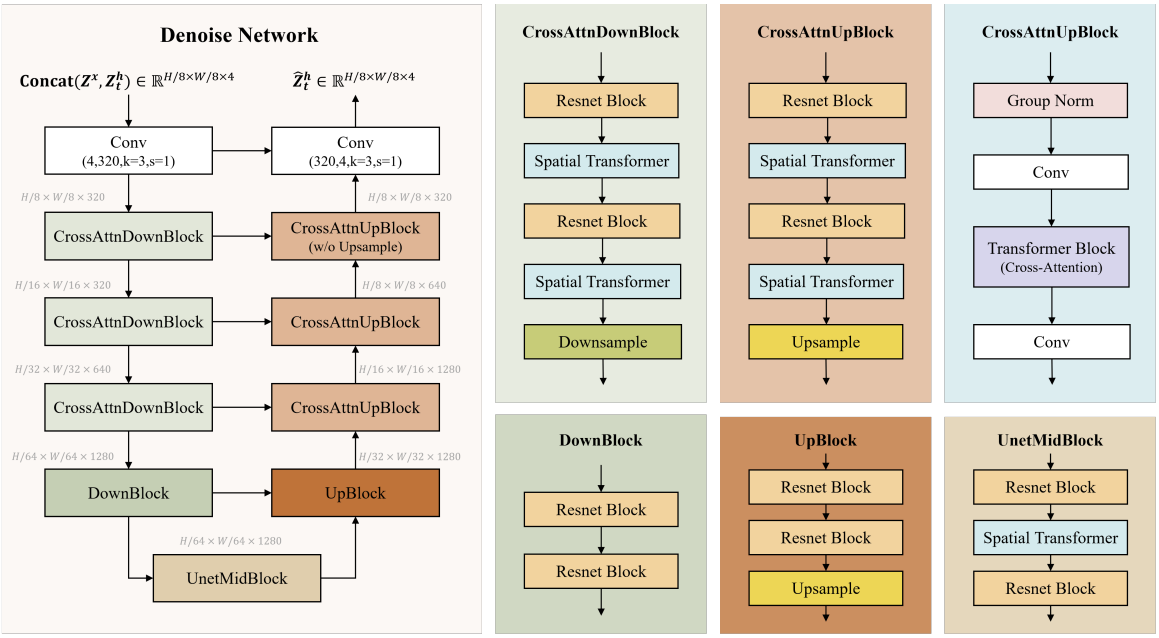

Figure 2. Overview of the denoising network architecture.

processes can be defined directly as

$$q_\sigma(\mathbf{z}_t^\mathbf{h} \mid \mathbf{z}_0^\mathbf{h}) = \mathcal{N}\left(\mathbf{z}_t^\mathbf{h}; \sqrt{\bar{\alpha}_t}\mathbf{z}_{t-1}^\mathbf{h}, (1-\bar{\alpha}_t)\mathbf{I}\right), \quad (1)$$

where β_t is the default noise variance schedule and $\alpha_t = 1-\beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. By employing a reparameterization trick, we can sample noise $\mathbf{z}_t^\mathbf{h}$ at any arbitrary timestep t. The formula for adding noise is

$$\mathbf{z}_t^\mathbf{h} = \sqrt{\bar{\alpha}_t}\mathbf{z}_0^\mathbf{h} + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (2)$$

where ϵ denotes standard Gaussian noise. As the timestep t increases, $\bar{\alpha}_t$ approaches zero, thereby $\mathbf{z}_t$ converging to pure noise.

2.2.2 Reverse Denoising Process The reverse process aims to reconstruct $\mathbf{z}_0^\mathbf{h}$ from a random noise sample $\mathbf{z}_T^\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, conditioned on the input image features $\mathbf{z}^\mathbf{x} = \mathcal{E}(\mathbf{x})$. The reverse process is parameterized as follows:

$$p_\theta(\mathbf{z}_{t-1}^\mathbf{h} \mid \mathbf{z}_t^\mathbf{h}, \mathbf{z}^\mathbf{x}) = \mathcal{N}\left(\mathbf{z}_{t-1}^\mathbf{h}; \mu_\theta(\mathbf{z}_t^\mathbf{h}, t, \mathbf{z}^\mathbf{x}), \sigma_t^2\mathbf{I}\right), \quad (3)$$

where $\mu_\theta = \sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1-\alpha_t}}$ σ_t is a configurable parameter to control the stochasticity. Unlike image generation tasks, in terrain reconstruction, each input image corresponds to a single, unique correct terrain. The reconstruction process requires a deterministic mapping. By setting $\sigma_t = 0$, the sampling becomes deterministic.

The reverse denoising process is unknown, so we use a denoise network $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t^\mathbf{h}, t, \mathbf{z}^\mathbf{x})$ with learnable parameters θ to fit an approximation of the inverse distribution $q_\sigma(\mathbf{z}_{t-1}^\mathbf{h} \mid \mathbf{z}_t^\mathbf{h})$. The corresponding DDIM sampling process is as follows:

$$\mathbf{z}_{t-1}^\mathbf{h} = \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{z}}_0^\mathbf{h} + \sqrt{1-\bar{\alpha}_{t-1}} \cdot \boldsymbol{\epsilon}_\theta(\mathbf{z}_t^\mathbf{h}, t, \mathbf{z}^\mathbf{x})) \quad (4)$$

During inference, the reverse process is initialized with pure noise $\mathbf{z}_T^\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively denoised using this DDIM sampler to obtain the terrain latent code $\hat{\mathbf{z}}_0^\mathbf{h}$. Using the sampling method described above, the denoise process can be performed with fewer steps without significant degradation in terrain reconstruction quality.

2.3 Network Architecture

This work builds upon a successful image generation architecture. The model has been trained on a mixture of real and synthetic data, endowing it with rich visual prior knowledge, powerful image generation capabilities, and strong generalization ability. This provides the foundation for the model to achieve high-quality terrain reconstruction. The entire model comprises three components: a multi-scale denoising U-Net augmented with a fusion attention mechanism as the core part, alongside a VAE encoder that compresses input data into a latent space and a VAE decoder that reconstructs the output from the denoised latent.

2.3.1 Denoising Network To enhance the fidelity of terrain reconstruction, we introduce a multi-scale U-Net augmented with a fusion attention mechanism. As illustrated in Fig. 2, the core denoising network $\boldsymbol{\epsilon}_\theta$ is a time-conditioned U-Net operating on the latent space. It processes the input latent through a series of downsampling and upsampling blocks, capturing both local details and global context.

Specifically, the input image latent $\mathbf{z}^\mathbf{x}$ and the noisy terrain latent $\mathbf{z}_t^\mathbf{h}$ are first concatenated along the channel dimension. So that image features can continue to guide the terrain reconstruction process. The timestep t is embedded via sinusoidal positional encoding and injected into each residual block via adaptive layer normalization. As the stable diffusion model is designed for text-to-image generation tasks, upblocks and downblocks at different resolution levels incorporate cross-attention mechanisms to align multi-model information. To better inherit the visual prior from large-scale models, these modules have been retained. During training or inference, zero embeddings are fed into these modules, ensuring that no information is provided to them and that the terrain reconstruction process remains unaffected. Thus, the denoise network can learn terrain geometry from visual cues such as shadows, slopes, and

textures and maintain strong performance throughout the denoising process.

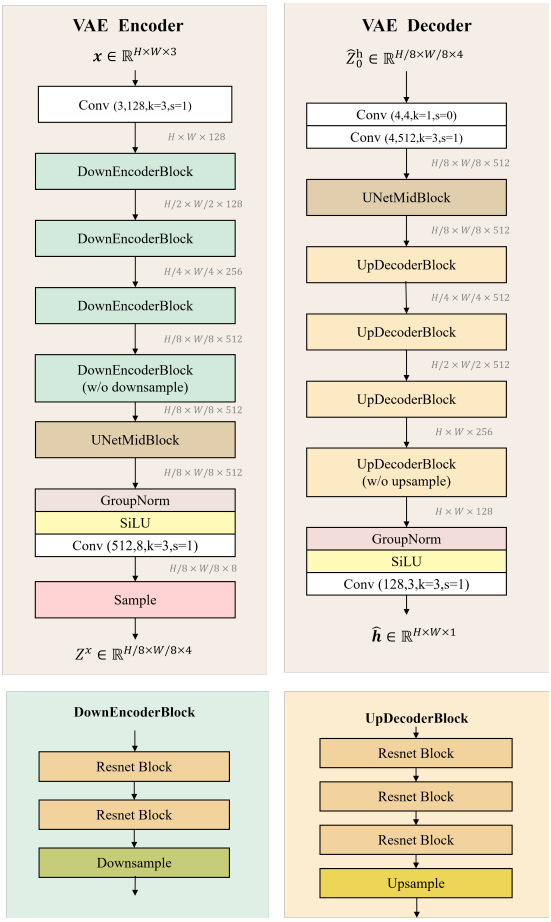

Figure 3. Overview of the VAE encoder and decoder architecture.

2.3.2 VAE Encoder and decoder The pre-trained VAE encoder $\mathcal{E}$ and decoder $\mathcal{D}$ are inherited from the original Stable Diffusion. They transform all operations into a latent space, reducing computational complexity while ensuring high-fidelity reconstruction. As illustrated in Fig. 3, the VAE encoder $\mathcal{E}$ is responsible for compressing the input image $\mathbf{x}$ and the reference terrain $\mathbf{h}$ into low-dimensional latent codes $\mathbf{z}^{(\mathbf{x})}$ and $\mathbf{z}_0^{\mathbf{h}}$, respectively. The VAE decoder $\mathcal{D}$ subsequently maps the denoised latent code $\mathbf{z}_0^{\mathbf{h}}$ back to pixel space, outputting the inference relative height $\hat{\mathbf{h}}$. Since the original VAE is designed for three-channel RGB images while the depth map is single-channel, the depth map is replicated along the channel dimension to form three identical channels before being fed into the encoder, ensuring compatibility. Both VAEs are pre-trained on large-scale datasets and remain frozen during training.

2.4 Loss Function

To achieve both perceptual quality and geometric accuracy, we train the model with a composite loss function that combines a latent-space noise prediction loss and an end-to-end reconstruction loss.

The noise prediction loss follows the standard denoising objective in latent space (Song et al., 2020). As mentioned above, the learned denoising distribution $p_\theta(\mathbf{z}_{t-1}^{\mathbf{h}} \mid \mathbf{z}_t^{\mathbf{h}})$ needs to fit the noisy adding inverse distribution $q_\sigma(\mathbf{z}_{t-1}^{\mathbf{h}} \mid \mathbf{z}_t^{\mathbf{h}})$. This requires bringing the means of the two normal distributions as close together as possible. By deriving this optimization objective, it can be simplified to making the prediction noise $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t^{\mathbf{h}}, t, \mathbf{z}^{\mathbf{x}})$ as close as possible to the random noise $\boldsymbol{\epsilon}$. Therefore, the latent noise prediction loss function for the denoising model is defined as follows:

$$\mathcal{L}_{\text{noise}} = \mathbb{E}_{t,\mathbf{z}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t^{\mathbf{h}}, t, \mathbf{z}^{\mathbf{x}})\|_2^2\right]. \tag{5}$$

This loss enforces a coherent denoising trajectory, ensuring that the reverse process progressively refines the latent representation.

To better perceive the overall trends in terrain changes and optimize the accuracy of the reconstructed results, we add an end-to-end reconstruction loss. Specifically, the denoised latent $\hat{\mathbf{z}}_0^{\mathbf{h}}$ obtained from the reverse process is decoded to the relative height $\hat{\mathbf{h}} = \mathcal{D}(\hat{\mathbf{z}}_0^{\mathbf{h}})$. We apply low-pass filtering to both this model output and the reference relative height. The difference between the two is calculated to obtain the end-to-end reconstruction loss:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{\mathbf{h},\hat{\mathbf{h}}}\|\text{Lowpass}(\mathbf{h}) - \text{Lowpass}(\hat{\mathbf{h}})\|_1. \tag{6}$$

This end-to-end loss effectively addresses the limitation of recovering global trends when the model processes in latent space, thereby improving the accuracy of the reconstruction results.

The total training objective is a weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{noise}} + \lambda_{\text{recon}} \cdot \mathcal{L}_{\text{recon}}, \tag{7}$$

where λ_{recon} balances the contribution of the reconstruction loss. This composite loss ensures that the denoising process remains physically coherent while also achieving high quantitative accuracy on the final height output.

3. Experiments and results

We trained the denoising network using a dataset comprising 30,000 pairs of uniformly downsampled HiRISE images and corresponding terrain samples at 2 m/pixel resolution. The training pairs were extracted from diverse Martian landscapes to ensure the model learns a broad range of topographic features, including craters, ridges, and smooth plains. For evaluation, we selected six test areas that are distinct from the training data and feature representative and diverse topography, thereby providing a rigorous assessment of the model's generalization capability. The reconstructed terrain results are shown in Figure 4, and the quantitative reconstruction accuracy relative to the reference HiRISE DTM is reported in Table 1.

Test Area	RMSE (m)	Abs Rel (m)
Area 1	1.37	0.26
Area 2	1.52	0.24
Area 3	1.04	0.24
Area 4	1.84	0.28
Area 5	1.64	0.31
Area 6	1.13	0.25

Table 1. Reconstruction errors in areas with different landforms.

It can be observed that the diffusion-based model effectively reconstructs a high-resolution terrain model, capturing the elevation trends consistent with the reference terrain. Craters of varying sizes, the undulating terrain within the delta, the raised hills,

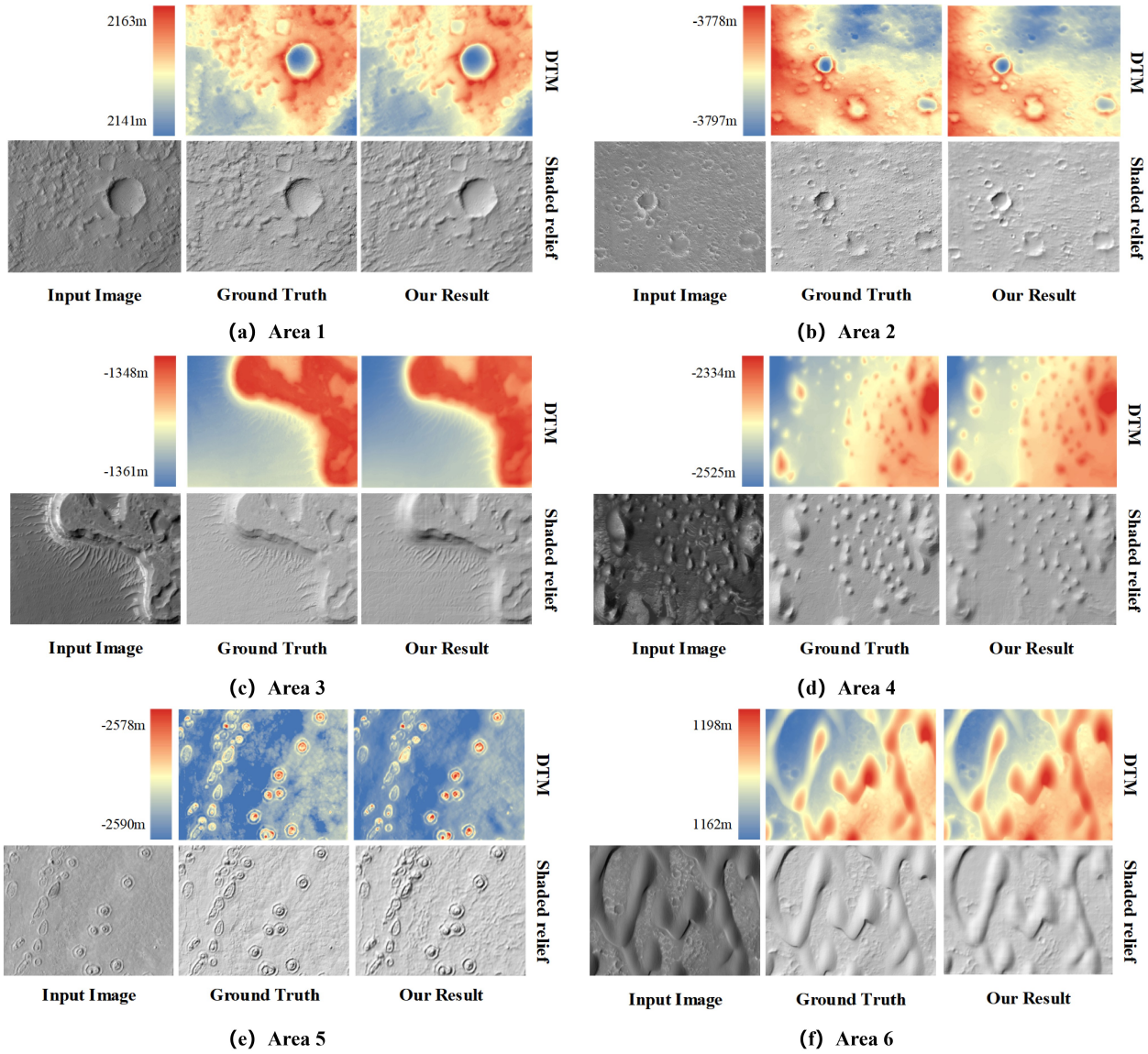

Figure 4. Different terrain reconstructions and their shaded reliefs (azimuth: 315°; elevation: 45°), where the first column shows the input HiRISE raw image, the second column displays the reference HiRISE DTM, and the third column presents our reconstruction results.

the dense small sand cones, and sand dunes have all been successfully reconstructed. The reconstruction accuracy, measured as the root-mean-square error (RMSE) between the predicted elevation and the ground-truth HiRISE DTM, is 2 meters. This demonstrates that the proposed method achieves faithful elevation recovery even in areas with complex topography.

Furthermore, we quantified the visual similarity between the hill-shade relief derived from the reconstructed DTMs and the input HiRISE imagery. Hill-shade rendering was performed under identical solar illumination conditions for both the reconstructed DTMs and the reference HiRISE DTMs. When comparing these hill-shade renderings with the original input images, our reconstruction achieved an average structural similarity index (SSIM) of 0.530. This is notably higher than the average SSIM of 0.384 obtained for the photogrammetrically derived HiRISE DTMs. The result indicates that, from a visual perspective, our reconstructed terrain exhibits stronger correspondence with the input imagery than the conventional photogrammetric products. Such improvement is attributed to the diffusion model’s ability to exploit image intensity patterns and texture cues as implicit constraints for terrain shape, leading to enhanced topographic-image consistency.

These findings collectively verify that the proposed method can retain detailed topographic features and achieve stronger alignment between the reconstructed terrain and the input image, compared with reference photogrammetric DTMs.

4. Conclusion

We proposed a diffusion-based framework for monocular Mars terrain reconstruction, enabling high-resolution DTMs to be generated without the need for stereo imagery or complex photometric modeling. The method retains fine-scale geomorphological structures and produces terrain surfaces that align more

closely with optical image details compared with conventional photogrammetric DTMs. Quantitative assessment shows a reconstruction accuracy of about 2 m and improved visual correspondence to imagery, indicating that diffusion probabilistic models can effectively mitigate detail loss present in existing approaches. These experimental results demonstrate the significant potential of diffusion models for three-dimensional reconstruction of Martian terrain, enabling subsequent large-scale, high-resolution terrain reconstruction.

Acknowledgment

The work described in this paper was supported by the National Natural Science Foundation of China (Project nos. 42571422, 42571518, and 42595541), and the Shanghai Pilot Program for Basic Research (22TQ1400300).

References

Cao, J., Huang, R., Xu, Y., Ye, Z., Qian, J., Hoegner, L., Tong, X., 2025. LoGAN: A novel local attentive generative adversarial resizable network for detailed 3D reconstruction of the Martian surface using monocular HiRISE images and DTMs. *ISPRS Journal of Photogrammetry and Remote Sensing*, 225, 302–327.

Chen, R., Mahmood, F., Yuille, A., Durr, N. J., 2018. Rethinking monocular depth estimation with adversarial training. *arXiv preprint arXiv:1808.07528*.

Chen, X., Zhang, Z., Li, J., Li, S., Xu, T., Zheng, B., Sun, X., Wu, Y., Diao, Y., Li, X., 2024. Formation of Tianwen-1 landing crater and mechanical properties of Martian soil near the landing site. *International Journal of Mining Science and Technology*, 34(9), 1293–1303.

Chen, Z., Wu, B., Liu, W. C., 2021. Mars3DNet: CNN-based high-resolution 3D reconstruction of the Martian surface from single images. *Remote Sensing*, 13(5), 839.

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.

Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R. C., Schindler, K., 2024. Repurposing diffusion-based image generators for monocular depth estimation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9492–9502.

Kirk, R. L., Mayer, D. P., Fergason, R. L., Redding, B. L., Galuszka, D. M., Hare, T. M., Gwinner, K., 2021. Evaluating stereo digital terrain model quality at Mars rover landing sites with HRSC, CTX, and HiRISE images. *Remote Sensing*, 13(17), 3511.

La Grassa, R., Gallo, I., Re, C., Cremonese, G., Landro, N., Pernechele, C., Simioni, E., Gatti, M., 2022. An adversarial generative network designed for high-resolution monocular depth estimation from 2D HiRISE images of Mars. *Remote Sensing*, 14(18), 4619.

La Grassa, R., Re, C., Tullo, A., Gallo, I., Cremonese, G., 2026. Transformer-driven monocular high-resolution DTM generation on mars via multimodal integration of CaSSIS imagery and MOLA altimetry. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 19, 100118.

Li, R., Hwangbo, J., Chen, Y., Di, K., 2011. Rigorous photogrammetric processing of HiRISE stereo imagery for Mars topographic mapping. *IEEE Transactions on Geoscience and Remote Sensing*, 49(7), 2558–2572.

Liu, W. C., Wu, B., 2023. Atmosphere-aware photoclinometry for pixel-wise 3D topographic mapping of Mars. *ISPRS Journal of Photogrammetry and Remote Sensing*, 204, 237–256. https://doi.org/10.1016/j.isprsjprs.2023.09.017.

Liu, Y., Wang, Y., Di, K., Peng, M., Wan, W., Liu, Z., 2022. A generative adversarial network for pixel-scale lunar DEM generation from high-resolution monocular imagery and low-resolution DEM. *Remote Sensing*, 14(21), 5420.

McEwen, A., Byrne, S., Hansen, C., Daubar, I., Sutton, S., Dundas, C., Bardabelias, N., Baugh, N., Bergstrom, J., Beyer, R. et al., 2023. The high-resolution imaging science experiment (HiRISE) in the MRO extended science phases (2009–2023). *Icarus*, 115795.

Neumann, G. A., Rowlands, D. D., Lemoine, F. G., Smith, D. E., Zuber, M. T., 2001. Crossover analysis of Mars orbiter laser altimeter data. *Journal of Geophysical Research: Planets*, 106(E10), 23753–23768.

Pham, D.-H., Do, T., Nguyen, P., Hua, B.-S., Nguyen, K., Nguyen, R., 2025. Sharpdepth: Sharpening metric depth predictions using diffusion distillation. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17060–17069.

Qian, J., Ye, Z., Xu, Y., You, Q., Huang, R., Liu, S., Xie, H., Feng, Y., Tong, X., 2025. Multi-Image Shape and Albedo From Shading With Atmospheric Correction for Precise Topographic Reconstruction on Mars. *IEEE Transactions on Geoscience and Remote Sensing*.

Sakai, S., Kushiki, K., Sawai, S., Fukuda, S., Miyazawa, Y., Ishida, T., Kariya, K., Ito, T., Ueda, S., Yokota, K. et al., 2025. Moon landing results of SLIM: A smart lander for investigating the Moon. *Acta Astronautica*, 235, 47–54.

Song, J., Meng, C., Ermon, S., 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Sun, L., Zhou, H., Yang, L., Zhao, D., Zhang, D., 2025. LD-DEM: Latent Diffusion with Conditional Decoding for High-Precision Planetary DEM Generation from RGB Satellite Images. *Aerospace*, 12(8), 658.

Sutton, S. S., Chojnacki, M., McEwen, A. S., Kirk, R. L., Dundas, C. M., Schaefer, E. I., Conway, S. J., Diniega, S., Portyankina, G., Landis, M. E. et al., 2022. Revealing active Mars with HiRISE digital terrain models. *Remote Sensing*, 14(10), 2403.

Tao, Y., Muller, J.-P., Xiong, S., Conway, S. J., 2021. MADNet 2.0: Pixel-scale topography retrieval from single-view orbital imagery of Mars using deep learning. *Remote Sensing*, 13(21), 4220.

Tao, Y., Walter, S. H., Muller, J.-P., Luo, Y., Xiong, S., 2023. A High-Resolution Digital Terrain Model Mosaic of the Mars 2020 Perseverance Rover Landing Site at Jezero Crater. *Earth and Space Science*, 10(10), e2023EA003045.

Yang, L., Zhu, Z., Sun, L., Zhang, D., 2024. Global attention-based DEM: A planet surface digital elevation model-generation method combined with a global attention mechanism. *Aerospace*, 11(7), 529.