Title of the manuscript: Monocular Depth Estimation of Planetary Landforms: A Diffusion Model Approach for Faster Inference

Complete list of authors:

1. Nethmi Jayakody
2. Poorna Cooray
3. Supun Dasanayake
4. Thanuja D. Ambegoda

Affiliations of all authors: Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka

Corresponding author's details (email, institution, etc.)

thanuja@cse.mrt.ac.lk

Department of Computer Science & Engineering, University of Moratuwa

Monocular Depth Estimation of Planetary Landforms: A Diffusion Model Approach for Faster Inference

Nethmi Jayakody, Poorna Cooray, Supun Dassanayke, Thanuja D. Ambegoda

Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka

Please use manuscript.pdf uploaded in Supplementary material for reviewing since this docx has wrongly rendered formulae

Abstract

Monocular depth estimation presents a significant challenge due to the inherent complexity of deriving three-dimensional structures from two-dimensional imagery. Additionally, in the context of Mars satellite imagery, further challenges need to be overcome given data limitations and the substantial demand for computational resources. Our research introduces a novel Conditional Diffusion Model to address efficient depth map generation from monocular images of planetary landforms. Leveraging the advanced feature extraction capabilities of the Swin Transformer, our approach generates depth maps accurately by incorporating rich contextual information. This study not only addresses the computational and data-related challenges of traditional depth estimation methods but also significantly improves inference times, making it highly applicable to remote sensing and planetary geosciences. By presenting a scalable and efficient solution for accurate depth perception from limited single-image inputs, this work contributes to advancements in both computer vision technology and the exploration of Martian topography. Our open-source software and dataset contribution can be found at https://monogeodepth.github.io/mono-geo-depth/.

Keywords: Monocular Depth Estimation, Planetary Landforms, Diffusion Model, Deep Learning, HiRISE

1. Introduction

Depth estimation is a critical task in planetary science, particularly in remote sensing and photogrammetry applications. Accurate depth maps are essential for a range of operations, including landing site selection, terrain navigation, and detailed geological studies. The ability to derive depth information quickly from satellite images of planetary surfaces, such as those of Mars, is particularly important given the time-sensitive nature of many missions and the challenging environments involved (Epp et al., 2014).

Traditional depth estimation techniques, including stereovision, structure from motion (SfM), and LiDAR data fusion, have been extensively studied and applied. However, these methods face significant limitations when it comes to comprehensive data collection on celestial bodies. For instance, stereovision and SfM require multiple viewpoints, which are not always feasible for planetary missions. Similarly, LiDAR, while highly accurate, is often constrained by the weight and power requirements of space missions, limiting its applicability for broad surface analysis.

Monocular depth estimation (MDE) emerges as a promising alternative, capable of generating depth maps from single images while avoiding the data collection issues associated with binocular or multi-view methods. Recent architectural innovations, such as AdaBins (Bhat et al., 2021), have made significant strides in improving the accuracy of MDE. Despite these advances, there is still a gap in applying these methods to the specific context of satellite imagery. Also, one of the significant challenges in this field is the scarcity of ground-truth depth data, which is crucial for training and validating depth estimation models (Hargitai et al., 2015).

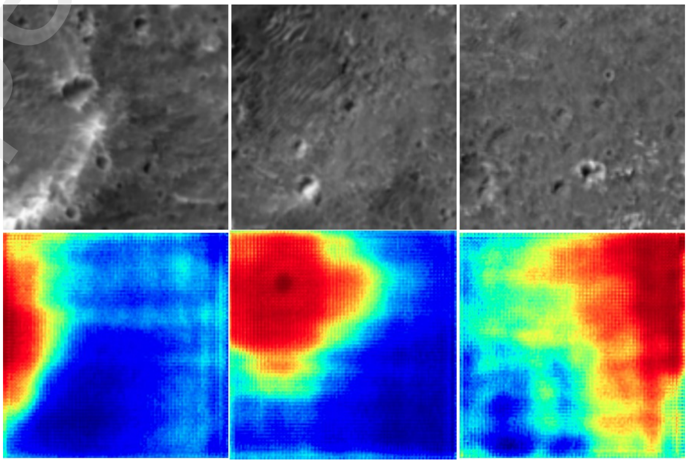


Figure 1: Illustration of MonoGEOdepth: Top: input satellite images. Bottom: depth predicted by our model

Our project focuses mainly on Mars due to its rich and simple geological diversity, availability of data, and relevance to space exploration. By leveraging the advancements in MDE, we aim to enhance the depth estimation techniques specifically for Mars satellite imagery, thereby contributing to more accurate and efficient data analysis in planetary science (Martin, 2021), (Xiao, 2023).

This research builds on the recent success of vision transformers (ViTs) (Dosovitskiy et al., 2020) and Diffusion Models (Ho et al., 2020), which have demonstrated superior performance over traditional Convolutional Neural Network (CNN) based models in feature extraction tasks (Dosovitskiy et al., 2020), (Liu et al., 2021). While approaches like the Generative Adversarial Network-based method proposed by La Grassa et al. (Grassa et al., 2022) have shown promise for monocular depth estimation from 2D HiRISE images, Diffusion Models are noted for their better noise adaptability, distribution coverage, and scalability (Dhariwal and Nichol, 2021), (A et al., 2018a), (I et al., 2020), (A et al., 2021).

The core innovation of our method lies in leveraging these advances to develop a monocular depth estimation technique specifically tailored for Mars satellite imagery, with a focus on improving inference time. By optimizing the inference process through a single-step forward and backward diffusion process, rather than an iterative process, our approach addresses the practical constraints of planetary missions, offering a more efficient solution for real-time depth estimation. This enhancement is crucial for operations that require timely data processing, such as autonomous navigation and landing site analysis (Subramanian et al., 2023).

1.1. Related Work

This paper addresses the critical need for a novel monocular depth estimation method specifically designed for analyzing Mars satellite images. In computer vision, traditional depth estimation is divided into three categories: monocular depth estimation (MDE), binocular depth estimation (BDE), and multiview depth estimation (MVDE) (Masoumian et al., 2022). Data collection feasibility issues limit the viability of approaches such as BDE, MVDE, and LiDAR data fusion for comprehensive analysis across entire planetary landscapes, despite their promising performance in a variety of applications. As a result, developing a monocular depth estimation method for satellite imagery has emerged as a critical domain at the intersection of computer vision and planetary geoscience.

Monocular depth estimation, the task of predicting depth maps from single RGB images, has seen significant advancements in recent years. Early work by Saxena et al. (2005a) adopted a supervised learning approach, collecting training data from monocular images of unstructured outdoor scenes and their corresponding ground-truth depth maps. Their model employed a discriminatively-trained Markov Random Field (MRF) that considered multiscale local and global image features. This approach demonstrated promising results in recovering accurate depth maps, even in challenging environments.

Introducing a two-step approach, Eigen et al. (2014) tackled the monocular depth estimation problem by utilizing two deep network stacks: one for coarse global predictions and another for local refinements. To mitigate the issue of scale ambiguity, they introduced a scale-invariant error metric. Eigen et al.'s method achieved state-of-the-art results on benchmark datasets without the need for super pixelation, showcasing its effectiveness in handling this challenging task.

Laina et al. (2016) proposed a fully convolutional architecture that incorporated residual learning for monocular depth estimation. Their network included a novel feature map upsampling technique and introduced the reverse Huber loss, specifically tailored for depth maps. One of the key advantages of their approach was its real-time performance on images and videos, coupled with the requirement of fewer parameters and less training data compared to the existing state-of-the-art.

In contrast to regression-based methods, Cao et al. (2017) and Fu et al. (2018) reformulated depth estimation as a pixelwise classification task. They discretized continuous ground-truth depths into bins and used fully convolutional deep residual networks for classification. This shift allowed them to estimate depth ranges and provide confidence levels for their predictions, which could be further improved using post-processing techniques such as Conditional Random Fields (CRF).

Yuan et al. (June 2022) introduced the neural window Fully Connected Conditional Random Fields (FC-CRFs) to optimize depth estimation. They leveraged the potential of fully connected CRFs by splitting the input into windows, reducing computation complexity. A multi-head attention mechanism was employed to compute potential functions, significantly improving performance on benchmark datasets.

Recent advancements in monocular depth estimation methods have demonstrated significantly improved results compared to the pre-deep learning era, as evidenced by various references focusing on CNN, deep learning, and transformer-based monocular depth estimation.

Among these methods, Yang et al. (October 2021) proposed TransDepth, an innovative architecture that combines the strengths of both CNNs and transformers. Their introduced decoder with attention mechanisms based on gates effectively captures both local and global information, achieving state-of-the-art performance on challenging datasets.

Adopting a cross-distillation approach, Shao et al. (2023) combined Transformer and CNN strengths, along with uncertainty modeling and data augmentation techniques. Their model, URCDC-Depth, outperformed previous state-of-the-art methods with no additional computational burden at inference time.

Furthermore, Bhat et al. (2021) addressed global information processing in depth estimation by introducing the AdaBins architecture, which utilizes a transformer-based approach to adaptively divide the depth range into bins and estimate bin centers for each image. This technique led to substantial improvements over the state-of-the-art on various depth datasets.

In recent years, there has been a notable increase in attempts to use generative methods for monocular depth estimation. Grassa et al. (2022) ventured into monocular depth estimation from satellite images, introducing SRDiNet, a GAN-based (Goodfellow et al., 2020), (Y et al., 2019 May), (B et al., 2021), (A et al., 2018b) solution that estimates digital terrain models (DTMs) at a higher resolution from a single image. This approach combines super-resolution and DTM estimation, enhancing fine details in the final output.

Additionally, Saxena et al. (Feb. 2023) investigated the application of Denoising Diffusion Models (Ho et al., 2020), (Song et al., 2021), (Song and Ermon., 2019), (Trippe et al., 2022), (Hoogeboom et al., 2022) to monocular depth estimation, addressing challenges arising from noisy and incomplete depth data. Their work illustrates the potential of enhancing the accuracy of depth estimation in satellite imagery, facilitating more precise mapping and analysis of terrestrial features and landscapes from space in our project.

Vision Transformers (ViTs) (Ranftl et al., 2021), (Dosovitskiy et al., 2020) have emerged as a compelling alternative to CNNs in various computer vision applications. Previous work has suggested that integrating GANs with ViTs can improve the performance of state-of-the-art methods for monocular depth estimation in both indoor and outdoor image scenarios. Lee et al. (2019) explored this integration, proposing techniques to stabilize the discriminator and modifying the generator's architecture to address the instability of GAN training with ViTs. (Gu¨ndu¨c¸, Oct. 2021) introduced Vit-Gan, a versatile architecture for image-to-image translation tasks, including semantic image segmentation and single-image depth perception. Leveraging vision transformers and Conditional GANs with a Markovian discriminator (PatchGAN), Vit-Gan aims to enhance realism in generated images. While traditional loss functions often lead to blurry results, optimization techniques like GANs focus on producing sharp and realistic outputs. However, despite these advancements, overall ViT integration with GANs does not outperform other state-of-the-art CNN-based GAN architectures.

Several of the noteworthy papers for the above, along with a few other papers, are summarized in Table 1.

1.2. Contributions

The domain of planetary depth estimation faces distinct challenges, most notably the scarcity of ground truth satellite depth data and the significant computational resources required for analysis. This study addresses these issues.

The contributions of this paper are:

- Proposing a novel approach with faster inference for monocular depth estimation with inherent adaptability of Diffusion Models and Vision Transformers to effectively mitigate the impact of noise in satellite imagery.
- Contributing to the existing largest HiRISE dataset, which consists of 679 DTMs for depth estimation, by expanding it to 983 DTMs.
- Developing an open-source application that seamlessly translates satellite images into accurate depth representations.

This paper aims to push the boundaries of monocular depth estimation in planetary imagery, thus opening the way for improved understanding and analysis of celestial bodies like Mars via innovative computer vision methodologies.

1.3. Paper structure

The remainder of the paper is organized as follows: Section 2 describes the methodology. It discusses the preprocessing steps, design decisions, details about the deep learning model, and inference time. Section 3 describes the dataset created and used and summarizes the experimental findings, including test results for model validation and comparison. It also contains the implementation details of our model. In section 4, we discuss the findings and provide context for them. Then in section 5 we provide our deliverables. Finally, in Section 7, we present a summary of the findings and draw our conclusions.

Reference	Year	Methodology
Saxena et al. (2005a)	2005	discriminatively-trained Markov Random Field (MRF)
Eigen et al. (2014)	2014	CNN
Laina et al. (2016)	2016	CNN
Cao et al. (2017)	2017	Classification using deep fully convolutional residual networks
Fu et al. (2018)	2018	Deep ordinal regression network
Gu¨ndu¨c¸ (Oct. 2021)	2019	Encoder-Decoder network
Bhat et al. (2021)	2021	Encoder-Decoder network
Yang et al. (October 2021)	2021	ViT
Yuan et al. (June 2022)	2022	CRF
Kim et al. (2022)	2022	Encoder-Decoder network
Patil et al. (2022)	2022	Combines a neural network architecture with a novel approach to predict dense plane coefficients and seed pixels
Shao et al. (2023)	2023	Both Transformer and CNN
Li et al. (March 2023)	2023	CNN-Transformer hybrid network

Table 1: Summary of Related Work

2. The MonoGEOdepth Architecture

Our project draws inspiration from the foundational Stable Diffusion architecture, as outlined in the original paper on the subject (Rombach et al., Apr. 2022). Diffusion Models are renowned for their performance in generative tasks, especially in text-to-image generation. In a Diffusion Model, there are two main processes:

1. Forward diffusion process.
2. Reverse diffusion process (Denoising process).

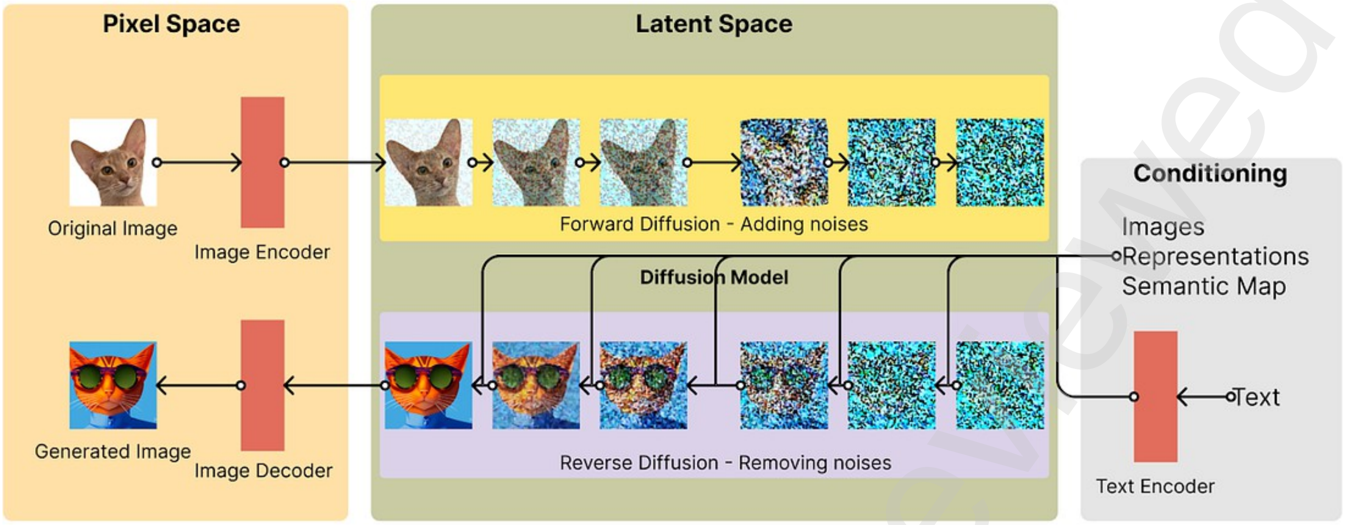


Figure 2: Architecture diagram of Stable Diffusion which is the main inspiration for our approach. The Stable Diffusion Model consists of an image encoder and an image decoder to convert between pixel space and the latent space. In the latent space, the forward diffusion process (noise addition) and the reverse diffusion process (noise reduction) happen to generate a new image. Images, representations, a semantic map, or text can be used as conditioning to remove the noise.

Furthermore, in Stable Diffusion as shown in Figure 2, these two processes occur in a lower-dimensional latent space. A Variational Autoencoder is used to convert images from pixel to latent space. By doing this, the Stable Diffusion Model can generate high-quality images with lesser computational requirements.

During the forward diffusion process, the model blurs the original image with Gaussian noise. In this process, the noise added to the image depends only on the previous image.

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

In Equation (1), x_0 is the original image, and x_t is the image with noise added after t timesteps. The way in which noise is sampled is described by the following formula. According to that, the noise is based on a specific variance t, where I is a unit matrix. t defines how much noise we need to add at each step (Ho et al., 2020).

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \tag{2}$$

In the reverse diffusion process, a U-Net neural network is trained to unblur this image, based on the conditioning provided to the Diffusion Model. Here the U-net is denoted as $u_\theta(x_t, t)$, and during this process, the network gradually denoises the image by predicting x_{t-1} using x_t.

$$p(x_{t-1}|x_t) = N(x_{t-1}; u_\theta(x_t, t), \beta_t I) \tag{3}$$

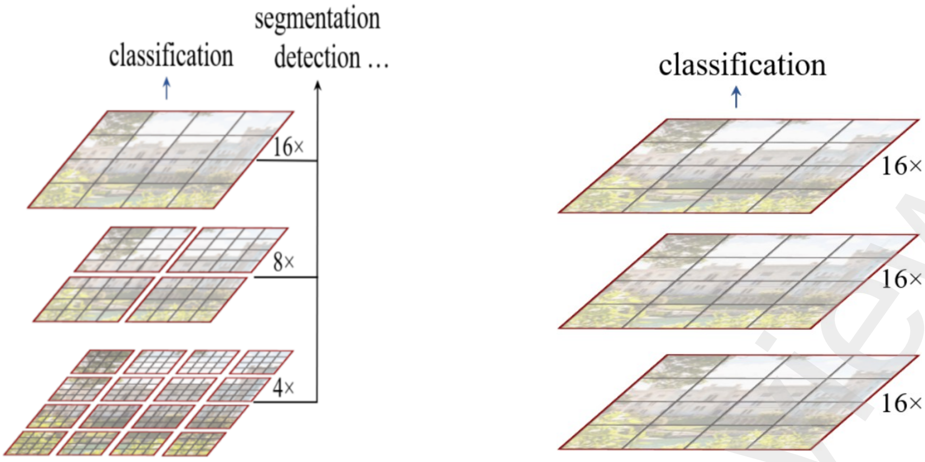
By iteratively applying the above process, the Diffusion Model can generate a noiseless image, which is a depth map in our domain.

In recent years, Diffusion Models have been used for other tasks such as gap filling of incomplete depth images (Saxena et al., Feb. 2023) and segmentation (Chen et al., A2022), (Dhariwal and Nichol, 2021). We recognized its potential to be modified as an image-to-image generation model for depth estimation tasks. However, to tailor this architecture to our specific goals, we proposed several strategic changes. Our adaptation relied heavily on the combination of conditioning, Variational Autoencoder (VAE), and A noise scheduler. As opposed to a text-to-image generation model like the stable Diffusion Model, our model is a conditional image-to-image generation model.

2.1. Data Preparation

Our dataset consists of images captured for the High-Resolution Imaging Science Experiment (HiRISE) (HiRISE Repository), (Mattson et al.) project. This project, which started in 2005, uses a powerful camera that can capture images on Mars's surface with a resolution of 0.25m per pixel. These cameras are onboard the NASA Mars Reconnaissance Orbiter, which is located 200km to 400km above the surface of Mars.

Due to the limitations of computational resources and time, we selected a set of 300 Digital Terrain Models (DTMs) for our training set. The three DTMs captured from the Oxia Planum site were used as the validation dataset. The resolution of the DTMs is 1 m pixel^{-1}, while the resolution of the associated satellite images can vary from 0.25 m pixel^{-1} to 2 m pixel^{-1}. During data preparation, we applied nearest-neighbor interpolation methods to convert the satellite images to the same resolution as the DTMs. Furthermore, since each pair of DTMs and satellite images has different dimensions, we generated 256 pixel x 256 pixel tiles from these pairs to train our model. In these DTMs, some depth values are recorded as 'NaN' due to certain anomalies. Therefore, we discard the tiles which have these 'NaN' values in their DTMs. To achieve better results, we normalized the values using a variation of min-max normalization.



(a)

(b)

Figure 3: The Swin Transformer (a) provides detailed, multi-resolution feature maps, capturing more nuanced features from the images. In contrast, the Vision Transformer (b) offers a single, less detailed feature map. This makes Swin Transformer more suitable for feature extraction, both locally and globally.

$$image_{normalized} = \frac{image - min\ val}{max\text{-}val - min\text{-}val} \times 2 - 1 \tag{4}$$

2.2. Design Choices

In this section, we discuss the major design choices we made during the project. We decided to implement our Diffusion Model architecture based on the Stable Diffusion architecture. However, to adapt the Stable Diffusion architecture, which is a text-to-image generation model, for the depth estimation task, we had to make a few modifications, including conditioning, Variational Autoencoder, and noise scheduler.

2.2.1. Condition for the Diffusion Model

We decided to utilize satellite image input for conditioning our model. For this, we had two possible design choices:

1. Use the satellite image directly as a conditioning signal (Saxena et al., Feb. 2023).
2. Use a vision transformer to extract features from the satellite image and use those features as a conditioning signal (Duan et al., 2023).

We selected the second method, which has shown better results for feature extraction in previous work (Liu et al., 2021). The main objective of using a transformer-based approach (Yuan et al., 2021), (Chu et al., 2021), (Vaswani et al., 2017) as the backbone is their innovative strategy to convert the image into a sequence of non-overlapping patches when extracting features (Liu et al., 2021) facilitating local and global feature extraction.

More specifically, we decided to use a Swin Transformer backbone as the transformer backbone instead of a Vision Transformer used by Liu et al. (2021) because Swin Transformer architecture employs a hierarchical structure of alternating stages of local and global self-attention mechanisms. This design has made the Swin Transformer outperform Vision Transformer/ DeiT (Dosovitskiy et al., 2020), (Touvron et al., 2020) and ResNe(X)t models (He et al., 2016), (Xie et al., 2017). Initially, the input image is divided into smaller patches of 4x4 pixels, which undergo multiple transformer layers to extract features at different scales. The resulting feature maps are then hierarchically aggregated across stages to generate a rich representation of the input image, which includes local and global features, in contrast to the Vision Transformer, which uses patches of 16x16 pixels and produces feature maps of single low resolution, as shown in Figure 3. Especially, the patches of 4x4 pixels in the Swin Tranformer allow the extraction of finer details, which are useful in depth estimation (Saxena et al., 2005b) than the patches of 16x16 in the Vision Transformer. All these approaches allow the Swin Transformer to effectively capture intricate patterns and relationships within the input data, making it suitable for tasks such as image understanding and feature extraction.

2.2.2. Variational Autoencoder

We are using a Variational Autoencoder in our model to convert high-dimensional images into latent space. The Diffusers library (Face, 2024) is the go-to library for state-of-the-art Diffusion Models because of the extensive community support and the modularity it provides. The default Variational Autoencoder in the Diffusers library is a relatively complex model that accepts a 3-channel input and converts it into a latent signal. However, in our case, we only need to convert single-channel 512x512 images into a 4-channel 64x64 latent signal. Therefore, we developed a simple autoencoder that has three 2D convolution layers and three transpose convolution layers. This autoencoder helped us increase accuracy and reduce computational requirements.

2.2.3. The starting point of the Denoising process

In general, the denoising process of the Diffusion Model starts with the noise-added latent signal of the original image. The amount of noise added to the original image determines

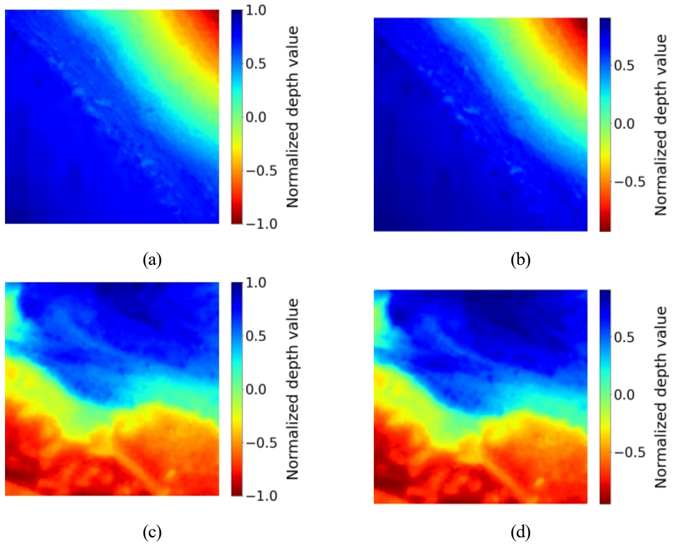


Figure 4: A comparison between the original images (a), (c), and the reconstructed images (b),(d) by our Variational Autoencoder which is used for converting pixel image into a latent signal and convert the latent signal back to a pixel image

how similar the generated image is to the original image. To generate a depth map using a Diffusion Model, we had to decide on a starting point for our model. Previous work done by Saxena et al (Saxena et al., Feb. 2023) used a noise-added, incomplete ground truth depth map as the input so that the Diffusion Model could complete the depth map while going through the denoising process. Since we didn't have such an original image, we had to decide on the signal we were going to use to start our denoising process. We experimentally checked the following starting points for our model:

1. Noise-added edge image latent signal of the satellite image
2. A random noise signal

We started with these two signals and trained our model to find which starting signal gave us better results. According to our experiment, as shown in Table 2, starting with a random noise signal outperformed starting with the noisy edge signal. Therefore, to train our model, we decided to use a random noise signal.

Starting Point	Training loss for 4 epochs
Edge images	0.41988
Random noise	**0.3842**

Table 2: MSE loss values between predicted noise from the U-Net and the actual noise for 4 epochs for 100 DTMs in the training process.

2.2.4. *Inference Time steps*

The reverse diffusion process of Stable Diffusion Model is an iterative process. As the number of iterations increases, both the refinement of the generated image and the inference time increase. Since one of our main goals is to develop an efficient model for depth estimation, we have decided to generate the depth image in a single iteration. This approach enables real-time depth estimation with entry-level GPUs that have a minimum of 6GB memory.

Method	Avg inference time in sec (CPU)	Avg PSNR	Avg SSIM
ESPCN (Shi, 2016)	**0.008795**	32.7059	0.9276
EDSR (Lee, 2017)	5.923450	**34.1300**	**0.9447**
FSRCNN (Tang, 2016)	0.021741	32.2681	0.9248
Bicubic	0.000208	32.1638	0.9305
Nearest neighbor	0.000114	29.1665	0.9049
Lanczos	0.001094	32.4687	0.9327

Table 3: Image super-resolution benchmarking results for the General-100 dataset are presented (Tang, 2016). PSNR, which stands for Peak Signal to Noise Ratio, and SSIM, which stands for Structural Similarity Index Measure, are used as metrics. The higher the PSNR and SSIM values, the better the results.

2.3. *Training Architecture*

Training the U-Net to accurately predict the noise in a latent signal of an image (in our case a depth map) is the primary goal of the training process. As in Figure 5, first the Variational Autoencoder will receive the ground truth tiles and convert them to a latent signal after using the FSRCNN model for resolution upscaling. According to previous works (Grassa et al., 2022), we can obtain better depth prediction results by upscaling resolution due to detail sharpening. The reason for selecting the FSRCNN model to upscale resolution is that, according to 3, the FSRCNN model shows high PSNR and SSIM values with a lower inference time. Therefore, we can maintain both the accuracy and efficiency of our architecture. While we maintain a constant random noise signal as the Z_t state of the diffusion process, the converted latent signal of the ground truth tiles becomes the Z_0 state. We define the the actual noise we should predict during the denoising process as:

$$\text{actual noise} = Z_0 - Z_t \tag{5}$$

In this case, t stands for a constant timestep that we define, which is 1000. Next, we feed the Z_t state random noise signal, the constant timestep, and the features that were extracted from the input grayscale tile using the Swin Transformer to the U-Net in order to predict noise. The goal of this denoising process is to obtain the Z_0 state, which is obtained by removing this predicted noise from Z_t. We define the loss function for the U-Net as:

$$\text{UNet loss} = \text{MSE}(\text{actual noise}, \text{predicted noise}) \tag{6}$$

where MSE is the mean square error. To reduce the loss between the predicted noise from the U-Net and the actual noise (Equation (4)), the U-Net will be trained for a number of iterations in this manner.

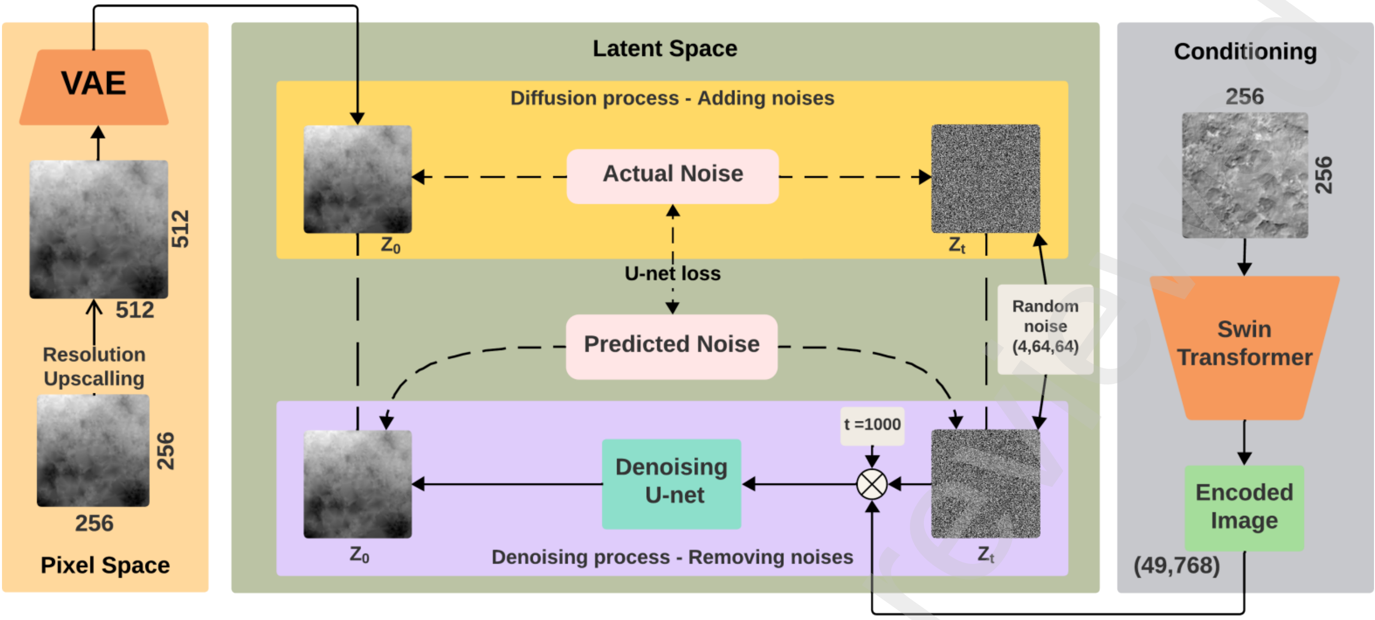


Figure 5: MonoGEODepth training architecture: During the training, a satellite image is being used as the conditioning. We start our diffusion process with the latent signal corresponding to the ground truth depth image, and use the Variational Autoencoder (VAE) we developed to convert images between pixel space and latent space. With the conditioning, we provide a constant timestep value to U-Net to generate the depth image during the denoising process.

2.4. *Inference Architecture*

A constant random noise signal, represented in Figure 6 as Z_t, will serve as the starting point during the inference time. Here, we use the same constant noise signal that we used during the training process. Consequently, the Z_t signal and the constant timestep containing the features that were extracted from the input image are used by the denoising U-Net to predict the noise. The Z_0 signal, which is the latent form of the input image's depth map, is then obtained by subtracting the predicted noise from the Z_t signal. Ultimately, the Z_0 signal will be fed into the variational auto encoder's decoder and mapped into the pixel space. The final result will be an upscaled 512x512 depth map of the original picture.

3. Experiments and Results

We first trained our model for 10 epochs with 100 DTMs and further trained it for another 20 epochs with 300 DTMs. The loss curves for both the training and validation process are plotted in Figure 7. We conducted a series of experiments to evaluate our model's accuracy in estimating monocular depth from satellite images. In this section, we first discuss the dataset and the evaluation metrics we used, followed by the implementation details of our model. We then compare the results from our model with the ground truth values. Furthermore, we benchmark our model against the state-of-the-art methods for depth estimation from satellite images (Grassa et al., 2022).

3.1. *Datasets and evaluation metrics*

HiRISE DTMs are digital terrain models made for the Mars surface. These models indicate the elevation of a particular point using the value of that data point. DTMs are created using stereo-matching techniques with two images taken from different angles (HiRISE Repository). the pixel resolution of HiRISE images varies from 0.25m/pixel to 0.5m/pixel. The terrain models can be derived at a post spacing approximately 4 times the pixel scale of the input images, resulting in post spacings of 1m - 2m for DTMs. HiRISE images and DTMs are available on the HiRISE website of the University of Arizona (HiRISE Repository).

We created a dataset consisting of 983 stereo pairs and DTMs using this HiRISE archive. Currently, it is the largest HiRISE DTM dataset available for open access. However, we only use 300DTMs to train our model. To account for the different resolutions of HiRISE images, we divided them into 256x256 tiles for training and validation of our model. Our model takes the left stereo image as input and considers the relative depth values obtained from DTMs as the ground truth depth values for training.

For the evaluation metrics, we use the Root Mean Squared Error (RMSE)(Equation 7) and Absolute Relative Error (Equation 8) for the HiRISE dataset. These are the metrics used in the prior work done by Grassa et al. (2022). In addition to those metrics, we also use Squared Relative Error (Equation 9), RMSE-Log Error (Equation 10), and Delta Error (Equation 11) to evaluate the results of our model. These five metrics are standard evaluation metrics used in prior work. (Eigen et al., 2014)

$$\mathrm{RMSE} = \sqrt{\frac{1}{|D|}\sum_{p\in D}(g-p)^2} \qquad (7)$$

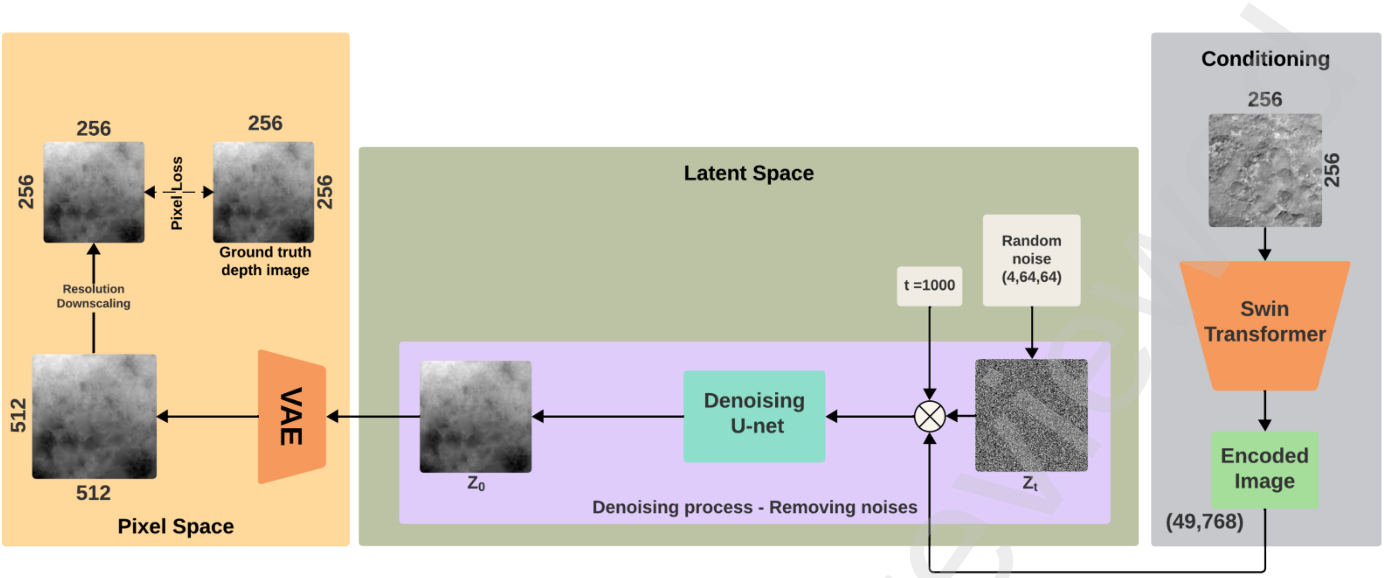


Figure 6: MonoGeoDepth inference architecture: During the inference time, we used the trained denoising U-Net to generate the depth image in latent space based on the satellite image conditioning. VAE is used to convert the latent signal into a depth image.

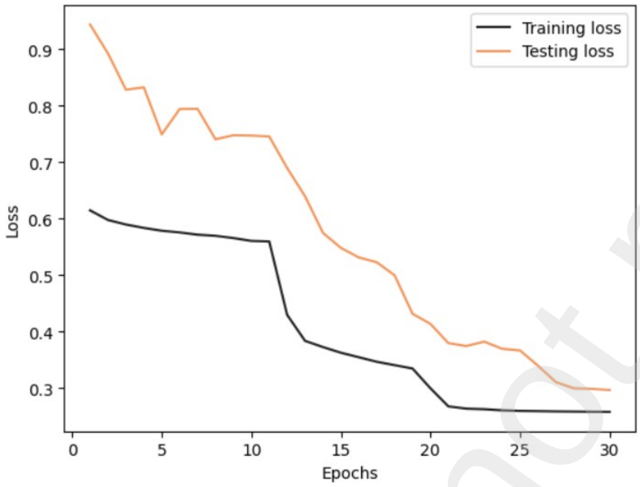


Figure 7: RMSE loss values of training and validation epochs. We trained the model for 30 epochs using 300 DTMs. The training loss line plots the RMSE losses between the predicted noise from U-Net and the actual noise while the testing loss line plots the RMSE losses between the predicted depth image and ground truth depth image in each epoch.

$$\text{Abs-Rel} = \frac{1}{|D|} \times_{p\in D} \frac{|g-p|}{g} \tag{8}$$

$$\text{Sq-Rel} = \frac{1}{|D|} \times_{p\in D} \frac{(g-p)^{2}}{g} \tag{9}$$

$$\text{RMSE-Log} = \sqrt{\frac{1}{|D|} \times_{p\in D} (\log(g) - \log(p))^{2}} \tag{10}$$

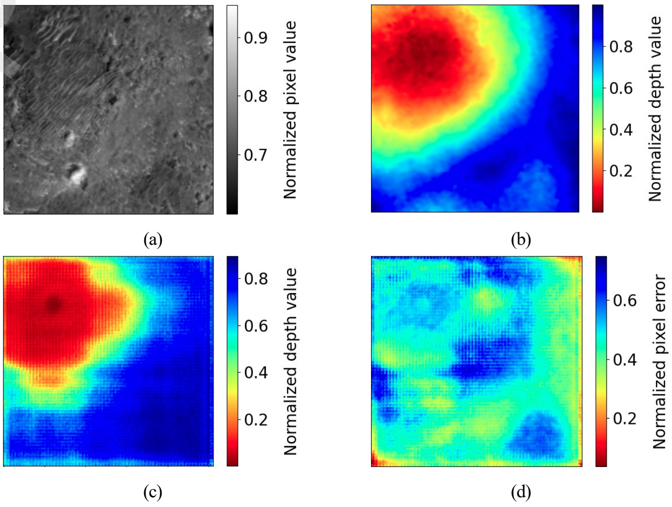
$$\delta t = \frac{1}{|D|} \times_{p\in D} \text{"} \max \text{!} \frac{g}{p}, \frac{p}{g} \text{\#} < 1.25^{t} \times 100\% \tag{11}$$

Where:

p - predicted depth

g - ground truth

D - set of all predicted depth values



(a)

(b)

(c)

(d)

Figure 8: Comparison of (a) input satellite image, (b) ground truth depth image, (c) generated depth image by our model, and the (d) pixel difference between ground truth and predicted depth images

3.2. Implementation details

We use Python with the PyTorch library to train our model. Our batch size is 8, and we perform 2 steps for gradient accumulation. The training process runs for 30 epochs, utilizing the Adam optimizer. The learning rate for training is 5 × e^{-4}.

Our model is trained using a single RTX 3090 24GB GPU. A single epoch takes 9 hours to train with 300DTMs and for inferencing 500 patches of 256x256 pixels, it takes 35 seconds with a rate of 14 tiles per second.

3.3. Comparison of results with the ground truth values

Figure 8 shows a side-by-side comparison between the input satellite image, ground truth depth image, predicted depth image, and the difference between predicted and depth images in the pixel space. The red areas in Figure 8b and Figure 8c correspond to areas with relatively less elevation compared to the areas which are plotted in blue. These depth images correspond to the satellite images captured from the Oxia Planum site. According to the images, we can see that even though there is a difference between the predicted and ground truth depth images (Figure 8d, the model has predicted the relative depths successfully. Figure 9 mirrors the above-mentioned idea since we can observe less error in the relative difference between the values in latent space (Figure 9e, even though the values in the ground truth image and the predicted image (Figure 9f) have considerable differences.

3.4. Comparison to the state-of-the-art

We have selected the SRDiNet (Grassa et al., 2022) model as the most comparable previous work for our project. For the HiRISE dataset, SRDiNet is the only previous work that uses monocular depth estimation techniques. Therefore, we consider the SRDiNet model as the state-of-the-art for depth estimation for HiRISE satellite images.

Method	RMSE	Abs-Rel	Inference time (s)
SRDiNet (Grassa et al., 2022)	**0.2011**	**0.1697**	218
Ours	0.2976	0.2288	**35s**

Table 4: Comparison between models trained on HiRISE dataset. We used Oxia Planum site HiRISE images as the validation dataset, which was also used to validate the SRDiNET model. To achieve the given inference time of 218, SRDiNET used 4 RTX 5000 GPUs while our model used a single RTX 3090 GPU. SRDiNET model shows better RMSE and Abs-Rel values compared to our model. However, our model predicts depth image in significantly less time.

4. Discussion

One of the significant challenges in planetary landform analysis is the scarcity of ground-truth depth values (Hargitai et al., 2015). Training a Diffusion Model also requires a large amount of data. In our project, we overcame these problems by using a pre-trained model (Face, 2024) for text-to-image generation and then fine-tuning the model to generate depth maps using satellite images. Furthermore, we increased the dataset size by dividing the images into 256 pixel x 256 pixel tiles and applying augmentation techniques. This approach allowed us to generate a more robust training dataset, improving our model's performance and generalization capabilities.

We have performed experiments to compare the performance of our model with the state-of-the-art model (Grassa et al., 2022) using the Oxia Planum site HiRISE images as the validation dataset. The main objectives of our model are (a) to generate depth maps with faster inference times (b) to explore the benefits of the denoising capability of Diffusion Models and ViT for monocular depth estimation. As a result, we have developed a model that demonstrates superior performance in time-sensitive scenarios (Table 4). For processing a satellite image with 500 tiles, our model requires only approximately 35 seconds using a single Nvidia RTX 3090 GPU, compared to 218 seconds for the SRDiNet model with 4 Nvidia RTX 5000 GPUs. This equates to an impressive processing rate of approximately 14 tiles per second per GPU. Consequently, our model is adept at estimating depth in real-time for small areas, even under stringent time constraints, while demanding fewer computing resources.

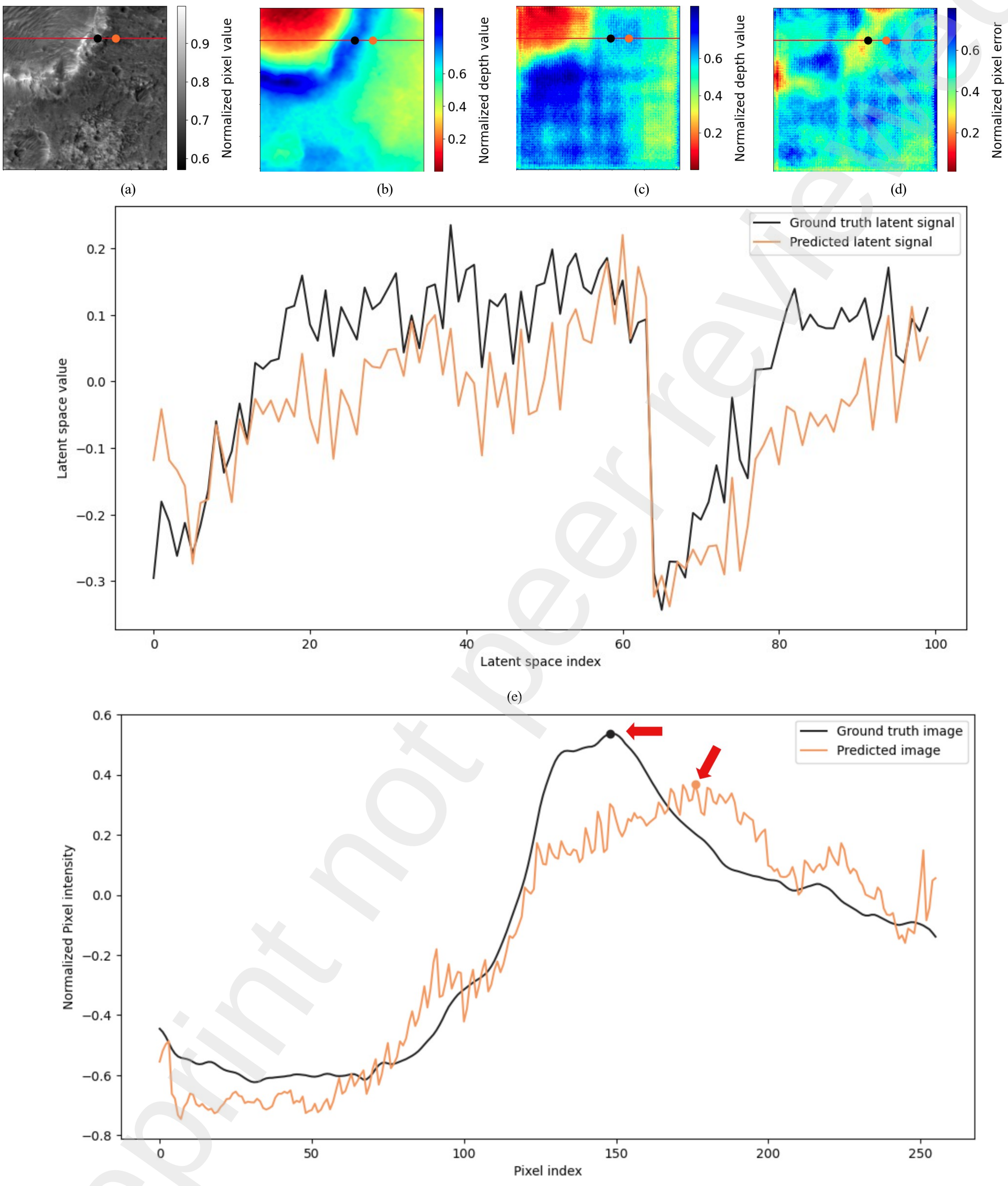
Even though the model performs well in achieving faster inference time, a small loss can be observed between the predicted depth image and the ground truth depth image, as plotted in Table 9f. Compared to Figure 9f, Figure 9e shows much fewer errors in the difference between the ground truth values and the predicted values. This difference can be mitigated through a more accurate Variational Autoencoder. In addition, since we normalize absolute depth values in ground truth DTMs into the [-1,1] scale in the preprocessing stage, our current model can estimate depth in pixel intensities. To estimate relative depth, we have to retrain the model without dividing it by the depth range in the preprocessing stage.

Furthermore, as shown in Table 2, contrary to intuitive assumptions, initializing the denoising process with random noise instead of edge images leads to superior denoising outcomes. This finding suggests that the random noise signal serves as a more effective starting point for the denoising process. Additionally, although traditional Conditional Diffusion Models utilize noise schedulers to control the addition and removal of noise, our architecture omits this component for the sake of time efficiency. However, future modifications could incorporate a noise scheduler, leveraging increased GPU power to enhance performance further.

5. Dataset Contribution and Source Code

As a contribution of this paper, we have created the largest Digital Terrain Model (DTM) dataset of Mars, extending the previous work done by (Grassa et al., 2022). This dataset includes a stereo pair of satellite images and the respective digital terrain models of those images. We collected 983 stereo pairs and DTMs from the HiRISE archive for this dataset, compared to the 679 stereo pairs and DTMs in the (Grassa et al., 2022). This dataset will be useful for training any machine learning model in its domain.

Furthermore, we have developed an open-source visualization tool to convert a given satellite image into a depth image. This tool can be extended as a real-time depth estimation tool, considering its lower inference time. Our dataset and the



(a) (b) (c) (d)

(e)

(f)

Figure 9: Difference between predicted and ground truth values. (a) Input satellite image, (b) ground truth depth image, (c) generated depth image by our model, (d) pixel difference between ground truth and predicted depth images, (e) predicted and ground truth values of a selected part of the latent signal, and (f) predicted and ground truth depth values along the red color line in the image (a). The blue color dots in images (a) - (d), and (f) represent the highest points of the ground truth depth image, while the yellow color dots in images (a) - (d), and (f) represent the highest points of the predicted depth image.

source code for our application are available at the following link: https://monogeodepth.github.io/mono-geo-depth/.

6. Limitation of the Study

Our study is limited by the use of a single-step Diffusion Model for monocular depth estimation, meaning that only one iteration is performed in both the forward and backward diffusion processes. This approach prioritizes low inference time, essential for time-critical scenarios like real-time Mars satellite image analysis. However, this results in comparatively lower accuracy. While increasing the steps in the diffusion process could improve accuracy, it would also significantly increase inference time, conflicting with our focus on efficiency. Future research should explore methods to better balance speed and accuracy.

7. Summary and Conclusion

We introduce the MonoGEOdepth model, a novel approach that addresses the challenge of generating high-quality depth maps from single satellite images of planetary landforms. Leveraging advanced feature extraction capabilities through the Swin Transformer, coupled with a Conditional Diffusion Model, our method significantly enhances depth estimation efficiency while overcoming the limitations of traditional depth estimation methods on celestial bodies like Mars. By addressing computational and data-related challenges and improving inference times, our model demonstrates the potential for applications in remote sensing and planetary geosciences. Experimental results indicate that the model performs better in time-sensitive scenarios, and further training of the U-Net and Variational Autoencoder can improve the performance even further. Through scalable and efficient depth perception from limited single-image inputs, our work contributes to advancements in both computer vision technology and the exploration of Martian topography, offering valuable insights for tasks ranging from landing site selection to detailed geological studies.

CRediT authorship contribution statement

Nethmi Jayakody: conceptualization, data curation, investigation, methodology, software, validation, writing – original draft, writing – review & editing. **Supun Dasanayake**: conceptualization, data curation, investigation, methodology, software, validation, writing – original draft, writing – review & editing. **Poorna Cooray**: conceptualization, investigation, project administration, resources, writing – review & editing. **Thanuja D. Ambegoda**: conceptualization, writing – original draft, writing – review & editing, supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

A, A., M, M., G, B., 2021. Generative adversarial network: An overview of theory and applications. International Journal of Information Management Data Insights .

A, C., T, W., V, D., K, A., B, S., AA., B., 2018a. Generative adversarial networks. An overview. IEEE signal processing magazine .

A, C.K., SM, B., M., P., 2018b. Monocular depth prediction using generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops .

B, H., JQ, Z., A, N., Tuch D, V.K., S, G., DS, E., 2021. Self-supervised generative adversarial network for depth estimation in laparoscopic images. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France .

Bhat, Farooq, S., Alhashim, I., Wonka, P., 2021. Adabins: Depth estimation using adaptive bins. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 4009–4018.

Cao, Y., Wu, Z., Shen, C., 2017. Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Transactions on Circuits and Systems for Video Technology 28, 3174–3182.

Chen, T., Li, L., Saxena, S., Hinton, G., Fleet, D.J., A2022. A generalist framework for panoptic segmentation of images and videos. arXiv preprint arXiv:2210.06366 .

Chu, X., Zhang, B., Tian, Z., Wei, X., Xia, H., 2021. Do we really need explicit position encodings for vision transformers? arXiv preprint arXiv:2102.10882 .

Dhariwal, P., Nichol, A., 2021. Diffusion models beat gans on image synthesis. Advances in neural information processing systems (NIPS/NeurIPS) , 8780–8794.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., vain Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Duan, Yiqun, Guo, X., Zhu, Z., 2023. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. arXiv preprint arXiv:2303.05021 .

Eigen, D., Puhrsch, C., Fergus, I., 2014. Depth map prediction from a single image using a multi-scale deep network. In Advances in Neural Information Processing Systems 14, 2366–2374.

Epp, C., Robertson, E., Carson, J., 2014. Real-time hazard detection and avoidance demonstration for a planetary lander. In Proceedings of the AIAA SPACE conference and exposition .

Face, H., 2024. Diffusers. URL: https://github.com/huggingface/diffusers. gitHub repository.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D., 2018. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , 2002–2011.

Goodfellow, Ian, et al., 2020. Generative adversarial networks. Communications of the ACM 63.11 , 139–144.

Grassa, L., Riccardo, 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.

Gu¨ndu¨c¸, Y., Oct. 2021. Vit-gan: Image-to-image translation with vision transformes and conditional gans. arXiv preprint arXiv:2110.09305 .

Hargitai, Henrik, A´ kos Kereszturi (Eds.), 2015. Encyclopedia of Planetary Landforms. Springer, New York.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition , 770–778.

HiRISE Repository, . Hirise repository. https://www.uahirise.org/dtm/. Online; accessed 20 April 2023.

Ho, J., Jain, A., , Abbeel, P., 2020. Denoising diffusion probabilistic models. Advances in neural information processing systems (NIPS/NeurIPS) , 6840–6851.

Hoogeboom, E., Satorras, V.G., Vignac, C., Welling, M., 2022. Equivariant diffusion for molecule generation in 3d. arXiv e-prints .

I, G., J, P.A., M, M., B, X., D, W.F., S, O., A., C., Y., B., 2020. Generative adversarial networks. Communications of the ACM .

Kim, Ga, D., Ahn, W., Joo, P., Chun, D., Kim, S., J., 2022. Global-local path networks for monocular depth estimation with vertical cutdepth. arXiv 2022, arXiv:2201.07436 .

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks. In 2016 Fourth International Conference on 3D Vision , 239–248.

Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H., 2019. From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 .

Lee, L.B..S.S..H.K..S.N..K.M., 2017. Enhanced deep residual networks for single image super-resolution. Proceedings of the IEEE conference on computer vision and pattern recognition workshops , 136–144.

Li, Z., Chen, Z., Liu, X., Jiang, J., March 2023. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. Machine Intelligence Research .

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision , 10012–10022.

Martin, N., 2021. Mars settlement likely by 2050, says unsw expert, but not at levels predicted by elon musk. https://www.unsw.edu.au/newsroom/news/2021/03/mars-settlemen\protect\@normalcr\ relax-likely-by-2050-says-unsw-expert-but-not-at-levels\protect\@normalcr\relax-predicted-by-elon-musk. Online; ac- cessed 27 May 2024.

Masoumian, Armin, Rashwan, H.A., Cristiano, J., Asif, M.S., Puig., D., 2022. Monocular depth estimation using deep learning: A review. Sensors 22 .

Mattson, Kirk, S.., Heyd, R.., McEwen, R.., Eliason, A.., Hare, E.., Beyer, T.., Howington-Kraus, R.., Okubo, E.., Herkenhoff, C.., K., . Release of hirise digital terrain models to the planetary data system. Lunar Planet .

Patil, V., Sakaridis, C., Liniger, A., Gool, L.V., 2022. P3depth: Monocular depth estimation with a piecewise planarity prior. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition .

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

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., Apr. 2022. High-resolution image synthesis with latent diffusion models. arXiv:2112.10752v2 .

Saxena, A., Chung, S.H., Ng, A.Y., 2005a. Learning depth from single monocular images. In Advances in Neural Information Processing Systems 18, 1–8.

Saxena, A., Chung, S.H., Ng, A.Y., 2005b. Learning depth from single monocular images. Advances in neural information processing systems .

Saxena, S., Kar, A., Norouzi, M., Fleet, D.J., Feb. 2023. Monocular depth estimation using diffusion models. arXiv:2302.14816v1 .

Shao, S., Pei, Z., Chen, W., Li, R., Liu, Z., Li, Z., 2023. Urcdc-depth: Uncertainty rectified cross-distillation with cutflip for monocular depth estimation. arXiv preprint arXiv:2302.08149 .

Shi, W..J.C..F.H..J.T..A.P.A..R.B..D.R..Z.W., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Proceedings of the IEEE conference on computer vision and pattern recognition , 1874–1883.

Song, J., Meng, C., Ermon, S., 2021. Denoising diffusion implicit models. In International Conference on Learning Representations (ICLR) .

Song, Y., Ermon., S., 2019. Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems (NIPS/NeurIPS) .

Subramanian, J., Asirvadam, V., Zulkifli, S., Singh, N.S., Shanthi, N., Lagisetty, R., 2023. Target localization for autonomous landing site detection: A review and preliminary result with static image photogrammetry. Drones .

Tang, D.C..C.C.L..X., 2016. Accelerating the super-resolution convolutional neural network. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands , 391–407.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Je´gou, H., 2020. Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877 .

Trippe, B.L., Yim, J., Tischer, D., Broderick, T., Baker, D., Barzilay, R., Jaakkola, T., 2022. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. arXiv preprint arXiv:2206.04119 .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Łukasz Kaiser, Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems (NIPS/NeurIPS) .

Xiao, L., 2023. Evolution of the geological environment and exploration for life on mars. Journal of Earth Science 34, 1626–1628. doi:10.1007/s12583-023-1929-7.

Xie, S., Girshick, R., Dolla´r, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , 1492–1500.

Y, A., MR, S., PP, D.G., A, M., N., T., 2019 May. Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. In 2019 International conference on robotics and automation (ICRA) .

Yang, G., Tang, H., Ding, M., Sebe, N., Ricci, E., October 2021. Transformer-based attention networks for continuous pixel-wise prediction. In Proceedings of the IEEE International Conference on Computer Vision , 16269–16279.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Tay, F.E., Feng, J., Yan, S., 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. arXiv preprint arXiv:2101.11986 .

Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P., June 2022. Neural window fully-connected crfs for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , 3916–3925.