Earthbender: An Interactive System for Stylistic Heightmap Generation using a Guided Diffusion Model

Danial Barazandeh
Computer Graphics and Virtual Reality Research Lab
University of Bremen
Bremen, Germany
barazand@uni-bremen.de

Gabriel Zachmann
Computer Graphics and Virtual Reality Research Lab
University of Bremen
Bremen, Germany
zach@cs.uni-bremen.de

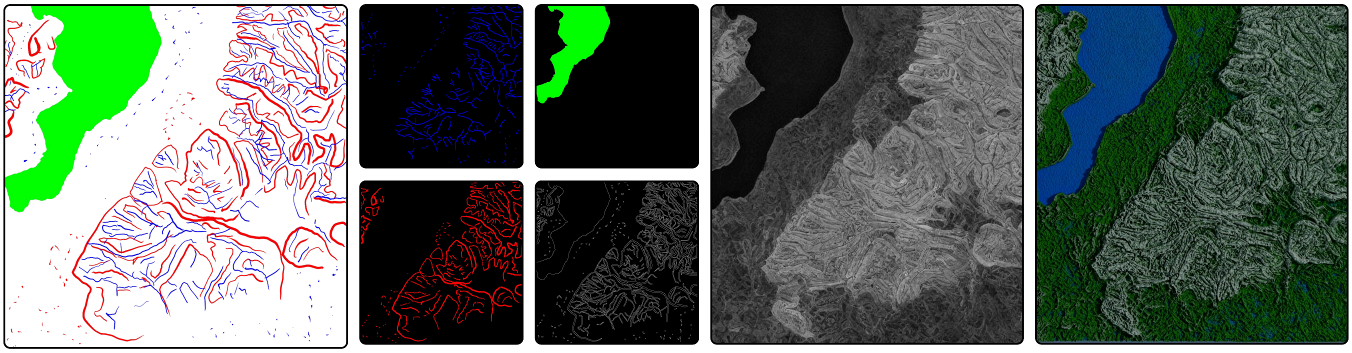


Figure 1: The Earthbender workflow: transforming a simple 2D sketch into a detailed heightmap. The final 3D terrain (right side) is rendered in Blender

Abstract

Games, 3D simulations, and cinematic pipelines depend on realistic 3D terrain for immersion. However, creating detailed 3D terrain is labour-intensive: artists sculpt elevation, iterate on mountains, rivers, lakes, and must often repeat the entire workflow when the design changes. Recent generative approaches are attempting to address this challenge, but they primarily focus on a single landform (typically mountains) and overlook structural features, such as river networks, roads, or lakes.

We propose a sketch-conditioned diffusion framework that generates depth maps representing complete landscapes, including mountains, river networks, and lakes. Our method extends Stable Diffusion with a ControlNet branch that takes multiple channel inputs: Canny edges for overall structure, red for mountains, green for lakes, and blue as a carving tool for painting roads and rivers onto the heightmap.

This approach addresses the technical challenges while prioritizing the artist's creative control. Our interactive system, Earthbender, gives the artist fine-grained control over every detail in the heightmap, demonstrating a collaborative model where the generative AI acts as a powerful assistant to achieve an artistic vision, rather than replacing the artist's creativity.

Our experiments show that our ControlNet-based approach significantly outperforms traditional GANs in both data efficiency and output quality. Furthermore, we present an analysis demonstrating that the choice of loss function acts as a powerful artistic control, allowing the user to select between a sharp, detailed style and a softer, more organic output better suited for downstream game engine workflows.

CCS Concepts

• **Human-centered computing → Graphical user interfaces**; • **Computing methodologies → Shape modeling**; **Generative and developmental approaches**.

Keywords

Interactive Systems, Sketch-Based Modeling, Terrain Generation, ControlNet, Generative Models, Diffusion Models

ACM Reference Format:
Danial Barazandeh and Gabriel Zachmann. 2025. Earthbender: An Interactive System for Stylistic Heightmap Generation using a Guided Diffusion Model. In *The 18th ACM SIGGRAPH Conference on Motion, Interaction, and Games (MIG '25), December 03–05, 2025, Zurich, Switzerland.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3769047.3769053

1 Introduction

In game development, 3D simulations, and cinematic pipelines, creating vast and believable environments is crucial for achieving

MIG '25, Zurich, Switzerland

ACM ISBN 979-8-4007-2236-3/25/12
https://doi.org/10.1145/3769047.3769053

an immersive experience. A fundamental component of these environments is the terrain, which is often represented by a large 3D mesh. However, the process of authoring high-quality terrain remains a significant bottleneck. Artists, level designers, and technical artists must spend hours manually sculpting digital surfaces or using procedural tools that often require indirect manipulation of complex parameters. Furthermore, any significant change in the design often requires repeating the entire laborious workflow. Using a heightmap to generate terrain can yield high-quality results. Still, the source of heightmaps is usually limited to satellite data, and altering the heightmaps is as time-consuming as sculpting the mesh itself.

To address this, researchers have explored the use of AI generative models to create heightmaps. While early approaches using Generative Adversarial Networks (GANs) [Goodfellow et al. 2020] demonstrated the potential for generating realistic textures, they often provide insufficient user control over the final output. They usually require large, specialized datasets to train effectively. The trade-off has consistently been between the quality of the generated heightmap [Voulgaris et al. 2021] and the artist's ability to guide the creation process. Even in newer models based on Diffusion [Ho et al. 2020] models, there is still a lack of control that can compromise artistic freedom [Löchner et al. 2023].

In this work, we argue for a different paradigm: generative AI as a collaborative tool that enhances, rather than replaces, human creativity. We present Earthbender, a novel interactive system for generating high-quality heightmaps through a direct, sketch-based workflow. Our approach is built on a diffusion framework that extends a pre-trained Stable Diffusion model with a ControlNet [Zhang et al. 2023] branch. This ControlNet is conditioned on a multi-channel semantic sketch created by the artist, where Canny edges provide the overall structure, and specific colours directly map to geographical features: red indicates regions of positive elevation displacement (mountains), blue marks carved depressions (rivers and roads), and green defines planar areas (lakes).

Our system is designed from the ground up to prioritize artistic control. It features a complete GUI that includes not only the core drawing tools but also a suite of real-time post-processing controls for fine-tuning the output. Through a comprehensive set of experiments, we analyze the effectiveness of our approach and present our findings. The primary contributions of this paper are:

- A novel, interactive system ("Earthbender") for the direct, sketch-based authoring of multi-feature terrain heightmaps using a guided diffusion model.
- A comparative study demonstrating that our ControlNet-based approach significantly outperforms a traditional GAN architecture (Pix2PixHD) [Wang et al. 2018] in both data efficiency and the structural fidelity of the generated output.
- An analysis of loss functions reveals that the choice of loss function acts as a robust artistic control, allowing the user to select between a sharp, detailed style and a softer, more organic output better suited for downstream game engine workflows.
- The results of a qualitative evaluation study that validates the usability and creative utility of our system in a practical, artist-centric workflow.

2 Related Work

2.1 Image-to-Image Translation

The field of conditional image-to-image translation was significantly advanced by the pix2pix framework [Isola et al. 2017]. This approach introduced a general-purpose solution using a Conditional Generative Adversarial Network (cGAN) [Mirza and Osindero 2014] capable of learning the mapping between various visual domains. Its key architectural innovations were a U-Net based generator [Ronneberger et al. 2015], which uses skip connections to pass low-level image information directly from the encoder to the decoder, and a PatchGAN discriminator that focuses on preserving high-frequency details by classifying local image patches. This combination proved highly effective for a wide range of tasks, such as translating sketches to photographs or satellite imagery to maps. However, while foundational, the original pix2pix architecture often struggled to produce high-quality results at higher resolutions, a challenge that was directly addressed by its successor.

A foundational approach in high-resolution, conditional image synthesis is the work of [Wang et al. 2018], often known as pix2pixHD. To overcome the instability of training GANs on high-resolution images, they introduced a novel coarse-to-fine generator and a multi-scale discriminator architecture. This, combined with a feature-matching loss, allowed for the generation of photorealistic images from semantic label maps at unprecedented resolutions. Their work established a powerful baseline for GAN-based, image-to-image translation and demonstrated a high degree of visual quality. However, the system's primary input is a dense semantic map, and it does not focus on the sparse, multi-channel, and artist-driven sketch-based workflow that we explore in our system.

A landmark in interactive image synthesis is GauGAN [Park et al. 2019], powered by the Spatially-Adaptive Denormalization (SPADE) architecture. This system demonstrated state-of-the-art results in transforming semantic layouts into stunningly photorealistic landscapes, offering users a high degree of control. However, its core interaction paradigm is fundamentally different from our approach. GauGAN operates on dense semantic segmentation maps, where every pixel in the input is assigned a specific class like "sky," "mountain," or "water." While incredibly powerful, this method is not directly comparable to our work. Our system, Earthbender, is explicitly designed to interpret sparse, multi-channel artistic sketches, which more closely mimic a traditional drawing workflow. Therefore, we do not include a quantitative comparison, as the vastly different input modalities would make such an analysis inequitable and unenlightening. Instead, we position our work as exploring a complementary, sketch-driven approach to terrain authoring.

Another powerful, artist-centric approach was presented by [Perche et al. 2023], who developed a system for terrain authoring based on the StyleGAN2 architecture. Their key contribution is the concept of "spatialised style," where they use a sophisticated encoding process to allow artists to mix the styles of different terrains in specific, user-defined regions. While their work also focuses on providing a suite of interactive tools, their underlying technical approach is fundamentally different from ours. Their system is built on a GAN that is trained from scratch and is centered on the manipulation of style in a latent space. In contrast, our work leverages the powerful generative prior of a large, pre-trained diffusion model and focuses on direct, semantic control of geological features through a multi-channel sketch, representing a complementary philosophy for interactive terrain generation.

It is also important to differentiate our work from the extensive research in neural style transfer, which, while related to image synthesis, addresses a fundamentally different problem. State-of-the-art methods have evolved significantly from the original optimization-based approach, with techniques like Adaptive Instance Normalization (AdaIN) [Huang 2017] enabling real-time, arbitrary style transfer. More recent transformer-based models like StyleFormer [Park and Kim 2022] and Stytr^2 [Deng et al. 2022] have pushed the boundaries even further, offering enhanced control over the stylization process and producing remarkably high-quality artistic images. While these methods are powerful and might seem applicable, they are fundamentally designed to transfer textural and color properties from a style image onto the global structure of a content image. They are not suited for our task, which requires the model to interpret sparse semantic inputs from a sketch and generate new, corresponding geometric structures (a heightmap), rather than simply re-texturing an existing one.

2.2 Diffusion Models

Our work is built upon the foundation of Denoising Diffusion Probabilistic Models (DDPMs) [Ho et al. 2020], a class of generative models that have recently demonstrated state-of-the-art image synthesis quality. The core mechanism of a DDPM involves a two-stage process: a fixed forward process that incrementally adds Gaussian noise to an image over a series of timesteps until it becomes pure noise, and a learned reverse process that iteratively denoises the image to reconstruct a clean sample. The key insight presented by [Ho et al. 2020] is that this reverse process can be effectively modeled by training a neural network—typically a U-Net—on a simplified objective: predicting the noise that was added at each timestep. This approach not only proved capable of generating images with higher fidelity than many contemporary GANs but also offered a more stable and straightforward training regime. We chose this architecture as the basis for our system due to its proven ability to generate high-quality, detailed outputs and its inherent structure, where the step-by-step denoising process provides a robust framework for injecting the strong spatial conditioning required for our sketch-based control.

While the aforementioned Denoising Diffusion Probabilistic Models (DDPMs) produce high-quality results, they are computationally intensive as they operate directly in the high-dimensional space of pixels. A significant breakthrough in addressing this limitation came with the introduction of Latent Diffusion Models (LDMs) [Rombach et al. 2022], the core technology behind the popular Stable Diffusion model. The key insight of LDMs is to perform the iterative denoising process not in pixel space, but in a much smaller, perceptually-equivalent latent space learned by a powerful autoencoder. By first compressing the image into this lower-dimensional space, the model can learn the primary semantic and conceptual composition of the data with far greater computational efficiency. We selected a pre-trained LDM, specifically Stable Diffusion v1.5, as the foundation for our system because it provides a robust, state-of-the-art generative base without the prohibitive training costs of earlier diffusion models, allowing us to focus our efforts on training a novel control mechanism.

A pivotal development in guiding large-scale diffusion models is ControlNet [Zhang et al. 2023], a neural architecture designed to add robust spatial conditioning to pre-trained text-to-image models. The authors address the challenge of finetuning massive models on smaller, task-specific datasets without catastrophic forgetting. The core idea is to lock the original model's weights, preserving its vast knowledge, while creating a trainable copy of its encoding layers. This trainable copy learns the specific input condition (e.g., edges, human pose, depth maps) and feeds its output back into the frozen model. Crucially, the connection between the two is made with "zero convolution" layers—weights initialized to zero—which prevents harmful noise from corrupting the powerful pre-trained backbone during the initial stages of training. This elegant approach allows for efficient and stable training of a wide variety of spatial controls. Our work directly leverages this architecture; we adopt the ControlNet framework to train our own specialized model that learns to interpret multi-channel semantic sketches for terrain generation.

More recently, the Terrain Diffusion Network (TDN) by [Hu et al. 2024] also explored sketch-guided terrain generation, introducing a complex, multi-level denoising scheme to achieve a high degree of geological realism. While their work focuses on producing geologically plausible terrain influenced by climatic patterns, our work differs in its primary goal of enhancing direct artistic control. This difference in focus is reflected in our architectural choice: instead of building a complex model from scratch, we leverage a large, pre-trained model via a single ControlNet. This 'simplicity-for-interactivity' approach allows us to focus our contribution on the artist-centric system and its user-facing controls.

The most relevant prior work to our own is the excellent system presented by [Löchner et al. 2023], which also utilizes a diffusion model for interactive terrain authoring, featuring controls for creating ridges, erosion, and flat areas. However, while the high-level goals are similar, our work differs fundamentally in three key areas. First, in our architectural approach, we leverage the powerful prior of a large, pre-trained text-to-image model (Stable Diffusion) via a ControlNet. In contrast, their system is trained from scratch on a smaller, domain-specific diffusion model. Second, in terms of data efficiency, this architectural choice enables us to achieve high-fidelity results by training on a small, hand-drawn dataset of only 400 images, in stark contrast to the 6 million images required by their system. Third, in our primary contribution, a core finding of our work is the analysis of the loss function itself as a form of artistic control, a nuanced aspect not explored in their paper. Our work, therefore, demonstrates a significantly more data-efficient and flexible methodology for achieving fine-grained, stylistic control in terrain authoring.

2.3 Procedural and Sketch-Based Terrain Generation

The challenges of manual authoring have long been addressed by traditional procedural content generation (PCG) tools that use algorithms like fractal noise [Perlin 1985] and physics-based simulations, such as hydraulic erosion [Génevaux et al. 2013] [Mei et al.

2007]. While powerful, these methods present their own significant hurdles for artistic expression. They typically require the indirect manipulation of numerous abstract parameters, leading to a workflow that can feel more like trial-and-error than direct creation. Even approaches that add a layer of control, for example by using software agents to guide the generation [Doran and Parberry 2010], still rely on a fundamentally indirect authoring process. A key limitation of these approaches is that they are based on mathematical functions that simulate, rather than learn, natural processes. This can make it difficult to generate the complex, subtle, and diverse geological features found in real-world data. Our work builds on this history by seeking to combine the generative power of modern AI with a more direct, artist-centric control paradigm.

Generative Adversarial Networks (GANs) have been successfully applied to terrain generation. [Spick et al. 2019], for example, used a Spatial GAN to learn the features of specific real-world regions from satellite data, enabling the generation of new, stylistically similar heightmaps. However, as an unconditional generative method, this approach does not provide the direct, sketch-based control over landscape features that is the primary focus of our work. Other research has focused on comparing different generative architectures for terrain synthesis. [Demergis 2021], for example, conducted a comparative analysis of VAEs [Kingma and Welling 2014], GANs, and PixelCNNs [van den Oord et al. 2016] for the specific task of generating island heightmaps. The study concluded that the GAN-based approach provided the best trade-off between visual quality and generation speed. However, the work focused exclusively on unconditional generation, where the models learn to produce random islands from a learned distribution. This highlights the need for conditional methods, like our ControlNet-based system, that allow for direct artistic control over the specific features and layout of the generated terrain.

Other GAN-based approaches have focused on a two-stage pipeline for terrain generation. [Voulgaris et al. 2021], for instance, first used an unconditional GAN to generate a random, realistic satellite image and then employed a separate conditional GAN (pix2pix) [Isola et al. 2017] to translate that satellite image into a corresponding heightmap. While this method can produce a wide variety of plausible terrains, the creative process is indirect and lacks user control; the system generates a random landscape rather than allowing an artist to author a specific one. In contrast, our work focuses on a direct, single-stage pipeline where the artist's sketch provides explicit, fine-grained control over the final output.

Other research has directly tackled the problem of sketch-based control for terrain authoring. [Ramos et al. 2023], for example, proposed the Dual Critic Conditional Wasserstein GAN (DCCWGAN) [Radford et al. 2015], a novel architecture designed to transform low-fidelity sketches into realistic heightmaps. Their system cleverly uses two separate discriminators: one to enforce the realism of the output and a second to ensure the generated terrain is faithful to the user's input sketch. While this work validates the demand for artist-centric tools and shows a successful GAN-based implementation, our approach differs by leveraging a pre-trained diffusion model. This allows us to use a more direct, multi-channel semantic sketch for conditioning and avoids the need to train a complex generator and discriminator from scratch.

Concurrent to our work, the TerraFusion system, presented in a recent pre-print by [Higo et al. 2025], also explores the use of latent diffusion models for sketch-based terrain authoring. Their primary contribution is a framework for the joint generation of both a heightmap and a corresponding color texture, which they achieve by concatenating the latent representations of both modalities. Their sketch-based control, similar to ours, uses colored lines to define features like ridgelines and valleys. While this work further validates the power of guided diffusion for this task, our approach differs in its focus: we concentrate exclusively on the high-fidelity generation of the heightmap itself and introduce a more granular, four-channel semantic input for finer control. Furthermore, we provide a complete, interactive system with real-time post-processing, which is a central component of our contribution.

3 Methodology

3.1 System Overview

The Earthbender system is an end-to-end interactive pipeline designed to translate an artist's high-level semantic sketch into a detailed terrain heightmap. The entire workflow, illustrated in Figure 2, emphasizes a rapid, iterative creative process by providing granular artistic control at every stage.

The workflow begins in a custom GUI, where the artist authors a multi-channel semantic sketch. This sketch is then pre-processed into a 4-channel tensor representing feature masks and structural edges. This tensor serves as conditional guidance for our custom-trained ControlNet [Zhang et al. 2023], which in turn steers a pre-trained Latent Diffusion Model (Stable Diffusion v1.5) [Rombach et al. 2022]. The model operates in a compressed latent space to efficiently generate a raw grayscale heightmap that conforms to the artist's input.

Finally, the raw heightmap is presented back to the artist in the GUI for an interactive refinement stage. Here, a suite of real-time post-processing filters allows for fine-tuning the terrain's visual characteristics. To complete the feedback loop, the final result can also be inspected in an integrated 3D pre-visualization viewport.

3.2 Interactive Sketch Input and Pre-processing

The artist's interaction with the Earthbender system is mediated through a custom graphical user interface built with PyQt6. This interface is designed to provide a seamless and expressive design experience, allowing for both expressive input and granular control over the generation parameters before inference begins.

3.2.1 *Input Tools.*

The primary input is a canvas where the artist authors a multi-channel semantic sketch. The system provides several tools to facilitate this process, including semantic brushes for mountains, rivers, and lakes; a pressure-sensitive brush for expressive strokes; and a spray paint tool with controllable spread and density for creating more organic features. A standard eraser tool is also provided for refinement.

3.2.2 *Pre-Inference Parameter Control.*

Before running the model, the artist can adjust two sets of critical parameters via sliders in the GUI. The individual influence of mountains, rivers, and lakes can be independently weighted, acting as scalar multipliers on their respective semantic masks. Additionally, a global ControlNet

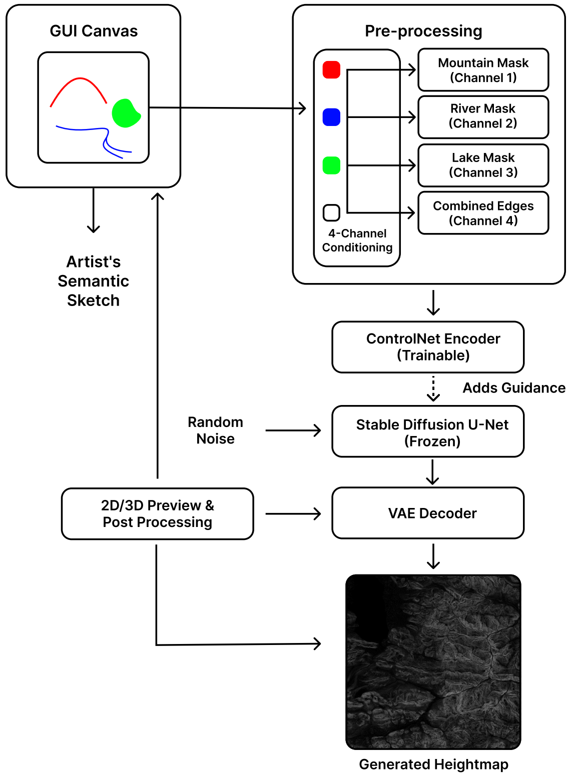


Figure 2: An overview of the Earthbender system architecture. An artist's semantic sketch is pre-processed into a 4-channel conditioning tensor. This tensor guides a ControlNet, which steers a frozen Stable Diffusion U-Net to generate a raw heightmap. The result is then presented back to the artist for interactive post-processing and 3D visualization.

conditioning_scale parameter determines the overall fidelity of the generated image to the input sketch.

3.2.3 Conditioning Tensor Pre-processing. Once the artist initiates the generation, the sketched image is processed into a 4-channel conditioning tensor. This pipeline involves generating binary masks from the sketch's HSV color values, applying the user-defined feature weights, and creating a comprehensive structural map by combining Canny edges from both the overall sketch and the individual feature masks. The three weighted masks and the combined edge map are then stacked to form the final [mountains, rivers, lakes, edges] tensor that is fed into the ControlNet.

3.3 Model Architecture and Training

This section details the specifics of the model architecture, the training procedure, and our analysis of the loss functions used.

3.3.1 Model Architecture. We adopt the ControlNet architecture, a now-standard method for adding spatial conditioning to large, pre-trained diffusion models [Zhang et al. 2023]. The core of our system retains the original, frozen weights of the Stable Diffusion v1.5 U-Net and VAE, preserving the powerful, general-purpose prior learned from billions of images. To introduce our custom control mechanism, we create a trainable copy of the weights of the Stable Diffusion U-Net's twelve encoder blocks and its middle block. This trainable copy is designed to learn the relationship between our 4-channel semantic sketch and the desired output structure. Following the ControlNet methodology, the output of each trainable block is added back to the corresponding skip-connection of the frozen U-Net. This connection is mediated by "zero convolution" layers—1x1 convolutions with weights and biases initialized to zero. This ensures that at the beginning of training, no noise is added to the U-Net's features, thereby protecting the robust pre-trained backbone from being corrupted and allowing for stable and efficient fine-tuning.

3.3.2 Training Procedure. The ControlNet was trained on our custom dataset of paired sketch-and-heightmap images. We utilized the Hugging Face Accelerate library for efficient training. The model was trained for a maximum of 50,000 steps with a batch size of 4, employing an early stopping criterion to prevent significant overfitting. We used the AdamW optimizer with a constant learning rate of 2e-5 and a cosine learning rate scheduler with 500 warmup steps. The entire training process was conducted on a single consumer-grade GPU, demonstrating the efficiency of the ControlNet fine-tuning approach.

3.3.3 Loss Function as Artistic Control. A key part of our investigation involved analyzing how the choice of loss function could act as a form of artistic control over the final output style. We explored two primary objectives. The first was the standard denoising objective from Latent Diffusion, a simplified mean squared error (MSE) in the latent space [Zhang et al. 2023]:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t,t)||_2^2\right] \tag{1}$$

where z_0 is the initial latent, ϵ is the sampled noise, t is the timestep, and $\epsilon_\theta(z_t, t)$ is the model's noise prediction.

The second objective was a custom hybrid loss function designed to exert more explicit control over the pixel-space output. This loss combines the standard latent-space MSE with a weighted L1 pixel-space loss and a smoothness term for lake regions:

$$\mathcal{L}_{Hybrid} = \mathcal{L}_{LDM} + \lambda_{L1}\mathcal{L}_{L1,weighted} + \lambda_{smooth}\mathcal{L}_{smooth} \tag{2}$$

Here, $\mathcal{L}_{L1,weighted}$ is an L1 loss between the decoded prediction and the ground truth, with a higher weight on lake regions. $\mathcal{L}_{smooth}$ penalizes image gradients within lake regions to encourage flat surfaces.

The smoothness term, $\mathcal{L}_{smooth}$, is specifically designed to address the challenge of generating flat, featureless surfaces for lake regions. It operates in pixel space by penalizing the image gradients—the rate of change between adjacent pixels—but only within the areas defined by the lake mask. This discourages the model from creating noisy textures or unwanted bumps on lake surfaces. The loss is formally defined as the mean L1 norm of the masked image gradients:

$$\mathcal{L}_{\text{smooth}} = \mathbb{E}\left[\left|\nabla_x \hat{Y}\right| \odot M_{\text{lake}} + \left|\nabla_y \hat{Y}\right| \odot M_{\text{lake}}\right] \tag{3}$$

where $\hat{Y}$ is the generated output image, M_{lake} is the binary mask for the lake regions, $\nabla_x \hat{Y}$ and $\nabla_y \hat{Y}$ are the horizontal and vertical image gradients, and $\odot$ is the element-wise multiplication. This formulation directly corresponds to our implementation, which calculates the L1 difference between adjacent pixels and applies the lake mask before taking the mean.

Our experiments showed that neither loss is definitively superior; rather, they produce distinct and valid artistic styles. The standard $\mathcal{L}_{LDM}$ yields a terrain with sharp, high-frequency details, while our $\mathcal{L}_{Hybrid}$ produces a smoother, more organic output. Notably, the smoothness and weighted L1 terms in our hybrid loss were particularly effective at generating the flat, featureless surfaces we desired for lake regions. Based on this specific improvement, our artistic preferences, and the goal of creating terrains well-suited for downstream game engine workflows, we selected the **custom hybrid loss** for the final "Earthbender" system.

3.4 Interactive Post-Processing and 3D Visualization

To complete the artist-centric workflow, the Earthbender system includes a final stage for the interactive refinement and visualization of the generated heightmap. This stage is designed to be entirely real-time and non-destructive, allowing for rapid iteration and fine-tuning of the final output.

3.4.1 2D Post-Processing Filters. Once the raw grayscale heightmap is generated by the model, a suite of post-processing filters, controlled by sliders in the GUI, can be applied. These filters operate on the 2D image data and are designed to give the artist precise control over the final look and feel of the terrain.

- Per-Feature Brightness Control: The artist can independently adjust the brightness of mountain, river, and lake regions, effectively increasing or decreasing their elevation. This is implemented as a simple additive operation, where the slider value is applied to pixel intensities within the corresponding feature's semantic mask. This enables subtle strengthening or weakening of specific terrain elements.
- Distance-Based Blending: To soften the integration of features into the surrounding landscape, the system includes a sophisticated blending tool. This is implemented using a distance transform (cv2.distanceTransform) on the inverse of a feature's mask. The resulting distance map is used to create a smooth falloff gradient around the feature's outer edge. The artist can control both the width of this gradient (Outer Blur) and its intensity (Blur Brightness). This allows for the creation of soft transitions, halos, or subtle shadows around features, with the brightness of the blended region being controlled independently from the feature itself.

3.4.2 3D Pre-visualization. During our initial qualitative evaluation, a consistent piece of feedback was the difficulty users had in mentally translating the 2D grayscale heightmap into a three-dimensional form. To address this directly, we developed and integrated an interactive 3D pre-visualization viewport into our system. This feature is not intended to produce a final, game-ready mesh, but rather to serve as an immediate and intuitive guide for the artist.

The viewer takes the final post-processed heightmap and generates a 3D surface mesh. To overcome the limited dynamic range often present in raw VAE outputs, the viewport provides the artist with a set of essential real-time controls:

Level Clamping: Sliders for adjusting the black and white points allow the artist to remap the height data, effectively increasing the contrast and defining which grayscale value corresponds to the lowest and highest points on the terrain.

- Gamma Correction: A mid-level gamma control allows for the non-linear adjustment of the height curve, enabling the artist to make the terrain feel flatter or steeper.
- Vertical Scale: A Z-Scale slider acts as a global multiplier on the final height data, allowing the artist to exaggerate the verticality of the terrain for dramatic effect.

These controls, combined with standard 3D camera navigation, provide a powerful and immediate feedback loop.

4 Experiments and Results

To validate the effectiveness of our proposed system, we conducted a series of quantitative and qualitative experiments. This section details the custom dataset created for this task, the experimental setup for our comparative analysis, and the results of our evaluations. We aim to demonstrate the superiority of our diffusion-based approach over traditional GANs and to analyze the artistic impact of our custom loss function.

4.1 Dataset

To train and evaluate our models, we created a custom dataset of 400 paired sketch-and-heightmap images. The foundation of our dataset is a collection of high-resolution digital elevation models (DEMs) from NASA SRTM Digital Elevation 30m [Farr et al. 2007]. We use Google Earth Engine [Gorelick et al. 2017] to select our area of interest in the .tif format and then export 1024x1024 tiles of the region.

To create the paired sketch for each heightmap, every sketch in our dataset was manually hand-drawn. This process involved tracing key geological features from the ground-truth DEMs, such as ridgelines, rivers, and lakes, and translating them into our semantic color language. While labor-intensive, this approach ensures that our training data accurately reflects the natural variations and imperfections of a real artistic workflow, providing a robust foundation for training a model that is responsive to genuine user input. The sketches shown in our qualitative results (Figure 3) were taken from the test set, which was created using the same manual process.

4.2 Experimental Setup

To provide a comprehensive evaluation, we compare four different models, representing two distinct architectures and two different loss functions. For all experiments, we use a suite of standard quantitative metrics to evaluate the fidelity and realism of the generated heightmaps against the ground truth test set. This includes a traditional reconstruction metric (PSNR), a modern perceptual metric (LPIPS) [Zhang et al. 2018], and two distributional metrics (FID and KID) [Binkowski et al. 2018] [Zhang et al. 2018]. Both FID and KID measure the statistical similarity between the distributions of generated and real images, where lower scores indicate a more realistic and diverse output

The models in comparison are as follows:

- **ControlNet-Default:** Our ControlNet architecture was trained using the standard latent-space MSE loss ($\mathcal{L}_{LDM}$) common to Latent Diffusion Models.
- **ControlNet-Hybrid (Ours):** Our final ControlNet model trained with our custom hybrid loss function ($\mathcal{L}_{Hybrid}$), which includes pixel-space L1 and smoothness terms. This is the model used in the final Earthbender system.
- **Pix2PixHD-Default:** A baseline implementation of the Pix2PixHD architecture trained on our dataset using its standard objective function.
- **Pix2PixHD-Hybrid:** To ensure a fair comparison of loss functions between architectures, we also trained the Pix2PixHD model with our custom hybrid loss function.

All models were trained for a maximum of 50,000 steps using the same training and validation data splits (10% validation), with checkpoints saved periodically. For our final evaluation, we selected the checkpoint for each model that demonstrated the best performance on the validation set.

4.3 Quantitative Analysis

We evaluated the performance of all four models on our held-out test set of 40 images. The results, calculated in a series of paired and distributional metrics, are presented in Table 1.

Table 1: Quantitative comparison of models on the test set. "Default" refers to models trained with their standard loss functions, while "Hybrid" refers to models trained with our custom loss function. For PSNR, higher is better. For LPIPS/FID/KID, lower is better. Best scores are in bold.

Model	PSNR ↑	LPIPS ↓	FID ↓	KID (x100) ↓
Pix2PixHD (Default)	**19.74**	**0.5398**	407.15	38.20 ± 0.00
Pix2PixHD (Hybrid)	19.17	0.8304	356.47	28.63 ± 0.00
ControlNet (Default)	12.08	0.5424	**280.93**	15.46 ± 0.00
ControlNet (Hybrid)	11.91	0.5587	287.13	**14.52 ± 0.00**

It is important to note that the absolute values of FID and KID are dataset-dependent and cannot be directly compared to scores reported on other domains (e.g., ImageNet, FFHQ). In our setting, these metrics serve primarily as relative indicators of performance across the baselines evaluated on the same dataset.

At first glance, the results present a paradox. The traditional, pixel-wise reconstruction metric, PSNR, overwhelmingly favors the Pix2PixHD models, with the default GAN achieving the highest score of 19.74. However, a closer look at the more advanced, perceptually-aligned metrics reveals a different and more accurate story.

The distributional metrics, FID and KID, which measure the statistical similarity between the generated images and the ground truth, decisively favor our ControlNet-based approach. The ControlNet model with the default loss achieved the best FID score (280.93), while our hybrid loss variant achieved the best KID score (14.52). Both significantly outperform the best Pix2PixHD model (FID of 356.47 and KID of 28.63). This discrepancy highlights a well-known limitation of pixel-wise metrics in evaluating generative tasks. PSNR rewards the blurry, low-frequency outputs of the GAN because they are "less wrong" on a pixel-by-pixel average. In contrast, FID and KID correctly identify that the detailed, high-frequency textures generated by our ControlNet models are far more realistic and representative of the true data distribution.

The comparison between our two ControlNet loss functions further reinforces our methodology. The scores across all metrics are remarkably close, indicating that both models produce outputs of a similar overall quality. The default loss model is slightly better in terms of FID, while our hybrid loss model is slightly better in terms of KID. This confirms that the choice between them is primarily an artistic one, based on the desired output style (sharp and detailed vs. smooth and organic), rather than a clear difference in quantitative performance.

4.4 Qualitative Analysis

To resolve the paradox presented by the quantitative metrics, a qualitative visual comparison is necessary. Figure 3 shows a representative example from our test set.

While the quantitative metrics provide a valuable overview, a qualitative visual analysis is essential for understanding the practical and artistic differences between the models. The numerical paradox, where the Pix2PixHD models achieve superior PSNR scores, is immediately resolved upon visual inspection of the outputs, as shown in Figure 3.

Across all test cases, the Pix2PixHD models consistently fail to produce usable or structurally coherent results. The outputs are characterized by a low-frequency, blurry texture that lacks any of the fine details or sharp features present in the ground truth. As seen in the examples, the GAN baseline fails to interpret the specific semantic meaning of the input sketch; mountains, rivers, and lakes all dissolve into a noisy, indistinct pattern. The model is unable to generate the sharp coastlines or intricate riverbeds specified in the input, rendering its output unusable for any practical application.

In stark contrast, both ControlNet models demonstrate a profound understanding of the input sketch, generating complex, detailed, and structurally faithful heightmaps. The models successfully interpret the semantic colors, creating elevated mountain ridges, carved river valleys, and flat lake beds that correspond directly to the artist's input. The high-frequency textural detail is not only present but also geologically plausible, aligning with the superior FID and KID scores that these models achieved.

The comparison between the two ControlNet loss functions reveals the stylistic trade-off discussed in our methodology. The model trained with the default latent denoising loss produces a terrain with a very high frequency of sharp, "etched" detail, excellent for raw realism. The model trained with our custom hybrid loss, however, produces a slightly smoother and more organic result, with more consolidated shapes that are often more desirable for downstream applications like game engines. This visual evidence confirms that our primary quantitative metrics (FID/KID) are far more indicative of perceptual quality than traditional metrics

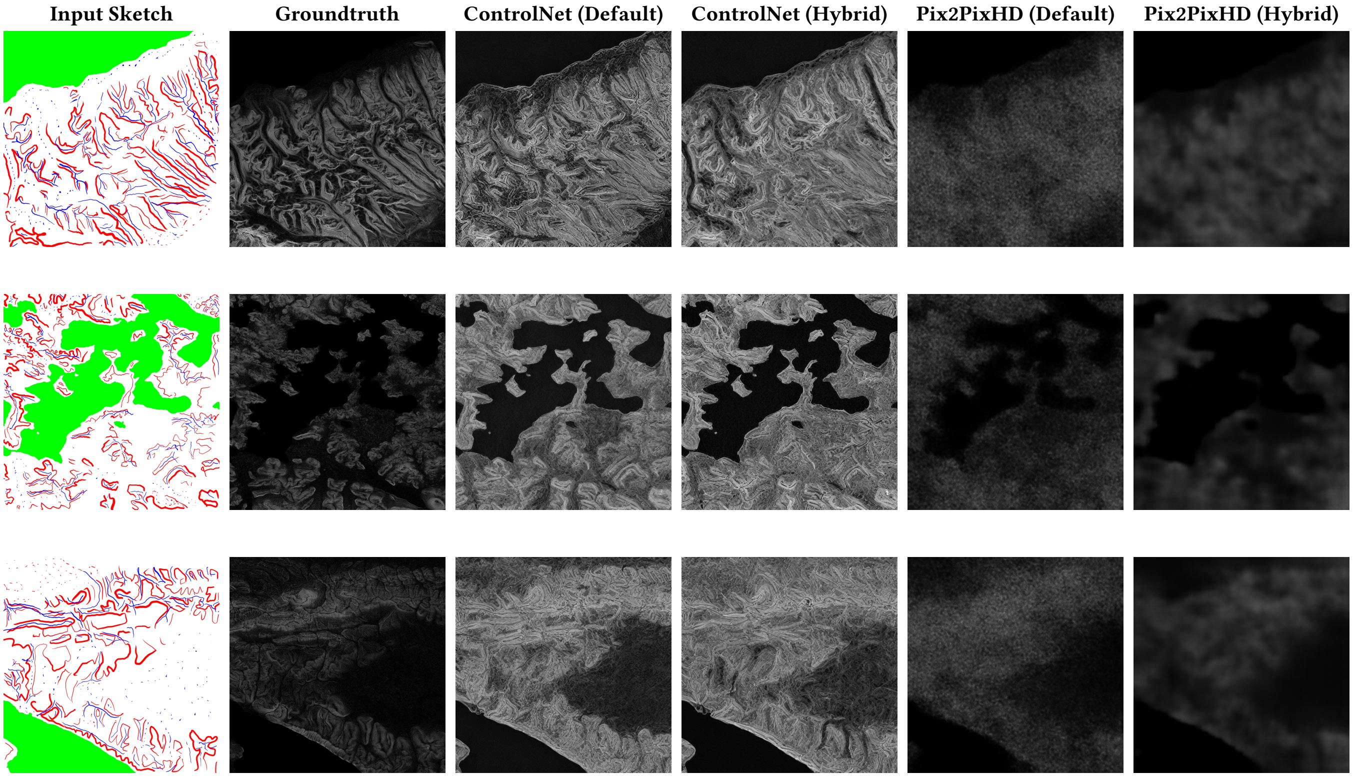


Figure 3: Qualitative comparison on three examples from the test set. For each example (row), we show (from left to right): the input sketch, the ground truth heightmap, and the outputs from our four trained models. This figure clearly illustrates the superior detail and structural fidelity of the ControlNet-based models compared to the Pix2PixHD baselines, which consistently produce blurry and unusable results.

(PSNR) and validates our choice of the hybrid loss as a means of achieving a specific, artistically preferable style.

4.5 Training Dynamics

To provide insight into our model development process, Figure 4 displays the training and validation loss curves for our two ControlNet models. Both models exhibit stable training dynamics, with the training loss consistently decreasing over time. The validation loss, which was calculated every 1,000 steps, was used to monitor for overfitting and to select the final model checkpoints for evaluation.

For the model trained with the default latent denoising loss, the validation loss reached its minimum at approximately 30,000 steps, after which it began to show signs of overfitting. We therefore selected the 30,000-step checkpoint for this model. For our custom hybrid loss model, the validation loss was lowest and most stable around the 20,000-step mark, and we selected this checkpoint for all subsequent experiments. These plots confirm that our training procedure was stable and that our final models were selected based on their optimal performance on unseen validation data.

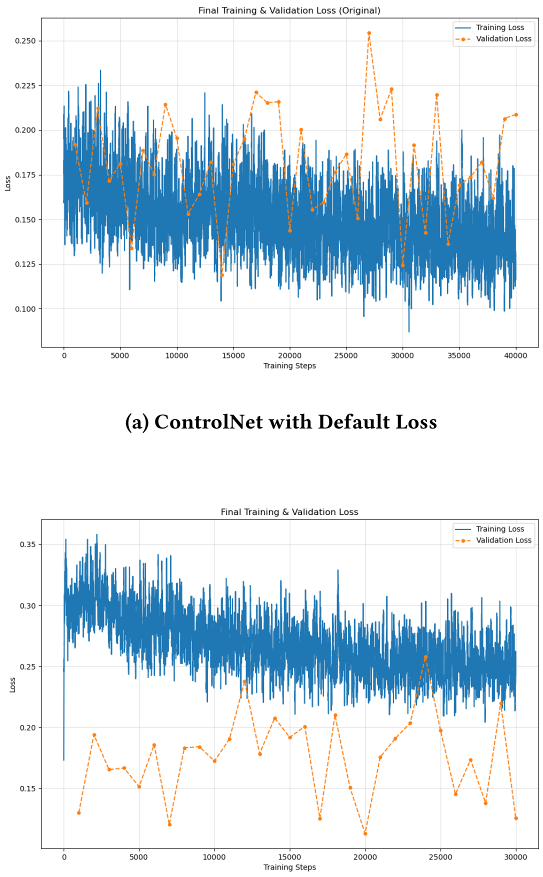
4.6 Qualitative Evaluation (User Feedback)

To complement our quantitative metrics and validate the practical usability of the Earthbender system, we conducted a preliminary participant feedback study. This evaluation was designed not as a formal expert review, but to gather initial user feedback and assess how the tool performs in the hands of its target audience.

4.6.1 Methodology. The study involved 15 participants, primarily computer science students with a diverse range of prior experience in game development and digital art, from novice to experienced. Each session began with a five-minute guided introduction to the system's interface, features, and workflow. Following the tutorial, participants were given ten minutes for a hands-on, free-form creative task where they were encouraged to explore the tool and create a terrain of their own design.

Upon completion of the task, each participant was asked to fill out two questionnaires. The first was the industry-standard System Usability Scale (SUS) [Brooke 1996], a 10-item questionnaire that provides a reliable, quantitative measure of a system's overall usability. The second was a custom 10-item questionnaire, using a 7-point Likert scale, designed to gather specific feedback on the core features of Earthbender, such as the sense of creative control and its potential utility in an artistic pipeline.

4.6.2 Results. The feedback from the study was highly positive. The Earthbender system achieved an average SUS score of 86.33, which corresponds to a grade of "A" and is considered an "Excellent"



(a) ControlNet with Default Loss

(b) ControlNet with Hybrid Loss

Figure 4: Training and validation loss curves for the two ControlNet models. The validation loss (orange, dashed line) was used to select the optimal checkpoint for each model before significant overfitting occurred.

result. This strong score indicates that users found the interface to be highly usable and easy to learn.

While the SUS score provides a robust measure of overall usability, it does not capture the specific nuances of the creative experience. To gain deeper insights, we also analyzed the results from our custom-designed questionnaire, visualized as a violin plot in Figure 5. The responses were overwhelmingly positive, with the distribution for most questions heavily concentrated in the "Agree" (6) and "Strongly Agree" (7) range. Participants reported a strong sense of overall control and a high degree of fidelity to their artistic intent. The core drawing tools for mountains, lakes, and carving also received very high and consistent ratings, validating our sketch-based interaction paradigm.

Notably, the response to the "Necessity of Post-Processing Utility" was the most positive of all, with nearly all participants rating it as essential. This provides strong evidence that the post-processing filters can help the creative workflow and provide more control. The only area with significant variance in feedback was the "Guidance Scale Utility," suggesting that while some users found it powerful, its function may be less intuitive for others. Overall, this initial feedback validates our artist-centric design philosophy, confirming that Earthbender is not only a functional tool but also an effective creative partner.

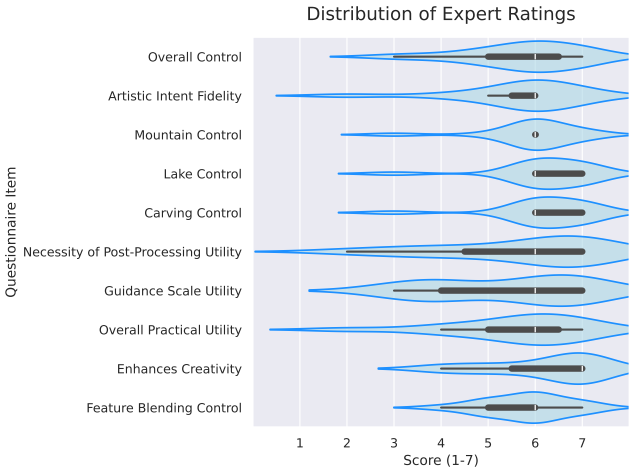


Figure 5: Distribution of participant responses (1–7 Likert scale) to our 10-item qualitative questionnaire (N = 15 participants). Each violin shows the distribution of individual participant responses for that item. The questionnaire items (y-axis labels) were:

- **Overall Control:** "I felt in control of the creative process."
- **Artistic Intent Fidelity:** "The final heightmap accurately reflected my artistic intent."
- **Mountain Control:** "I had fine-grained control over the placement and shape of the mountains (red)."
- **Lake Control:** "I had fine-grained control over the placement and shape of the lakes (green)."
- **Carving Control:** "I had fine-grained control over my carving tool (blue)."
- **Post-Processing Utility:** "The post-processing sliders (brightness, blur) were a necessary and useful feature for refining the final image."
- **Guidance Scale Utility:** "The Guidance Scale slider gave me meaningful control over the output's detail."
- **Practical Utility:** "I would find this tool useful in a real game development or artistic pipeline."
- **Enhances Creativity:** "This tool enhances my creativity rather than replacing it."
- **Feature Blending Control:** "I had control over blending the features with the combination of colors."

5 Discussion

Our experiments demonstrate that a ControlNet-based approach, guided by a multi-channel semantic sketch, is a highly effective method for interactive terrain authoring. This section interprets the broader implications of our results, discusses the limitations of our system, and proposes avenues for future work.

While our user study indicated a wide variance in how participants initially perceived the guidance scale feature (as seen in Figure 5), we observed that users typically settled on a preferred value and did not change it later. Each user, based on their own artistic vision, had a preference for a certain level of creative deviation versus strict adherence to their sketch. To quantify the impact of this parameter, we analyzed its effect on our 40-image test set. As shown in Table 2, increasing the guidance scale from 1.0 to 1.7 resulted in a significant and consistent improvement in the reconstruction and perceptual metrics (PSNR and LPIPS) for both ControlNet models. This confirms that the guidance scale is a powerful parameter that provides meaningful control over the output's quality and visual fidelity.

Table 2: Analysis of the effect of the ControlNet Guidance Scale. We compare the performance of our two ControlNet models on the test set using the default scale (1.0) versus a higher value (1.7). The results show that a higher guidance scale consistently improves the reconstruction and perceptual metrics (PSNR and LPIPS), while the distributional metrics (FID and KID) show a more complex trade-off. Best scores are highlighted in bold.

Model	PSNR ↑	LPIPS ↓	FID ↓	KID (x100) ↓
Default (Guidance Scale 1.7)	**13.36**	**0.5098**	297.39	**13.56 ± 0.00**
Hybrid (Guidance Scale 1.7)	12.96	0.5226	305.47	15.73 ± 0.00
Default (Guidance Scale 1)	12.08	0.5424	**280.93**	15.46 ± 0.00
Hybrid (Guidance Scale 1)	11.91	0.5587	287.13	14.52 ± 0.00

Limitations and User Feedback. While the feedback from our participant study was overwhelmingly positive, it also highlighted areas for improvement. Some of our users with more experience in the field of game design provided quotes that clearly stated some of the weaknesses of our approach. One user said, *"I can't choose how tall each mountain is"*. Another critical observation was, *"I don't have any idea on what scale we are working on, and I don't know anything about the size of these features on the map."*. These problems have been a constant limitation in all previous works that have attempted to generate a heightmap using deep generative models. However, hearing these quotes from our users provided valuable information about possible roadmaps for future feature work.

These comments on explicit control naturally lead to a related consideration: the system's handling of implicit variations in an artist's input style. It is expected that leveraging a pre-trained stable diffusion model makes the model robust, even when facing a less-detailed sketch as input. The robustness of the model with different levels of detailed input is an interesting research question that can be explored, particularly in terms of how much the generated output differs from the artist's expectation with respect to each level of input detail.

Furthermore, our research revealed several extensions that highlighted the current limitations of our approach. An attempt to train the model on a combination of our sketch and a text prompt was unsuccessful; we found that the strong, explicit spatial guidance from the ControlNet sketch consistently overpowered the weaker, more abstract guidance from the text. We also attempted to train a version for 1024x1024 output, but fine-tuning a model pre-trained at 512x512 did not successfully generalize to the higher resolution. The small size of our dataset likely exacerbated these challenges. We expect that changing the pre-trained model to one that has been trained with higher resolution will give the Earthbender the ability to produce higher-resolution output. Still, we were unable to conduct the experiment due to the hardware limitations necessary for both training and inference. To achieve higher resolution with the current pre-trained model, Earthbender can utilize up-scaling; however, selecting the correct algorithm for up-scaling our heightmaps is an interesting research question that we aim to explore in our feature work.

6 Conclusion and Future Work

In this work, we presented Earthbender, a novel interactive system that pushes the boundaries of sketch-based terrain authoring. We have shown that it is a viable and effective tool, capable of producing high-quality, detailed heightmaps that are qualitatively superior to traditional GAN-based approaches. Our core philosophy is that generative AI models will only be viewed as a positive force by artists if they serve to enhance, rather than replace, creative control. Earthbender is a step in this direction, demonstrating an artist-centric workflow that is both powerful and intuitive.

Our findings and limitations point to several clear directions for future work. A critical next step is to tackle higher-resolution generation, perhaps by exploring new U-Net backbones or cascaded refinement models. Further studies are needed to find more effective methods for combining sketch and text conditioning, which may require novel datasets and architectures. To enhance its practical utility, the system could be trained to understand real-world scales and to provide per-feature height control, allowing artists to specify the exact elevation of individual mountains. Finally, to better align the model's output with its intended application, future iterations could be trained on datasets where the ground truth heightmaps are already specialized and optimized for use in game development pipelines.

Acknowledgments

This work was partially supported by a stipend from the University of Bremen. Our complete source code and dataset can be found at: https://danial-barazandeh.github.io/Earthbender/

References

Mikolaj Binkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*. arXiv:1710.07626 [cs.LG]

John Brooke. 1996. *SUS: A "Quick and Dirty" Usability Scale.* Taylor and Francis, London. 189–194 pages.

Dimitri Demergis. 2021. Comparative Analysis of Machine Learning Techniques for Island Heightmap Generation. In *2021 International Joint Conference on Neural Networks (IJCNN)*. 1–8. doi:10.1109/IJCNN52387.2021.9533580

Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. 2022. StyTr2: Image Style Transfer with Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jonathon Doran and Ian Parberry. 2010. Controlled Procedural Terrain Generation Using Software Agents. *IEEE Transactions on Computational Intelligence and AI in Games* 2, 2 (2010), 111–119. doi:10.1109/TCIAIG.2010.2049020

T. G. Farr, P. A. Rosen, E. Caro, R. Crippen, R. Duren, S. Hensley, M. Kobrick, M. Paller, E. Rodriguez, L. Roth, D. Seal, S. Shaffer, J. Shimada, J. Umland, M. Werner, M. Oskin, D. Burbank, and D. E. Alsdorf. 2007. The Shuttle Radar Topography Mission. *Reviews of Geophysics* 45, 2 (2007), RG2004. doi:10.1029/2005RG000183

Jean-David Génevaux, Éric Galin, Eric Guérin, Adrien Peytavie, and Bedrich Benes. 2013. Terrain generation using procedural models based on hydrology. *ACM Trans. Graph.* 32, 4, Article 143 (July 2013), 13 pages. doi:10.1145/2461912.2461996

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (Oct. 2020), 139–144. doi:10.1145/3422622

Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* (2017). doi:10.1016/j.rse.2017.06.031

Kazuki Higo, Toshiki Kanai, Yuki Endo, and Yoshihiro Kanamori. 2025. TerraFusion: Joint Generation of Terrain Geometry and Texture Using Latent Diffusion Models. arXiv:2505.04050 [cs.GR] https://arxiv.org/abs/2505.04050

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 574, 12 pages.

Zexin Hu, Kun Hu, Clinton Mo, Lei Pan, and Zhiyong Wang. 2024. Terrain diffusion network: climatic-aware terrain generation with geological sketch guidance. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAAI'24)*. AAAI Press, Article 1402, 9 pages. doi:10.1609/aaai.v38i11.29150

Xun Huang. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1–10.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR* (2017).

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. arXiv:http://arxiv.org/abs/1312.6114v10 [stat.ML]

J. Löchner, J. Gain, S. Perche, A. Peytavie, É. Galin, and É. Guérin. 2023. Interactive authoring of terrain using diffusion models. *Computer Graphics Forum* 42 (2023). Issue 7. doi:10.1111/cgf.14941

Xing Mei, Philippe Decaudin, and Bao-Gang Hu. 2007. Fast Hydraulic Erosion Simulation and Visualization on GPU. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*. 47–56. doi:10.1109/PG.2007.15

Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. arXiv:1411.1784 [cs.LG] https://arxiv.org/abs/1411.1784

Park, Taesung, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic Image Synthesis with Spatially-Adaptive Normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Jeeseung Park and Younggeun Kim. 2022. Styleformer: Transformer based Generative Adversarial Networks with Style Vector. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8973–8982. doi:10.1109/CVPR52688.2022.00878

Simon Perche, Adrien Peytavie, Bedrich Benes, Eric Galin, and Eric Guérin. 2023. Authoring Terrains with Spatialised Style. *Computer Graphics Forum* 42, 7 (2023), e14936. doi:10.1111/cgf.14936 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14936

Ken Perlin. 1985. An image synthesizer. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '85)*. Association for Computing Machinery, New York, NY, USA, 287–296. doi:10.1145/325334.325247

Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv preprint arXiv:1511.06434* (2015).

Nuno Ramos, Pedro Santos, and João Dias. 2023. Dual Critic Conditional Wasserstein GAN for Height-Map Generation. In *Proceedings of the 18th International Conference on the Foundations of Digital Games* (Lisbon, Portugal) *(FDG '23)*. Association for Computing Machinery, New York, NY, USA, Article 45, 4 pages. doi:10.1145/3582437.3587183

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer, Cham, 234–241. doi:10.1007/978-3-319-24574-4_28

Ryan J. Spick, Peter Cowling, and James Alfred Walker. 2019. Procedural Generation using Spatial GANs for Region-Specific Learning of Elevation Data. In *2019 IEEE Conference on Games (CoG)*. 1–8. doi:10.1109/CIG.2019.8848120

Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. 2016. Conditional Image Generation with PixelCNN Decoders. arXiv:1606.05328 [cs.CV] https://arxiv.org/abs/1606.05328

Georgios Voulgaris, Ioannis Mademlis, and Ioannis Pitas. 2021. Procedural Terrain Generation Using Generative Adversarial Networks. In *2021 29th European Signal Processing Conference (EUSIPCO)*. 686–690. doi:10.23919/EUSIPCO54536.2021.9616151

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, and Alexei A. Efros. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 3813–3824. doi:10.1109/ICCV51070.2023.00355

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.