INFINITE LEAGUES UNDER THE SEA: PHOTOREALISTIC 3D UNDERWATER TERRAIN GENERATION AUGMENTED BY VISUAL FOUNDATION MODELS

Tianyi Zhang **Weiming Zhi** **Joshua Mangelson** **Matthew Johnson-Roberson**
Carnegie Mellon University Brigham Young University
tianyiz4@andrew.cmu.edu

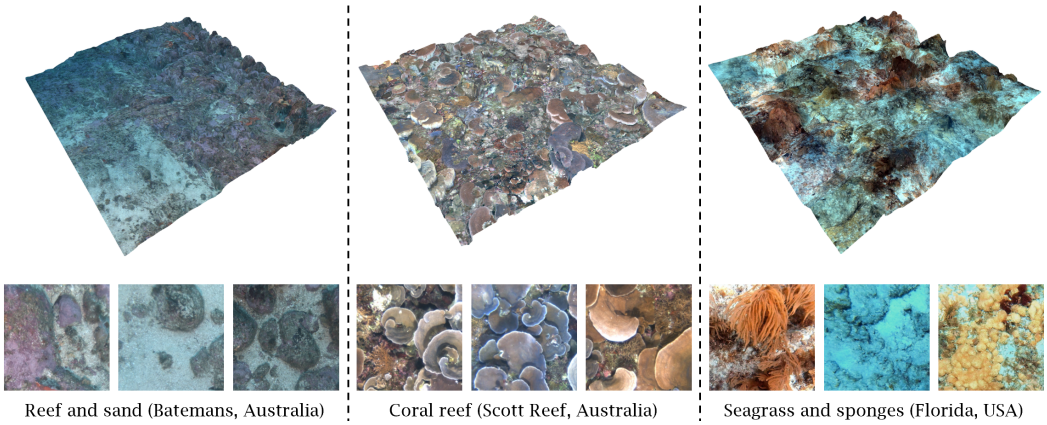

Figure 1: **Underwater 3D terrain generation:** Given 2D images of the real world seafloor collected by robots, we distill 3D geometry and semantic information from visual foundation models and train diffusion model that generate realistic 3D underwater scenes. (All visual contents shown in this figure are generated)

ABSTRACT

This paper tackles the problem of generating representations of underwater 3D terrain. Off-the-shelf generative models, trained on Internet-scale data but not on specialized underwater images, exhibit downgraded realism, as images of the seafloor are relatively uncommon. To this end, we introduce *DreamSea*, a generative model to generate hyper-realistic underwater scenes. DreamSea is trained on real-world image databases collected from underwater robot surveys. Images from these surveys contain massive real seafloor observations and covering large areas, but are prone to noise and artifacts from the real world. We extract 3D geometry and semantics from the data with visual foundation models, and train a diffusion model that generates realistic seafloor images in RGBD channels, conditioned on novel *fractal* distribution-based latent embeddings. We then fuse the generated images into a 3D map, building a 3D Gaussian Splatting (3DGS) model supervised by 2D diffusion priors which allows photorealistic novel view rendering. DreamSea is rigorously evaluated, demonstrating the ability to robustly generate large-scale underwater scenes that are consistent, diverse, and photorealistic. Our work drives impact in multiple domains, spanning filming, gaming, and robot simulation.

1 INTRODUCTION

Scene generation is widely studied today, with deep neural networks capable of creating realistic 3D environments trained on large-scale visual data. This technology has a significant impact across various fields, including the film and gaming industries, as well as robotics and autonomous vehicle simulations. In this paper, we explore the application of deep generative models to the unique setting

of underwater environments. Without sufficient data and annotations, the following questions for underwater scene generation remains open:

- What kind of data can we use to train an underwater generative model?
- How can we train the underwater 3D generative model without 3D scans?
- How can we control the sampling process while the data come with no captions or annotations?
- How can we generate underwater terrain with natural-looking variation in appearance?
- What techniques can we use from off-the-shelf 3D generative models and what is lacking in current open-source models?

In this work, we tackle the problem from the perspective of robot perception. Underwater robots and autonomous underwater vehicles (AUVs) are designed to travel long distances under the sea, maintaining altitude and route to survey the designated area autonomously Iscar et al. (2018); Zheng et al. (2025). Compared to typical images and videos on the Internet, underwater robotic images cover much larger areas of the terrain. However, the massive amounts of data collected by underwater robots present unique challenges: It is difficult to acquire 3D information directly from sensory streams, as depth sensors and LiDARs commonly do not work well underwater. In addition, natural water bodies are highly dynamic, and visibility is low as a result of light scattering and absorption in the medium. Therefore, Structure-from-Motion (SfM) Agarwal et al. (2011) and Simultaneous Localization and Mapping (SLAM) Mur-Artal et al. (2015); Song et al. (2024) solutions have unstable performance. As a result, a significant amount of robotic data comes with no camera poses, and the cost of expert annotation is extremely high.

This paper introduces *DreamSea*, a diffusion-based generative model that can infinitely generate photorealistic 3D underwater scenes. **DreamSea is trained on RGB images captured by underwater robots without any 3D sensory information, SfM poses or human annotations.** After training, scenes generated by DreamSea are spatially consistent in geometry with natural-looking variations in appearance. The contributions of this paper are as follows:

1. A novel approach that leverages a *fractal* distribution of latent embeddings to control the appearance of generated terrains;
2. Integration of visual foundation models (VFMs) on unseen underwater images to exploit semantic and 3D geometric information for scene generation; and
3. A pipeline that integrates the state-of-the-art developments image diffusion, inpainting, VFMs and 3DGS Kerbl et al. (2023), to allow the generation of photorealistic 3D terrains from unannotated images.

2 Related Work

2.1 Procedural Terrain Generation

Early studies on procedural terrain generation focus on generating elevation maps that resemble the 3D structure of real-world terrain Miller (1986). In particular, explicit mathematical models such as fractional Brownian motion (fBm) Mandelbrot (1983), the diamond square algorithm Fournier et al. (1998), and Perlin noise Perlin (1985) are commonly used to approximate natural variations. Modern approaches have enabled the generation of 3D scenes consisting of a variety of assets procedurally and rendered with photorealistic quality Raistrick et al. (2023). Similar procedural strategies have also been applied to generate room layouts Deitke et al. (2022) and object-level Greff et al. (2022) layouts that can be used to train embodied AI algorithms. However, those modern approaches are based on pre-modeled 3D assets. While it is feasible to specify these assets in advance for commonly seen objects and scenes, e.g. indoor environment, this is not the case for unseen environments such as the deep sea. When applying the contemporary procedural generator Infingen Raistrick et al. (2023) to the underwater domain, the resulting generated scenes are filled with repeated assets with lower rendering quality than scenes generated in more typical domains. We illustrate attempts to generate underwater scenes using large off-the-shelf models in Figure 2.

Figure 2: Off-the-shelf solution for generating underwater scenes: ChatGPT and SORA are able to generate scenes with diverse appearances, but present heavy artificial effects even though prompted with the "photorealistic style" keyword. Simulation environments Song et al. (2025) based on classic rendering pipelines, e.g. UNav-Sim Amer et al. (2023) and Infinigen Raistrick et al. (2023), present limited performance when generating diverse and uncommon 3D assets.

2.2 Deep Generative Models

Given an image dataset, an image generation model learns the distribution of this dataset. Unseen image samples can be generated as samples drawn from this distribution. Early techniques such as Variational Autoencoders (VAEs) Kingma & Welling (2014) and Generative Adversarial Networks (GANs) Goodfellow et al. (2020) are able to generate realistic images. In recent years, models such as DDPM Ho et al. (2020), Stable Diffusion Rombach et al. (2022) and DiT Peebles & Xie (2023) allow high-quality generation that can be conditioned on language inputs. These technologies have also led to commercialized models such as ChatGPT and SORA. While these models are capable of creating arbitrary scenes, we find, empirically, that the quality of generated underwater scenes is significantly lower than other more common environments. It can be hypothesized that the training data for underwater scenes is scarce and unbalanced. The development of specialized models with curated data for underwater scenes is still an open problem. In this work, our DreamSea model leverages a DDPM Ho et al. (2020) network with the RePaint Lugmayr et al. (2022) framework as a backbone image generation and inpainting model.

2.3 3D Scene Representation and Generation

Three-dimensional scenes are often represented as point clouds, meshes or implicit functions, and generative models can be trained on 3D datasets such as ScanNet Dai et al. (2017) to create 3D assets and scenes. Recent advancements in neural radiance fields (NeRFs) Mildenhall et al. (2021) techniques enable 3D scene reconstruction with photorealistic quality by optimizing directly over photometric loss. Building upon NeRFs, 3DGS Kerbl et al. (2023) developed an explicit representation which enables efficient training and rendering at 100+ fps, making it a great fit for creating 3D scenes and simulating robot perception Yuan et al. (2024). It is common to use 2D diffusion priors to support generation of 3D assets either using NeRFs Poole et al. or 3DGS Tang et al.; Yi et al. (2024).

2.4 Visual Foundation Models

Underwater robotic field tests typically result in massive amounts of images that are extremely challenging to annotate and often lack 3D information. In this work, we leverage visual foundation models, which are trained on internet-scale data to infer semantic and geometric information by the images collected by our robots. CLIP Radford et al. (2021) is a vision-language model (VLM) trained on internet-scale image-caption pairs and generalizes to unseen images. DINOv2 Oquab et al. (2024) is another foundation model that encodes an RGB image in a vector representation. In this work, we train the image diffusion model conditioned on DINO v2 representations, so the diffusion can be controlled in the latent space. Depth Anything v2 Yang et al. (2024) is a depth foundation model that predicts depth from RGB images. In many cases this is used to generate RGB+Depth (RGBD) images from RGB image inputs. Using foundation models in a zero-shot manner is widespread in fields such as robotics Zhi et al. (2024a;b), where labels are not abundant.

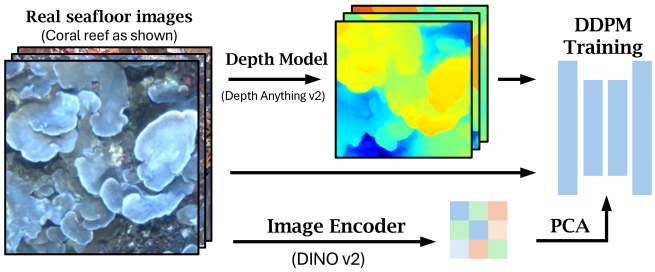

Figure 3: **Overview of Training:** Given RGB-only images collected from underwater surveys, we generate depth channels and embeddings with visual foundation models Yang et al. (2024); Oquab et al. (2024). A DDPM network is then trained with an RGBD image as input conditioned on embeddings.

3 DreamSea

At the center of DreamSea is a terrain generation model that varies in spatial coordinates. This model can then generate a set of consistent images spanning a desired spatial region, which can be used to construct 3DGS representations. Particular care needs to be taken to ensure that the generated images reflect both the biological and landscape diversity of marine environments, while being spatially-consistent.

This section elaborates on the design consideration and methodology details of DreamSea, and is structured as follows. In section 3.1, we outline the extraction of relative depth from diverse underwater data from different expeditions. In section 3.2, we introduce our diffusion-based generative model that is conditioned on *zero-shot visual features*, enabling the controlled generation on varied underwater environments. In section 3.3, we introduce our novel fractal-based generation approach, which enables a set of spatially *consistent* underwater images to be generated and allows explicit control of the diversity of the generated terrain. Finally, in section 3.4, we leverage the terrain generated by our generative model to construct a 3DGS representation supervised by the 2D diffusion prior. An overview of our training procedure is outlined in Figure 3, and the generation procedure is sketched in Figure 4.

3.1 3D Structure from Depth Foundation Model

To build more consistent 3D structures underwater, we seek to incorporate depth into the diffusion-based generative model. This, however, can be challenging. While traditional 3D reconstruction and mapping methods such as SfM and SLAM have been demonstrated on underwater data, the community struggles to scale up the application of these methods due to challenging underwater environments. These challenges often manifest via low visibility, dynamic surroundings, heavy motion blur under low light, and different sensor set-ups between expeditions to collect data. In this paper, we use the depth foundation model, Depth Anything v2 Yang et al. (2024), to generate a depth map from 2D image data. Depth foundation models are good at predicting the relative depth distribution in single frames. We normalize this prediction to $[0, 1]$. In this work, we consider depths up to a scale factor, and do not require absolute metric depth. The metric scale can be recovered with additional sensors or classic stereo-matching methods. Estimated depths are used as additional channels for the real-world training data.

3.2 Conditional Diffusion on Zero-shot Features

Underwater robotic images do not come with captions. Additionally, annotating underwater data is also exceedingly challenging and requires a massive expert-level effort. Relying on manual labels would both be costly and difficult to scale. In light of this, we leverage the foundation visual model, DINO v2 Oquab et al. (2024), to extract zero-shot features from underwater images: for the image data set, we first generate DINO v2 features and then apply Principal Component Analysis (PCA) on the feature set to project high-dimensional features to the low-dimensional space. This reduced dimensional feature vector then acts as a descriptor of the contents within the image. Sim-

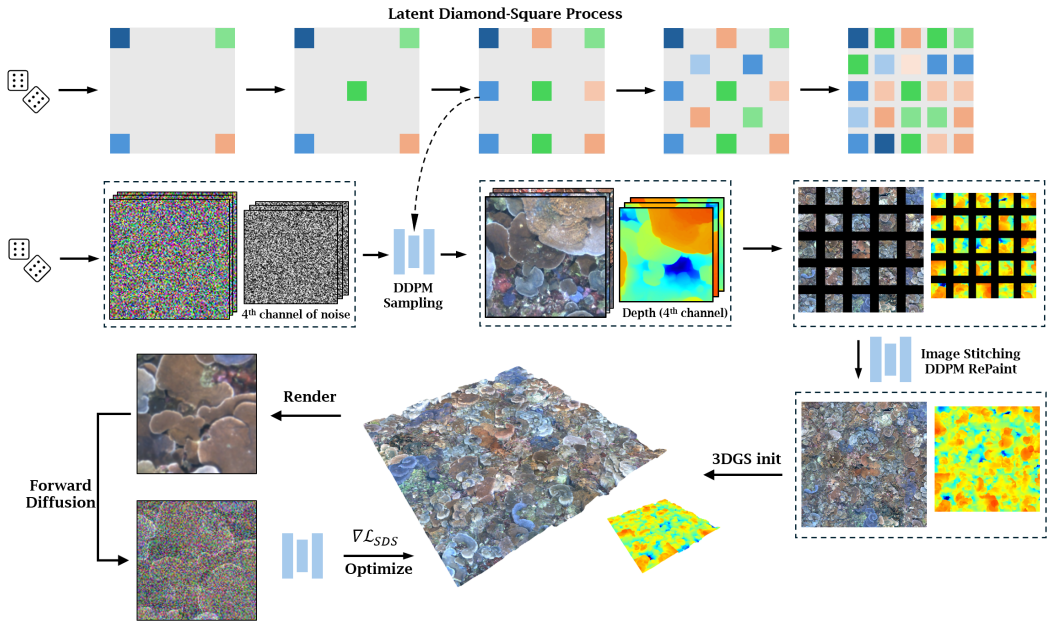

Figure 4: **Overview of Generation:** Our approach generates fractal embedding with the diamond-square method first, then generates images conditioned on these embeddings. We use RePaint Lugmayr et al. (2022) to stitch the images together into a dense RGBD map. The RGBD map can be converted into a 3D point cloud and initialized as a 3DGS model Kerbl et al. (2023). The 3DGS model is further refined with 2D diffusion priors using Score Distillation Sampling (SDS) loss allowing realistic rendering from novel views.

ilar ideas have been explored in LangSplat Qin et al. (2024) in which a Variational Autoencoder (VAE) Kingma & Welling (2014) is trained to project CLIP Radford et al. (2021) features onto a low-dimension space. Early work by Zhang et al. Zhang & Johnson-Roberson (2022) takes a similar approach on seafloor mapping data with self-supervised training. However, here, by integrating foundation models, we are not required to train large neural networks from scratch to extract features, and can instead apply weights pre-trained on Internet-scale data.

After obtaining a reduced-dimensional feature vector for each image, we train a diffusion model conditional on feature vectors, to generate both RGB and depth images.

Let us denote the feature vector as

$$\boldsymbol{\phi} \leftarrow \texttt{PCA}(\texttt{DINOv2}(\mathbf{I})), \tag{1}$$

where $\mathbf{I}$ is an image and $\texttt{PCA}(\texttt{DINOv2}(\cdot))$ indicates applying PCA to the feature vector outputted by the DINO model, reducing dimensionality. During inference, our conditional generative model can be expressed as, $\mathbf{I} \sim P(\mathbf{I}|\boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is a visual feature vector we condition upon. Generating spatially-consistent and yet diverse landscape images, requires controlling the evolution of $\boldsymbol{\phi}$ over the spatial domain, which alters the generative distribution of the terrain.

3.3 Fractal Latent Terrain Generation

An inherent property of naturally-occurring terrains is that coordinate points that are close in geometric distance should have similar attributes. The spatial distribution of natural terrain is often modeled using fractal processes to approximate natural-looking variations. We imbue this inductive bias into DreamSea through a novel **fractal embeddings framework**, which assumes that the latent vectors over the spatial domain follow fractal processes.

We begin by initializing the latent vectors at the corners of an arbitrary square region for which we seek to generate terrain. We seek to sample a latent function $\boldsymbol{\Phi} : \mathbb{R}^2 \rightarrow \mathbb{R}^d$, where d is the dimensionality of the latent vector after PCA reduction. Specifically, $\boldsymbol{\Phi}(\cdot)$ outputs a latent vector $\boldsymbol{\phi}$ for a given coordinate (x, y), which can then be used to control the image generation.

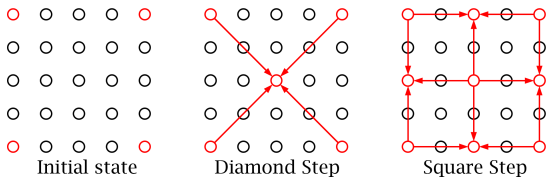

Figure 5: The Diamond-Square algorithm, which recursively interpolates on a spatial grid, is used to generate latent embeddings in our approach. The red arrows start from the vertices of the existing square and diamond shapes from the previous iteration, and point towards the new center points.

The latent function can be seen as a sample from a fractal process, generated from the *Diamond-Square* Algorithm applied to estimate the function output over a dense grid that covers the desired region. Here, the outputs are estimated recursively through a recursive two step process. First, in the *diamond step* we estimate the function value at the spatial mid-points of each square regions using the four corners of each square - forming four new diamonds. Next, we apply a *square step*, to estimate the mid-points of diamond regions from the corner points of each diamond — forming squares that subdivided the original square. In each step, we compute the latent vector values at the centers of square and diamond shape patterns as the mean of the corner points of the regions plus some random noise. Let us denote the set of vertices of a square or diamond shape as the set K, and the center point of the square or diamond as $\mathbf{r}_c$, the latent vector value at the center is given by

$$\mathbf{\Phi}(\mathbf{r}_c) = \frac{1}{|K|} \sum_{\mathbf{r} \in K} \mathbf{\Phi}(\mathbf{r}) + s\boldsymbol{\sigma}, \qquad \boldsymbol{\sigma} \sim \mathcal{N}(0, \mathbf{I}). \tag{2}$$

Here, s is a scaling factor that controls the variability of the landscape. This factor s is gradually decayed. Therefore, starting with latent vector values at the vertices of a square, we can recursively estimate latent vector values over the entire square region.

A single iteration of this process, along with illustrated vertices, is shown in Figure 5. The end result of this step is a 2D spatial field of latent fractal embeddings that can be used to conditionally generate a set of images with strong spatial dependency.

To accomplish this, we train a diffusion model using RGB images from real underwater imagery augmented with depth generated using Depth Anything v2 Yang et al. (2024). The resulting model is used to generate an RGBD image for each vertex in the spatial latent field and then RePaint Lugmayr et al. (2022) is used to in-fill any gaps between each pair of neighboring images, to form a spatially consistent map in the form of an RGBD point cloud.

Here, we highlight that the function of images over the 2D spatial domain is drawn from a *doubly stochastic process*. The set of generated images, $\{\mathbf{I}_\mathbf{x}\}_{\mathbf{x} \in \mathbb{R}^2}$, can be considered as a function drawn from the conditional diffusion model, which itself is dependent on a latent function, $\mathbf{\Phi}(\mathbf{x})$, drawn from a fractal process, governed by the scale factor s. Specifically,

$$\{\mathbf{I}_\mathbf{x}\}_{\mathbf{x} \in \mathbb{R}^2} \sim \underbrace{P(\mathbf{I}|\mathbf{\Phi}(\mathbf{x}))}_{\text{Diffusion Model}}, \qquad \mathbf{\Phi}(\mathbf{x}) \sim \underbrace{P(\mathbf{\Phi}|s)}_{\text{Fractal Process}}. \tag{3}$$

We note that the doubly stochastic nature of our image generation enables highly diverse terrains to be generated.

3.4 3D Scene Generation via Gaussian Splatting

In this section, we convert the RGBD point cloud generated in the previous step into a geometrically-consistent 3DGS model that uses the generated images as a strong prior. The resulting model provides us with a 3D structure that is dense and allows for the generation of novel images from arbitrary viewing poses.

We begin by using the depth channels from the generated images to initialize 3D Gaussians following the default method Kerbl et al. (2023). Then we freeze the 3D positions of the Gaussian cloud and refine the appearance with 2D diffusion priors. Given a cloud of Gaussians $\mathbf{G}$ initialized, each Gaussian g_i includes the following attributes: position $\mathbf{p}_i$, covariance Σ_i, opacity α_i and radiance

$\mathbf{c}_i$, that $g_i = \{\mathbf{p}_i, \Sigma_i, \alpha_i, \mathbf{c}_i\} \in \mathbf{G}$. With a subset of Gaussians $\mathcal{N} \in G$ ordered along a camera ray, the pixel value in an image can be rendered from 3DGS models with the following rendering equation:

$$C = \sum_{i \in \mathcal{N}} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{4}$$

Here $\mathbf{p}_i$ is initialized from the depths of the generated images and frozen when optimizing the Gaussians. Our reasons for doing so are three-fold: 1. Our point cloud is already sufficiently dense; 2. optimizing position often comes with Gaussian duplication operations leading to memory overflow for large generated scenes; and 3. supervising the geometry with an up-to-scale depth diffusion model is not well studied. We use the *Score Distillation Sampling* (SDS) loss introduced in DreamFusion Poole et al. to optimize the 3D Gaussian model from 2D diffusion prior:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\mathbf{I}^r) \triangleq \mathbb{E}_{t,\epsilon} \left[w(t) \left(\hat{\epsilon}(t) - \epsilon \right) \frac{\partial \mathbf{I}^r}{\partial \theta} \right] \tag{5}$$

here θ is the parameters of Gaussian cloud $\mathbf{G}$ to be optimized, $\mathbf{I}^r$ is the rendered image; $\hat{\epsilon}$ and ϵ are predicted noise and added noise; t is the timestep in the diffusion process and $w(t)$ is the weighting function following the implementation in Poole et al. (parameter y and $\mathbf{z}_t$ in the original paper are omitted here for brevity).

4 Experiments

4.1 Datasets

The results presented throughout the paper are trained on real-world data collected from four different locations with three different robot platforms, spanning a time from 2009 to 2024 (see Figure. 6). The *Scott Reef* and *Batemans datasets* were collected from 2009 to 2015 with a Seabed-class AUV, Sirius, which features a dual-hull design for stabilized imaging underwater. We post-process the raw images, hosted on `Squidle.org`, to have normal exposure. The *Hawaii dataset* was collected in April 2024 with an Iver AUV, the torpedo design allowed it to travel long distances and sample images from the seafloor. The *Florida dataset* was collected in August 2023 with a customized remotely operated vehicle (ROV) equipped with ZED cameras. Each location presents a unique benthic appearance and is reflected in our model.

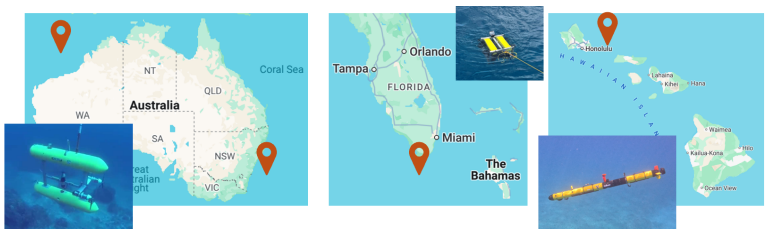

Figure 6: Results demonstrated in this paper are trained on data collected from 4 different sites with 3 different robot platforms.

4.2 Implementation Details

Our model's implementation is adapted from DDPM networks. We train each model on a single NVIDIA RTX4090 GPU with 24GB VRAM for 2000 epochs, with a batch size of 12. Although the size of each data set differs, it usually takes $\sim$ 200 hours to train on a dataset with 10k images, at the resolution of 224×224. We use the first two main components from PCA results on DINO v2 embeddings. From our empirical study, we find it to be sufficient to describe the variation in appearance of underwater environments. This is consistent with the practice in Zhang & Johnson-Roberson (2022); Qin et al. (2024).

4.3 QUALITATIVE EVALUATION

We train the model on the dataset collected from various locations capturing diverse underwater appearances. At a glance, the generated images resemble the real images well, as shown in Figure 7. The generated relative depth also aligns well with human perception, indicating that our training pipeline successfully learns the visual distribution of real underwater datasets and distills the 3D information from the depth foundation model.

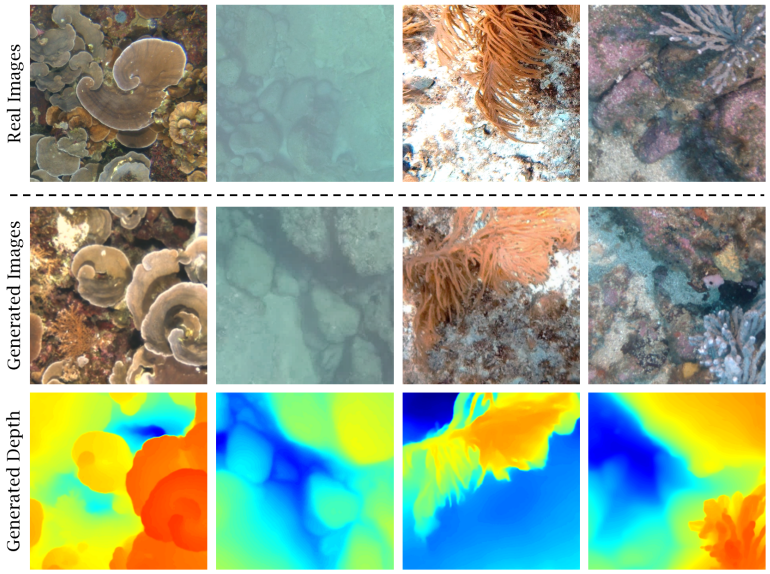

Figure 7: Our diffusion model is able to output realistic images as well as depth estimation distilled from depth anything v2 Yang et al. (2024).

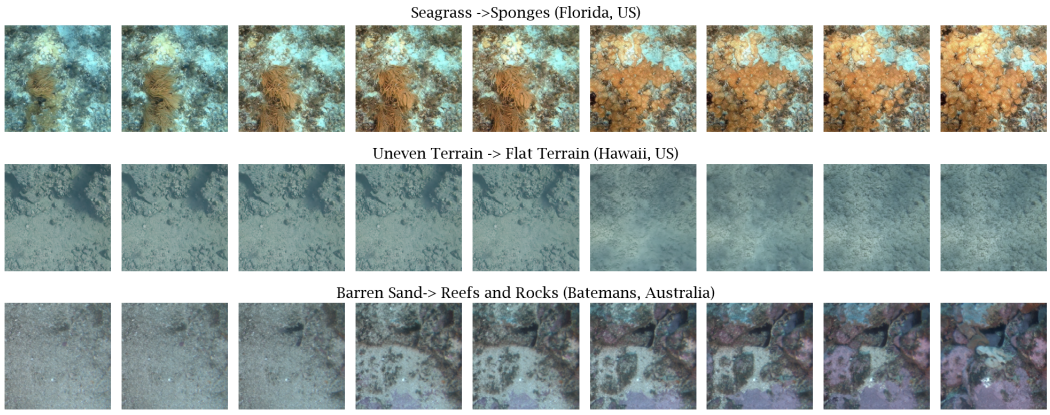

Figure 8: Examples of image generation conditioned on interpolated DINO embeddings. A smooth transition can be observed.

4.4 LATENT CONTROLLED GENERATION

Generating images and maps with latent embedding control plays a critical role in creating terrain with appearance aligned with human preference and natural variation. We demonstrate a smooth image transition over the latent space: Figure 8 shows images generated with latent embedding interpolated in latent space. We can see how the appearance of the images smoothly transits and we can recognize how the content of the image shifts smoothly, demonstrating that latent embeddings from VFMs controls underwater image generation smoothly and can be well aligned with human perception.

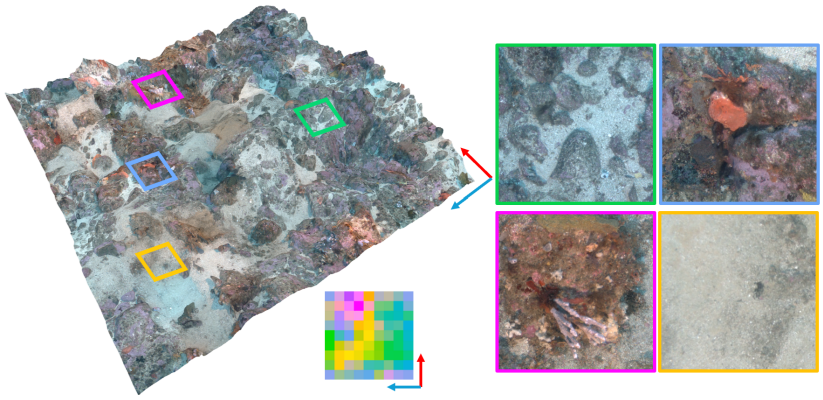

Figure 9: Latent Controlled Generation on fractal embeddings, with $s = 0.6$. Diversity observed even locally.

We further show the 2D map generated from a fractal latent field. In Figure 9, where the latent field is generated with $s = 0.6$. We observe that the added stochasticity injected into the latent process visibly enhances the diversity of the generated terrain. In the $s = 0$ case, the generated patterns repeat locally, while when $s = 0.6$, we observe diverse patterns and elevations even when considering a local region. This locally diversity can be governed by tuning the scale factor s, further motivating our doubly stochastic formulation.

5 Limitations and Opportunities

Our current model only estimates relative as opposed to metric scale. The metric scale could optionally be acquired by auxiliary sensors such as IMUs, calibrated cameras, calibration targets, or single/multi-beam acoustic sensors.

Viewing angles are only from the top down. Although the datasets we use are collected with different robot platforms, they are all from top-down view. This is constrained by the fact that each robot is designed to be passively stable in a hydrodynamic environment. This work further motivates the design of new robot and perception systems to allow for more diverse viewing angles Liu et al. (2024).

It will also be useful to generate images which can integrate partial expert annotations to semi-supervise DreamSea. Determining how to bridge such a system with broader marine science, biography and geography community is still an open problem.

6 Conclusion

Generating realistic and diverse underwater terrains and scene representations has a wide variety of applications, spanning video games, movies, robotics, and marine science. Existing generative methods struggle to generate sufficiently varied and physically accurate underwater images. To tackle this, we introduce *DreamSea*, a diffusion-based generative model which we train on a collection of large-scale unannotated underwater imagery collected by robots at different locations. Our approach conditions generation upon visual latent embeddings extracted using foundation models. Furthermore, DreamSea imbues spatial awareness into the generative model via a novel fractal embedding algorithm. The resulting terrain generation allows for the generation of highly diverse underwater environments, while considering spatial-dependencies. The resulting terrain visuals and estimated depths are integrated as priors to construct 3DGS models, which provide 3D geometry and enable novel-view images to be produced. DreamSea is rigorously evaluated and demonstrates the capability to generate large-scale hyper-realistic underwater scenes.

REFERENCES

Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M. Seitz, and Rick Szeliski. Building rome in a day. *Communications of the ACM*, 54:105–112, 2011.

Abdelhakim Amer, Olaya Álvarez-Tuñón, Halil İbrahim Uğurlu, Jonas Le Fevre Sejersen, Yury Brodskiy, and Erdal Kayacan. Unav-sim: A visually realistic underwater robotics simulator and synthetic data-generation framework. In *2023 21st International Conference on Advanced Robotics (ICAR)*, pp. 570–576. IEEE, 2023.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.

Alain Fournier, Don Fussell, and Loren Carpenter. *Computer rendering of stochastic models*, pp. 189–202. Association for Computing Machinery, New York, NY, USA, 1998. ISBN 158113052X. URL https://doi.org/10.1145/280811.280993.

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

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3749–3761, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Eduardo Iscar, Corina Barbalata, Nicholas Goumas, and Matthew Johnson-Roberson. Towards low cost, deep water auv optical mapping. In *OCEANS 2018 MTS/IEEE Charleston*, pp. 1–6, 2018. doi: 10.1109/OCEANS.2018.8604772.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:1, 2014.

Xinyi Liu, Tianyi Zhang, Matthew Johnson-Roberson, and Weiming Zhi. Splatraj: Camera trajectory generation with semantic gaussian splatting, 2024. URL https://arxiv.org/abs/2410.06014.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11461–11471, June 2022.

B.B. Mandelbrot. *The Fractal Geometry of Nature*. Einaudi paperbacks. Henry Holt and Company, 1983. ISBN 9780716711865. URL https://books.google.com/books?id=0R2LkE3N7-oC.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, December 2021. ISSN 0001-0782. doi: 10.1145/3503250. URL https://doi.org/10.1145/3503250.

Gavin S P Miller. The definition and rendering of terrain maps. In *Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '86, pp. 39–48, New York, NY, USA, 1986. Association for Computing Machinery. ISBN 0897911962. doi: 10.1145/15922.15890. URL https://doi.org/10.1145/15922.15890.

Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. doi: 10.1109/TRO.2015.2463671.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Ken Perlin. An image synthesizer. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '85, pp. 287–296, New York, NY, USA, 1985. Association for Computing Machinery. ISBN 0897911660. doi: 10.1145/325334.325247. URL https://doi.org/10.1145/325334.325247.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20051–20060, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12630–12641, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Jingyu Song, Onur Bagoren, Razan Andigani, Advaith Sethuraman, and Katherine A. Skinner. Turtlmap: Real-time localization and dense mapping of low-texture underwater environments with a low-cost unmanned underwater vehicle. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1191–1198, 2024. doi: 10.1109/IROS58592.2024.10801692.

Jingyu Song, Haoyu Ma, Onur Bagoren, Advaith V. Sethuraman, Yiting Zhang, and Katherine A. Skinner. Oceansim: A gpu-accelerated underwater robot perception simulation framework, 2025. URL https://arxiv.org/abs/2503.01074.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *The Twelfth International Conference on Learning Representations*.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *CVPR*, 2024.

Ziwen Yuan, Tianyi Zhang, Matthew Johnson-Roberson, and Weiming Zhi. Photoreg: Photometrically registering 3d gaussian splatting models. *CoRR*, abs/2410.05044, 2024. doi: 10.48550/ARXIV.2410.05044. URL https://doi.org/10.48550/arXiv.2410.05044.

Tianyi Zhang and Matthew Johnson-Roberson. Learning cross-scale visual representations for real-time image geo-localization. *IEEE Robotics and Automation Letters*, 7(2):5087–5094, 2022. doi: 10.1109/LRA.2022.3154035.

Jiaxi Zheng, Guangmin Dai, Botao He, Zhaoyang Mu, Zhaochen Meng, Tianyi Zhang, Weiming Zhi, and Dixia Fan. Rs-modcubes: Self-reconfigurable, scalable, modular cubic robots for underwater operations. *IEEE Robotics and Automation Letters*, 10(4):3534–3541, 2025. doi: 10.1109/LRA.2025.3543139.

Weiming Zhi, Haozhan Tang, Tianyi Zhang, and Matthew Johnson-Roberson. Unifying representation and calibration with 3d foundation models. *IEEE Robotics and Automation Letters*, 9(10), 2024a. doi: 10.1109/LRA.2024.3451396.

Weiming Zhi, Haozhan Tang, Tianyi Zhang, and Matthew Johnson-Roberson. Simultaneous geometry and pose estimation of held objects via 3d foundation models. *IEEE Robotics and Automation Letters*, 9(12), 2024b. doi: 10.1109/LRA.2024.3501677.