GENERATIVE ADVERSARIAL NETWORKS VS. STABLE DIFFUSION: A COMPARATIVE ANALYSIS FOR IMAGE GENERATION

Sage Woodard
Kansas State University
sagewoodard@ksu.edu

INTRODUCTION

The field of Generative AI has seen rapid advancements, particularly with the development of deep learning models capable of generating realistic images, videos, and other forms of data. Among these, Generative Adversarial Networks (GANs) and Stable Diffusion Models (SDMs) have emerged as leading approaches for image generation tasks [5], [7].

Satellite imagery generation poses unique challenges due to its reliance on geospatial and spectral precision. Models must not only produce visually realistic outputs but also adhere to specific land-use and land-cover characteristics present in real-world data [3]. While GANs are widely used for generative tasks, they often suffer from instability during training and difficulties in producing fine details [2], [5]. Stable Diffusion, a newer latent diffusion-based approach, offers a potentially more robust alternative by operating in latent space to reduce computational complexity while maintaining image quality [7].

This comparison is driven by the need to identify the most effective generative model for use in video game terrain generation and VR applications, where realistic and dynamic imagery is essential. Fréchet Inception Distance (FID) is employed as the primary metric to evaluate image quality [8], supplemented by qualitative visual inspection. These metrics provide insights into how closely the generated images resemble the original EuroSAT dataset [3].

The broader objective of this research is to pave the way for future applications in geospatial deep learning, including 3D terrain reconstruction and real-time rendering. This study lays the foundation for integrating GANs and Stable Diffusion Models with high-resolution satellite data and advanced geospatial techniques, aligning closely with data science objectives and the evolving demands of AI-driven industries [13], [15].

BACKGROUND AND RELATED WORK

Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. in 2014, are a class of generative models that consist of two neural networks: a Generator and a Discriminator [5]. The Generator creates synthetic data, while the Discriminator evaluates its authenticity by distinguishing real from fake samples. These networks are trained adversarially, with the Generator improving its outputs based on feedback from the Discriminator. GANs have found applications in image synthesis, data augmentation, and style transfer.

However, traditional GANs often encounter challenges such as instability during training and mode collapse. Variants like Wasserstein GAN (WGAN) and its improved version, WGAN-GP, address these issues by using a Wasserstein loss function with gradient penalties to stabilize the training process and ensure continuity [2], [10]. This approach has proven particularly effective for generating high-quality images.

Stable Diffusion Models (SDMs)

Stable Diffusion, introduced by Rombach et al. in 2022, represents a shift from pixel-based generative models to latent diffusion models [7]. Unlike GANs, which directly generate images, Stable Diffusion operates in a latent space—a compressed representation of image data. This approach significantly reduces computational overhead while enabling the generation of high-resolution images.

Stable Diffusion uses a diffusion process that iteratively denoises a latent representation, guided by input conditions such as text prompts or images. This method has proven effective for tasks like text-to-image generation, image-to-image translation, and inpainting. Its computational efficiency and versatility make it a strong contender for applications requiring high-resolution outputs, including geospatial image synthesis [7], [15].

EuroSAT Dataset

The EuroSAT dataset is a benchmark for land-use and land-cover classification, comprising 27,000 labeled satellite images across 10 classes, including urban, agricultural, and forested regions [3]. Derived from Sentinel-2 satellite data, the dataset features 13 spectral bands, enabling robust analysis of land-use patterns. For this project, only the RGB bands were utilized,

as they align with the capabilities of GANs and Stable Diffusion. The EuroSAT dataset has been widely used in geospatial machine learning, serving as a standard for evaluating image classification and generative models [3], [13].

RELATED STUDIES

Several prior studies have laid the groundwork for this research:

- Wasserstein GAN: Arjovsky et al. (2017) demonstrated that WGANs outperform traditional GANs in terms of stability and convergence for image generation tasks [2].
- FID Metric: Heusel et al. (2017) introduced Fréchet Inception Distance (FID) as a robust metric to evaluate the similarity between generated and real images [8]. FID has since become a standard for assessing generative model performance.
- Latent Diffusion: Rombach et al. (2022) showcased the effectiveness of latent diffusion models for generating high-quality, high-resolution images across various domains, including geospatial applications [7].
- Hugging Face Diffusers: The Hugging Face library provides predefined pipelines for Stable Diffusion, facilitating its application to diverse datasets, including EuroSAT [15].

METHODOLOGY

Dataset Preparation

The EuroSAT dataset was selected as the benchmark for this study, with the analysis focused solely on the forest class. This decision ensures a consistent evaluation of generative models within a specific land-use category while leveraging the dataset's detailed satellite imagery. The dataset consists of 27,000 labeled images, with RGB bands extracted from Sentinel-2 satellite data [3].

For this project:

- Image Preprocessing: All images were resized to 256×256 pixels, normalized to a range of [0,1], and divided into training (80%), validation (10%), and test (10%) sets.
- Format and Augmentation: Images were prepared in formats compatible with TensorFlow for GANs and PyTorch for Stable Diffusion. Basic augmentation techniques, such as rotation and horizontal flipping, were applied to enhance model robustness during training.

Model Implementation

Generative Adversarial Networks (GANs)

The GAN model used a WGAN-GP framework, implemented in TensorFlow. The architecture included:

1. Generator:
    - A convolutional neural network designed to synthesize 256×256 images from random noise vectors.
    - Optimized using Wasserstein loss, encouraging the production of realistic images by maximizing the Discriminator's evaluation [2], [10].
2. Discriminator:
    - A convolutional network tasked with distinguishing between real and generated images.
    - Trained using Wasserstein loss with gradient penalties to improve stability and continuity during training [2].

Hyperparameters:

- Learning rates: 1×10^{-5} (Generator), 1×10^{-6} (Discriminator).
- Optimizer: Adam with $\beta 1=0.5$ and $\beta 2=0.9$.
- Epochs: 50 with a batch size of 16.

Stable Diffusion Models (SDMs)

Stable Diffusion was implemented using the Hugging Face Diffusers library, leveraging pre-trained pipelines to fine-tune the model on the EuroSAT dataset. The approach involved:

1. Latent Space Operation:
    - Stable Diffusion operates in compressed latent space rather than pixel space, improving computational efficiency and enabling high-resolution synthesis [7].
2. Denoising Process:
    - The model iteratively denoises a random latent vector, guided by input conditions, to produce coherent outputs [15].

Hyperparameters:

- Pre-trained weights from Hugging Face's Stable Diffusion implementation.
- Fine-tuning for 10 epochs on the EuroSAT dataset with a learning rate of 5×10^{-5}.
- Batch size: 8.

EVALUATION METRICS

Fréchet Inception Distance (FID)

The FID metric is a standard method for evaluating the quality of generated images by comparing the statistical similarity between real and generated image features. It calculates the mean and covariance of features extracted from a pre-trained Inception model for both datasets. Lower FID scores indicate a closer match between the generated and real data distributions [8].

The formula for Fréchet Inception Distance (FID) is:

$$FID = \left|\left|\mu_r - \mu_g\right|\right|^2 + Tr(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{\frac{1}{2}})$$

Where:

- μ_r and Σ_r: Mean and covariance of features for real images.
- μ_g and Σ_g: Mean and covariance of features for generated images.

This metric is particularly effective in detecting subtle differences in high-dimensional data, making it ideal for tasks involving satellite imagery [8], [12].

Visual Inspection

While quantitative metrics like FID provide a numerical measure of performance, visual inspection offers a qualitative evaluation of generated images. It involves analyzing the images for:

- Realism: Do the images resemble real-world satellite data?
- Fidelity: Are the images free from visual artifacts?
- Class Accuracy: Do the generated images correspond correctly to their intended land-use class?

Generated images from the GAN and Stable Diffusion models were compared side-by-side with original EuroSAT samples to evaluate their alignment with the dataset's characteristics [3], [7].

Computing Environment

All experiments were conducted using Google Colab with an NVIDIA A100 GPU. Colab provided sufficient computational resources for training GANs and fine-tuning Stable Diffusion while maintaining accessibility.

EXPERIMENT DESIGN AND RESULTS

Experimental Setup

The experiment involved training and evaluating both GAN and Stable Diffusion models on the EuroSAT dataset. All experiments were conducted in a controlled environment using Google Colab with an NVIDIA A100 GPU for efficient computation. The dataset was preprocessed and split into training (80%), validation (10%), and test (10%) sets to ensure robust evaluation [3].

Generative Adversarial Networks (GANs)

The GAN model was implemented using a WGAN-GP framework with TensorFlow. Key training parameters included:

- Loss Function: Wasserstein loss with gradient penalties, which ensured stability and continuity during training [2], [10].
- Hyperparameters:
    - Generator learning rate: 1×10^{-5}, Discriminator learning rate: 1×10^{-6}.
    - Optimizer: Adam with $\beta 1=0.5$, $\beta 2=0.9$.
    - Batch size: 16; Epochs: 50.

Stable Diffusion Models (SDMs)

The Stable Diffusion Model was implemented using the Hugging Face Diffusers library, leveraging pre-trained pipelines fine-tuned on the EuroSAT dataset. Training specifics included:

- Latent Space Operation: Reduced computational requirements while maintaining high-resolution synthesis [7].
- Hyperparameters:
    - Fine-tuning learning rate: 5×10^{-5}.
    - Batch size: 8; Epochs: 10.

RESULTS

Quantitative Results

The Fréchet Inception Distance (FID) was calculated exclusively for the forest class in the EuroSAT dataset to provide a focused comparison between GAN and Stable Diffusion outputs. The results are as follows:

- GAN: 1846.37
- Stable Diffusion: 165.74

The significant difference in FID scores demonstrates the superior performance of Stable Diffusion, which produced images with much closer feature distributions to the real dataset. The s-value analysis further highlights this gap:

- s-value (Difference): 1680.63
- s-value (Ratio): 11.14

Qualitative Results

Visual inspection focused solely on forest images from the EuroSAT dataset to maintain consistency and evaluate the models in a specific land-use class. The results are as follows:

- GAN: Generated images displayed noticeable artifacts, including visual imperfections such as blurriness, inconsistent textures, and irregular pixel patterns. These artifacts resulted in images that lacked fine details and failed to accurately replicate the natural textures typically found in forested areas.
- Stable Diffusion: Produced sharper, more coherent images that captured the texture and overall characteristics of forest regions more effectively than the GAN model. The latent diffusion approach enabled finer details and better color consistency.

Sample comparisons of generated and original images illustrate these differences (refer to Figures 1, 2a, and 2b).

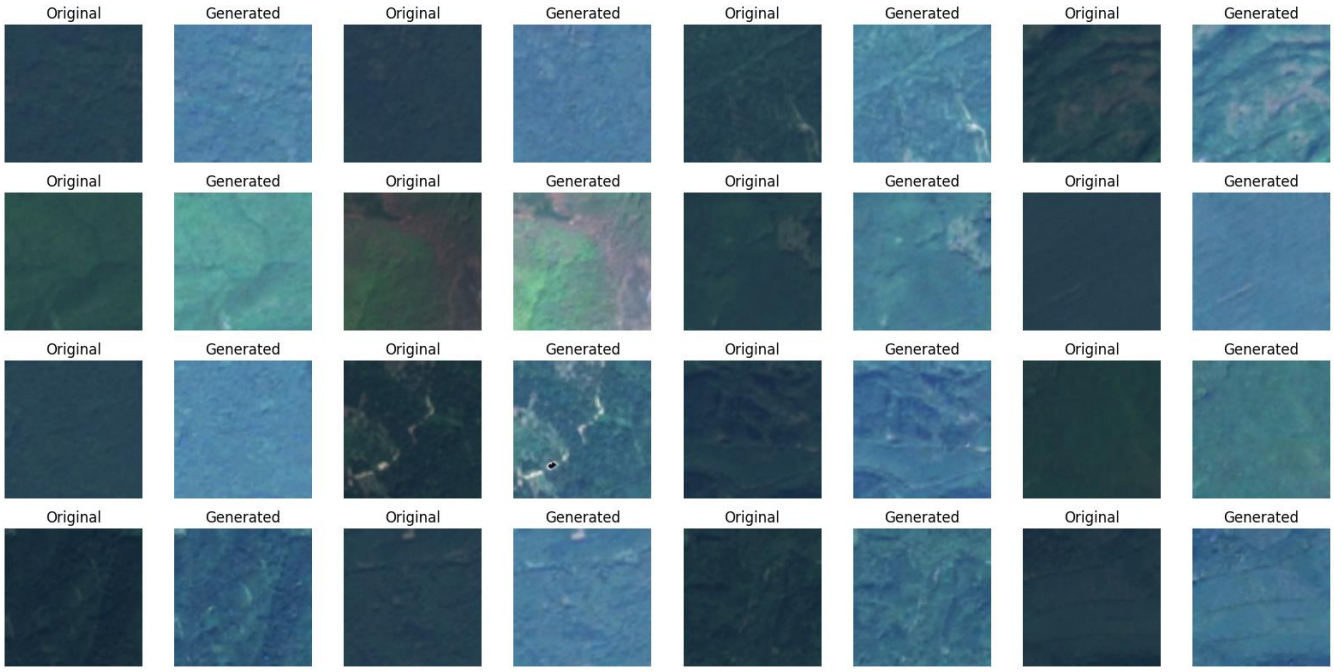
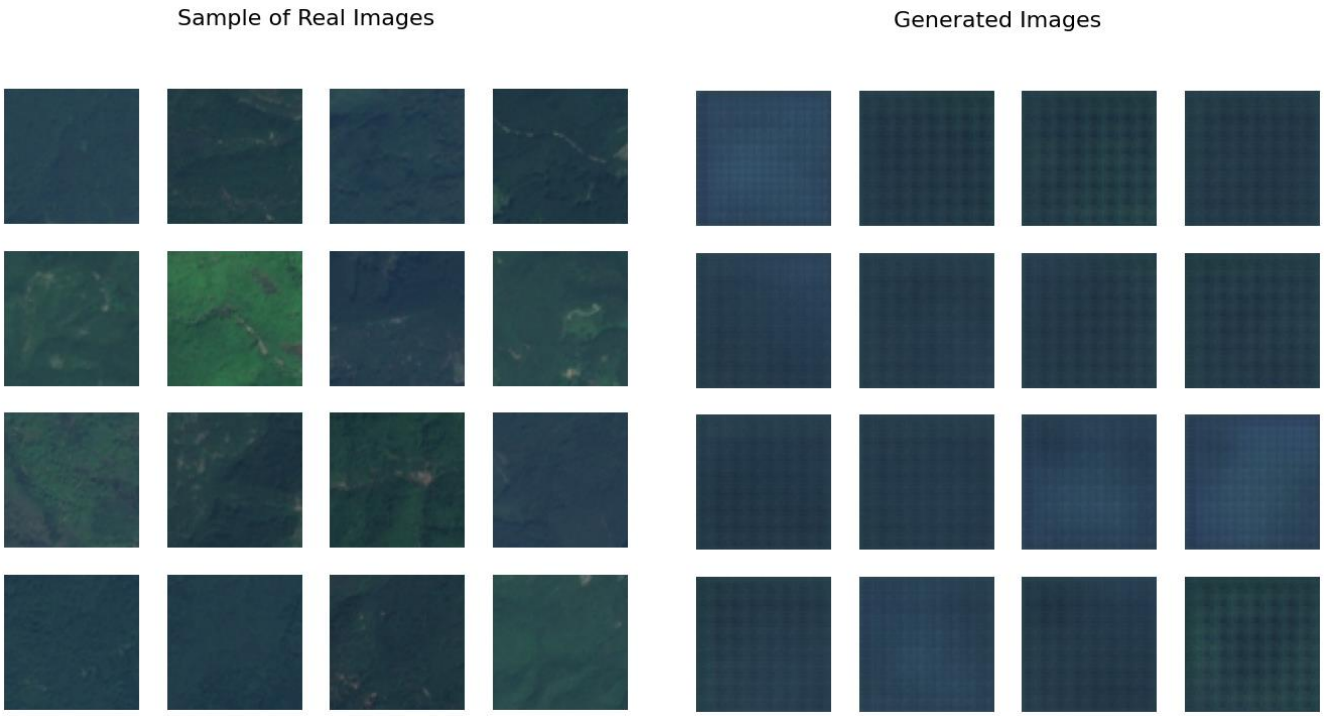


Figure 1: Comparative sample of real EuroSAT images and GAN generated images.



Figures 2a & 2b: Comparative samples of real EuroSAT images and Stable Diffusion generated images.

Discussion of Qualitative and Quantitative Alignment

The experiment highlights the limitations of GANs for satellite imagery generation, particularly in capturing high-resolution details and avoiding artifacts. In contrast, Stable Diffusion's latent space approach allowed for more robust and realistic outputs, making it better suited for tasks involving satellite imagery [3], [7].

SUMMARY AND FUTURE WORK

Summary

This study evaluated the performance of Generative Adversarial Networks (GANs) and Stable Diffusion Models (SDMs) for generating synthetic satellite images using the EuroSAT dataset. Through a combination of quantitative analysis using the Fréchet Inception Distance (FID) and qualitative evaluation via visual inspection, the findings highlight significant differences between the two approaches.

- GANs:
  - Despite being a well-established generative model, GANs struggled with generating high-resolution images of satellite data.
  - The FID score of 1846.37 indicates a considerable gap between real and generated image distributions [2], [8].
- Stable Diffusion:
  - Demonstrated a clear advantage with its latent diffusion process, producing sharper, more coherent images that closely resembled real satellite data.
  - Achieved an FID score of 165.74, over 11 times better than the GAN model [7], [15].

The qualitative results also validate these findings, as Stable Diffusion consistently outperformed GANs in generating images that aligned well with the real-world classes in the EuroSAT dataset [3], [7].

These results demonstrate the potential of Stable Diffusion for satellite imagery generation, particularly for applications requiring high resolution and fine detail, such as terrain generation in video games and VR environments.

Future Work

Building on this research, the following directions are proposed to expand the application of generative AI in geospatial tasks:

1. Integration of Geospatial Deep Learning:
   - Utilize tools like TorchGeo to incorporate contextual geospatial data, such as elevation, slope, and land cover, into the generative models [13].
2. High-Resolution Data:
   - Expand the dataset to include high-resolution satellite imagery from platforms like Google Earth Engine to improve the detail and applicability of generated outputs [3].
3. Hybrid Model Approaches:
   - Investigate combining the strengths of GANs and Stable Diffusion to leverage the complementary advantages of both models, particularly for generating realistic satellite data with high variability.
4. 3D Terrain Reconstruction:
   - Explore deep learning-based 3D reconstruction techniques to transform generated 2D images into dynamic 3D terrains suitable for real-time rendering in video games and VR applications [7], [15].
5. Real-World Applications:
   - Extend the applicability of these models to broader geospatial tasks, including urban planning, environmental monitoring, and disaster response scenarios.

These advancements will enhance satellite image realism and align with current research goals.

References

[1] M. Abadi, et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," *Software Available from tensorflow.org*, 2016. Available: https://www.tensorflow.org/.

[2] M. Arjovsky, et al., "Wasserstein GAN," *arXiv preprint*, arXiv:1701.07875, 2017.

[3] P. Helber, et al., "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019. Available: https://arxiv.org/abs/1709.00029.

[4] I. Goodfellow, et al., "Generative Adversarial Networks," *Communications of the ACM*, 2014.

[5] I. Goodfellow, et al., *Deep Learning*. MIT Press, 2016.

[6] I. Gulrajani, et al., "Improved Training of Wasserstein GANs," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5767–5777, 2017.

[7] R. Rombach, et al., "High-Resolution Image Synthesis with Latent Diffusion Models," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. Available: https://arxiv.org/abs/2112.10752.

[8] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6626–6637, 2017.

[9] A. Paszke, et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Advances in Neural Information Processing Systems*, 2019. Available: https://pytorch.org/.

[10] A. Radford, et al., "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *arXiv preprint*, arXiv:1511.06434, 2015.

[11] M. Seitzer, "Pytorch-FID: Fréchet Inception Distance for PyTorch," *GitHub*, 2018. Available: https://github.com/mseitzer/pytorch-fid.

[12] TensorFlow, "Generate Images with Stable Diffusion," TensorFlow. Available: https://www.tensorflow.org/tutorials/generative/generate_images_with_stable_diffusion.

[13] TorchGeo Documentation. Available: https://torchgeo.readthedocs.io/en/latest.

[14] P. Von Platen, et al., "Diffusers: State-of-the-art Diffusion Models for Image Generation," Hugging Face, 2022. Available: https://huggingface.co/docs/diffusers.

[15] Z. Wang, et al., "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. Available: https://ieeexplore.ieee.org/document/1284395.