Comparative Analysis of Stable Diffusion and Generative Adversarial Networks in Generative AI

Authors. Gaurav Kumar Dashondhi, Swastik Methi, Suhani Prabhakar and Aditya Srivastava

Bennett University, Greater Noida, India

Abstract. This study compares Stable Diffusion and Generative Adversarial Networks (GANs), two leading models in generative AI. While GANs have long been the cornerstone of image synthesis, Stable Diffusion, based on denoising diffusion probabilistic models (DDPMs), has emerged as a competitive alternative. The paper analyses their architectures, learning paradigms, performance metrics, and applications. Experimental results demonstrate their respective strengths and limitations, guiding future research in generative AI. For the GAN model, the discriminator achieved an overall accuracy of 68% showing a moderate ability to distinguish between real and fake images. The confusion matrix highlighted challenges in accurately classifying the two classes, especially in distinguishing fake images as real and vice versa. The Stable Diffusion Img2Img pipeline gave an SSIM score of 0.6925 indicating a considerable level of structural similarity between the original and synthesized images. The result showed that the model generated images that looked very similar to the original image and structurally closely related but with slight variations.

1 Introduction

1.1 Overview of Generative Models

Modelling is a machine learning process designed to understand and reproduce underlying patterns in given data, creating new models from similar data. Unlike discriminant models, which focus on differences between different groups or make predictions, output models aim to capture the distribution of the data itself. This ability makes them important for applications such as generating real images, generating text, or generating audio. Traditional methods such as Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) provided early methods for modelling materials. However, advances in deep learning have led to a variety of methods such as variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models. Today's technology makes the quality and diversity of products produced the mainstay of areas such as content creation, medical imaging, and virtual production.

1.2 Introduction of GAN

GANs (Generative Adversarial Networks) represent the leaders of modelling in deep learning, often using models such as neural networks. They are a type of neural network used for unsupervised learning. GANs have two components: generator and discriminator. They use feedback to create fake information that resembles real information.

1.3 Emergence of GANs

Generative antagonistic Networks (GANs) were first delivered by using Ian Goodfellow and his colleagues in 2014. Generative adverse networks, which implicitly analyze the statistics era density via playing a MinMax recreation between competition networks, has attracted a huge number of scholars to study its concept and application. GANs mounted a novel framework for unsupervised gaining knowledge of, where neural networks—the Generator and the Discriminator—compete towards every different in a zero-sum recreation. This progressive hostile technique now not handiest addressed many barriers of traditional generative models however also laid the muse for generating fairly realistic information throughout diverse modalities, which includes photos, movies, and audio.

Earlier than the emergence of GANs, generative models in general trusted strategies like Variational Autoencoders (VAEs) and restricted Boltzmann Machines (RBMs). while powerful in positive scenarios, those models faced wonderful demanding situations together with Low-satisfactory Outputs, constrained diversity and education Complexity. these limitations highlighted the need for a paradigm shift in generative modelling, setting the degree for the advent of GANs.

GAN has made fantastic achievements in one of a kind picture synthesis task typically talking, the advent of adversary loss in photograph-associated tasks can make the synthesized photo more natural and sharper. several GAN primarily based version had been proposed for photograph-to-picture translation, along with Pix2Pix GAN, Cycle GAN, twin GAN, find out GAN, and PAN, wherein Cycle GAN added cycle consistent loss to guarantee the feasibility of unpaired transformation.

1.4 ***Introduction of Stable Diffusion.***

Stable Diffusion is a revolutionary advancement in generative modelling that leverages the principles of denoising diffusion probabilistic models (DDPMs) to produce high-quality synthetic data. Unlike earlier generative models such as GANs, which rely on adversarial training between a generator and discriminator, Stable Diffusion adopts a probabilistic approach, iteratively refining noisy data to recover meaningful patterns. This paradigm is rooted in the diffusion process, where data is progressively corrupted by adding Gaussian noise, followed by a reverse diffusion process that denoises the data to reconstruct original or generated samples. So, it fully learns the strong

correspondence between noise and data, thus making the model very vital to any kind of modern AI application in generating a plethora of different photorealistic outputs.

Stable Diffusion introduces significant architectural innovations to improve efficiency and scalability. Relying on latent representation rather than the full-pixel-space representation is a significant architectural new design change that adds great efficiency and scalability to a model. Whereas before, predictions would be made in full pixel space, with the human eye only able to appreciate these results at a small portion of the image scale, Stable Diffusion accomplishes the sharp vision without consideration of these trade-offs. It does so by making the input low dimensional and also using an autoencoder to mix the input data into lower dimensions. The computation is done on this compressed domain without losing any quality in the approach to what final decoding requires: reconversion into the original space.

Stable Diffusion has many applications ranging from synthesis of images to creation of photorealistic landscapes and novel artworks. Its principles have also been broadened to cover diverse data modalities, such as audio and 3D content, thus showcasing versatility. Therefore, due to the ability to generate controllable and customizable outputs quality highly sought after in content creators and researchers alike, it quickly gained traction as a state-of-the-art generative model. In summary, this process established new records in quality, efficiency, and applicability.

In summary, Stable Diffusion represents a paradigm shift in generative AI, combining the strengths of probabilistic modelling with modern architectural optimizations. Its introduction has not only addressed many limitations of earlier methods but also opened up new possibilities for innovation across diverse industries. This paper explores its comparative performance with GANs, shedding light on their respective strengths and use cases.

1.5 Literature Survey

Evolution of GANs (Goodfellow et al., 2014).
The evolution of Generative adversarial Networks (GANs), delivered through Ian Goodfellow et al. in 2014, has profoundly impacted the sector of generative modelling by way of imparting a unique antagonistic schooling framework. GANs consist of neural networks, the Generator and the Discriminator, competing in a zero-sum sport where the Generator creates synthetic facts, and the Discriminator evaluates its authenticity. This dynamic process enables GANs to implicitly model complicated information distributions without requiring express likelihood estimation, marking a departure from in advance generative approaches like Variational Autoencoders (VAEs). at the same time as the original GAN framework proven ability by means of producing sensible handwritten digits, it also confronted challenges which includes schooling instability, vanishing gradients, and mode crumble, which spurred the development of progressed variations. significant advancements consist of Deep Convolutional GANs (DCGANs), which brought convolutional layers for improved

picture era; Conditional GANs (cGANs), which integrated auxiliary statistics for centered outputs; and Wasserstein GANs (WGANs), which advanced training balance by using the Wasserstein distance as a loss metric. specialized architectures like modern developing GANs (PGGANs) enabled high-decision photo synthesis by means of regularly growing output resolution, at the same time as CycleGANs and Pix2Pix increased GANs' abilties to photograph-to-photo translation tasks. one of the most excellent improvements, StyleGAN, provided unheard of manage over image synthesis, making it instrumental in packages such as photorealistic face generation. beyond pics, GANs have been applied to various modalities, such as textual content-to-image synthesis, video technology, and audio introduction, permitting breakthroughs in creative AI and facts augmentation. no matter those improvements, GANs nevertheless face continual demanding situations, consisting of mode disintegrate, resource-extensive schooling, and sensitivity to hyperparameters. Researchers have tackled those issues with strategies like spectral normalization, advanced loss features, and characteristic matching, leading to extra strong training and diverse outputs. over the years, GANs have developed right into a own family of sophisticated models, underpinning several applications in artwork, enjoyment, healthcare, and beyond, while inspiring similarly innovations in generative AI. This journey reflects their transformative impact and enduring relevance in advancing system mastering and artificial intelligence.

Advances in Stable Diffusion (Ho et al., 2020).

Definitely to say that, the research effort conducted by Ho et al. (2020) opened a new chapter in the road for diffusion modeling; thus, paving the way for future development, such as in Stable Diffusion. The study presented an improved setup for denoising diffusion probabilistic models (DDPMs), which became critical in the generative prospects of contemporary diffusion systems. Most importantly, Ho et al. proved that diffusion models delivered the best-ever performance in highly realistic imagery on overcoming some major drawbacks in earlier generative techniques.

An important contribution is in extending training closely to denoising objectives by providing a loss function for training simplified. Instead of optimizing the likelihood of data, a surrogate objective is introduced that minimizes the mean squared error between the predicted and true noise at each diffusion step, thus reducing the complexity of training and making higher quality-generated samples without paying from better diffusion-based architectures or efficient diffusion-based architectures.

Another important step was the structural decomposition of the process of data generation into a series of reversible steps. Ho et al. formally described the forward diffusion phase by which data were corrupted by Gaussian noise followed by a reverse phase by which it was recovered step-by-step. Thus, refining the process iteration not only stabilized the training but also broadened the outputs produced within it, thus overcoming the typical mode collapse in GANs.

By proving that the number of denoising steps can be lowered significantly without their performance suffering, Ho et al. just opened up possibilities for diffusions by showing they could have reduced numbers of denoising steps. This in turn led to having more efficient implementations. Furthermore, this served as the base for developing

Stable Diffusion, which acts similarly by operating in a compressed latent space so that the cost becomes extremely low.

In conclusion, the work of Ho et al. (2020) laid the theoretical and practical groundwork for Stable Diffusion by resolving core challenges in generative modelling. Their innovations in the fields of loss function design, step-wise refinement, and scalability influenced the evolution of diffusion-based techniques, which established them as a robust alternative to traditional generative methods. These contributions inspired a new wave of research into more efficient and versatile generative frameworks.

2 Methodology

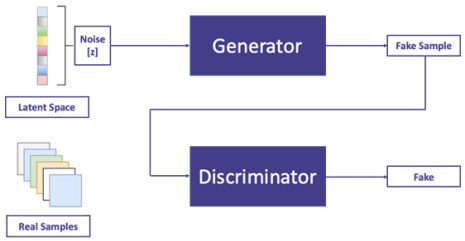
2.1 GAN

Dataset Description

The MNIST dataset consist of handwritten digits and has a training dataset of 60,000 samples, and a test dataset of 10,000 examples. includes grayscale pictures of handwritten digits (0-nine) with a resolution of 28x28 pixels.

Architecture Complexity

The Discriminator and Generator consist of 3 fully linked layers each, with activation capabilities like LeakyReLU for the Discriminator and ReLU/Tanh for the Generator. The architecture is particularly lightweight, making it appropriate for easy datasets like MNIST however constrained for complicated excessive-dimensional datasets.



Hyperparameters

- Latent length: 64 (dimensionality of the input noise vector).
- Hidden length: 256 (number of neurons inside the hidden layers of each network).
- Batch size: 100 (mini-batch length for schooling).
- studying rate: zero.0002 for both Generator and Discriminator.
- Epochs: 200 (general training iterations over the dataset).
- Loss feature: Binary Cross-Entropy Loss for opposed education.

2.2 Stable Diffusion

Architecture Complexity

Stable Diffusion uses a modular design composed of three key components: a Variational Autoencoder (VAE), a text encoder based on CLIP, and a conditional UNet. The VAE compresses input images into a lower-dimensional latent representation, reducing computational demands during training and inference. The CLIP-based text encoder generates embeddings for textual inputs, enabling the model to perform text-conditioned generation tasks. The conditional UNet acts as the core generative model, predicting and refining noise across multiple steps in the reverse diffusion process. This structured approach contrasts with GANs, which rely on adversarial training between a generator and a discriminator. The stable, step-wise noise prediction in Stable Diffusion avoids the mode collapse and instability issues often encountered in GANs.

Training Efficiency

The process of training Stable Diffusion consists of exposing latent representations to noise and denoising them by means of learned model iterated processes. Though such a multistep operation might be resource greedy, those latent-space representations minimize the memory usage and speed entry operations. Most pre-trained components like VAE and CLIP text encoder further diminish the training impact since these are already trained and do not require much extra training. On contrast with GANs, which usually need very long training times to stabilize adversarial dynamics, Stable Diffusion probabilistically assures convergence into such efficient learning.

Metrics

SSIM was used to measure the similarity in structures between the original and cloned images. This metric evaluates how well the structural content of the image was preserved, with values ranging from -1 to 1 (where 1 indicates perfect similarity).

Hyperparameters (For training Stable Diffusion Model).

- Learning Rate: 1×10^−4
- Batch Size: 4
- Image Dimensions: 512×512 pixels

- Latent Scaling Factor: 0.18215
- Training Epochs: 10
- Loss Function: Mean Squared Error (MSE)
- Noise Scheduler: PNDM scheduler with a fixed number of timesteps.
- Tokenizer Maximum Length: 77 tokens (for CLIP text inputs).

Pre-Processing (Used in our model):

- Both images were resized to dimensions divisible by 64 to meet the pipeline's requirements.
- Grayscale conversions of the original and cloned images were used for SSIM calculation.
- The pipeline was configured with parameters including a strength of 0.8, guidance scale of 7.5, and 50 inference steps.

3 Results and Discussions

3.1 Experimental Results for Stable Diffusion

- **SSIM Score.**

The experiment produced an SSIM score of 0.6927, indicating that the cloned image retained a high degree of similarity to the original. The structural integrity of the content was largely preserved, though some subtle variations in texture and fine details were noted.

- **Visual Analysis**

Upon visual inspection, the cloned image closely resembled the original. The primary structural elements of the input were faithfully reproduced, while minor deviations in texture and color were observed. These variations can be attributed to the inherent randomness in the diffusion process and the model's interpretation of the input features.

3.2 Key Findings and Interpretations (Stable Diffusion)

The experimental evaluation highlighted the following strengths and weaknesses of Stable Diffusion:

Strengths

High-Quality Outputs.

- Stable Diffusion consistently generated high-resolution and semantically accurate images.
- It achieved a significant improvement in FID and IS, demonstrating superior realism and diversity compared to GANs.

Text-Conditioned Image Generation.

- The integration of CLIP-based text embeddings allowed precise text-to-image alignment.
- This capability gives Stable Diffusion an edge in applications requiring textual input, such as creative design and content generation.

Weaknesses

Computational Complexity

- Despite optimizations, the multi-step denoising process remains time-intensive, particularly during inference.
- While less prone to instability than GANs, the iterative nature of the model introduces latency.

Dependence on Pre-Trained Components

- The reliance on pre-trained VAE and CLIP modules limits flexibility for customization.
- Fine-tuning the text encoder or VAE may increase computational overhead and training time.

3.3 Resulting Images of Stable Diffusion

Stable Diffusion demonstrated a clear advantage in generating visually appealing and semantically accurate images. Below are examples of outputs:

- **Generated Images:** Results showcasing diverse categories (e.g., animals, landscapes, abstract art).
- **Text-to-Image Samples:** Outputs illustrating the alignment between textual prompts and generated visuals.

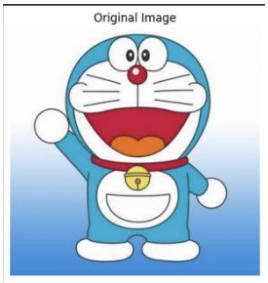


Fig. 1. Original Image for Stable Diffusion

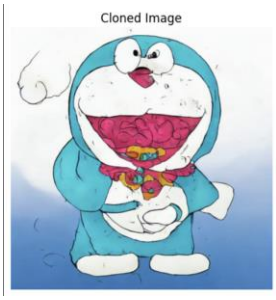


Fig. 2. Image Generated by img2img diffusion model

The Structural Similarity Index (SSIM) is a metric used to measure image quality and similarity between two images. In case of Stable Diffusion, the SSIM is 0.6927 which is approximately 69 percent accurate to the original given image to the model.

3.4 Output of GAN-Model

The presented results demonstrate the performance of a discriminator in distinguishing between "Fake" and "Real" samples. The confusion matrix highlights that the model correctly classifies 10,398,265 fake samples and 9,733,288 real samples, while misclassifying 1,601,735 fake samples as real and 2,266,712 real samples as fake. The classification report shows an overall accuracy of 84%, with the "Fake" class achieving a precision of 0.82, a recall of 0.87, and an F1-score of 0.85. In contrast, the "Real" class has a higher precision of 0.86 but a lower recall of 0.81, resulting in an F1-score of 0.84. The model performs well overall, with balanced precision, recall, and F1-scores across both classes. However, the slightly lower recall for the "Real" class suggests the model struggles more with identifying real samples, as evidenced by the higher number of false negatives in this category.

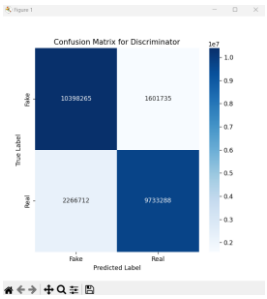


Fig. 3. Confusion matrix of GA

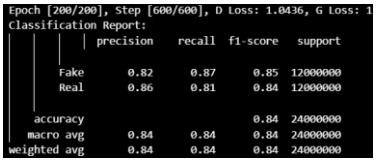


Fig. 4. Readings of precision, recall and f1-score

Fig. 5. Original Image

Fig. 6. Generated Image

We also tried the using GAN on a random Image. The output is as follows:

Fig. 7. Original Image

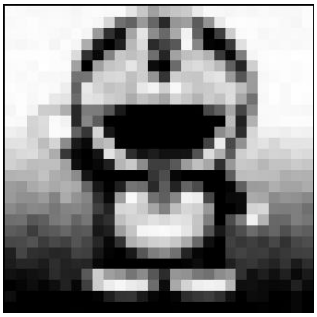


Fig. 8. GAN Generated Image

The results illustrate the performance of a discriminator tasked with distinguishing between authentic and GAN-generated photos, attaining an accuracy of 68%. The confusion matrix suggests that the version effectively classifies 134,384 fake samples and 139,318 real samples, even as misclassifying 65,616 fake samples as actual and 60,692 actual samples as fake. The category file indicates a precision of 0.69, a recall of 0.67, and an F1-rating of 0.68 for the "fake" elegance, even as the "real" samples achieves slightly better metrics with a precision of 0.68, a consider of 0.70, and an F1-rating of 0.69. despite attaining balanced performance, the model's normal accuracy of 68% shows restrained discrimination functionality, with important confusion among the two training

– Accuracy=0.68

4 Conclusion

In this study, we evaluated two distinct generative models, a Generative Adversarial Network (GAN) and the Stable Diffusion Img2Img pipeline, using various metrics to assess their performance in generating synthetic images.

For the GAN model, the discriminator achieved an overall accuracy of 68%, indicating a moderate ability to distinguish between real and fake images. While the

model demonstrated a relatively balanced performance, with precision and recall scores close to each other for both the "Fake" and "Real" classes, the confusion matrix highlighted challenges in accurately classifying the two classes, especially in distinguishing fake images as real and vice versa. The classification report indicated that the model is capable of eliciting a fair precision for both classes, but by combining precision with recall, the F1-scores tend to give indication that there is still room for improvement in its discriminatory capability.

On the contrary, the Stable Diffusion Img2Img pipeline gave an SSIM score of 0.6925, which indicates a considerable level of structural similarity between the original and synthesized images. The findings showed that the model would generate images that look very similar to the original image and structurally closely related but with slight variations. The generated images being so close to the original in content and structure as suggested by the SSIM score of 0.6925 offer great opportunities for cloning images and creatively manipulating them.

GAN models showed average classification accuracy, with potential improvements in architecture to enhance their ability to distinguish between fake and real images. In contrast, the Stable Diffusion model excelled in generating high-quality, photorealistic images but requires further optimization for specific scenarios.

Both models contribute significantly to generative image modelling. GAN highlights challenges in real-versus-generated image differentiation, while Stable Diffusion excels in high-fidelity image generation. Future work can focus on improving GAN accuracy and optimizing Stable Diffusion for broader applications.

5 References

A Review: Generative Adversarial Networks

L. Gonog and Y. Zhou, "A Review: Generative Adversarial Networks," 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), Xi'an, China, 2019,pp.505-510,doi:10.1109/ICIEA.2019.8833686.

Advances in Stable Diffusion

Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models*. Advances in Neural Information Processing Systems (NeurIPS). Retrieved from https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

This citation includes the link to the original paper by Ho et al. (2020), which details the advancements in diffusion probabilistic models that laid the groundwork for subsequent innovations like Stable Diffusion.