Comparative Analysis of GANs and Diffusion Models in Image Generation

Haotian Wang

School of Electronics and Computer Science, University of Southampton, Southampton, UK

hw26n22@soton.ac.uk

Abstract. Image generation has emerged as a rapidly evolving field with transformative applications in entertainment, medical imaging, virtual environments, and various other industries. This paper presents a thorough analysis of technological developments and current mainstream models in image generation, focusing on Generative Adversarial Networks (GANs) and diffusion models. It explores their theoretical underpinnings, advantages, limitations, and practical applications. The study highlights that GANs are particularly effective in producing diverse and high-quality images with notable speed but are often hindered by issues such as mode collapse and training instability. In contrast, diffusion models are praised for their ability to generate high-fidelity and detailed images, though they require extensive computational resources, making them less suitable for resource-constrained environments. The paper also discusses the current challenges faced by these models, such as biases and inefficiencies, and proposes potential solutions and future research directions. By addressing these issues and exploring ways to enhance model efficiency and multi-modal generation capabilities, this study aims to provide valuable insights that will drive further innovation and practical application in the field of image generation.

Keywords: Image Generation, Generative Adversarial Networks (GANs), Diffusion Models, Computational Resources.

1. Introduction

The generation of images encompasses both the creation of novel visuals and the modification of existing ones through algorithms that analyze extensive datasets to detect underlying patterns [1]. This technology serves diverse fields such as entertainment, art, healthcare imaging, and virtual environments, significantly boosting efficiency in these domains [2]. Recent years have witnessed rapid progress in image generation methods, driven by advancements in deep learning and neural network technologies [3]. Given the broad applications and swift evolution of these techniques, it is imperative to comprehensively assess their development, understand their latest achievements, and forecast their future trajectories.

The evolution of image generation technology traces its roots back to early rule-based systems and basic probability models. However, the advent of deep learning has revolutionized this field [2]. Variational Autoencoders (VAEs), introduced by Kingma and Welling in 2013, provided a novel approach to encoding images into latent spaces and generating new images from them [4]. In contrast, Goodfellow et al.'s Generative Adversarial Networks (GANs), proposed in 2014, have had a more profound impact compared to VAEs. The inception of this framework has spurred considerable exploration into enhancing visual accuracy and variety [1]. The diffusion framework, pioneered by Dhariwal and Nichol in 2021, has progressively improved noise structure, achieving notable results in generating highly detailed and realistic images [5]. Subsequent advancements by Ho and Salimans have further refined this approach through successive diffusion frameworks focused on high-fidelity image generation [6]. Transformer-driven models, such as OpenAI's DALL-E and Contrastive Language-Image Pre-Training (CLIP), have extended capabilities to synthesize visual content from textual descriptions by integrating language processing abilities [7, 8]. These models underscore the significance of multimodal image synthesis, where the system acquires proficiency in producing images from various input modes, including text, speech, and diverse non-visual data forms [8].

Each model exhibits distinct strengths and weaknesses. GANs, for instance, are widely applied in tasks like enhancing image resolution, inpainting missing parts, and altering styles due to their ability

to produce high-quality visuals [2]. Nevertheless, GANs commonly encounter challenges such as unstable training dynamics and pattern repetition, where the model generates a restricted range of images [9]. VAEs excel in generating a variety of visuals, yet their output often lacks the high fidelity seen in GANs [4]. Diffusion models have emerged as potent contenders capable of creating highly realistic images, albeit at the cost of substantial computational resources [5]. Transformer architectures such as DALL-E and CLIP push the boundaries further by enabling image synthesis directly from textual prompts, thus democratizing and simplifying the process of creating visuals [8]. Simultaneously, they impose rigorous requirements on computational capacity and memory, greatly complicating the processes of training and deployment [9]. Furthermore, iterative refinement techniques for image super-resolution have proven effective in enhancing both image quality and resolution, enhancing their utility in image synthesis [10].

The main aim of this research is to conduct a comprehensive analysis of current advances in image generation techniques and to explore potential future research directions. This study will delve into the fundamental principles and contextual foundations of image generation, focusing on the evolution of methodologies such as GANs and diffusion models. By analyzing these techniques in depth, the research will assess their strengths and limitations through empirical evaluations across diverse scenarios. Looking ahead, the study aims to identify potential areas for enhancement and breakthrough, thereby guiding future research efforts in advancing image generation technologies.

This section introduces the subject matter, outlining its historical context, significance, and research goals. Section 2 delves into the theoretical underpinnings of image generation methods. Section 3 evaluates empirical findings from different approaches, comparing their effectiveness. Finally, Section 4 concludes the investigation, discussing further investigation and potential innovations in the area of image generation.

2. Methodology

2.1. Dataset Description and Preprocessing

In image generation research, datasets such as the Canadian Institute for Advanced Study 10 (CIFAR-10), CelebFaces Attribute Dataset (CelebA), and Large-scale Scene Understanding (LSUN) are often used to train GAN and diffusion models. CIFAR-10 contains 32x32 images in 10 categories, totaling 60,000 images [5]. CelebA includes more than 200,000 images of celebrity faces, each labeled with 40 facial attributes [6], and LSUN provides millions of images in different categories, such as bedrooms and churches [7] (see Fig. 1, Fig. 2 and Fig. 3). These datasets are sourced from a common repository and require pre-processing steps such as normalization, resizing, and enhancement to ensure consistency and improve model performance. This preprocessing helps address variations in image quality and size, ultimately improving the effectiveness of the generated model in diverse scenarios, including image synthesis, super resolution, and style conversion [11, 12].

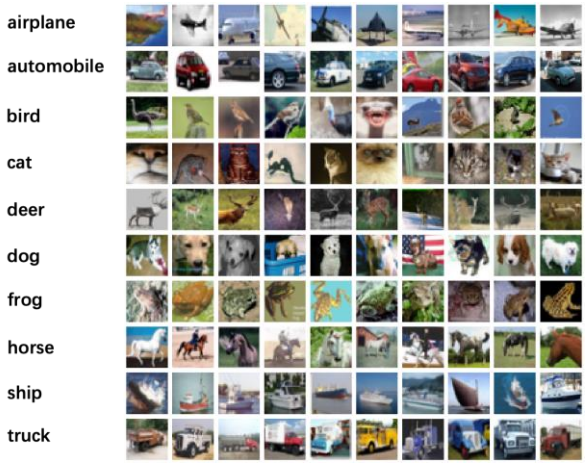

Figure 1. CIFAR-10 dataset sample images

Figure 2. CelebA dataset sample images

Figure 3. LSUN dataset sample images

2.2. Proposed Approach

This paper explores recent advancements in image generation, focusing on the principles, frameworks, and processes of GANs and diffusion models. The research aims to conduct an in-depth analysis of these core methods, examining their foundational principles and frameworks. It discusses the practical applications of GANs and diffusion models, evaluating their advantages, disadvantages, and development trajectories. Additionally, the paper assesses the comparative strengths and weaknesses of these technologies, explores their areas of application, and identifies areas needing improvement. Future prospects for GANs and diffusion models are also explored, presenting innovative insights and potential advancements. The structure of the paper, as illustrated in Fig. 4, begins with a thorough analysis of the concept of image generation, followed by a detailed discussion of core methods. It then reviews the practical applications of these models, provides a comparative analysis, and offers a forward-looking perspective. This comprehensive approach aims to deliver a deep understanding of current developments in image generation, offering valuable insights into the evolution and future directions of GANs and diffusion models.

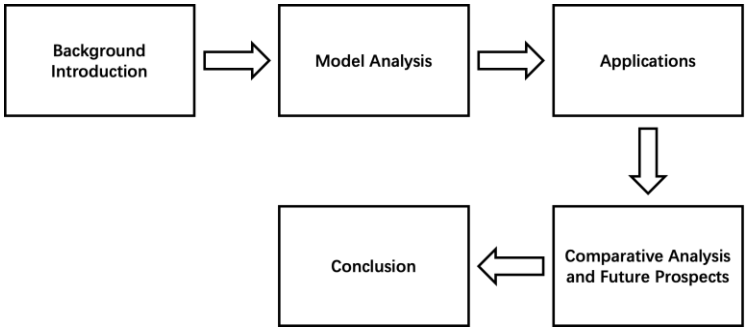

Figure 4. LSUN dataset sample images

2.2.1 Generative Adversarial Networks (GANs)

GANs represents a major step forward in the direction of image generation. In 2014, Goodfellow and his team introduced this model, which comprises two neural networks, a generator and a discriminator, functioning oppositely. The generator creates plausible images from random input, and the discriminator's task is to distinguish between actual images and those produced by the generator, which drives the continuous enhancement of the generator's performance. GANs are known for their capacity to produce diverse, realistic images by learning complex data distributions, which makes them invaluable in applications such as image synthesis, super resolution, and style transfer. The core principle underlying GANs is an adversarial min-max game, where the iterative process leads to the generation of increasingly realistic images [11]. The implementation process starts with a generator generating an image from random noise, which is then evaluated by a discriminator, as shown in Fig. 5. Both networks are trained simultaneously, using techniques such as spectral normalization and gradient penalties to ensure stable training and high-quality results. The influence of GANs extends to various fields, creating new styles in art; Improved resolution in medical imaging; Enrich the training data set in terms of data enhancement. Gans provide a powerful tool for generating high-fidelity synthetic data, and have also laid the foundation for the birth and development of improved models.

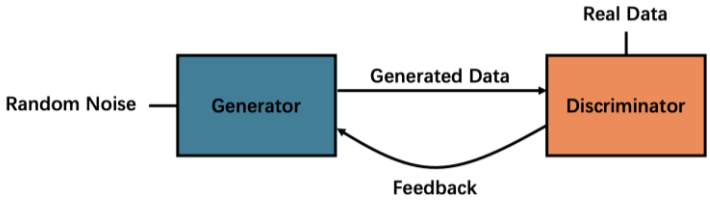

Figure 5. GANs Model Process

GAN-based models have diverse applications, with Style-Based GAN (StyleGAN), introduced by Karras et al. in 2019, being a notable example. StyleGAN allows for precise control over generated images by manipulating styles at various network levels, resulting in high-resolution, realistic images [9]. Its generator architecture separates latent spaces into intermediate styles, which are applied to different layers, enabling detailed image synthesis. This model is particularly valuable for high-quality image generation, including realistic faces and artistic content. Another significant model is Cycle-Consistent GAN (CycleGAN), proposed by Zhu et al. in 2017 [13]. CycleGAN facilitates image-to-image translation without paired examples by learning cycle-consistent mappings between domains, meaning that converting the image from one domain to another and returning to rebuild the original image. It employs two generators and two discriminators to ensure this consistency. CycleGAN is essential for applications like style transfer, where paired training data is unavailable, such as converting photos into paintings.

2.2.2 Diffusion Models

Diffusion model is a new image generation method proposed by Dhariwal and Nichol in 2021 [5], which aims to create high-quality images by iteratively refining random noise. The concept involves gradually transforming a simple noise distribution into a complex data distribution through a series

of steps. Diffusion models excel in generating highly realistic and detailed images, making them particularly valuable for applications that require high fidelity, such as medical imaging and art of realism. The principle of diffusion model is to start with random noise vector, apply the de-noising process iteratively, and gradually restore the original image data. This process can be seen as the opposite of the diffusion process in physics, in which particles diffuse away from their initial state. In the context of image generation, the model learns to reverse this diffusion by progressively removing noise. A diffusion model's framework usually consists of two processes, as Fig. 6 illustrates. A reverse process learns to eliminate the noise that a forward process adds to the data in multiple steps. During training, the training image with Gaussian noise is destroyed first, and then the neural network predicts and removes the noise at each step. This iterative de-noising continues until the model produces a clear, high-quality image. The significance of the diffusion model lies in its robustness and the high quality of the generated images. In some cases, they outperform traditional generative models like GANs, especially when generating images with fine detail and texture. In medical imaging, diffusion model can improve the clarity of diagnostic images and improve the accuracy of medical diagnosis. In art and entertainment, they can create realistic visuals and special effects. Overall, diffusion models represent a promising direction for generative modeling, providing a powerful tool for applications that require high-quality synthetic data.

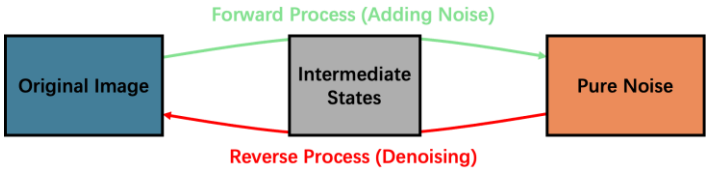

Figure 6. Diffusion Model Process

Diffusion-based techniques are widely employed, mostly for producing high-quality images and various other types of data. For instance, Denoising Diffusion Probabilistic Models (DDPMs), introduced by Ho et al in 2020, iteratively de-noises randomly sampled noise vectors to generate images [6]. The model is excellent at producing high-quality images with fine detail and stable training. Gaussian noise is progressively added to the data by DDPMs using a forward process, and the data is then de-noised using a reverse process. After training, the model can reverse the noise process step by step. The training involves adding noise to the image in several steps and then training the neural network to iteratively predict and subtract that noise. DDPMs are significant for its robustness and high-quality output, suitable for high-resolution image synthesis and noise reduction applications in medical imaging. The Denoising Diffusion Implicit Models (DDIMs) proposed by Song et al in 2020 is an extension of DDPMs, aiming to improve the efficiency and quality of image generation [14]. The approach lowers the number of necessary denoising steps by introducing an implicit sampling method. DDIMs allow deterministic sampling by modifying the reverse diffusion process to improve speed and quality. The model learns to de-noise in fewer steps while preserving image fidelity when the implementation first adds noise to the training image. DDIMs provide a more efficient method of image generation, making it valuable for real-time applications and scenarios that require fast image synthesis. It is also important to discuss Score-based Generative Models (SGMs), which employ score matching technology to drive the generation process from noise to data by learning the gradient of the data distribution [15]. They are good at producing high-resolution images and are suitable for a variety of data types. SGMs learn to estimate the fractional function, which is the gradient of the logarithmic probability of the data, and uses it to iteratively refine the noise into a high-quality image. Training involves learning a fractional function from noisy data and then using this function to de-noise the sample step by step. SGMs are particularly effective in applications that require high fidelity and resolution, such as detailed scientific visualizations and high-end art generation.

3. Result and Discussion

3.1. Performance Evaluation

Table 1 compares the performance of four models: StyleGAN3, U-Net GAN, DDPM++, and Latent Diffusion. These metrics include Average Generation Time, Fréchet Inception Distance (FID), Inception Score (IS), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR). StyleGAN3 had the shortest average generation time at 0.06 seconds, followed by U-Net GAN with an time of 0.08 seconds. In contrast, DDPM++ and Latent Diffusion had significantly longer average generation times of 4.50 seconds and 4.20 seconds respectively. This shows that the GAN model has a significant advantage in terms of generation speed. When it comes to FID scores, StyleGAN3 has the lowest score of 3.9, meaning that the images it generates are the closest to the original, and U-Net GAN has a slightly higher score of 4.5. DDPM++ and Latent Diffusion have relatively high FID scores of 5.8 and 6.0, indicating that these models produce relatively low image quality. The IS scores of all models are close, with StyleGAN3 slightly ahead at 4.8, while other models are between 4.6 and 4.7. It can be seen that there is a slight difference in the diversity of generated images. DDPM++ and Latent Diffusion excel on the SSIM measure at 0.96 and 0.95, illustrating their benefits in producing images that resemble the original image structurally. StyleGAN3 and U-Net GAN are slightly less impressive at 0.94 and 0.93, respectively. In terms of PSNR, DDPM++ and Latent Diffusion performed equally well, with scores of 31.8 dB and 31.5dB. It is significantly higher than the 30.0dB of StyleGAN and 29.8dB of U-Net GAN, reflecting the advantages of diffusion in image fidelity and noise suppression. These results show that GAN models, especially StyleGAN3, perform well on generation speed and specific quality metrics (FID and IS), while diffusion-based models are superior in image structural similarity and fidelity (SSIM and PSNR). These differences may be due to the different architectural characteristics of GAN and diffusion models. GAN models employ adversarial training mechanisms, where the discriminator network aids the generator in incrementally creating more realistic images. So, they perform particularly well on FID and IS metrics. In contrast, Diffusion models, such as DDPM++ and Latent Diffusion, synthesize images by iteratively refining a noisy input to approximate the real data distribution. This process, though computationally intensive, maintains high-frequency details and structural information, leading to superior performance on metrics like SSIM and PSNR.

Table 1. Comparative Performance of GAN-based and Diffusion-based Models

Method	Average Generation Time (s/image)	FID	IS	SSIM	PSNR (dB)
StyleGAN3	0.06	3.9	4.8	0.94	30.0
U-Net GAN	0.08	4.5	4.6	0.93	29.8
DDPM++	4.50	5.8	4.7	0.96	31.8
Latent Diffusion	4.20	6.0	4.6	0.95	31.5

3.2. Discussion

GANs, especially models like StyleGAN3, stand out for their capacity to generate high-resolution and photorealistic images, which is reflected in metrics such as FID and IS. These metrics demonstrate the model's capacity to create images with an outstanding level of visual diversity and fidelity. Critical to this performance is the adversarial training mechanism in GANs, where discriminators help improve the generator's output, pushing the generator to create increasingly realistic images. Diffusion models, such as DDPM++ and Latent Diffusion, focus more on iteratively denoising process to approximate the true data distribution from initial noise. Although more computationally intensive and slower, these methods excel at preserving fine detail and structural integrity in the images, and they can therefore achieve higher SSIM and PSNR scores. The architectural differences between GANs and diffusion models give them different advantages and disadvantages. GANs, while fast, can experience mode collapse, i.e. they are unable to capture the full diversity of training data, resulting in less variation in the resulting images. The diffusion model

avoids this problem through its design, which essentially maintains a more comprehensive representation of the data distribution, resulting in a more diverse and detailed output. In practical applications, these differences suggest that GANs are better suited for use cases that prioritize speed and image diversity, such as entertainment and art. In contrast, diffusion models are better suited for scenarios that require high precision and detail like medical diagnosis and scientific research, where the quality of each generated image is critical.

The field of image generation continues to offer extensive research opportunities. Future research should focus on enhancing model efficiency and multi-modal generation capabilities. This includes developing model compression and acceleration techniques to enable high-quality image generation on mobile and edge devices, which is crucial for real-time processing in mobile apps, augmented reality (AR), and low-bandwidth video streaming. Improving training and sampling algorithms is also essential to reduce computational resource needs and speed up model deployment.

Multimodal and cross-modal generation, which involves seamless transitions between text, images, audio, and video, as well as integrating 3D content with 2D images, presents significant research value. This technology could revolutionize film and TV production, game development, and multimedia content creation by allowing rich content generation from textual descriptions. Additionally, advancing controllable and editable generation technologies can offer users finer control mechanisms and interactive editing capabilities, benefiting personalized advertising, virtual fitting, and intelligent design tools. Temporal consistency in video generation remains a critical challenge, requiring specialized spatial-temporal attention mechanisms and consistency loss functions. This is vital for film and TV post-production, animation, and virtual reality content creation. Current issues such as bias, copyright, and handling long-tail distributions also need addressing. Bias can be mitigated by diversifying training datasets and incorporating debiasing mechanisms. Copyright and ethical concerns necessitate technologies for content identification and moderation. Challenges with long-tail distributions and novel scenarios can be tackled using small-sample learning and meta-learning techniques to enhance data augmentation and synthetic data generation. Addressing these issues will enhance the applicability and reliability of generative models.

4. Conclusion

This paper reviews the latest advancements and challenges in image generation, highlighting future research directions and proposing solutions to existing issues. The study centers on GANs and diffusion models, examining their theoretical foundations, advantages, limitations, and practical applications across entertainment, multimedia content creation, and medical and scientific research. The findings indicate that GANs excel in generation speed and diversity, making them ideal for scenarios requiring rapid output. However, they often face issues with training instability and mode collapse. Conversely, diffusion models are better suited for applications demanding high detail and structural consistency, such as medical imaging and scientific visualization, though they require significant computational resources. Future research will focus on enhancing model efficiency and multi-modal generation capabilities. This includes exploring model compression and acceleration techniques to optimize high-quality image generation on mobile and edge devices. Additionally, improving the accuracy of multimodal content generation and resolving information conflicts between different modalities will be key areas of investigation. These efforts aim to advance the practical application and innovative development of image generation technologies across various fields.

References

[1] Goodfellow I. Pouget-Abadie J. Mirza M. et al. Generative adversarial nets. Advances in neural information processing systems, 2014, 27.

[2] Brock A. Donahue J. Simonyan K. Large scale GAN training for high fidelity natural image synthesis. 2018, arXiv preprint: 1809.11096.

[3] Chen M. Radford A. Child R. et al. Generative pretraining from pixels. International conference on machine learning. PMLR, 2020: 1691-1703.

[4] Kingma D.P. Welling M. Auto-encoding variational bayes. 2013, arXiv preprint: 1312.6114.

[5] Dhariwal P. Nichol A. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 2021, 34: 8780-8794.

[6] Ho J. Saharia C. Chan W. et al. Cascaded diffusion models for high fidelity image generation. Journal of Machine Learning Research, 2022, 23 (47): 1-33.

[7] Ramesh A. Pavlov M. Goh G. et al. Zero-shot text-to-image generation. International conference on machine learning. Pmlr, 2021: 8821-8831.

[8] Radford A. Kim J.W. Hallacy C. et al. Learning transferable visual models from natural language supervision. International conference on machine learning. PMLR, 2021: 8748-8763.

[9] Karras T. Laine S. Aila T. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4401-4410.

[10] Saharia C. Ho J. Chan W. et al. Image super-resolution via iterative refinement. IEEE transactions on pattern analysis and machine intelligence, 2022, 45 (4): 4713-4726.

[11] Karras T. Aila T. Laine S. et al. Progressive growing of gans for improved quality, stability, and variation. 2017, arXiv preprint: 1710.10196.

[12] Nichol A.Q. Dhariwal P. Improved denoising diffusion probabilistic models. International conference on machine learning. PMLR, 2021: 8162-8171.

[13] Zhu J.Y. Park T. Isola P. et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision. 2017: 2223-2232.

[14] Song J. Meng C. Ermon S. Denoising diffusion implicit models. 2020, arXiv preprint: 2010.02502.

[15] Song Y. Ermon S. Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems, 2019, 32.