# Research and Analysis of Image Generation Technology Based on Deep Learning

*Zhiming* Hou[1*]

[1]Beijing-Dublin International College, Beijing University of Technology, 100124 Beijing, China

**Abstract.** Deep generative models are the key driving force behind the breakthroughs in image synthesis in the field of artificial intelligence, aiming to learn complex data distributions and sample from them to generate realistic and diverse new images. This article aims to systematically trace the evolution process of this field from its pioneering models to the current cutting-edge technologies. The paper clearly presents this development process, the core review logic of this paper is to divide the mainstream deep generative models into three major frameworks based on the core modeling ideas and learning paradigms of the models, namely Generative Adversarial Network (GAN), Variational AutoEncoder (VAE), and Diffusion Model (DM). Based on this logic, this paper first expounds the basic principles of various frameworks, representative derivation methods and the inheritance relationships among them; Furthermore, the system compared the generation performance of different models on multiple benchmark datasets; Finally, the challenges currently faced by deep generation technology were deeply discussed, and its future research directions were prospected. Through the review, this article hopes to provide researchers with a clear panoramic view of technological development and inspire new research ideas.

## 1 Introduction

Image generation, as one of the core tasks in computer vision, has long attracted much attention. Its goal is to generate images that conform to the visual laws of the real world from random noise or specific input conditions. This task not only holds significant theoretical value, being used to explore the intrinsic distribution of data and representation learning, but also demonstrates great potential in numerous application scenarios, such as image editing, data augmentation, artistic creation, and medical image analysis. With the rapid development of deep learning, generative models have witnessed a revolutionary breakthrough. The quality of the generated images has risen from the initial blurry distortion to the current level where they can be "so realistic as to be indistinguishable from the real ones".

Although technological advancements are constantly evolving, the underlying technical paths reveal a clear pattern. In order to systematically review this process, this article adopts a review logic based on the core modeling concept. This article contends that the current research on deep generative models mainly follows three distinct paradigms, namely the

* Corresponding author's email: hou_zhiming@outlook.com

Generative Adversarial Network, the Variational AutoEncoder, and the Diffusion Model. This division is not arbitrary; rather, it is based on their different responses to the fundamental question of "how to map from simple data distributions to complex ones. "The GAN framework directly learns the deterministic or stochastic mapping from the latent space to the data space through a competitive game process of "generator - discriminator". The core lies in the mutual stimulation between "generation" and "discrimination". The VAE framework is based on probabilistic graphical models and variational inference, aiming to learn a generative model with latent variables, and approximate the intractable data likelihood by maximizing the evidence lower bound. Its process places greater emphasis on exploring the underlying structure of the data. The diffusion model takes a different approach. It transforms the data generation process into a sequential and progressively refined decision-making problem through a fixed forward noise addition process and a learnable reverse denoising process. Its concept is derived from non-equilibrium thermodynamics.

These three types of methods have fundamental differences in their thinking, forming three parallel and sometimes intersecting technical routes in the field of image generation. Each of them has its own focus and addresses issues at different levels, jointly promoting the progress of the field. Therefore, by using these three types of frameworks as the main thread of the review, it is possible to clearly and comprehensively present the internal logic and diversity of technological development.

Based on the above-mentioned review logic, the main contributions of this paper can be summarized as follows, firstly, a systematic review and summary were conducted on the basic principles, representative derivative methods, as well as the advantages and disadvantages of Generative Adversarial Network (GAN), Variational AutoEncoders (VAE), and Diffusion Model (DM). Secondly, performance results from public benchmark datasets were compiled, and a horizontal comparison was conducted among different models regarding their generation quality, diversity, and stability. Finally, based on an in-depth analysis of the current situation, several key challenges faced by the technology were pointed out, and the possible future research directions were also envisioned.

## 2 Method

Based on the core modeling concept, this article introduces deep generative models into three major categories for presentation. In terms of writing, the principle of proceeding from the overall to the specific should be followed. For each type of method, first, its basic process and the core problem it solves should be outlined, then its representative derivative methods should be introduced, and finally, its advantages and disadvantages should be summarized. At the same time, pay attention to the connections and contrasts among different methods to demonstrate the continuity of technological development.

### 2.1 Generative Adversarial Network

GAN have proposed a framework to estimate the generative model through an adversarial process [1]. The core idea is to train two neural networks simultaneously, namely “Generator” and “Discriminator”. The generator is dedicated to capturing the distribution of real data and mapping random noise to the data space to generate fake samples. The discriminator, however, acts as a binary classifier, attempting to distinguish whether the input samples come from real data or are generated by the generator. The two evolved together in the "min-max game", with the ultimate goal of enabling the generator to produce samples that are convincingly realistic, which makes it impossible for the discriminator to effectively distinguish them.

The introduction of GAN has sparked a research craze, and a series of derivative works are dedicated to enhancing the training stability and generation quality of GAN. DCGAN introduced convolutional neural networks into the GAN framework, proposing a stable structural design principle, which was a crucial step for the successful application of GAN in image generation [2]. WGAN theoretically analyzed the root cause of the instability in the original GAN training [3]. Replaced the JS divergence with the Wasserstein distance and imposing a Lipschitz constraint, it significantly improved the training stability. LSGAN achieves this by modifying the loss function of the discriminator to the least squares loss, providing the generator with smoother and more guided gradients, thereby alleviating the mode collapse problem [4].

GAN's main advantage is in their ability to generate extremely sharp and high-quality images. Its adversarial training mechanism forces the generator to continuously approach the high-probability regions of the actual data distribution. However, its drawbacks are also quite prominent. Specifically, the training process may be unstable and difficult to converge, prone to "mode collapse" (where the generator can only produce a limited variety of samples), and lacks explicit modeling of the latent space, resulting in weak interpretability and controllability of the generation process.

## 2.2 Variational AutoEncoder

Unlike the adversarial idea of GAN, Variational Autoencoder (VAE) are based on variational inference and aim to construct a probabilistic generative model [5]. VAE views the data generation process as being generated by a continuous latent variable. Its basic process includes an encoder and a decoder. The encoder maps the input data to the posterior distribution of the latent variables (usually assumed to be a Gaussian distribution), and samples to obtain the latent code; The decoder, thereafter reconstructs the data space from this latent code. VAE jointly trains the encoder and decoder by maximizing the evidence lower bound (ELBO), and its loss function includes reconstruction loss and regularization terms of the latent space.

The VAE framework has also given rise to various improved models. The $\beta$-VAE introduces a hyperparameter $\beta$ into the ELBO objective, which enhances the regularization strength of the latent space, encouraging the model to learn more decoupled and interpretable latent variable representations [6]. VQ-VAE employs vector quantization technology to discretize the output of the encoder onto a codebook, thereby learning a discrete latent space, which facilitates the subsequent generation based on the autoregressive model [7].

The main advantage of VAE lies in their having an explicit and structured latent space, and their training process is relatively stable and less prone to mode collapse. It provides a solid probabilistic framework, facilitating data interpolation and meaningful operations in the latent space. The drawback is that, due to the use of approximate inference and the optimization of ELBO, the generated images tend to be blurry, and are not as sharp and realistic as those generated by GAN. This is believed to be caused by the inconsistency between the optimization objective and human perception.

## 2.3 Diffusion Model

The diffusion model is a type of generative model that has emerged in recent years, and its inspiration comes from non-equilibrium thermodynamics. Its core idea is to learn the data distribution through a gradual "add-noise-remove-noise" process. There are 2 two processes of the framework, the forward diffusion process and the reverse generation process. The forward process eventually transforms a real image into pure Gaussian noise by adding Gaussian noise multiple times. The reverse process is a parameterized Markov chain that

learns how to start from pure noise, gradually remove the noise, and ultimately restore a clear image [8].

The success of the diffusion model is inseparable from the impetus of a series of key technologies. Denoising Diffusion Probabilistic Models (DDPM) simplifies the training objective, establishes the basic paradigm of using U-Net to predict noise and combining gradual denoising, and achieves high-quality generation results [9]. Stable Diffusion shifts the diffusion process to the latent space, significantly reducing computational costs and introducing cross-attention mechanisms to achieve efficient text-to-image generation, thereby promoting the widespread application of diffusion models [10].

The main advantage of the diffusion model lies in its powerful generation capability, and it has currently achieved a leading position in multiple benchmark tests. Its training objective is simple and stable, and there is no problem of model collapse. Meanwhile, the gradual generation process is combined with conditional guidance techniques (such as Classifier-Free Guidance), providing a high degree of controllability. Its most significant drawback lies in the slow generation speed, as it usually requires hundreds or even thousands of denoising sampling steps. Although some studies have focused on accelerating the sampling process, this remains a bottleneck in its practical application.

## 3 Experimental Results and Analysis

To evaluate and compare objectively the performance of various deep generative models, this section will introduce the commonly used evaluation datasets and metrics, and conduct a horizontal comparison analysis of the performance of representative methods.

### 3.1 Dataset and Evaluation Metrics

The CIFAR-10 dataset includes 60,000 32x32 color images, which belong to 10 different categories, and is commonly used for rapid prototype verification and preliminary comparisons.

ImageNet, a large-scale dataset which contains 1000 categories. Usually, branches with a resolution of 128x128 or higher (such as ImageNet-1k 128x128) are used to evaluate the capability of the model's handling complex and diverse scenarios.

LSUN Bedroom is a category of bedrooms in the large-scale scene understanding dataset, and is often used to evaluate the quality of high-resolution (such as 256x256) image generation by models in specific scenarios.

FFHQ contains 70,000 high-quality facial images, which are widely used to evaluate the fidelity and diversity of face generation.

### 3.2 Evaluation Index

Fréchet Inception Distance (FID) is used to measure the similarity between the feature space distributions of generated images and real images [11]. The lower the value, the higher the quality and diversity of the generation. It is currently the most widely used evaluation metric.

Inception Score (IS) calculates the score by evaluating the clarity and diversity of categories of the generated images [12]. The higher the value, the better. However, it is insensitive to mode collapse and relies on the pre-classification model.

Precision & Recall decouples the assessment of generation quality into two metrics, precision (measuring how many of the generated images are of high quality) and recall rate (measuring how many patterns in the actual data distribution are covered by the generation model) [13].

### 3.3 Performance Evaluation of Algorithms

This article summarizes the results from multiple literature reports and conducts a comprehensive comparative analysis of the performance of three types of models on the mainstream datasets (with the FID value serving as the primary criterion).

On the CIFAR-10 dataset, early GAN (such as WGAN-GP) and VAE achieved satisfactory results. However, in recent years, advanced diffusion models (such as DDPM, DDIM) and enhanced GAN (such as StyleGAN2-ADA) have pushed the FID value to single digits or even lower, demonstrating superior performance.

On more challenging high-resolution datasets such as ImageNet 128x128 and LSUN Bedroom 256x256, diffusion models (such as ADM) and style-based GANs (such as the StyleGAN series) consistently reported the lowest FID scores, representing the current state-of-the-art level. Especially for diffusion models, they tend to perform better in terms of the diversity of generated images and the richness of details. In contrast, the FID values of VAEs in these tasks are usually higher than those of the former two, indicating their limitations in generating high-fidelity images.

From the perspective of generation speed, GANs have an absolute advantage in the inference process. They can generate images with just one forward propagation. The standard diffusion model, however, requires hundreds of iterations and is several orders of magnitude slower. Even though methods like DDIM are available to accelerate sampling, it still cannot compete with GAN.

Experimental analysis shows that each framework has its own advantages. The diffusion model and modern GAN both excel in generation quality. Among them, the diffusion model may have an edge in terms of diversity and stability, while GAN are unrivaled in terms of inference efficiency. VAE, on the other hand, maintain unique value in terms of their stable training process and structured latent space. The choice of which model to use depends on the trade-offs made by the specific application scenario in terms of quality, speed, and controllability.

## 4 Discussion and Future Prospects

Although deep generative models have achieved remarkable results, there are still many challenges and development opportunities in this field. Based on the review of existing research, this article summarizes the following three key issues and then outlines the future research directions accordingly.

### 4.1 Controllable Generation and Interpretability

Most current models lack the ability to provide fine-grained and semantically meaningful control. It is difficult for users to precisely specify the posture, spatial arrangement or specific attributes of the objects in the generated image. Future research could focus on decoupling representation learning, so that the latent space dimension of the model aligns with meaningful semantic factors in the real world. Furthermore, integrating external knowledge (such as language and knowledge graphs) to guide the generation process and achieve higher-level semantic understanding and controllable generation is an important direction.

### 4.2 Efficiency and Real-time Applications

The sampling speed of the diffusion model is the main obstacle preventing it from being applied in real-time scenarios (such as video generation and interactive design). Future work will need to continue to explore efficient sampler designs (such as distillation techniques,

consistency models) and novel model architectures, while maintaining the quality of generation and reducing the number of sampling steps to just ten or even one. Meanwhile, model lightweighting and adaptive computing are also key approaches to enhancing efficiency.

### 4.3 Ethical Security and Reliability

As the capabilities of generative models have improved, their misuse for creating false information and deepfake technology has become increasingly prominent. It is of utmost importance to ensure that technological development is safe, reliable and ethical. In the future, it is necessary to vigorously develop content generation detection technology, traceable digital watermarks, and a sound governance framework. Meanwhile, it is also crucial to explore ways to enhance the model's ability to identify and eliminate biases, and to generate more fair and diverse content.

## 5 Conclusion

This article follows the review logic based on the "core modeling concept" and systematically summarizes the development process of deep generative models, from GAN, VAE to Diffusion Models. In the methodology section, this paper provides a detailed explanation of the basic principles and representative derivative methods of these three types of frameworks, and objectively analyzes their respective advantages and limitations. In the experimental results section, by integrating the performance data from public benchmarks, a comprehensive comparison and analysis was conducted among different models in terms of generation quality, diversity, and efficiency. Finally, in the discussion section, this paper identifies the current challenges in the field regarding controllable generation, efficiency, ethical security and proposes potential research directions for the future. It is hoped that this review can provide researchers in the related fields with a clear technical roadmap and help the deep generative model technology achieve new breakthroughs in both theory and application.

## References

1. W. Xia, Y. Zhang, Y. Yang, J. H. Xue, B. Zhou, and M. H. Yang. Gan inversion: A survey. IEEE transactions on pattern analysis and machine intelligence, 45(3), pp.3121-3138 (2022)
2. J. Jenkins, and K. Roy. Exploring deep convolutional generative adversarial networks (DCGAN) in biometric systems: a survey study. Discover Artificial Intelligence, 4(1), p.42 (2024)
3. Q. Wang, X. Zhou, C. Wang, Z. Liu, J. Huang, Y. Zhou, C. Li, H. Zhuang, and J. Z. Cheng. WGAN-based synthetic minority over-sampling technique: Improving semantic fine-grained classification for lung nodules in CT images. IEEE Access, 7, pp.18450-18463 (2019)
4. T. Xia, and L. Liu. LSN-GAN: A Novel Least Square Gradient Normalization for Generative Adversarial Networks. In 2024 IEEE 4th International Conference on Software Engineering and Artificial Intelligence (SEAI) (pp. 343-347). IEEE (2024)
5. S. Odaibo. Tutorial: Deriving the standard variational autoencoder (vae) loss function. arXiv preprint arXiv:1907.08956 (2019)

6. C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. Understanding disentangling in $\beta $-VAE. arXiv preprint arXiv:1804.03599 (2018)
7. A. Razavi, A. Van den Oord, and O. Vinyals. Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems, 32 (2019)
8. F. A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion models in vision: A survey. IEEE transactions on pattern analysis and machine intelligence, 45(9), pp.10850-10869 (2023)
9. A. Alblwi, S. Makkawy, and K. E. Barner. D-DDPM: Deep Denoising Diffusion Probabilistic Models for Lesion Segmentation and Data Generation in Ultrasound Imaging. IEEE Access (2025)
10. D. Bolya, and J. Hoffman. Token merging for fast stable diffusion. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4599-4603) (2023)
11. Y. Yu, W. Zhang, and Y. Deng. Frechet inception distance (fid) for evaluating gans. China University of Mining Technology Beijing Graduate School, 3(11) (2021)
12. S. Barratt, and R. Sharma. A note on the inception score. arXiv preprint arXiv:1801.01973 (2018)
13. M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing generative models via precision and recall. Advances in neural information processing systems, 31 (2018)