Advances and Challenges in Deep Learning Based Image Generation Models

Ruizhi Liu *

School of Fuzhou University, Fuzhou,350108, China

* Corresponding Author Email: RUIZHI.LIIU.2025@mumail.ie

Abstract. A hot topic in current computer vision research is image generation, which essentially refers to the use of computer algorithms and models to generate artistic and creative images. With the development of artificial intelligence and deep learning, numerous image generation models have emerged, and image generation technology is gradually becoming an important research direction in the fields of computer graphics and art. This has driven the widespread application of image generation in areas such as film, gaming, and painting. This paper provides a preliminary introduction to the basic principles of image generation models, including popular models, their evolution and variants. It then discusses some of the application areas in which image generation can be used. Finally, by analysing the challenges and issues faced by image generation models, this paper looks ahead to some of the work that future researchers will undertake and the direction in which image generation models will develop.

Keywords: Computer Vision, Image Generation, Deep Learning.

1. Introduction

Image generation technology has advanced significantly over the last several years. It's a significant step forward in AI for computer vision, and as a result, it's now employed in various areas, including art, entertainment, and scientific visualisation. However, early image generation models had obvious shortcomings in terms of generation speed, resolution, and image quality. The generated results often lacked detailed information, exhibited blurry textures, and displayed structural distortions, which failed to meet the requirements of practical applications for diversity, realism, and manageability. These shortcomings not only limit the promotion of generative models in scientific research as well as industrial scenarios but also affect their reliability and practicality in high-precision tasks (such as medical image synthesis and virtual reality content generation).

To address these issues, researchers have proposed various deep learning-based probabilistic generative models that explicitly or implicitly model the distribution of latent variables, thereby improving diversity while ensuring image realism. Representative models include Variational Autoencoders (VAE), Generative Adversarial Networks (GAN), and Diffusion Models (DM) [1-3]. Among them, VAE separates and reorganizes the style and content features of images through encoding and decoding processes, and has good potential space controllability; GAN uses a competitive game mechanism between generators and discriminators to improve image detail and realism significantly; DM generates high-quality images through an iterative process of gradual noise addition and noise removal, and performs well in terms of texture rendering and stability.

The importance of image generation technology is reflected in many aspects. First, it provides efficient tools for digital content production, reducing creation costs and improving production efficiency. Second, it offers diverse, high-quality training samples for machine learning tasks such as data augmentation and cross-modal conversion. Third, it provides innovative solutions for high-precision applications, including scientific research, medical imaging, and the restoration of cultural relics. Compared with traditional methods, these generative models not only have stronger feature expression capabilities and generative freedom but can also adapt to diverse application needs.

From a historical perspective, VAE and GAN emerged almost simultaneously, marking a milestone in the field of image generation and significantly improving the quality of generated images. Subsequently, research focus gradually shifted from resolution enhancement to diversity control,

conditional generation, and cross-modal fusion. In recent years, the rise of diffusion models has brought a new round of breakthroughs in this field, particularly in terms of their advantages in terms of generation quality, stability, and multimodal adaptability, making them a current research hotspot.

This paper will systematically review the principles, technological evolution, and advantages of three types of models, including VAEs, GANs, and DMs. It will analyse typical application cases and their performance in different tasks, and discuss current challenges and future development directions, to provide a systematic reference for the research and application of image generation technology.

2. Fundamental Theory of Image Generation Models Based on Deep Learning

2.1. Core Model Architecture

VAE decomposes images using an encoder to obtain style features and content features, then recombines them through self-decoding to generate the final image. GAN itself has a generator and a discriminator. The generator and discriminator continuously engage in a competitive game, resulting in increasingly higher image quality and ultimately producing the ideal image people desire. DM generates images by gradually adding small amounts of noise to its own images and then iteratively removing the noise in reverse. The generated images are rich in detail and high in quality.

2.2. Evaluation Indicators and Datasets

The evaluation of images generated by image generation models is primarily divided into two aspects: quantitative and qualitative. Quantitative evaluation involves calculating scores using specific formulas, while qualitative evaluation involves estimating scores through subjective manual scoring and comparing different models on the same theme.

In terms of quantitative metrics, general research can utilise the Inception Score (IS) to measure the diversity and recognition of generated images, and the FID score to calculate the "feature distance" between generated images and authentic images. In terms of qualitative assessment, it mainly relies on manual evaluation based on human scoring and model comparisons using different models to create images on the same theme.

There are many datasets for image generation. For simple testing, you can use CIFAR-10 for small datasets and COCO for large datasets. Alternatively, you can generate specific targets, such as faces, using a dedicated face dataset.

3. Typical Generative Models and Recent Technological Advances

3.1. Core Principles and Technological Iterations of Variational Autoencoders

3.1.1 Core Principles

The core idea of VAE is that it no longer encodes input data into a fixed point, but rather into a probability distribution, usually a Gaussian distribution [1]. By introducing the Evidence Lower Bound (ELBO), VAE transforms the problem of "maximising p(x)" into maximising the ELBO. The encoder outputs the parameters (mean and variance) of this distribution, then randomly samples a point from this distribution and sends it to the decoder to generate data. This design makes the potential space continuous and structured, enabling the generation of new samples.

3.1.2 Technological iteration

Conditional VAE (CVAE) introduces conditional information (such as category labels) into the encoder and decoder, enabling the model to generate specific types of samples based on specified conditions [4]. For example, the model can be instructed to create only images of the number "7."

Next is β-VAE, a variant that introduces an adjustable hyperparameter β in front of the KL divergence term [5]. When $\beta > 1$, the model places greater emphasis on learning disentangled

representations, i.e., each dimension of the latent space corresponds to an independent factor of variation in the data (such as the position, size, or colour of an object).

At the same time, VAE-GAN hybrid models have also emerged. To combine the stability of VAE training with the clarity of GAN-generated images, researchers have proposed the VAE-GAN [6]. These models typically utilise the discriminator of the GAN to replace the reconstruction loss function of the VAE, thereby encouraging the decoder to generate sharper, more realistic images.

In recent years, DMs have surpassed GANs and VAEs in terms of generation quality, emerging as a new research hotspot. This has also given rise to models that fuse VAEs with diffusion models. Many advanced diffusion models, such as the well-known Stable Diffusion, still use VAE in their architecture. This type of method, known as latent diffusion, models first uses a pre-trained VAE to compress high-dimensional image data into a smaller, more computationally efficient latent space, and then performs time-consuming diffusion and denoising processes in that latent space, significantly improving training and generation efficiency [7].

3.1.3 Limitations and Breakthroughs

Although the images generated by VAE are relatively blurry, the key breakthrough of VAE lies in no longer encoding the input data as a point, but rather as a probability distribution. The decoder generates data using the parameters of the distribution and a randomly sampled point from the distribution. This design makes the latent space continuous, enabling the generation of new samples. Subsequent variants and hybrid models of VAE continue to evolve, ensuring that the VAE remains a crucial component.

3.2. Evolution and Development of Generative Adversarial Networks

3.2.1 Theoretical breakthroughs

Before GANs, most generative models relied on explicit modelling of data distributions, such as VAEs. The generator and discriminator are the two primary components of GANs, which were inspired by the idea of zero-sum games in game theory. The discriminator's job is to discern between created and actual samples, whereas the generator's job is to develop real samples to trick the discriminator. By continuously improving their respective abilities through adversarial training, they ultimately reach a Nash equilibrium state. The images generated by this form of generation surpass those of previous models and have quickly gained the attention of researchers.

3.2.2 Evolutionary development

GAN was first proposed in 2014. In 2015, CNN was introduced into GAN, and the emergence of DCGAN significantly improved the quality of image generation [2,8]. Subsequently, in 2017, conditional labels were introduced to the Conditional GAN (cGAN) to enable controllable generation. Meanwhile, the emergence of Wasserstein GAN (WGAN) and Style-GAN achieved stable training and high-fidelity face generation, respectively [9, 10]. Currently, the GAN architecture has been combined with DALL·E, developed by OpenAI, and Stable Diffusion, along with the CLIP model, to achieve the functionality of generating images from text input [7, 11].

3.2.3 Technical Challenges

Mode collapse and training instability are the two primary issues that generative adversarial networks (GANs) encounter. Mode collapse refers to the generator being constrained to generate similar samples, resulting in a lack of diversity in the generated samples. Generator collapse can be divided into two categories: partial collapse, where the generator produces only one sample, and complete collapse, where the generator limits the generation of a series of samples. Another problem that GAN models encounter during training is training instability, which results from performance disparities between the discriminator and generator networks, or from an imbalance in their capacities, as the generator struggles to match the discriminator's performance.

3.3. Historic breakthroughs and cutting-edge advances in diffusion models

3.3.1 A historic breakthrough

DMs are undoubtedly the hottest topic in the field of generative models today. Their emergence has not only addressed the bottleneck of blurry images generated by VAEs but also mitigated the drawbacks of GANs, such as training mode collapse and instability. Diffusion models decompose the generation process into two stages: progressive noise addition (forward) and denoising (backwards) — utilising Markov chains. They employ score matching and ELBO optimisation to achieve explicit probabilistic modelling, further enhancing the manageability of the generated output. Even more surprising is the astonishing accuracy of diffusion models in video generation.

3.3.2 Recent Developments

Stable Diffusion embeds a VAE into the diffusion model, supplemented by the CLIP text editor and a cross-attention mechanism, making it one of the most widely used diffusion models today [7]. Compared to traditional diffusion models, Stable Diffusion significantly reduces the configuration required for generation, resulting in high-resolution and clear images. In terms of video generation, Sora's video generation AI utilises diffusion models to achieve high-precision video generation. Sora achieves video temporal consistency through a diffusion Transformer, reducing physical simulation errors by 70%. This represents a breakthrough in the field of video generation.

3.3.3 Historical Significance

The diffusion model has significant historical significance. It has not only achieved technological innovation and improved generative capabilities but has also restructured the industrial ecosystem in the field of generation, exerting a profound influence and marking a milestone event in the history of artificial intelligence development. Furthermore, the complete open-source nature of Stable Diffusion [7] has spurred global collaboration among researchers, leading to the development of various models adapted to different scenarios. Whether artists or individual creators, everyone can now fully participate in AI art creation, significantly advancing the process of AI civilianization.

4. Application scenarios for image generation models

4.1. Generating Creation

Currently, image generation models can be used for various types of creative work. Image generation models can produce multiple kinds of illustrations that meet human requirements, and they can already replace up to 30% of the work done by illustrators, significantly improving their work efficiency. Individuals can also utilise generative models to obtain the desired images and publish their artistic creations quickly.

4.2. Super Resolution

Using image generation technology, people can input low-resolution images and output clearer, high-resolution images, thereby generating super-resolution photos. This has broad application potential in restoring low-quality photos and videos.

4.3. Use in the medical field

In the field of medical imaging, people have been committed to reducing radiation doses because excessive radiation can cause harm to the human body. However, reducing the dose means that image noise will increase, leading to a decrease in image quality and the loss of some medical information. Currently, image conversion methods can be used to establish a mapping between noisy images and denoised images, thereby eliminating this problem and generating high-quality images.

4.4. Data Enhancement

The excellent performance of deep learning is inseparable from improvements in computing power and data. However, in the scientific research process, datasets for image categories often encounter the problem of small sample sizes, making it difficult for researchers to ensure that there are sufficient samples to support learning tasks. Currently, geometric transformations and colour transformations can be performed through image generation to generate new samples from small samples, thereby obtaining sufficient data to support research and expanding small sample sets. Another significant issue is data imbalance, where different categories within the same datasets may have different sample sizes. Previously, random sampling methods were used to address this issue through up-sampling or down-sampling. However, up-sampling involves the use of duplicate data, while down-sampling results in data loss. Expanding imbalanced data through image generation can effectively resolve this issue.

5. Conclusion

This paper presents the three leading models of VAEs, GANs, and DMs, based on the evolution of image generation models. It introduces breakthrough innovations, fundamental theories, and the development history of the three types of models, as well as some of the problems encountered, and discusses the rich application environment of image generation models today, providing proper references. Although image generation has been developing for decades, numerous issues remain to be solved to achieve better and more optimal models. Slow inference speed. Although diffusion models have significant advantages in terms of generation quality, their generation inference speed is currently one of their main development bottlenecks. Optimising model inference speed while reducing model complexity through quantisation pruning is a significant task for the further development of diffusion models in the future, aiming to achieve the goal of edge deployment. Long video generation. Current generative models are already capable of generating short videos using AI. However, the quality of longer videos, such as those exceeding 60 seconds, remains subpar. This is primarily due to the scarcity of long video datasets, resulting in a limited number of trainable samples and weak logical connections between frames over extended time spans. Improving the quality of long videos may be a key focus for future video generation efforts. Privacy and security issues. With the continuous development of generative models, the generation effects in some scenarios have already reached a level where they are indistinguishable from reality. Addressing the security risks posed by AI face swapping and similar issues is also an area that researchers need to consider further.

References

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

[2] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139–144.

[3] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 2020, 33: 6840–6851.

[4] Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. Advances in Neural Information Processing Systems, 2015, 28: 3483–3491.

[5] Higgins I, Matthey L, Pal A, et al. β-VAE: Learning basic visual concepts with a constrained variational framework. International Conference on Learning Representations, 2017.

[6] Larsen A B L, Sønderby S K, Larochelle H, et al. Autoencoding beyond pixels using a learned similarity metric. Proceedings of the International Conference on Machine Learning, 2016, 48: 1558–1566.

[7] Rombach R, Blattmann A, Lorenz D, et al. High-Resolution Image Synthesis with Latent Diffusion Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10684–10695.

[8] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.

[9] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of Wasserstein GANs. Advances in Neural Information Processing Systems, 2017, 30: 5767–5777.

[10] 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.

[11] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation. Proceedings of the International Conference on Machine Learning, 2021, 139: 8821–8831.