Dean&Francis
ISSN 2959-6157

A Study of Image Generation Methods that are based on Deep Learning

Haojun Hu

School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin, China
*Corresponding author: 0802220407@tute.edu.cn

Abstract:

In recent years, significant advancements have been made in the field of deep learning-driven image generation technology. Initially, the technology was characterized by the generation of low-quality image samples. However, it has since evolved to produce highly realistic and diverse images, which have found widespread applications in domains such as computer vision, art creation, medical imaging, and virtual reality. In this paper, we systematically study the current mainstream deep generative models, including Generative Adversarial Networks (GAN), Variational Auto-Encoders (VAE), and Diffusion Models. We focus on their core principles, representative results, performance characteristics, and application scenarios. We also analyze the strengths and shortcomings of the various models in depth. The paper undertakes a systematic comparison and summarization of the current state of research in the field, identifying the predominant challenges and anticipating future developments. The objective is to furnish researchers in the domain of image generation with a coherent technological trajectory and a comprehensive theoretical framework.

Keywords: Deep Learning; Image Generation; Generative Adversarial Network.

1.Introduction

The field has undergone three paradigm shifts with the rise of deep generative modeling, from breakthroughs in adversarial training in generative adversarial networks (GANs) to hidden-variable probabilistic modeling in variational autoencoders (VAEs) to diffusion model asymptotic noise mechanisms [1-3]

Domestic teams have achieved remarkable results in efficiency optimization, such as Zhang et al.'s lightweight GAN to achieve real-time generation on mobile; international frontiers focus on multimodal control and computational efficiency [4]; Rombach et al.'s Stable Diffusion achieves a computational breakthrough through potential space compression[5], however, Ho et al.'s DDPM reveals that diffusion models have a bottlenecks in sampling efficiency at the hundred-step level[6].

Despite technological innovation, current methods face core conflicts, such as quality vs. efficiency, controllability vs. generalizability, and cost vs. demand. This paper systematically reviews the evolution of deep learning image generation technology, focus-

ISSN 2959-6157

ing on the principles and technological breakthroughs of Generative Adversarial Networks (GANs), Variational Auto-Encoders (VAEs), and Diffusion Models. It quantitatively compares their performance boundaries in terms of generation quality, sampling efficiency, and conditional control ability. Ultimately, it puts forward a fusion architecture and ethical protection solution to resolve the conflicts between efficiency and controllability.

2.Core Generative Modeling Methodology

2.1 Generative Adversarial Networks

GANs, proposed in 2014 by Goodfellow et al., are a landmark work in image generation. They center on an adversarial training mechanism. The framework contains two deep neural networks. As shown in Fig. 1: Generator: Input hidden random noise vectors $\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})$ and use a transposed convolutional neural network for stepwise upsampling to generate fake samples $G(\mathbf{z})$.Discriminator: As a binary classifier, it distinguishes whether the input image comes from the real data distribution $\mathbf{x} \sim p_{\text{data}}(\mathbf{x})$ or the generated fake samples $D(x)$. It then outputs a probability value and calculates the loss function, which guides the optimization of the two networks. The optimization of the objective function through the minimax game is shown in equation (1). In the ideal case, when training reaches Nash equilibrium and the generating distribution is close to the real one $p_g(\mathbf{x}) \approx p_{\text{data}}(\mathbf{x})$.The discriminator output is $D(G(\mathbf{z})) \to 0.5$ when the generated and real samples reach equilibrium and become indistinguishable

$$\min_G \max_D V(D,G)=E_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}\left[\log D(x)\right]+E_{z\sim p_z(z)}\left[\log(1-D(G(z)))\right]$$

Where the generator minimization objective is $\min_G$, the discriminator maximization objective is $\max_D$, the value function is $V(D,G)$, the real data sample and distribution is $E_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}$, the discriminator output is $D(z)$, the generator processing is $G(z)$, and the loss term is $\log(1-D(G(z)))$

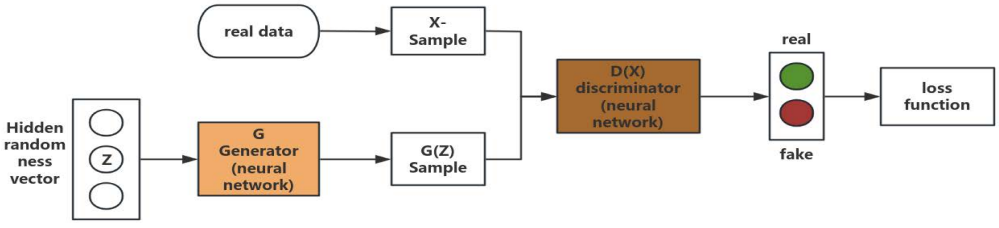

Fig. 1 GANs basic flow chart

GANs are driving technological progress. DCGAN improved quality by adding convolutional neural networks, stepwise convolution, batch normalization, and transposed convolutional up sampling[7]. However, it also suffers from pattern collapse, which can exceed 30%. This causes a decrease in sample diversity. In the CLEVR multi-object scene dataset, DCGAN often leads to an overpowered discriminator and a decrease in the generator gradient paradigm, resulting in a limited number of generated samples and exacerbating mode collapse. Major existing solutions usually increase the training time to reduce the crash rate. The StyleGAN family uses mapping networks and AdaIN to inject style vectors, with pixel-level noise to enhance detail performance. styleGAN2 and styleGAN3 build on this foundation to improve issues like artifacts and motion blur[8]. However, this family of models has a high training cost: at 1024×1024 resolution, a complete training process typically requires 256 V100 GPUs running for about 32 days. Under current hardware conditions, the model demands significant video memory usage, training time, and computational resources, making large-scale feasibility difficult to meet. GANs models generate samples with high visual fidelity and fast inference speed, but face limitations like unstable training, pattern collapse, lack of probability density estimation, and hyper parameter sensitivity. Cutting-edge solutions include spectral normalization, global dependency modeling by ViTGAN, and latent space decoupling by InfoGAN and StyleSpace.

2.2 Variational Autoencoders

Kingma and Welling's 2013 proposal of the variational auto-encoder (VAE) is a generative model based on probabilistic graphical models and variational inference. Fig. 2 shows the core architecture, which consists of an encoder and a decoder. The encoder maps input data X to a probability distribution in the latent space, learns to generate the mean $(\boldsymbol{\mu}_\phi)$ and variance $(\boldsymbol{\sigma}_\phi^2)$ of the key data, and introduces noise to the variance. The encoder outputs a random variable whose intensity is close to the standard normal distribution as 1. The mean plus the square root of the variance multiplied by the noise yields the variable Z* in the latent space. Subsequent access to the decoder samples from the latent distribution $(\mathbf{z} \sim q_\phi(\mathbf{z}\,|\,\mathbf{x}))$. It reconstructs data X' similar to the original input data after learning the output distribution $(p_\theta(\mathbf{x}\,|\,\mathbf{z}))$.The optimization objective

is to minimize the negative value of the evidence lower bound, as shown in equation (2).

$$\mathcal{L}_{\text{ELBO}} = E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}\mid\mathbf{z})\right] - D_{\text{KL}}\left(q_\phi(\mathbf{z}\mid\mathbf{x})\,\|\,p(\mathbf{z})\right) \qquad \text{X} \quad ,$$

should be close to X to preserve the data's fidelity. KL constraints on potential space regularization promote a standard normal distribution, gradually weakening noise influence. This regularizes the potential space, improving the model's performance and its ability to generalize.

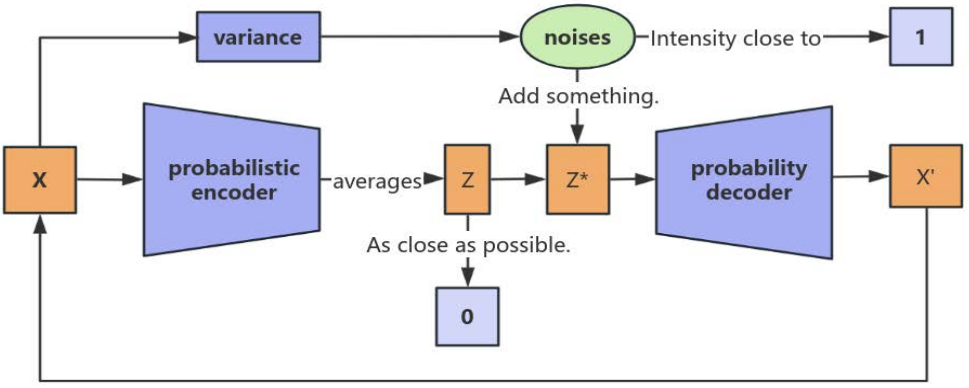

Fig. 2 VAE Core Architecture Flowchart

In the subsequent model evolution, β-VAE enhances the decoupling ability of potential features by introducing the coefficient β[8]. This enhancement achieves a success rate of ove-r 75% in feature editing tasks. However, the model's generated results are ambiguous, e.g., its structural similarity index (SSIM) is usually lower than 0.65 at a 512×512 resolution. This is far from the image quality standard required for practical applications. Additionally, current methods struggle to meet industrial-grade requirements for SSIM and KL dispersion when adjusting the β parameter. VQ-VAE and VQ-VAE2 discretize the continuous latent space through vector quantization. VQ-VAE2 uses a hierarchical structure to enhance generation resolution[9]. Training stability can be enhanced with codebooks of size K=512, but quantization loss is introduced. Using larger codebook sizes increases the risk of pattern collapse. VAEs have stable training, a spatial structure, and estimable data likelihood, but blurred images, lower sampling quality, and bias are challenges. Solutions include VQ discretization, hierarchical architecture, and adversarial training.

2.3 Diffusion Models

Diffusion modeling is the mainstream technique in current image generation (2023-2025), inspired by nonequilibrium thermodynamics. It learns data distribution through gradual noise addition and denoising. Fig. 3 shows the process starts with the raw data $\mathbf{x}_0$ on the left. Real data coresponds to model inputs. In the forward process, the data is progressively noised in t steps: Gaussian noise is introduced at each step according to a preset variance scheduling parameter. As the number of steps increases, the percentage of noise gradually increases until it is transformed into pure noise with an approximate standard normal distribution at step xt. As shown in equation (3), the per-step noise addition constitutes a strict Markov chain, in which the state of each step depends only on the result of the previous step.

$$\mathbf{x}_t = \sqrt{1-\beta_t}\,\mathbf{x}_{t-1} + \sqrt{\beta_t}\,\boldsymbol{\epsilon}_t \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0},\mathbf{I})$$

Where $\beta_t \in (0,1)$ denotes a predefined or learnable variance scheduling parameter and $\boldsymbol{\epsilon}_t$ denotes standard Gaussian noise that is independently and identically distributed, the process constitutes a Markov chain.

The reverse denoising process is presented on the right. Starting from pure noise, the model (usually with a U-Net or Transformer backbone) learns to predict the target by estimating and removing the noise from the currently noisy data at each step. As the denoising steps progress, the data gradually regains its true texture and structure. Eventually, new samples are generated that are consistent with the original data distribution. The figure clearly distinguishes the inverse processes of noise addition and denoising through visual symbols indicating arrow direction and noise intensity change. The figure may be labeled with time step t throughout, reflecting the dependence of the two processes on the time variable. The noise scheduling of the forward process and the denoising modeling of the inverse process both require time steps as key inputs. The reverse process is defined as shown in equation (4).

$$p_\theta(\mathbf{x}_{t-1}\mid\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1};\boldsymbol{\mu}_\theta(\mathbf{x}_t,t),\boldsymbol{\Sigma}_\theta(\mathbf{x}_t,t)\right)$$

Maximize the likelihood or minimize the variational lower bound (VLB), which is usually simplified by minimizing the mean square error between the predicted and true noises, as shown in equation (5).

ISSN 2959-6157

$$\mathcal{L}_{\text{simple}} = E_{t,\mathbf{x}_0,\epsilon}[\| \epsilon - \epsilon_\theta(\mathbf{x}_t, t) \|^2] \quad (5)$$

where the key parameters are defined as: $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$; α_t denotes the noise retention factor at the time step , and $\bar{\alpha}_t$ denotes the α_i cumulative product from step one to step t .

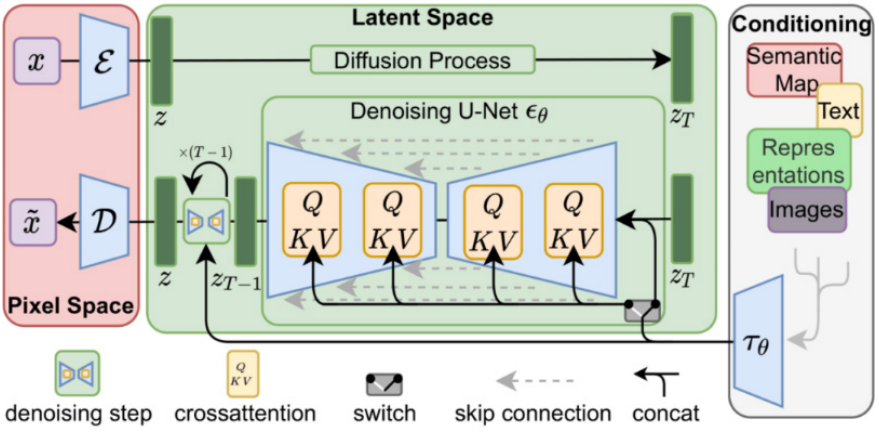

Fig.3 Deeper Understanding of Diffusion Models [5]

Diffusion models are the mainstream generative technique nowadays. DDPM (Denoising Diffusion Probabilistic Model) is the theoretical foundation of denoising diffusion probabilistic models. It proposes a denoising architecture and noise prediction objective function based on U-Net. Improved DDPM builds on this by optimizing the noise scheduling strategy and introducing a learnable variance design. However, DDPM's Markov chain sampling mechanism is slower than GAN's. Generating large-scale high-quality data may take days or months under the existing hardware conditions. LDM/Stable Diffusion diffuses in a low-resolution latent space, enabling consumer-grade GPUs to generate high-resolution images 96% more efficiently. This has greatly facilitated the development of AIGC applications, but also causes a loss of textual detail expression, leading to a semantic attenuation problem in text-to-image generation tasks. The generated results often fail to accurately match the description of the input text. Diffusion modeling techniques have the advantages of SOTA generation quality, training stability, and flexible condition control, but still face challenges such as slower sampling speed than GAN and high cost of pixel space. Optimization directions include distillation to accelerate less-step sampling, hidden space compression to reduce redundancy, and non-Markovian process to achieve jump denoising to further unleash the potential in the field of image generation and content creation.

Table 1 classifies and summarizes image generation models, including GANs, VAEs, diffusion models, and frontier hybrid models. GANs like DCGAN inference are fast, but prone to pattern collapse. StyleGAN2/3 excels at attribute editing, but costly to train. VAEs have a highly interpretable β-VAE latent space but generate fuzzy images. VQ-VAE2 uses discrete modeling, resulting in a loss of codebook quantization. DDPM is theoretically rigorous but slow to sample. LDM is efficient but loses text details. DDPM is theoretically rigorous but slow in sampling. LDM is computationally efficient but loses text details. Table 1 lists the advantages and limitations of different image generation models.

Table 1. Comparison of the main limitations of the model's core strengths

Model Category	representative model	Core Advantages	Major limitations
GANs	DCGAN	Millisecond reasoning speed	Mode crashes (crash rate >30%)
	StyleGAN2/3	Fine-grained attribute editing (PSNR>32dB)	High training cost (1024×1024 requires 256 V100 days)
VAEs	β-VAE	Interpretability of latent space (editing success rate >75%)	Generate blur (SSIM<0.65 @512×512)
	VQ-VAE2	Discrete latent space modeling (codebook size K=512)	Codebook quantization loss (PSNR↓2dB)

diffusion model	DDPM	Theoretical Rigor (VLB Optimization)	Slow sampling rate (vs GAN: ×1000)
	LDM	96% increase in computational efficiency (64×64 latent space)	Loss of textual detail (BLEU<0.4)

3. Challenges and future directions

Deep learning image generation techniques have evolved. GANs open a new era of generative AI with excellent visual fidelity, VAE builds a probabilistic modeling framework to achieve potential spatial controllability, and diffusion models have become the mainstream due to their training stability and modulation flexibility. Potential diffusion models have driven the industrialization of AIGC in the fields of art creation and scientific visualization. These techniques have been widely used in art creation, visual design, image editing, scientific research, and many other fields. However, challenges remain, including improving computational efficiency, fine-grained controllability and interpretability, solving the long-tail problem and low-sample generation, overcoming high-quality 3D and video generation, and addressing ethical and security risks.

Future research will focus on creating more efficient, controllable, and responsible image generation models, deepening the theoretical foundation, and exploring innovative applications in various fields. The future of deep learning image generation is not only about generating more realistic images, but also about safely, reliably, and creatively meeting the diverse needs of human society. The ultimate goal is a breakthrough in pixel-level fidelity and to build a creative engine for human-aligned values, which will safely and reliably empower cross-domain innovation.

Table 2 compares the efficiency, controllability, and security of GAN, VAE, and diffusion models. GAN training is unstable, VAE generates fuzzy samples, and the diffusion model is slow. The study proposes a GAN-diffusion relay architecture. StyleGAN3 generates the initial composition, and the diffusion model restores it in fewer steps to achieve high-quality output in real time while preserving rhythm. To address the model's inability to understand complex commands, ControlNet and semantic decoupling technology transform abstract concepts into visual units to promote the transformation of the generation model from "command execution" to "intent understanding." For security risks, blockchain watermarking and traceability technology injects invisible fingerprints or other identifiable information into the generation stage. Combined with a distributed depository and a mobile second traceability mechanism, front-end trust management of design-as-compliance is realized. The table presents multidimensional comparisons and scenarios in a structured form, providing a systematic reference for subsequent research..

Table 2.Challenges and solutions to the three models

Challenge Type	GAN	VAE	Diffusion Modeling	Fusion Solutions
Efficiency bottlenecks	Training instability	Fuzzy samples require iterative optimization	Hundred-step sampling latency	GAN-Diffusion Hybrid Architecture
Controllability flaws	Weak layout control	High crypto spatial coupling	Dependence on textual cues	ControlNet + Semantic Decoupling
Security risks	Depth falsification vulnerabilities	Privacy leakage risks	Copyright disputes	Blockchain Watermarking Traceability

4.Conclusion

This thesis analyzes the evolution of deep learning image generation technology over the past ten years, focusing on potential diffusion models like Stable Diffusion and their industrial applications in art creation and scientific visualization. It discusses challenges in efficiency, control, and safety and the technology's impact on various fields. By systematically studying the technology's development, this work provides a reference for related research and clarifies the research direction of developing efficient, controllable, and responsible image generation models. It analyzes existing problems and paves the way for a creativity engine aligned with human values and cross-domain innovation in a safe manner. The thesis encourages image generation technology to evolve beyond pixel-level realism to meet society's diverse needs. Overall, this research is significant in understanding the

ISSN 2959-6157

field's development, overcoming technical challenges, and expanding industrial applications.

References

[1] T. Karras et al., Alias-Free Generative Adversarial Networks, in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, pp. 852–863, Dec. 2021.

[2] A. Vahdat and J. Kautz, NVAE: A Deep Hierarchical Variational Autoencoder, in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 19667–19679, Dec. 2020.

[3] J. Ho, A. Jain, and P. Abbeel, Denoising Diffusion Probabilistic Models, in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 6840–6851, Dec. 2020.

[4] D. Kang et al., MobileStyleGAN: A Lightweight StyleGAN for Mobile Devices, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 22838–22847, Jun. 2023.

[5] R. Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10684–10695, Jun. 2022.

[6] Yang Song et al., Consistency Models, in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 32211–32228, Jul. 2023.

[7] T. Karras et al., Training Generative Adversarial Networks with Limited Data, in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.

[8] T. Chen, L. Li, and R. Grosse, Isolating Sources of Disentanglement in Variational Autoencoders, in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, pp. 11526–11538, Dec. 2021.

[9] P. Esser, R. Rombach, and B. Ommer, Taming Transformers for High-Resolution Image Synthesis, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 12873–12883, Jun. 2021.