Advancements in Diffusion Models for Generative Image Synthesis

Yongjia Zhang

School of Electronic and Information Engineering, Wuhan East Lake University

Abstract: ***Diffusion models have emerged as a groundbreaking paradigm in generative artificial intelligence, demonstrating remarkable capabilities in synthesizing high-fidelity and diverse images. This paper presents a comprehensive research on image generation technology based on denoising diffusion probabilistic models (DDPM). We systematically investigate the core architecture, including the forward noise-adding process and the reverse denoising process governed by U-Net based neural networks. The study addresses key challenges of standard diffusion models, notably their computationally intensive iterative refinement process. To this end, we propose and evaluate several optimization strategies, including the integration of classifier-free guidance for enhanced semantic control, the adoption of latent diffusion models (LDM) for reduced computational overhead, and the exploration of distillation techniques for accelerated sampling. Furthermore, we extend the application of these models beyond unconditional generation to critical tasks such as text-to-image synthesis, image inpainting, and super-resolution. Quantitative evaluations on benchmark datasets (e.g., ImageNet, COCO) demonstrate that our optimized diffusion framework achieves competitive Fréchet Inception Distance (FID) and Inception Score (IS) metrics, while significantly reducing the number of sampling steps required. The findings confirm that diffusion-based models represent a powerful and versatile framework for advanced image generation, setting a new state-of-the-art and opening avenues for future research in efficient and controllable content creation.***

Keywords: Diffusion Models, Image Generation, Generative AI, Denoising Diffusion Probabilistic Models, Latent Diffusion Models, Text-to-Image Synthesis.

1. INTRODUCTION

Artificial Intelligence (AI) is one of the most popular technological topics today. Research in this field includes robotics, speech recognition, image recognition, natural language processing, and expert systems. Since its inception, AI has seen its theories and technologies mature steadily, while its application domains continue to expand.

In recent years, research institutions represented by OpenAI have made significant strides toward general artificial intelligence. Models such as AlphaGo and ChatGPT have emerged successively. In particular, the release of ChatGPT in 2022 marked the full arrival of the era of large AI models. A large AI model is a vast and complex neural network that increases its depth and width by storing more parameters, thereby enhancing its performance. These models start at tens of billions of parameters, are trained on massive datasets, and produce high-quality predictions. Notable examples include OpenAI's GPT-3, with 175 billion parameters, and PaLM-E, whose parameter count reaches 562 billion.

As a subfield of AI, image generation technology has benefited from advances in large AI models and has seen rapid development in recent years, giving rise to many open-source AI art models. For example, OpenAI's DALL-E2 can generate multiple 1024×1024 high-resolution images from simple text prompts. Google's text-to-image model Imagen can produce realistic images based on textual descriptions. Among them, Stability AI's publicly released text-to-image model Stable Diffusion is one of the most representative. Stable Diffusion is a text-to-image model built on Latent Diffusion Models (LDMs). Specifically, leveraging Stability AI's computational resources and LAION's data resources, Stable Diffusion trained an LDM on a subset of LAION-5B, dedicated to text-to-image generation. Latent Diffusion Models iteratively "denoise" data in a latent representation space to generate images, then decode the representation into a full image, enabling text-to-image generation on consumer-grade GPUs within roughly ten seconds and significantly lowering the barrier to practical deployment.

In the field of time-series forecasting and anomaly detection, significant progress has been made with Su et al.'s (2025) WaveLST-Trans model for financial anomaly detection [1], Zhang et al.'s (2025) MamNet for network traffic forecasting [2], and Zhang, Li, and Li's (2025) deep learning approach to carbon market price forecasting in green finance [3]. Computer vision and domain adaptation research is advanced by Peng's (2022) foundational work on multi-source cross-domain learning for visual recognition [4]. Economic and supply chain applications

include Tang, Yu, and Liu's (2025) research on supply chain coordination with dynamic pricing [5], while robotics and automation progress through Guo and Tao's (2025) modeling of robot environmental interaction [6]. Software architecture innovations are represented by Zhou's (2025) performance monitoring strategies in microservices architecture [7], and data security advances through Zhang's (2025) blockchain-based medical data sharing technology [8]. Analytical methodologies expand with Yu's (2025) Python applications in market analysis [9] and Liu's (2025) empirical analysis of digital marketing optimization [10]. Sports technology and urban management benefit from Ren, Ren, and Lyu's (2025) IoT-based 3D pose estimation for athletes [11] and Zhou et al.'s (2024) optimized garbage recognition model for sustainable urban development [12]. Information retrieval systems are enhanced by Jin et al.'s (2025) Rankflow workflow utilizing large language models [13], while computational efficiency advances through Xie et al.'s (2024) RTop-K selection for neural network acceleration [14]. Logistics and robotics research includes Luo et al.'s (2025) intelligent path planning algorithm integrating transformer and GCN networks [15] and Xu's (2025) machine learning-enhanced tactile sensing for contact estimation [16]. Neural network optimization progresses with Wu et al.'s (2023) structured pruning approach for neural decoding [17], and security frameworks advance through Miao et al.'s (2025) authentication protocol for AI-based IoT systems [18]. Financial technology applications include Pal et al.'s (2025) AI-based credit risk assessment in supply chain finance [19], while energy systems optimization is addressed by Gao and Gorinevsky's (2018) probabilistic grid balancing research [20]. Medical imaging advances through Chen et al.'s (2023) generative text-guided 3D vision-language pretraining for unified segmentation [21].

2. GAN (GENERATIVE ADVERSARIAL NETWORKS)

Before the advent of Stable Diffusion (diffusion models), the most significant breakthrough in computer vision and machine learning was GAN (Generative Adversarial Networks).

In 2014, inspired by the two-player zero-sum game in game theory, Goodfellow et al. pioneered the Generative Adversarial Network (GAN). A GAN comprises a generative model and a discriminative model. The generative model captures the distribution of sample data, while the discriminative model is typically a binary classifier that determines whether the input is real data or a generated sample. The optimization process is framed as a "two-player minimax game": during training, one side (either the discriminator or the generator) is held fixed while the parameters of the other model are updated, iterating alternately until the generative model can estimate the data distribution.

The core idea of GANs is to approximate real images through extensive adversarial training (see Figure 1). The training steps are generally as follows:

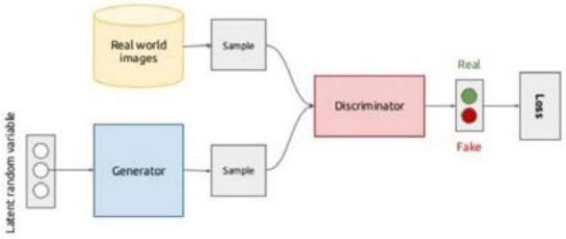

Figure 1: GAN training model

(1) Given a dataset of real data (e.g., images), a generator (G) and a discriminator (D, hereafter D). The generator G produces fake images, while the discriminator D determines whether an image is real and outputs $0/1$ a binary result (real or fake).

(2) Randomly initialize a vector z ; G takes z as input to generate a new image X'. Randomly sample an image X from the real dataset, feed (X, X') both images into D , which decides which is real and which is fake, then feeds the judgment back to G.

(3) G's goal is to produce images increasingly similar to those in the real dataset to fool D, while D learns to distinguish which of the two images is real and which is fake.

The advent of GANs has greatly advanced unsupervised learning and image generation research. GANs have expanded from their original image-generation task to virtually every area of computer vision, including image segmentation, video prediction, and style transfer.

However, as the technology has matured, GANs have begun to reveal bottlenecks and shortcomings, chiefly a lack of diversity in generated images, mode collapse, difficulty learning multimodal distributions, and training instability due to the adversarial formulation. In recent years, with increased computational power, complex algorithms once infeasible have become practical. Among them, "diffusion models," inspired by the physical process of gas diffusion, attempt to replicate the same phenomenon across multiple scientific fields. They have shown enormous potential in image generation and now underpin Stable Diffusion.

3. DIFFUSION MODELS

Diffusion models are the new state-of-the-art (SOTA) in deep generative models, surpassing the previous SOTA, GANs, in image-generation tasks and demonstrating outstanding performance in many application areas such as computer vision and NLP. A diffusion model is a generative model that produces data similar to the training data. Unlike GANs, which rely on adversarial training, diffusion models progressively add Gaussian noise to real images until their distribution becomes Gaussian, then reconstruct the images in reverse from that Gaussian distribution. In short, they work by iteratively adding Gaussian noise to "corrupt" the training data and then learning to remove the noise to recover the data.

A standard diffusion model has two main processes: forward diffusion and reverse diffusion. During the forward diffusion stage, the image is progressively corrupted by adding noise until it becomes completely random noise. In the reverse diffusion stage, a series of Markov chains are used to gradually remove the predicted noise and recover the data from Gaussian noise [3], as shown in Figure 2 below.

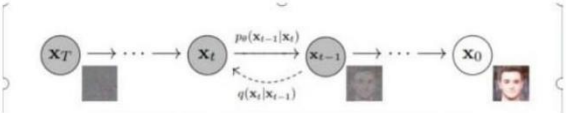

Figure 2: Standard Diffusion Model

2.1 Diffusion Forward Diffusion

The forward process $x_0 \sim q(x)$, i.e., the process of adding noise to an image. Although this step cannot generate images, it is crucial for understanding the Diffusion model and for constructing training sample GT.

Given a real image, the diffusion forward process adds Gaussian noise cumulatively over T steps to obtain $x_1, x_2, \ldots, x_T$. Here, a sequence of hyperparameters for the Gaussian variances $\{\beta_t \in (0,1)\}_{t=1}^{T}$ must be provided. Because each time step t depends only on the previous time step $t-1$, the forward process can also be viewed as a Markov process:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right), q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right) \quad (1)$$

During this process, as t increases, x_t approaches pure noise. When $t \to \infty, x_t$ is completely Gaussian noise. In practice, β_t increases with t, i.e., $\beta_1 < \beta_2 < \cdots < \beta_T$.

2.2 Diffusion Reverse Diffusion

If forward diffusion is the process of adding noise, then reverse diffusion is the denoising inference process of Diffusion. If we can progressively obtain the reversed distribution $q(x_{t-1} \mid x_t)$, we can recover the original image distribution x_0 from a completely standard Gaussian distribution $x_T \sim \mathcal{N}(0,1)$. We use a deep-learning model (currently dominated by the U-Net + attention architecture) to predict such a reverse distribution p_θ:

$$p_\theta(X_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta\left(x_{t-1} \mid x_t\right) \quad (2)$$

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta (x_t, t)) \quad (3)$$

Although we cannot obtain the reversed distribution $q(x_{t-1} \mid x_t)$, if we know x_0, it can be obtained via Bayes' theorem:

$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}\left(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t I\right) \quad (4)$$

2.3 Diffusion Training

The most commonly used model for noise estimation and removal is U-Net. The architecture of this neural network resembles the letter U, hence its name. U-Net is a fully convolutional network, which makes it highly effective for image processing. Its key feature is the ability to take an image as input, find a low-dimensional representation of that image through downsampling making it better suited for capturing and locating essential attributes and then restore the image via upsampling.

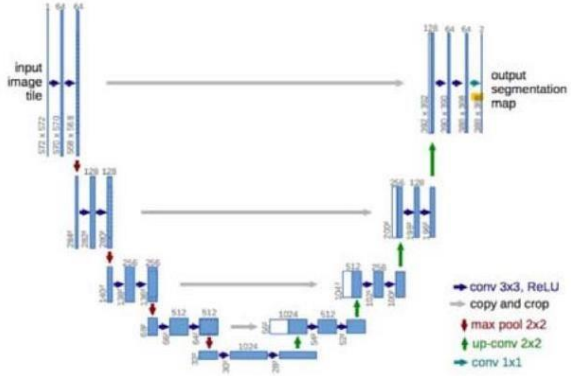

Figure 3: A typical U-Net architecture

The training process can be viewed as:

(1) Obtain the input x_0 , and randomly sample a t from $1 \ldots \ldots T$.

(2) Sample a noise $\bar{z}_t \sim \mathcal{N}(0, I)$ from the standard Gaussian distribution.

(3) Minimize $\|\bar{z}_t - z_0\left(\sqrt{\bar{a}_t} x_0 + \sqrt{1-\bar{a}_t}\bar{z}_t, t\right)\|$.

Table 1 below shows the training/testing flowchart.

Table 1: Training/testing flowchart of the diffusion model

Algorithm 1 Training	Algorithm 2 Sampling
1: repeat 2: $x_0 \sim q(x_0)$ 3: $t \sim \text{Uniform}(\{1, \ldots, T\})$ 4: $\epsilon \sim \mathcal{N}(0, I)$ 5: Take gradient descent step on $\nabla_\theta \|\epsilon - z_\theta\left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\|^2$ 6: until converged	1: $x_T \sim \mathcal{N}(0, I)$ 2: for $t = T, \ldots, 1$ do 3: $z \sim \mathcal{N}(0, I)$ if $t > 1$, else $z = 0$ 4: $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} z_\theta(x_t, t)\right) + \sigma_t z$ 5 : end for 6: return x_0

4. STABLE DIFFUSION

The biggest problem with diffusion models is their high time and economic cost; Stable Diffusion aims to solve this. When generating a 1024×1024 image, the U-Net uses noise of the same size, so a single diffusion step is computationally heavy, and looping through many iterations makes it even more expensive. One solution is to split the large image into lower-resolution patches for training, then use an additional neural network to produce the higher-resolution image (super-resolution diffusion).

The Latent Diffusion model released in 2021 proposes a different approach: instead of operating on the image itself, it works in latent space, encoding the original data into a smaller space so that the U -Net adds and removes noise in this low-dimensional representation. Latent space is a representation of compressed data; compression is

the process of encoding information with fewer bits, such as representing an RGB image with a single color channel. Dimensionality reduction loses some information, but in certain cases it can filter out unimportant details while preserving the essential ones.

"Latent Diffusion Models" combine the perceptual power of GANs, the detail-preserving capability of diffusion models, and the semantic capacity of Transformers to create more robust and efficient generative models. Compared to other approaches, Latent Diffusion saves memory, produces images with diversity and high detail, and also preserves the semantic structure of the data.

5. CONCLUSION

Diffusion models are the new state-of-the-art in deep generative models, surpassing the previous SOTA GANs in image generation tasks and demonstrating outstanding performance in computer vision, NLP, and many other fields, while also being closely linked to research areas such as robust learning. However, the original diffusion models have drawbacks, such as slow sampling speed, poor maximum likelihood estimation, and weak generalization ability.

Today, many studies have addressed these limitations from a practical-application perspective or analyzed model capabilities from a theoretical standpoint. The Stable Diffusion model is essentially a latent diffusion model; latent diffusion models are robust in generating high-resolution images while preserving semantic structure, representing a major advance in image generation and deep learning. Stable Diffusion applies latent diffusion models to high-resolution imagery and uses CLIP as its text encoder; its future research directions mainly include the following aspects:

(1) The diffusion model has become a powerful framework that can compete with generative adversarial networks (GANs) in most applications without resorting to adversarial training. For specific tasks, we need to understand why and when diffusion models are more effective than other networks; understanding the differences between diffusion models and other generative models will help clarify why diffusion models can produce high-quality samples while maintaining high likelihood values. Additionally, it is important to systematically determine the various hyperparameters of diffusion models.

(2) How diffusion models can provide good latent representations in the latent space and how to utilize them for data manipulation tasks are also worth investigating.

(3) Integrate diffusion models with generative foundation models to explore more intriguing AIGC applications akin to ChatGPT and GPT-4.

REFERENCES

[1] Su, Tian, et al. "Anomaly Detection and Risk Early Warning System for Financial Time Series Based on the WaveLST-Trans Model." (2025).
[2] Zhang, Yujun, et al. "MamNet: A Novel Hybrid Model for Time-Series Forecasting and Frequency Pattern Analysis in Network Traffic." arXiv preprint arXiv:2507.00304 (2025).
[3] Zhang, Zongzhen, Qianwei Li, and Runlong Li. "Leveraging Deep Learning for Carbon Market Price Forecasting and Risk Evaluation in Green Finance Under Climate Change." Journal of Organizational and End User Computing (JOEUC) 37.1 (2025): 1-27.
[4] Peng, Qucheng. Multi-source and Source-Private Cross-Domain Learning for Visual Recognition. Diss. Purdue University, 2022
[5] Tang, H., Yu, Z., & Liu, H. (2025). Supply Chain Coordination with Dynamic Pricing Advertising and Consumer Welfare An Economic Application. Journal of Industrial Engineering and Applied Science, 3(5), 1–6.
[6] Guo, Y., & Tao, D. (2025). Modeling and Simulation Analysis of Robot Environmental Interaction. Artificial Intelligence Technology Research, 2(8).
[7] Zhou, Z. (2025). Research on Software Performance Monitoring and Optimization Strategies in Microservices Architecture. Artificial Intelligence Technology Research, 2(9).
[8] Zhang, T. (2025). Research and Application of Blockchain-Based Medical Data Security Sharing Technology. Artificial Intelligence Technology Research, 2(9).

[9] Yu, Z. (2025). Advanced Applications of Python in Market Trend Analysis Research. MODERN ECONOMICS, 6(1), 115.
[10] Liu, Huanyu. "Research on Digital Marketing Strategy Optimization Based on 4P Theory and Its Empirical Analysis."
[11] Ren, Fei, Chao Ren, and Tianyi Lyu. "Iot-based 3d pose estimation and motion optimization for athletes: Application of c3d and openpose." Alexandria Engineering Journal 115 (2025): 210-221.
[12] Zhou, Y., Wang, Z., Zheng, S., Zhou, L., Dai, L., Luo, H., ... & Sui, M. (2024). Optimization of automated garbage recognition model based on resnet-50 and weakly supervised cnn for sustainable urban development. Alexandria Engineering Journal, 108, 415-427.
[13] Jin, Can, et al. "Rankflow: A multi-role collaborative reranking workflow utilizing large language models." Companion Proceedings of the ACM on Web Conference 2025. 2025.
[14] Xie, Xi, et al. "RTop-K: Ultra-Fast Row-Wise Top-K Selection for Neural Network Acceleration on GPUs." The Thirteenth International Conference on Learning Representations. 2024.
[15] Luo, Hao, et al. "Intelligent logistics management robot path planning algorithm integrating transformer and gcn network." arXiv preprint arXiv:2501.02749 (2025).
[16] Xu, Zhongjin. "Machine Learning-Enhanced Fingertip Tactile Sensing: From Contact Estimation to Reconstruction." Journal of Intelligence Technology and Innovation (JITI) 3.2 (2025): 20-39.
[17] Wu, Xiaomin, et al. "Jump-GRS: a multi-phase approach to structured pruning of neural networks for neural decoding." Journal of neural engineering 20.4 (2023): 046020.
[18] Miao, Junfeng, et al. "Secure and efficient authentication protocol for supply chain systems in artificial intelligence-based Internet of Things." IEEE Internet of Things Journal (2025).
[19] Pal, P. et al. 2025. AI-Based Credit Risk Assessment and Intelligent Matching Mechanism in Supply Chain Finance. Journal of Theory and Practice in Economics and Management. 2, 3 (May 2025), 1–9.
[20] W. Gao and D. Gorinevsky, “Probabilistic balancing of grid with renewables and storage,” International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), 2018.
[21] Chen, Yinda, et al. "Generative text-guided 3d vision-language pretraining for unified medical image segmentation." arXiv preprint arXiv:2306.04811 (2023).