Exploring Diffusion Models: A Comprehensive Survey from Theory to Applications

Authors: Liu Yiran

Date: 2024-01-06T00:00:00+00:00

Abstract

Diffusion models constitute a powerful class of generative models capable of producing high-quality results across multiple domains, including images, text, and audio. This survey aims to consolidate and analyze the latest research advances in diffusion-based generative models applied to the visual domain, encompassing both theoretical and practical contributions to the field. This paper first examines the characteristics and principles of three mainstream paradigms: Denoising Diffusion Probabilistic Models, score-based diffusion generative models, and diffusion generative models based on stochastic differential equations, and analyzes relevant derivative models designed to optimize internal algorithms and improve sampling efficiency. Secondly, it provides a comprehensive review of the current application landscape of diffusion models, including practical deployments in computer vision, natural language processing, time series analysis, multimodal research, and interdisciplinary domains. Finally, based on current research trends and challenges, it outlines prospective future development directions for diffusion models, aiming to provide guidance and inspiration for research in this field. This paper seeks to furnish researchers with a comprehensive overview of diffusion model research and applications, emphasizing its significant position and future potential in the Artificial Intelligence Generated Content (AIGC) domain.

Full Text

Preamble

Exploring Diffusion Models: A Comprehensive Review from Theory to Application

YiRan Liu[1]
[1]College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, 100029

Abstract: Diffusion models represent a powerful class of generative models capable of producing high-quality results across multiple domains including images, text, and audio. This review aims to summarize and analyze the latest research progress in diffusion generative models applied to the vision domain, encompassing both theoretical and practical contributions in the field. Initially, the article discusses the characteristics and principles of three mainstream models: denoising diffusion probabilistic models, score-based diffusion generative models, and diffusion generative models based on stochastic differential equations. It also analyzes derivative models aimed at optimizing internal algorithms and improving sampling efficiency. Furthermore, the review provides a comprehensive summary of current applications of diffusion models, including computer vision, natural language processing, time series analysis, multimodal research, and interdisciplinary fields. Finally, based on current trends and challenges, it offers a forecast for the future direction of diffusion models, aiming to guide and inspire research in the field. This article is intended to provide researchers with a comprehensive overview of diffusion model research and application, emphasizing its significant role and potential in the field of Artificial Intelligence Generated Content (AIGC).

Keywords: Deep Learning; Diffusion Models; Artificial Intelligence Generated Content; Generative Models; Multimodal Application

---

1 Introduction to Diffusion Models

Against the backdrop of rapid technological development, computer vision and artificial intelligence have become key driving forces advancing numerous fields. Particularly in the domain of generative models, the evolution from basic models to today' s highly complex and sophisticated architectures has not only broadened our theoretical horizons but also demonstrated tremendous potential in practical applications. As shown in [Figure 1: see original paper], among the many generative models, diffusion models have emerged prominently due to their unique generation methodology and high-quality output, rapidly becoming a focal point in both academia and industry.

With the advent of the big data era, effectively processing and utilizing massive amounts of data has become a significant challenge. In this context, generative models—especially diffusion models—have demonstrated remarkable capabilities in handling and generating high-quality data. Diffusion models, with their high flexibility and powerful generation capacity, have found applications across numerous fields, including but not limited to image generation, super-resolution, image restoration and editing, as well as natural language processing and multimodal learning.

Moreover, as computational power continues to grow, diffusion models are gradually overcoming limitations such as high computational costs, continuously enhancing their practicality and efficiency. Simultaneously, the academic community is actively exploring ways to further improve these models' performance and generalization capabilities through algorithmic optimization and novel architectural designs. However, research and application of diffusion models remain in a developmental stage, with many potential applications and improvement opportunities yet to be fully explored. Consequently, this article aims to review the latest research findings and progress in diffusion models, analyze current application scenarios and challenges, and predict future development trends. We begin with the fundamental theories of these models, progressively delve into various model designs and applications, and ultimately explore how to translate these theories and techniques into effective tools for solving real-world problems. Through this approach, we hope to provide a comprehensive and in-depth reference for diffusion model research and application.

As illustrated in [Figure 2: see original paper], generative models are essentially a set of probability distributions. On the left is a training dataset where all data are independent and identically distributed random samples drawn from some data distribution pdata. On the right is the generative model (probability distribution), which seeks to find a distribution p that is closest to pdata. By sampling new instances from p , we can obtain a continuous stream of new data. However, pdata often has a highly complex form, and images have high dimensionality, making it difficult to traverse the entire space while the observable data samples remain limited.

Currently, generative models have four main branches, as shown in [Figure 3: see original paper]: Generative Adversarial Networks (GAN) proposed by Ian Goodfellow et al. in 2014 [1], which operates through adversarial training between a discriminator and generator to produce images realistic enough to deceive the discriminator. GAN has achieved remarkable success in image generation, super-resolution, and image editing. Variational Auto-Encoder (VAE) [2] is a probability distribution-based generative model proposed by Kingma and Welling in 2013. VAE' s core idea involves encoding input images into feature vectors through an encoder that learns the mean and variance of a Gaussian distribution, while a decoder transforms these feature vectors back into generated images, focusing on learning generation capabilities. VAE finds extensive application in generating interpretable samples and samples with specific attributes. Normalizing Flow (NF) [3,4] is a generative model that transforms simple distributions into complex ones through a series of invertible transformation functions. These functions modify the shape of the probability density function to gradually approximate the target distribution. Normalizing flow models excel at generating high-dimensional data and samples with multimodal distributions. A key advantage is the ability to generate samples through stepwise transformations, making the generation process interpretable and controllable. Diffusion Models (DM) [5] are generative models that utilize forward and reverse processes to generate samples. In the forward process, noise is gradually added to data, while in the reverse process, the model attempts to predict the noise added at each step, thereby gradually recovering noise-free samples. Diffusion models employ deep learning backpropagation algorithms for training but are essentially Markov models. A key advantage of diffusion models is their ability to generate high-quality samples with solid theoretical foundations including probability models and stochastic differential equations. They have also gained favor for generating highly realistic and diverse samples, breaking GAN' s long-standing dominance in challenging image synthesis tasks and demonstrating impressive performance in computer vision [6–16], natural language processing [17–22], time series [23–25], multimodal [26–33], and interdisciplinary integration with traditional subjects [34–42].

However, these models typically require long training times and substantial computational resources because the reverse diffusion process involves numerous iterative steps. Nevertheless, as research progresses, many methods are being proposed to optimize these models'efficiency and performance. In comprehensive comparison, diffusion models are considered the current state-of-the-art generative models due to their superior performance and continuous technological optimization.

In summary, diffusion models represent a research hotspot in deep learning, holding not only theoretical significance but also demonstrating tremendous application potential in practice, leading the future development of generative models. Currently, diffusion models can be mainly categorized into three types: Denoising Diffusion Probabilistic Models (DDPM) [5,43,44], score-based generative models [45–47], and diffusion generative models based on Stochastic Differential Equations (SDEs) [48]. The following sections will provide in-depth discussion and analysis of the construction, theoretical foundations, and differences in generation processes among these three model categories.

1.1 Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models are deep generative models inspired by non-equilibrium thermodynamics that have demonstrated exceptional performance in generating high-quality images, audio, and other complex data distributions in recent years. The core idea of DDPM is to simulate the diffusion process, as shown in [Figure 4: see original paper], gradually transforming structured data into unstructured noise data, then recovering the original data through an inverse process.

In the forward diffusion process, the model progressively adds Gaussian noise to the original data, forming a series of data states x . This process can be defined through the following Markov chain: q x x. Here, Î represents diffusion steps, (1.1) tb is the variance hyperparameter associated with each step, I is the identity matrix. q x x - allows us to directly sample any noise version tx from the original data tx, i.e., tb can be directly sampled.

In the reverse generation process, DDPM' s core task is to gradually remove the noise introduced in the forward process and recover clear data. This process aims to train a neural network to simulate the inverse process of generating original data from noise. In practice, this is typically achieved by minimizing the variational lower bound, where the negative log-likelihood variational lower bound vlbL proposed by Sohl-Dickstein et al. [49] usually contains multiple KL divergence terms: log (KL p x (1.2) KL p x 0) ||, making it as close as possible to the true posterior distribution. Here, KL represents the Kullback-Leibler divergence between two probability distributions.

Through this iterative denoising and reconstruction process, DDPM can generate samples extremely similar to the original data with very high quality and detail fidelity. This method' s success depends on carefully designed diffusion steps and efficient network architectures, as well as the ability to accurately estimate probability distributions during training. With continuous technological advancement, DDPM has demonstrated powerful potential in image synthesis, audio generation, and other fields, becoming an important research direction in deep learning and generative models.

1.2 Score-Based Generative Diffusion Models

Score-based generative diffusion models, also known as Score-based Generative Models, are advanced deep generative models that generate data by manipulating the score function of data distributions, which is the gradient of the log probability density function. The core idea of these models is that by gradually adjusting the noise added to data, one can guide the data transformation from a simple Gaussian distribution to a complex target distribution. In these models, the score function represents the rate of change of the data distribution at each point, providing guidance for the data generation path.

During the training phase, the objective of score-based models is to learn a neural network that can accurately estimate the score at given data points. Typically, this network is trained to minimize the difference between the predicted score and the true score, which can be expressed using squared error. Here, E denotes expectation, indicating the average error over all data points and noise instances, sq is the score function learned by the model, s+ ò is the true score at data point x, pdata is the true data distribution, ò represents noise sampled from the standard normal distribution N(0,I), and s is the standard deviation of the noise used to adjust the noise level.

In the data generation process, the model starts from a simple Gaussian distribution and then gradually applies reverse diffusion steps to generate data. This process is typically modeled as a continuous stochastic process that can be represented by the following Stochastic Differential Equation (SDE): (,) (1.4). Here, dx represents the infinitesimal change in data x over a small time interval, s x t is the score function at time t, representing the direction and rate of probability density change of the data distribution at x. dt is the infinitesimal time increment, ( )g t is the noise coefficient, and dW is the noise term. By solving this SDE, the model can gradually generate samples from the target data distribution starting from noise samples.

Although score-based generative models have shown tremendous potential in generating high-quality and diverse samples, they have high computational costs and require substantial data and time for training. Additionally, designing and optimizing the score function for these models is challenging because it requires accurate estimation of the data distribution. However, despite these challenges, score-based generative models remain an important and active research direction in the generative model field due to their generation quality and theoretical advantages. With improvements in computational resources and algorithmic advancements, these challenges are expected to be gradually overcome, and score-based models are poised to play a more important role in various data generation tasks in the future.

1.3 Diffusion Generative Models Based on Stochastic Differential Equations

The development of diffusion generative models based on Stochastic Differential Equations (SDEs) originated from research on random diffusion processes in the physical world, which are typically mathematically described using SDEs. These equations have wide applications in physics, chemistry, biology, and finance, providing powerful tools for describing the time evolution of systems under random forces. In the early days of deep learning, these principles were not widely applied, but with growing demand for generative models and deepening theoretical research, researchers began exploring the application of these mathematical tools to data generation. Since increased noise quantity accompanies improved sample generation quality, Song proposed a method for infinite random noise generation to enhance sample quality.

The basic form of stochastic differential equations can be expressed as: g t dB (1.5). Here, tdx is the infinitesimal change of variable tx at time t, f x t is the deterministic drift term describing the expected change trend of the system state, ( )g t is the intensity of the noise term, and tdB is Brownian motion representing random noise. This equation describes a continuous-time random process that can be used to simulate the transition from simple distributions (such as Gaussian distributions) to complex data distributions.

In deep generative models, SDEs are used to construct a process that gradually "diffuses" from the target data distribution to a simple distribution (such as Gaussian noise), and then reverses this process to generate new data points. In practical applications, this involves accurately simulating and reversing the diffusion path, typically requiring the model to learn and approximate the solution of the SDE. In this process, it is crucial to determine appropriate drift and diffusion terms, which are usually implemented through deep neural networks trained to predict data distribution changes at given time steps.

The development of these models has opened new possibilities for high-quality data generation, particularly demonstrating advantages when handling continuous and complex data distributions. However, they also introduce new computational challenges because accurately solving and simulating SDEs is typically computationally intensive. Additionally, designing effective numerical methods to approximate these continuous processes is crucial for achieving efficient and stable generation.

With advances in computational technology and numerical methods, SDE-based diffusion generative models are becoming an important branch in the generative model field. They not only provide new perspectives for understanding and simulating data distributions but also offer new tools for generating high-quality, diverse data, indicating they will play a greater role in future data generation and simulation applications.

2 Development and Derivative Models of Diffusion Models

In diffusion model research and application, the dependence on Markov chains leads to computational burdens during sample generation, particularly when dealing with long-step diffusion processes. As the number of diffusion steps increases, estimating accurate score functions becomes more complex and computationally intensive. These challenges have motivated researchers to explore various innovative methods to optimize and improve diffusion model efficiency and performance. For instance, some research focuses on reducing the number of required diffusion steps or developing more efficient numerical methods to accelerate continuous-time diffusion process simulation. Other studies aim to improve model architectures or training strategies to enhance learning efficiency and generation quality.

Current research remains broadly based on three core diffusion model concepts: denoising diffusion probabilistic models, score-based generative models, and SDE-based generative models. These derivative models not only extend and optimize the original models theoretically but also demonstrate broader potential in practical applications, such as high-quality image and audio generation and simulation of complex data structures.

This chapter aims to deeply explore the latest research progress in diffusion models and their derivatives, including innovations in model architecture, optimization of training methods, and application cases across various fields. Through this review, the article reveals the important position of diffusion models in the generative model field and how they drive the field toward greater efficiency and precision. Simultaneously, it discusses existing challenges and future development directions, providing guidance and inspiration for future research. Through this comprehensive and in-depth analysis, this chapter aims to provide readers with a panoramic view of the current research status and future prospects of diffusion models and their derivatives.

2.1 Optimization Based on Probabilistic Denoising Diffusion Models

In contemporary research on probabilistic denoising diffusion model optimization, the dependence of model performance on parameter selection has attracted widespread attention. Since probabilistic denoising diffusion models contain multiple adjustable parameters, including noise levels, sampling steps, and network structures, their performance and efficiency are significantly influenced by parameter settings. To address this challenge, researchers have been dedicated to exploring and developing various strategies to optimize these key parameters to improve model generation quality and computational efficiency. Consequently, derivative models targeting parameter optimization have gradually been proposed.

2.1.1 Noise Optimization In the field of noise optimization for probabilistic denoising diffusion models, recent research has achieved significant progress by finely controlling noise parameters and improving noise injection methods, substantially enhancing model generation performance and efficiency. Nichol et al. [50] introduced specific cosine noise in the forward noising process to optimize the model' s log-likelihood performance, while incorporating learnable variance parameters in the reverse denoising process to effectively reduce the required sampling steps and improve overall sampling efficiency. This approach of optimizing model performance through fine control of the noise process opens a new direction for noise optimization.

Kingma et al. [51] explored the introduction of Fourier features into network inputs for noise prediction and, through in-depth analysis of the diffusion model' s Variational Lower Bound (VLB), revealed the decisive influence of signal-to-noise ratio function extremes on diffusion loss. This discovery not only deepens our understanding of diffusion model loss structures but also provides important theoretical foundations for model optimization.

Furthermore, research on dynamically adjusting noise parameters [52] continues to advance. One novel method employs a VGG-11 convolutional neural network to train optimal noise parameters for generating image samples with higher FID values, demonstrating that optimizing noise parameters can directly affect generated sample quality. Additionally, to address adversarial attacks that may be encountered during sample generation, the GDMP [53] purification noise framework improves model robustness and classification performance in practical applications by adding a purification mechanism during the denoising process and selecting appropriate diffusion time steps to overwhelm adversarial perturbations while preserving the main content of input images.

2.1.2 Improving Markov Chains In the optimization of probabilistic denoising diffusion models, strategies for improving Markov chains have significantly enhanced model performance in sample generation speed and quality. The proposal of Denoising Diffusion Implicit Models (DDIM) [54] represents a major improvement over traditional forward Markov processes. DDIM implements a non-Markovian process, utilizing predicted normal samples at each denoising step for the next estimation, greatly accelerating sampling speed while maintaining sample quality. This improvement is significant for reducing the model' s dependence on numerous iterations and enhancing generation efficiency.

Following DDIM, more research has focused on further improving these processes. The gDDIM model proposed by Zhang et al. [55] optimizes the denoising process from a numerical perspective. They discovered that when solving the corresponding stochastic differential equations, specific score approximations can be used to implement DDIM, and noted that deterministic sampling schemes can sample more rapidly than stochastic ones. This finding not only reduces the computational burden of the model but also provides new perspectives for understanding and applying deterministic processes in diffusion models.

The common goal of these studies is to optimize Markov chain design for faster and more efficient generation of high-quality data. This involves not only improvements in algorithms and model structures but also fine adjustments to model sampling processes and parameter settings. Through these improvements, models can generate samples in shorter time while ensuring diversity and quality, providing powerful tools for handling complex and high-dimensional data.

2.1.3 Multi-Model Integration In multi-model integration, combining diffusion models with other classical generative models has become an important research direction for improving generation performance and broadening application scope. Researchers have aimed to enhance model generation quality, sample diversity, and training efficiency by integrating diffusion models with different learning strategies and network architectures.

The diffusion decoding model with contrastive representation learning proposed by Sinha et al. [56] exemplifies the combination of diffusion models and representation learning. By introducing contrastive self-supervised learning into the diffusion process, this model not only learns to generate samples from a diffusion prior distribution but also further optimizes sample representation quality through contrastive learning, significantly improving performance on generation tasks and surpassing contemporary VAE models in multiple aspects.

The research by Peebles et al. [57] combines diffusion models with the latest Transformer models. By replacing the commonly used U-Net network in image generation tasks with Transformer networks, they explored the impact of strategies such as increasing network depth and width and adding more tokens on model performance. This combination leverages Transformers' advantages in long-range dependencies and parallel computation, further enhancing diffusion models' capabilities in handling high-dimensional data like images.

Additionally, the integration of GAN models with diffusion models [58] has shown innovative research directions. In these combined models, the robustness and noise injection mechanisms of diffusion models are used to improve GAN's shortcomings in sample generation stability. By introducing diffusion models into GAN's discriminator, researchers can better handle the non-overlapping problem between input data and generated data distributions, thereby improving the stability and quality of generated samples.

2.1.4 Handling Special Data Types In handling special data types, optimization and adaptability research of diffusion models has achieved significant progress. For unconventional data where features are primarily concentrated in low-density regions and scenarios with small sample sizes, a series of improved diffusion models have been proposed to better adapt to these data characteristics, thereby generating high-quality samples on these challenging data distributions.

The research by Sehwag et al. [59] optimized the sampling process of diffusion models by introducing two additional classifiers at each time step, shifting generation focus from high-density to low-density regions and ensuring that attention remains on low-sample-density data manifolds. This strategy allows the model to more effectively generate high-quality samples in low-density regions, improving generation performance on unconventional data.

In few-shot data generation, the FSDM framework utilizes conditional DDPM for small-scale image generation. By combining the VIT [60] framework to aggregate image patch information, it effectively learns the generation process of existing categories and can generate richer and more complex samples to compensate for insufficient sample quantities. Meanwhile, the DAG model [61] focuses on image generation with geometric properties, proposing a method for depth-aware image generation using internal representations, further broadening diffusion model applications on special data types.

For discrete data processing, Austin et al. [62] proposed discrete diffusion models. This model adds multiple transition matrices in the forward process and introduces a new loss function that combines the variational lower bound with auxiliary cross-entropy loss, effectively improving model performance on log-likelihood for image generation and demonstrating the adaptability and potential of continuous diffusion models on discrete data.

2.1.5 Hyperparameter Optimization In the development of diffusion models, hyperparameter optimization has become an important research area for improving model efficiency and generation quality. Considering that diffusion models' dependence on Markov chains in forward and reverse processes may lead to low processing efficiency, researchers have dedicated themselves to optimizing model sampling and training processes through various methods to achieve faster sample generation and more precise model training.

Watson' s research [63] employed reparameterization and repeated gradient computation strategies to optimize fast samplers for diffusion models, effectively reducing computational burden during sampling. By introducing the KID divergence metric as a loss function and using stochastic gradient descent for optimization, this method not only improved sampling efficiency but also maintained generated sample quality. Additionally, through special sampling parameterization techniques, this method can significantly reduce the required sampling steps, further accelerating model operation.

The bilateral denoising diffusion model proposed by Lam et al. [43] optimizes diffusion models from a different perspective. This model introduces scheduling networks and scoring networks to parameterize both forward and reverse processes. Through fine control of these network parameters, the bilateral denoising diffusion model can more effectively learn and simulate data generation processes, significantly reducing the steps required for sample generation while improving overall sample quality.

Through these hyperparameter optimization methods, models can generate required samples with fewer steps, substantially improving overall efficiency. The proposal and application of these methods not only deepen our understanding of diffusion models but also drive model performance and efficiency in practical applications. Through continuous technological innovation and research, diffusion models' performance in processing efficiency and generation quality will continue to improve. In the future, as more efficient hyperparameter optimization strategies are developed, we have reason to believe diffusion models will play a more critical and extensive role in the generative model field, providing powerful and flexible tools for solving various complex data generation tasks.

2.1.6 Reducing KL Divergence In optimizing the reverse denoising process by reducing KL divergence, recent research has achieved significant progress. By deeply understanding the impact of KL divergence in diffusion models on model performance, researchers have dedicated themselves to developing new methods to minimize KL divergence, thereby improving model inference efficiency and generation quality.

The research by Watson et al. [64] incorporated dynamic programming algorithms into the model to optimize the reverse denoising process. They utilized the characteristic that the Evidence Lower Bound (ELBO) can be decomposed into individual relative entropy terms (KL divergence terms), maximizing ELBO by minimizing these KL divergence terms. This approach allows the model to find optimal inference paths while maintaining generation quality, greatly improving inference process efficiency and effectiveness. This fine control of KL divergence demonstrates the importance of deeply understanding model statistical properties for optimization algorithms.

In Xiao' s research [65], a new method for minimizing KL divergence was proposed by integrating Generative Adversarial Networks (GAN) into the reverse denoising process. This method uses GAN to distinguish between real samples and denoised samples, further optimizing the denoising process through adversarial training, minimizing KL divergence and improving inference efficiency. This GAN-combined approach not only improves inference accuracy but also enhances the model' s adaptation to real data distributions through adversarial training.

2.1.7 Reducing Sampling Steps In diffusion model applications, reducing sampling steps to improve generation efficiency is a current research focus. Since traditional diffusion models require iterative generation throughout the entire time horizon, this process is typically time-consuming and inefficient. Therefore, researchers have dedicated themselves to optimizing time steps and sampling processes through various innovative methods to reduce required sampling steps and accelerate model generation speed.

Bao' s research [66] optimized time steps by introducing diagonal and full covariance, achieving significant improvement in DDPM generation efficiency. This optimization method not only accelerates the sampling process but also maintains generated sample quality. Through precise control and optimization of time steps, the model can generate high-quality samples more quickly.

Chung [67] approached the problem from the perspective of stochastic differential equations, using contraction theory for stochastic difference equations to optimize diffusion model sampling steps. By optimizing the initialization images in the forward process, the research found that steps in the reverse denoising process could be significantly reduced. This method provides an effective approach for reducing sampling steps through theoretical in-depth analysis and rigorous mathematical derivation, thereby improving overall generation efficiency.

2.2 Improvements to Score-Based Models

2.2.1 Improved Sampling Algorithms In score-based model generation tasks, improving sampling algorithms is key to generating high-resolution and stable images. Recent research has focused on developing new strategies and techniques to optimize the sampling process and improve the stability and quality of generated samples.

The work of Song et al. [45] represents innovation and progress in this field. They adopted new strategies for determining noise generation scales and suggested applying exponential moving averages to parameters during sampling to improve generation process stability and coherence. Additionally, they minimized weighted combinations of score and loss matching to optimize approximate maximum likelihood training for score diffusion models [46]. These methods not only enhance the model' s ability to generate high-quality images but also provide new perspectives for understanding and optimizing sampling algorithms for score-based models.

Jolicoeur-Martineau et al. [7] focused on improving annealed sampling methods, introducing a more stable consistency annealing sampling scheme. Furthermore, they proposed a hybrid training formula composed of denoising scores and adversarial objectives, aiming to further improve sampling stability and efficiency. Through this hybrid training approach, the model can achieve more efficient and stable training while ensuring generation quality.

2.2.2 Training Gradient Optimization In the training of score-based generative models, gradient optimization is key to improving model efficiency and accelerating inference processes. Since score-based generative models typically involve multiple iterations of sequential computation, traditional methods may lead to slow inference processes, prompting researchers to develop new strategies and techniques to optimize gradient computation during training.

LSGM [68] proposed an innovative variational autoencoder framework designed to train score-based generative models in latent space. The core of this method is applying score-based generative models to non-continuous data and learning smoother models in smaller spaces. By training in lower-dimensional spaces, LSGM can reduce network evaluation times and achieve faster sampling processes. This approach not only improves training and inference efficiency but also enhances generated sample quality by learning smooth score functions.

The Preconditioned Diffusion Sampling (PDS) model [69] optimizes gradient computation from another perspective. PDS reformulates the diffusion process through matrix preconditioning, effectively avoiding the ill-conditioned curvature problems present in traditional diffusion processes. This improvement not only maintains target distribution quality but also significantly enhances model efficiency and stability in practical applications. Through in-depth analysis and optimization of the mathematical formulation of the diffusion process, PDS provides an effective approach for training gradient optimization.

2.2.3 Other Improvement Aspects In score-based generative model research, besides sampling algorithm and training gradient optimization, there have been other multi-faceted innovative attempts to further enhance model effectiveness and applicability. Currently, the forward process largely relies on manual design, limiting model flexibility and adaptability. To address this issue, researchers have dedicated themselves to exploring new theories and methods to understand and optimize score-based generative models more deeply.

The research by Du et al. [70] conducted in-depth analysis and improvement of the forward process of score-based generative models by combining concepts from Riemannian geometry and Monte Carlo methods. They proposed a general framework for parameterized diffusion models based on forward processes, aiming to provide more flexible and efficient ways to design and implement forward processes for score-based generative models. Through testing on standard datasets, they demonstrated the effectiveness of this new framework, achieving results not only in improving model performance but also providing new possibilities for understanding and optimizing forward processes.

This research direction of applying advanced mathematical theories and methods to score-based generative models provides new possibilities for model design and optimization. By gaining deeper understanding of the mathematical nature and structure of models, researchers can design more accurate and efficient models that not only better adapt to various complex data generation tasks but also provide richer theoretical insights.

2.3 Improvements to SDE-Based Models

2.3.1 Multi-Model Integration In the multi-model integration of diffusion models, researchers have continuously expanded and optimized the functionality and performance of base models by introducing new theories and techniques. The work of Zhang et al. [71] proposed a modeling method combining normalizing flows with Stochastic Differential Equations (SDEs) based on differential equations. This method effectively simulates the generation process of complex data distributions by jointly training forward and reverse SDE neural networks and minimizing a common cost function for their differences. The innovation of this method lies in its use of backward SDE diffusion processes that start from Gaussian distributions and end at desired data distributions, providing an effective path for high-quality data generation.

Additionally, Kim et al. [72] proposed a nonlinear diffusion model based on SDE models. This model builds upon the standard SDE model of linear diffusion patterns, combining trainable normalizing flows with diffusion processes. It uses flow networks to learn noise distributions through linear diffusion in latent space, then performs nonlinear diffusion in data space. The innovation of this method lies in improving model flexibility and generation capability while maintaining structural simplicity and ease of training.

Ho et al. [73] introduced a classifier-free guidance method aimed at overcoming the problem that traditional models using classifier guidance limit sampling results to local domains of data distribution. Their method, based on implicit classifiers derived from Bayes' rule, can generate extremely high-fidelity samples using only one conditional diffusion model and one unconditional diffusion model. This classifier-free guidance method provides greater freedom and broader applicability for generative models while maintaining high quality of generated samples.

2.3.2 Improved Sampling Algorithms In current diffusion model research, improving sampling algorithms is key to enhancing model efficiency and generation quality. Particularly in the application of numerical SDE solvers, traditional methods typically require numerous score network evaluations, causing efficiency issues in practical applications. Therefore, researchers have dedicated themselves to developing new strategies and techniques to optimize the sampling process and improve generation efficiency and quality.

Jolicoeur-Martineau et al. [74] designed an optimized SDE solver whose innovation lies in having adaptive step sizes that can be customized individually for score-based generative models, requiring only two score function evaluations. This optimized solver greatly reduces computational burden and improves generation process efficiency while maintaining high quality of generated samples.

Bortoli et al. [75] addressed the problem that the forward generation process requires substantial time to transform noise distributions into Gaussian distributions. They improved generation efficiency by solving entropy-regularized optimal transport problems on path spaces (also known as Schrödinger bridge problems). This method achieves faster forward generation processes by optimizing path distribution transformations while maintaining the quality of generated sample distributions.

Dockhorn et al. [76] connected diffusion models with statistical mechanics and proposed a new critically-damped Langevin diffusion model (CLD). This model simplifies the complexity of directly learning data scores by adding a velocity variable that needs to be learned to the data and learning the conditional distribution function of velocity given data, making it easier to generate high-resolution images.

Liu et al. [77] approached diffusion model processes from the perspective of solving differential equations on manifolds. They found that using conventional numerical methods to solve reverse processes returns low-quality samples, while pseudo-numerical methods are fast. To improve sample quality, they divided numerical methods into gradient and transport parts, aiming to make the transport part as close as possible to the target manifold, thereby improving the quality and efficiency of generated samples.

2.3.3 Other Improvement Aspects In diffusion model research, significant progress has been made in improvements targeting high-dimensional data and model decoupling. These innovations not only enhance model generation quality under specific conditions but also increase model flexibility and controllability.

Deasy et al. [78] addressed challenges with high-dimensional data by proposing a noise-introduction Gaussian denoising score matching method to achieve controllable diffusion intensity. By introducing heavy-tailed distributions, this method improves score estimation and sampling convergence, significantly enhancing generation performance on unconditional imbalanced datasets. This method not only improves the performance of original SDE models on high-dimensional data generation but also provides new possibilities for diffusion models in handling more complex data structures.

Karras et al. [79] focused on solving the black-box problem between diffusion model step units, proposing a strategy to decompose models into independent units. This approach increases model interpretability and flexibility because modifications to individual units do not affect the status of other units. Based on this, Karras made two main contributions: first, using the Heun method as an ODE solver for the sampling process, improving sampling accuracy and efficiency; second, preprocessing neural network inputs and their corresponding labels to train score-based models, a strategy that improves model learning efficiency and generated sample quality.

3 Applications of Diffusion Models

3.1 Computer Vision

3.1.1 Image Super-Resolution In the field of Single Image Super-Resolution (SISR), diffusion generative models have been extensively studied to address issues such as over-smoothing, mode collapse, and high memory consumption, while improving generated image resolution. Researchers have continuously enhanced super-resolution image generation quality and efficiency by introducing new concepts and methods.

The SRDiff [80] model utilizes diffusion generative models and Markov chain characteristics to transform high-resolution images into simple latent distributions, then gradually generates ultra-high-resolution images during the reverse process. In this process, low-resolution image information is encoded as conditional noise and used for denoising, effectively improving super-resolution image quality.

The SR3 [13] model employs an iterative refinement strategy to improve image resolution, addressing single-pass defects. It combines DDPM' s stochastic denoising process and achieves iterative optimization denoising by training U-Net models for different noise levels, effectively realizing ultra-high-resolution image generation. This iterative refinement strategy not only improves generated image resolution but also optimizes overall generation quality and stability.

The CDM [81] model adopts a cascaded approach to compose multiple diffusion models into a pipeline. This cascade uses different generation models at different spatial resolutions, including a base diffusion model for generating low-resolution data and SR3 models for raising images to ultra-high resolution. Through this multi-cascade strategy, the CDM model not only improves image resolution but also optimizes image details and quality at each resolution level.

3.1.2 Image Synthesis In the image synthesis domain, diffusion models are gradually becoming an important alternative approach, particularly in addressing issues such as GAN model training instability and incomplete data coverage. Researchers are innovating image synthesis methods and frameworks by combining the characteristics and advantages of diffusion models.

The UNIT-DDPM [82] model exemplifies applying diffusion models to image synthesis, particularly in unpaired image-to-image tasks. This model combines DDPM and introduces metadata and target domains, forming joint distributions by minimizing denoising score matching in one domain. Then, it uses this joint distribution for Markov chain updates and finally generates denoised final samples through Markov Chain Monte Carlo methods. This approach overcomes some limitations of GANs in image synthesis and provides an effective pathway for generating richer and more diverse image samples.

The research by Wang et al. [83] applied DDPM to semantic image synthesis. Their model provides noisy images to U-Net structure encoders, while semantic layouts are provided to decoders through multi-layer spatial adaptive normalization operators. Additionally, by introducing classifier-free guided sampling strategies, they further improved sampling quality and semantic interpretability. This method effectively enhances the performance and flexibility of semantic image synthesis.

Inspired by the BART [84] model in the natural language domain, the ImageBART [14] model solves autoregressive image synthesis problems by learning to reverse polynomial diffusion processes. This model introduces contextual information to mitigate exposure bias in autoregressive models and solves free-form image inpainting problems without requiring specific mask training. This approach not only improves autoregressive image synthesis efficiency but also provides new solutions for complex and free-form image synthesis tasks.

3.1.3 Multi-Dimensional Image Domain In the 3D image generation domain, diffusion model applications are rapidly developing, providing new solutions for high-fidelity 3D shape synthesis, point cloud processing, and scene-scale categorical distribution learning. Researchers have continuously improved 3D image generation quality and efficiency through innovative methods and theories.

The shape generation and completion unified framework (PVD) proposed by Zhou et al. [85] can synthesize high-fidelity shapes, complete partial point clouds, and generate multiple completion results from single-view depth scans of real objects. By combining diffusion generative model characteristics, this framework achieves efficient generation and completion of complex 3D shapes, providing new tools for processing and understanding 3D shapes.

Luo et al. [86] proposed a probabilistic model for point cloud generation. They treat point cloud generation as learning a reverse diffusion process that transforms noise distributions into desired shape distributions. This model can be applied not only to point cloud shape completion, upsampling, and synthesis but also to data augmentation, as shown in [Figure 5: see original paper], greatly expanding the application scope and efficiency of point cloud data. This approach of applying reverse diffusion processes to point cloud generation provides new perspectives and methods for 3D data processing.

Lee et al. [87] used discrete diffusion models to learn scene-scale categorical distributions and used the resulting categorical distributions to represent scenes, thereby assigning multiple objects to corresponding semantic categories. This method not only improves scene understanding and classification accuracy but also provides an effective method for generating multiple 3D images simultaneously. Through this discrete diffusion model, researchers can better understand and generate complex 3D scenes, offering new possibilities for 3D image generation and processing.

3.2 Natural Language Processing

The widespread application of diffusion models in computer vision has sparked interest among Natural Language Processing (NLP) researchers, who have begun exploring the possibility of applying denoising diffusion models to text processing. However, unlike the continuous space of images, text sequences have discrete characteristics, posing challenges for directly applying diffusion models. To overcome this difficulty, researchers have proposed two main solutions.

3.2.1 Mapping Discrete Text to Continuous Representation Space In the research field of mapping discrete text to continuous representation space, models such as Difformer [17], DiffusionLM [18], and DiffuSeq [19] represent current technological progress and theoretical exploration. These models enhance diffusion models' capabilities in processing discrete text data through innovative structures and strategies, providing new methods and perspectives for text generation and processing.

The Difformer [17] model combines Transformer architecture with diffusion model characteristics, effectively transforming discrete data into continuous data for training by introducing additional anchor loss functions, normalization modules, and Gaussian noise factors. This combination improves model flexibility and generation capability in text processing while maintaining the powerful expression and understanding capabilities of Transformer architecture.

The DiffusionLM [18] model proposes a new continuous diffusion-based non-autoregressive language model that iteratively denoises Gaussian noise vectors into word vectors, creating hierarchical continuous latent relationships between vectors. This method not only improves text generation coherence and quality but also provides new theoretical pathways for understanding and optimizing text generation processes.

The DiffuSeq [19] model adds an embedding layer to map discrete text to continuous representation space. In the reverse process, the model is trained to find approximate text distribution sequences, thereby effectively generating high-quality text samples. This method enhances text generation efficiency and quality while maintaining the richness and diversity of text data.

3.2.2 Generalized Diffusion Models In diffusion model generalization research, compared to traditional methods of mapping discrete text to continuous spaces, the DiffusER [20] model proposes a novel approach focusing on directly generalizing the diffusion process on original text. The core of this model is to treat text editing operations such as deletion, addition, or modification as noising processes, constructing a diffusion model that more closely aligns with the essential nature of text data. In the reverse denoising modeling process, DiffusER focuses on learning the inverse transformation process of text to achieve efficient generation of target text, as shown in [Figure 6: see original paper]. This method not only maintains the discrete characteristics of text data but also provides an intuitive and effective text generation method.

On the other hand, the DiffusionBERT [21] model combines the advantages of the popular BERT [88] model and diffusion models, proposing a new time-step scheduling scheme during training. This scheme achieves more refined noise control and more effective learning processes by controlling the degree of noise added at each step based on information from each token. Through this BERT-combined approach, DiffusionBERT not only improves text generation quality but also significantly enhances model understanding and expression capabilities.

3.3 Time Series

In the time series analysis domain, diffusion model applications are becoming an emerging trend, aiming to address limitations of traditional autoregressive models in handling complex dependencies, missing data, and long-term forecasting. Diffusion models improve time series analysis accuracy and efficiency by introducing new structures and strategies.

The CSDI [23] model replaces traditional autoregressive models with conditional score-based diffusion models to learn conditional distributions. This model uses observed data as conditional input to the diffusion model, leveraging information from observations for denoising. Additionally, CSDI employs self-supervised methods during training, separating observations into conditional information and imputation targets to compensate for missing ground truth. This method shows significant advantages in handling missing data and forecasting in time series.

The SSSD [24] model integrates conditional diffusion models with structured state space models, excelling at capturing long-term dependencies in time series. This model demonstrates good performance in both time series imputation and forecasting tasks, particularly showing its superiority when handling complex and long-term dependent data structures.

The TimeGrad [25] model, based on energy-based generative models, combines the advantages of RNN [89] and diffusion models to capture time series characteristics. In this process, it learns gradients by optimizing the variational bound of data likelihood and uses Langevin sampling during inference to convert white noise into samples from the distribution of interest through Markov chains. This method shows excellent performance in multivariate probabilistic time series forecasting, particularly in long-term forecasting and learning complex data structures.

3.4 Multimodal Research

3.4.1 Text-to-Image In the text-to-image domain, diffusion models have become a key factor driving technological development, providing new possibilities for generating images corresponding to descriptive text. Diffusion model applications have shown significant advantages in improving image quality, solving generation bias, and enhancing generation efficiency.

The VQ-Diffusion [26] model solves the unidirectional bias problem present in previous generative models in text-to-image tasks. This model uses a masking mechanism to avoid error accumulation during inference, thereby improving generated image quality and accuracy. This method not only enhances the correlation between generated images and input text but also improves image details and quality.

The DALLE-2 [27] model adopts a two-stage approach for text-to-image tasks, as shown in [Figure 7: see original paper]. In the first stage, the CLIP [90] model transforms images and text into conditional embedding prior models; in the second stage, a diffusion model-based decoder completes image embedding work to generate final images. This two-stage approach fully leverages diffusion models' advantages in image generation while ensuring effective utilization of text information.

The Imagen [28] model consists of an encoder for text sequences and a cascaded diffusion model for generating high-resolution images. By improving the original U-Net model for efficiency enhancement, the Imagen model not only improves generated image resolution but also enhances overall generation process efficiency and quality.

In the text-to-3D image generation domain, the Point-E model proposed by OpenAI [29] represents an important technological breakthrough in this field. This model comprehensively employs strategies from two diffusion models to achieve 3D image generation from text descriptions. This method marks a new development direction for text-to-3D image generation technology, providing an efficient and precise generation strategy.

3.4.2 Text-to-Speech In the Text-to-Speech (TTS) domain, diffusion models are applied in innovative text-to-speech generation methods aimed at improving speech synthesis quality, efficiency, and naturalness. Researchers have improved traditional TTS system performance by introducing new architectures and strategies, providing new possibilities for speech synthesis technology.

The Grad-TTS [30] model proposes a novel text-to-speech solution, namely a score-based decoder. This model gradually transforms noise predicted by the encoder and generates mel-spectrograms aligned with text input through monotonic alignment search. This method effectively converts text information into high-quality speech, improving speech synthesis naturalness and accuracy.

The DiffTTS [31] model solves the problem of insufficient effective capacity caused by bijective constraints on model width by filling intermediate representations with noise increments, effectively improving model capacity and performance. This method not only improves speech synthesis efficiency but also enhances generated speech quality and naturalness.

The ResGrad [32] model, as a lightweight diffusion model, uses residuals as generation targets to improve the original process that required synthesizing speech from scratch. It transforms the inference process of existing TTS models into a plug-and-play manner, greatly improving speech synthesis flexibility and adaptability. This method improves generated speech quality while reducing model complexity and computational requirements.

3.4.3 Text-to-Video In the text-to-video generation domain, diffusion model applications provide new perspectives and methods for video editing and generation. As technology continues to develop, diffusion models demonstrate their unique advantages and potential in handling more complex multimedia tasks.

The Dreamix [91] model leverages diffusion model characteristics to combine low-resolution information with high-resolution information for video editing during the inference stage based on provided text information. By fine-tuning the model' s initial stages, this model effectively improves the fidelity and accuracy of edited videos. This method not only enhances video editing quality but also provides an efficient and flexible solution for text-based video content creation.

The Tune-A-Video [92] model treats text-to-video generation as generating a series of continuous images. By proposing a sparse causal attention mechanism, it extends spatial self-attention originally used in image generation to the spatiotemporal domain. This method effectively completes video generation work, improves video continuity and naturalness, and enhances the model' s understanding of video content and structure.

3.5 Interdisciplinary Applications

3.5.1 Medical Imaging In medical imaging applications, diffusion models provide new solutions for tasks such as image reconstruction and defect detection, demonstrating their potential and advantages in handling high-dimensional medical data and complex medical problems.

For inverse problems of reconstructing images from measurement data, researchers [37,38] utilized score-based generative models as advanced image reconstruction tools to reconstruct images through prior data consistency. This method leverages diffusion model characteristics to generate high-quality images consistent with actual medical images, providing an effective strategy for medical image reconstruction.

The DDM model proposed by Kim et al. [33], composed of diffusion modules and deformation modules, is specifically designed to learn spatial deformation information between source and target volumes. This model generates 4D (3D images plus time) cardiac data by generating images of the transformation process, effectively improving cardiac data generation quality and precision, providing important image support for cardiac disease diagnosis and treatment.

In medical defect detection tasks, DDPM models have been proposed to replace traditional autoencoder models [34-36] for training on healthy images. During inference, anomalies are detected by subtracting generated healthy image samples from original images, thereby achieving high-precision medical defect detection. This method not only improves defect detection accuracy but also provides a more effective and reliable technical means for medical diagnosis.

3.5.2 Molecular Modeling In the molecular modeling domain, particularly for protein molecules, diffusion models are applied to learn and generate dynamic structural information of proteins. These studies provide new strategies and tools for protein structure prediction and design.

The research by Anand et al. [39] utilizes diffusion generative models to learn dynamic structural information such as protein rotation and translation, thereby generating protein basic structures and sequences. By capturing protein dynamic characteristics, this method can generate more accurate and biologically feasible protein structures, providing important structural information for protein engineering and drug design.

The ProteinSGM [41] model formulates protein modeling as an image inpainting problem and performs precise protein structure modeling based on conditional diffusion generation methods. This approach provides a new precise modeling method for complex protein structures by analogy with image inpainting strategies, improving protein structure prediction accuracy and efficiency.

The DiffFolding [42] model treats protein backbone structures as a series of continuous angles to capture the relative orientations of constituent amino acid residues. Combined with diffusion generative models, this model generates new stable folded structures from random unfolded structures. This method not only can provide possible protein folding structures but also reveal details and patterns in protein folding processes, offering new perspectives for protein function research and drug development.

4 Existing Problems and Future Research Directions

Based on current research, diffusion models have demonstrated significant advantages and broad potential across multiple generation domains. However, their development still faces several key challenges and problems that require urgent solutions.

First, diffusion models' objective of transforming original images to full Gaussian noise maps in the forward process leads to inference processes involving multiple sampling steps and long sampling times, thereby increasing time costs. Research on how to optimize the stopping conditions of the forward noising process to converge to specific prior distributions within expected timeframes and incorporate adaptive mechanisms is a current key research focus.

Second, diffusion models' generation process relies on long Markov chains, making the entire process exhibit a black-box characteristic that limits the capture of dependency relationships and understanding of model optimization. Current research needs to focus on how to decompose diffusion models into independent units for white-box processing and how to optimize default Markov chains by adopting alternative models that are easier to capture and train.

Third, improved models derived from diffusion models are still mostly based on DDPM' s original settings. Future research could consider diffusion models as a generalized model type, conducting independent research based on core ideas such as sampling algorithms, diffusion schemes, and prior distribution construction, making them easier to combine with other existing models and expanding application scope.

Fourth, current evaluation of diffusion model generated samples is primarily based on FID scores, but this evaluation cannot comprehensively reflect sample restoration effects and diversity. Therefore, developing new evaluation metrics to more comprehensively assess model-generated sample quality is an important future research direction.

Finally, diffusion model training typically employs Evidence Lower Bound (ELBO) to minimize KL divergence between posterior and prior distributions. However, the theory for simultaneously optimizing ELBO and NLL has not been proven, leading to potential mismatch problems between actual and target samples. This issue concerns the actual reliability and practicality of models and requires in-depth research.

Overall, future diffusion model research will focus on optimizing sampling algorithms, reducing model complexity, and improving sampling efficiency. Specific considerations could include transforming traditional step-by-step sampling algorithms into more effective methods such as Hamiltonian Monte Carlo (HMC), introducing pretrained models for parameter initialization, and adopting better hyperparameters to accelerate training processes. These optimization directions are all worthy of further exploration and research.

References

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

[2] Kingma D P, Welling M. Auto-Encoding Variational Bayes[J]. arXiv: 1312.6114, 2022.

[3] Zhang M, Sun Y, McDonagh S, et al. Flow Based Models For Manifold Data[J]. arXiv: 2109.14216,

[4] Rezende D, Mohamed S. Variational Inference with Normalizing Flows[C]// Proceedings of the 32nd International Conference on Machine Learning. 2015:

[5] Ho J, Jain A, Abbeel P. Denoising Diffusion Probabilistic Models[C]//Advances Information Processing Systems. 2020: 6840–6851.

[6] Cheng S-I, Chen Y-J, Chiu W-C, Adaptively-Realistic Image Generation

From Stroke Sketch With Diffusion Model[C]// IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2023:4043-4051.

[7] Jolicoeur-Martineau A, Piché-Taillefer R, Combes R T des, et al. Adversarial score matching and improved sampling for image generation[J]. arXiv: 2009.05475, 2020.

[8] Chen T, Zhang R, Hinton G. Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning[J]. arXiv 2208.04202, 2022.

[9] Gu Z, Chen H, Xu Z, et al. DiffusionInst: Diffusion Model Instance Segmentation[J]. arXiv:2212.02773, 2022.

[10] Xu J, Wang X, Cheng W, et al. Dream3D: Zero-Shot Text-to-3D Synthesis Using 3D Shape Prior and Text-to-Image Diffusion Models[J]. arXiv:2212.14704, 2022.

[11] Ye M, Wu L, Liu Q. First Hitting Diffusion Models Categorical Data[J]. arXiv 2209.01170, 2022.

[12] Furusawa C, Kitaoka S, Li M, et al. Generative Probabilistic Image Colorization[J]. arXiv: 2109.14518,

[13] Saharia C, Ho J, Chan W, et al. Image Super-Resolution Via Iterative Refinement[J]. IEEE and Machine on Pattern Analysis Transactions Intelligence, 2022, 45(4):4713-4726.

[14] Esser P, Rombach R, Blattmann A, et al. ImageBART: Bidirectional Context with Multinomial Diffusion for Autoregressive Image Synthesis[C]// Advances in Neural Information Processing Systems. 2021:3518–3532.

[15] Batzolis G, Stanczuk J, Schönlieb C-B, et al. Non-Uniform Diffusion Models[J]. arXiv: 2207.09786,

[16] Lee S, Chung H, Kim J, et al. Progressive Deblurring of Diffusion Models for Coarse-to-Fine Image Synthesis[J]. arXiv: 2207.11192, 2022.

[17] Gao Z, Guo J, Tan X, et al. Difformer: Empowering Diffusion Models on the Embedding Space for Text Generation[J]. arXiv 2212.09412, 2023.

[18] Li X L, Thickstun J, Gulrajani I, et al. Diffusion-LM Improves Controllable Text Generation [J]. arXiv: 2205.14217, 2022.

[19] Gong S, Li M, Feng J, et al. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models[J]. arXiv: 2210.08933, 2022.

[20] Reid M, Hellendoorn V J, Neubig G. DiffusER: Discrete Diffusion via Edit-based Reconstruction[J]. arXiv: 2210.16886, 2022.

[21] He Z, Sun T, Wang K, et al. DiffusionBERT: Improving Generative Masked Language Models with Diffusion Models[J]. arXiv: 2211.15029, 2022.

[22] Lin Z, Gong Y, Shen Y, et al. GENIE: Large Scale Pre-training for Text Generation with Diffusion Model[J]. arXiv:2212.11685, 2022.

[23] Tashiro Y, Song J, Song Y, et al. CSDI: Conditional Probabilistic Time Series Imputation[C]//Advances in Neural Information Processing Systems. 2021:

[24] Alcaraz J M L, Strodthoff N. Diffusion-based Time Series Imputation and Forecasting with Structured State Space Models[J]. arXiv: 2208.09399,

[25] Rasul K, Seward C, Schuster I, et al. Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting [C]//Proceedings of the 38th International Conference on Machine Learning. 2021: 8857–8868.

[26] Gu S, Chen D, Bao J, et al. Vector Quantized Diffusion Model for Text-to-Image Synthesis[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022:10686-10696.

[27] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical Text-Conditional Image Generation with CLIP Latents[J]. arXiv: 2204.06125, 2022.

[28] Saharia C, Chan W, Saxena S, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding[J]. arXiv: 2205.11487, 2022.

[29] Nichol A, Jun H, Dhariwal P, et al. Point-E: A System for Generating 3D Point Clouds from Complex Prompts[J]. arXiv:2212.08751, 2022.

[30] Popov V, Vovk I, Gogoryan V, et al. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech [C]//Proceedings of the 38th International Conference on Machine Learning. 2021: 8599–8608.

[31] Jeong M, Kim H, Kim H, et al. Diff-TTS: A Denoising Diffusion Model for Text-to-Speech[J]. arXiv: Audio and Speech Processing, 2021.

[32] Chen Z, Wu Y, Leng Y, et al. ResGrad: Residual Denoising Diffusion Probabilistic Models for Text to Speech[J]. arXiv:2212.14518, 2022.

[33] Kim B, Ye J C. Diffusion Deformable Model for 4D Temporal Medical Image Generation[J]. arXiv: 2206.13295, 2022.

[34] Wolleb J, Bieder F, Sandkühler R, et al. Diffusion Models for Medical Anomaly Detection[J]. arXiv: 2203.04306, 2022.

[35] Sanchez P, Kascenas A, Liu X, et al. What is Healthy? Generative Counterfactual Diffusion for Lesion Localization[C]//Deep Generative Models. Cham: Springer Nature Switzerland. 2022: 34–44.

[36] Wyatt J, Leach A, Schmon S M, et al. AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models using Simplex Noise[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022:

[37] Song Y, Shen L, Xing L, et al. Solving Inverse Problems in Medical Imaging with Score-Based Generative Models[J]. arXiv: 2111.08005, 2021.

[38] Chung H, Ye J C. Score-based diffusion models for accelerated MRI[J]. arXiv:2110.05243, 2022.

[39] Anand N, Achim T. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models[J]. arXiv:2205.15019, 2022.

[40] Anand N, Achim T. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models[J]. arXiv:2205.15019, 2022.

[41] Lee J S, Kim P M. ProteinSGM: Score-based generative modeling for de novo protein design[J]. bioRxiv, 2022.

[42] Wu K E, Yang K K, Berg R van den, et al. Protein structure generation via folding diffusion[J]. arXiv: 2209.15611, 2022.

[43] Lam M W Y, Wang J, et al. Bilateral Denoising Diffusion Models[J]. arXiv:2108.11514, 2021.

[44] Giannone G, Nielsen D, Winther O. Few-Shot Diffusion Models[J]. arXiv: 2205.15463, 2022.

[45] Song Y, Ermon S. Improved Techniques for Training Score-Based Generative Models[C]//Advances in Neural Information Processing Systems. 2020: 12438–12448.

[46] Song Y, Durkan C, Murray I, et al. Maximum Likelihood Training of Score-Based Diffusion Models [C]//Advances in Neural Information Processing Systems. 2021:1415–1428.

[47] Song Y, Ermon S. Generative Modeling by Estimating Gradients of the Data Distribution [C]//Advances in Neural Information Processing Systems. 2019.

[48] Song Y, Sohl-Dickstein J, Kingma D P, et al. Score-Based Generative Modeling through Stochastic Differential Equations[J]. arXiv:2011.13456, 2021.

[49] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep Unsupervised Learning using Nonequilibrium Thermodynamics[C]// Proceedings of the 32nd International Conference on Machine Learning. PMLR. 2015:2256–2265.

[50] Nichol A Q, Dhariwal P. Improved Denoising Diffusion Probabilistic Models[C] //Proceedings of the International Conference on Machine Learning. 2021:8162–8171.

[51] Kingma D, Salimans T, Poole B, et al. Variational Diffusion Models[C]//Advances in Neural Information Processing Systems. 2021: 21696–21707.

[52] San-Roman R, Nachmani E, Wolf L. Noise Estimation for Generative Diffusion Models[J]. arXiv: 2104.02600, 2021.

[53] Wang J, Lyu Z, Lin D, et al. Guided Diffusion Model for Adversarial Purification[J]. arXiv:2205.14969, 2022.

[54] Song J, Meng C, Ermon S. Denoising Diffusion Implicit Models[J]. arXiv: 2010.02502, 2020.

[55] Zhang Q, Tao M, Chen Y. gDDIM: Generalized denoising diffusion implicit models[J]. arXiv:2206.05564, 2022.

[56] Sinha A, Song J, Meng C, et al. D2C: Diffusion-Denoising Models for Few-shot Conditional Generation[J]. arXiv: 2106.06819, 2021.

[57] Peebles W, Xie S. Scalable Diffusion Models with Transformers[J]. arXiv:2212.09748, 2022.

[58] Wang Z, Zheng H, He P, et al. Diffusion-GAN: Training GANs with Diffusion[J]. arXiv:2206.02262, 2022.

[59] Sehwag V, Hazirbas C, Gordo A, et al. Generating High Fidelity Data From Low-Density Regions Using Diffusion Models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022:11492–11501.

[60] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[J]. arXiv:2010.11929, 2021.

[61] Kim G, Jang W, Lee G, et al. DAG: Depth-Aware Guidance with Denoising Diffusion Probabilistic Models[J]. arXiv:2212.08861, 2023.

[62] Austin J, Johnson D D, Ho J, et al. Structured Denoising Diffusion Models in Discrete State-Spaces [C]//Advances in Neural Information Processing Systems. 2021:17981–17993.

[63] Watson D, Chan W, Ho J, et al. Learning Fast Samplers for Diffusion Models by Differentiating Through Sample Quality[J]. arXiv:2202.05830, 2022.

[64] Watson D, Ho J, Norouzi M, et al. Learning to Efficiently Sample from Diffusion Probabilistic Models[J]. arXiv:2106.03802, 2021.

[65] Xiao Z, Kreis K, Vahdat A. Tackling the Generative Learning Trilemma with Denoising Diffusion GANs[J]. arXiv:2112.07804, 2021.

[66] Bao F, Li C, Sun J, et al. Estimating the Optimal Covariance with Inexact Mean in Diffusion Probabilistic Models[J]. arXiv: 2212.08861, 2022.

[67] Chung H, Sim B, Ye J C. Come-Closer-Diffuse-Faster: Accelerating Conditional Diffusion Models for Inverse Problems Through Stochastic Contraction[C] //2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 12413–12422.

[68] Vahdat A, Kreis K, Kautz J. Score-based Generative Modeling in Latent Space[C]//Advances in Neural Information Processing Systems. 2021: 11287–11297.

[69] Zhang L, Zhu X, Feng Z, et al. Score-based Generative Models with Preconditioned Diffusion Sampling[J]. arXiv: 2207.02196, 2022.

[70] Du W, Yang T, Zhang H, et al. A Flexible Diffusion Model[J]. arXiv:2206.10365, 2022.

[71] Zhang Q, Chen Y. Diffusion Normalizing Flow[J]. arXiv:2110.07579, 2021.

[72] Kim D, Na B, Kwon S J, et al. Maximum Likelihood Training of Implicit Nonlinear Diffusion Models[J]. arXiv:2205.13699, 2022.

[73] Ho J, Salimans T. Classifier-Free Diffusion Guidance[J]. arXiv:2207.12598, 2022.

[74] Jolicoeur-Martineau A, Li K, et al. Gotta Go Fast When Generating Data with Score-Based Models[J]. arXiv:2105.14080, 2021.

[75] Bortoli V D, Thornton J, Heng J, et al. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling[J]. arXiv: 2106.01357, 2021.

[76] Dockhorn T, Vahdat A, Kreis K. Score-Based Generative Modeling with Critically-Damped Langevin Diffusion[J]. arXiv:2112.07068, 2022.

[77] Liu L, Ren Y, Lin Z, et al. Pseudo Numerical Methods for Diffusion Models on Manifolds[J]. arXiv: 2202.09778, 2022.

[78] Deasy J, Simidjievski N, Liò P. Heavy-tailed denoising score matching[J]. arXiv:2112.09788, 2022.

[79] Karras T, Aittala M, Aila T, et al. Elucidating the Design Space of Diffusion-Based Generative Models[J]. arXiv:2206.00364, 2022.

[80] Li H, Yifan Y, Chang M, et al. SRDiff: Single Image Super-Resolution with Diffusion Probabilistic Models[J]. arXiv:2104.14951, 2021.

[81] Ho J, Saharia C, et al. Cascaded Diffusion Models for High Fidelity Image Generation[J]. arXiv:2106.15282, 2021.

[82] Sasaki H, Willcocks C G, Breckon T P. UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models[J]. arXiv:2104.05358, 2021.

[83] Wang W, Bao J, Zhou W, et al. Semantic Image Synthesis via Diffusion Models[J]. arXiv:2207.00050, 2022.

[84] Lewis M, Liu Y, Goyal N, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:7871–7880.

[85] Zhou L, Du Y, Wu J. 3D Shape Generation and Completion through Point-Voxel Diffusion[J]. arXiv:2104.03670, 2021.

[86] Luo S, Hu W. Diffusion Probabilistic Models for 3D Point Cloud Generation[J]. arXiv: 2103.01458, 2021.

[87] Lee J, Im W, Lee S, et al. Diffusion Probabilistic Models for Scene-Scale 3D Categorical Data[J]. arXiv:2301.00527, 2023.

[88] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019:4171–4186.

[89] Schmidt R M. Recurrent Neural Networks (RNNs): A gentle Introduction and Overview[J]. arXiv:1912.05911, 2019.

[90] Radford A, Kim J W, Hallacy C, et al. Learning Transferable Visual Models From Natural Language Supervision[J]. arXiv: 2103.00020, 2021.

[91] Molad E, et al. Dreamix: Video Diffusion Models are General Video Editors[J]. arXiv: 2302.01329, 2023.

[92] Wu J Z, Ge Y, et al. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation[J]. arXiv:2212.11565, 2022.

Note: Figure translations are in progress. See original paper for figures.