TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 26/31 pp357–383
DOI: 10.26599/TST.2024.9010047
Volume 30, Number 1, February 2025

Diffusion Models for Medical Image Computing: A Survey

Yaqing Shi, Abudukelimu Abulizi*, Hao Wang, Ke Feng, Nihemaiti Abudukelimu, Youli Su, and Halidanmu Abudukelimu

Abstract: Diffusion models are a type of generative deep learning model that can process medical images more efficiently than traditional generative models. They have been applied to several medical image computing tasks. This paper aims to help researchers understand the advancements of diffusion models in medical image computing. It begins by describing the fundamental principles, sampling methods, and architecture of diffusion models. Subsequently, it discusses the application of diffusion models in five medical image computing tasks: image generation, modality conversion, image segmentation, image denoising, and anomaly detection. Additionally, this paper conducts fine-tuning of a large model for image generation tasks and comparative experiments between diffusion models and traditional generative models across these five tasks. The evaluation of the fine-tuned large model shows its potential for clinical applications. Comparative experiments demonstrate that diffusion models have a distinct advantage in tasks related to image generation, modality conversion, and image denoising. However, they require further optimization in image segmentation and anomaly detection tasks to match the efficacy of traditional models. Our codes are publicly available at: https://github.com/hiahub/CodeForDiffusion.

Key words: diffusion models; generative models; medical image; large model

1 Introduction

Medical image is an indispensable part of the modern healthcare system, playing a key role in the early detection, diagnosis, treatment planning, monitoring of treatment effects, and patient care[1]. As technology continues to advance, its role in enhancing the efficiency and accuracy of disease diagnosis and treatment will continue to grow. However, interpreting medical images is complex, requiring expertise and comprehensive analysis for accurate diagnosis, demanding high accuracy and consistency of diagnostic results[2]. This undoubtedly places higher requirements for the accuracy and consistency of diagnosis results.

Medical Imaging Computing (MIC), as a cutting-edge interdisciplinary research field, focuses on the analysis and processing of medical imaging data[3]. This field integrates knowledge from various disciplines, such as computer science, medical imaging, deep learning, medical physics, and biomedical engineering. Its goal is to improve disease detection and diagnostic accuracy by applying advanced computational methods to the acquisition, processing, analysis, and interpretation of medical images[4]. With the advancement of deep learning, MIC is bringing innovation and opportunities to medical research and clinical practice[5]. However, advancements in this field have been accompanied by a series of challenges,

• Yaqing Shi, Abudukelimu Abulizi, Hao Wang, Ke Feng, Youli Su, and Halidanmu Abudukelimu are with School of Information Management, Xinjiang University of Finance and Economics, Urumqi 830012, China. E-mail: syq220424@126.com; a_abliz@outlook.com; 2114273044@qq.com; f18963348096@163.com; hasuyouli@163.com; abdklmhldm@gmail.com.
• Nihemaiti Abudukelimu is with Yili Friendship Hospital, Yining 835000, China. E-mail: 13779121322@163.com.
* To whom correspondence should be addressed.
Manuscript received: 2023-09-17; revised: 2024-02-01; accepted: 2024-02-28

particularly in the area of deep learning model training. The various sources and modalities of medical image data, as well as patient privacy protection issues, pose significant obstacles to training efficient and accurate deep learning models. Deep learning models, such as autoregressive models[6], Generative Adversarial Networks (GANs)[7], and Variational AutoEncoders (VAEs)[8], have been used in MIC. However, these models still face challenges in improving sample quality and controllability. Recently, the successful application of diffusion models in computer vision has attracted the attention of MIC researchers, who have begun to explore these new techniques to address the current challenges[9].

The fundamental principle of diffusion models is to learn the relevant features of data by progressively adding noise and subsequently learning to denoise, effectively capturing the data distribution[10]. This ability of diffusion models has led to its unprecedented potential in areas, such as image generation[11]. For example, some large models, such as Dall-E 2[12], Stable Diffusion[13], Imagen[14], and Midjourney[†], utilize gradual addition and removal of noise in the diffusion models to learn the fine structure and patterns of generative data. This enables them to perform the task of text-to-image generation. These achievements represent a significant development in generative modeling technology, and demonstrate the potential and application value of diffusion models in various fields, such as computer vision[15], natural language processing[16, 17], reinforcement learning[18], and MIC[9].

Recently, there have been significant advancements in the application of diffusion models in the field of MIC. Therefore, a comprehensive review in this area is urgently needed. Figure 1 shows the application of diffusion models in various tasks. While there are already reviews summarizing the use of diffusion models in MIC[9, 19], this paper focuses on their application in five specific tasks: image generation, modality conversion, image segmentation, image denoising, and anomaly detection. The paper also addresses the challenges and issues associated with each task. Additionally, this paper discusses and conducts fine-tuning experiments on Stable Diffusion in the field of MIC, analyzing the advantages and disadvantages of diffusion models compared to other generative models through extensive comparative experiments, and proposing optimization strategies.

The remaining structure of the article is as follows: Section 2 introduces the main types of diffusion

[†]https://www.midjourney.com

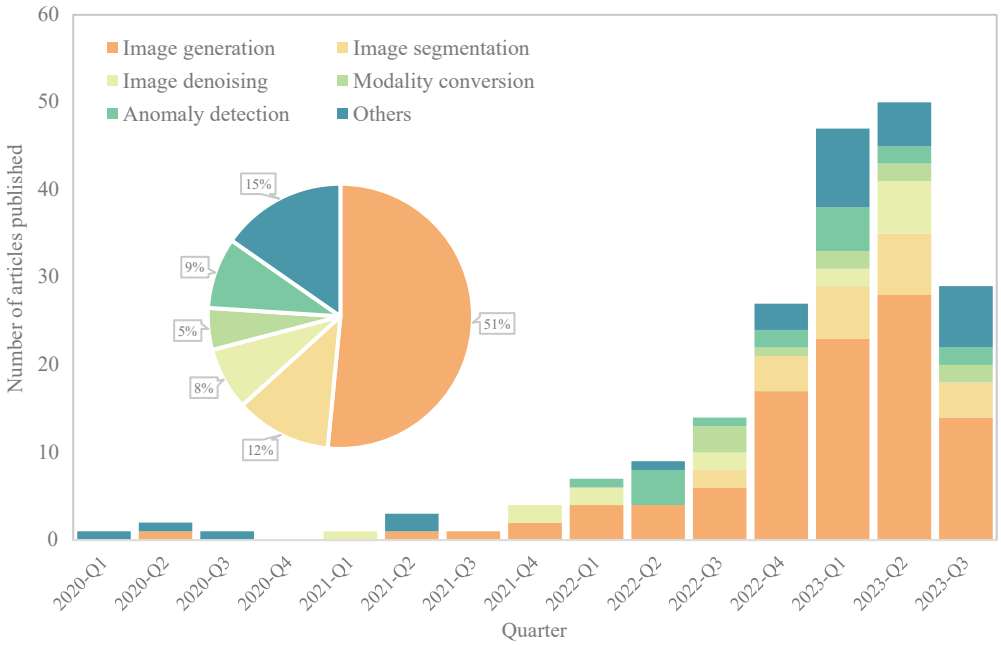


Fig. 1 Application of diffusion models in different tasks. Different colors represent different tasks. Pie chart and bar chart are used to illustrate the publication trends of papers related to diffusion models over the past three years from PubMed, Scopus, Web of Science, IEEE Xplore, Google Scholar, and ArXiv. The pie charts show the distribution of applications of diffusion models in the five tasks, while the bar charts reveal the trends of these applications based on the retrieved literature.

models, sampling methods, and basic architectures; Section 3 reviews the application of diffusion models in five MIC tasks and discusses the existing problems for each task; Section 4 analyzes the performance of diffusion models in different tasks through fine-tuning Stable Diffusion and comparative experiments, and proposes targeted improvement strategies based on experimental results; Section 5 summarizes the overall challenges faced by diffusion models in the field of MIC and discusses future development directions.

Our major contributions are as follows:

• Summarize the fundamentals of diffusion models, including sampling methods and infrastructure.

• Provide an overview of the clinical importance of diffusion models in five MIC tasks, research advances, and current issues.

• Perform fine-tuning Stable Diffusion to explore its potential in medical image generation tasks. Additionally, we invite 10 clinical physicians to participate in evaluating the fine-tuning results.

• Conduct a series of comparative experiments to analyze the advantages and improvement strategies of the diffusion models in five MIC tasks.

• Provide an outlook on future development directions of diffusion models after discussion on the challenges faced by this approach in performing MIC tasks.

2 Diffusion Model

2.1 Foundations for diffusion models

2.1.1 Denoising diffusion probabilistic models

Denoising Diffusion Probabilistic Models (DDPMs)[11] typically contain two Markov chains: a forward Markov chain and a reverse Markov chain. Figure 2 shows the process of adding noise and denoising using DDPM. The forward chain adds noise to the images, while the inverse chain removes noise. The core work of DDPMs is to train a neural network to learn the data distribution of the training dataset and generate new data.

The forward process involves gradually adding Gaussian noise to the original data until the data structure is corrupted and becomes random noise. Specifically, given the original data x_0, its distribution is denoted as $x_0 \sim q(x)$. Since the memoryless nature of the Markov chain, the probability distribution of the next state x_t in the diffusion process can only be determined by the current state x_{t-1}, i.e., x_t and x_{t-1} satisfy the following relation:

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\varepsilon \tag{1}$$

where β_t increases with timestep t. There can be several choices for the distribution of the noise ε, DDPMs use the standard normal distribution, denoted as $\varepsilon \sim \mathcal{N}(0,1)$.

Since the noise adding process of DDPMs follows a Gaussian distribution, the process from x_{t-1} to x_t can be described as

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{2}$$

Equation (2) is likewise the most common choice for transition kernels, in which $\sqrt{1-\beta_t}x_{t-1}$ is the mean and β_t is the variance. I represents the identity matrix, which is used to ensure that the noise is independent and has the same variance across all dimensions. In

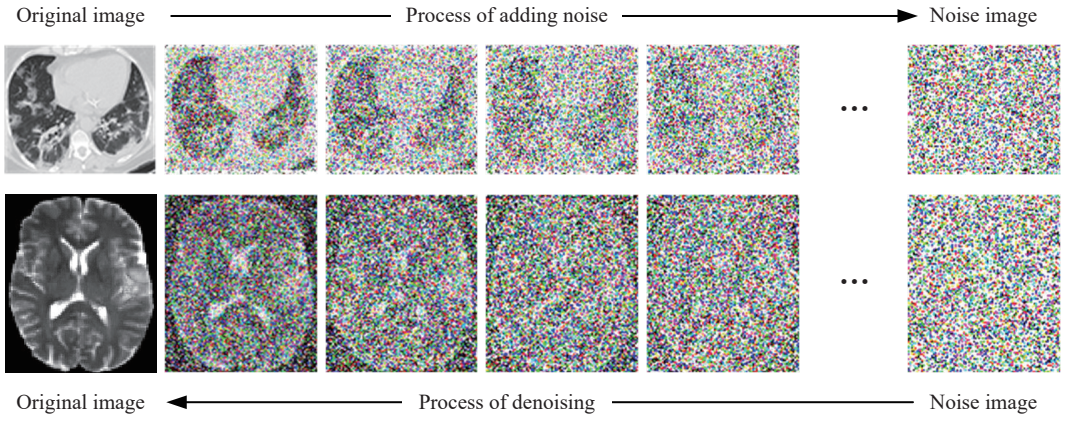


Fig. 2 Process of adding noise and denoising using DDPM. The process of adding noise uses a forward Markov chain to gradually add noise to the original image until the original image becomes purely noisy. The process of denoising uses a reverse Markov chain to gradually denoise the image until the image is restored to its original state.

addition, if we make $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$, in Eq. (1) by the reparameterization trick, the noise data x_t at timestep t can be expressed as

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon \tag{3}$$

This means that we can introduce the noise data x_t for any step only if we have x_0, and have determined β_t for each step.

The reverse process means that the real samples are generated by gradually denoising the noised data until the original data structure is restored. Specifically, given the noise data x_t, its distribution is denoted as $x_t \sim \mathcal{N}(0,1)$. Then, the DDPMs denoise the data using the reverse Markov chain.

Since the denoising process of DDPMs also follows a Gaussian distribution, the denoising process from x_t to x_{t-1} can be described as

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1};\, \mu_\theta(x_t,\, t),\, \sum_\theta(x_t,\, t)) \tag{4}$$

where $p_\theta(x_{t-1}|x_t)$ is a probability density function, $\mu_\theta(x_t,\, t)$ and $\sum_\theta(x_t,\, t)$ are the mean and variance of x_{t-1}, respectively, and θ denotes a neural network parameter. It can be seen from the above equation that the neural network needs to learn how to best adjust the mean and variance at each step of the denoising process to make it closer to the original data.

According to the above, DDPMs need to learn the data distribution by adding noise and denoising to generate new data. However, the noise of the forward process and the reverse process are not fixed at the same value. Therefore, the key to the reverse process is to extract the noise ε from x_t by neural network, and make it similar to the noise used in the forward process to reduce the gap between the generated data and the real data. The loss in training is defined as follows:

$$\text{Loss} = \|\varepsilon - \varepsilon_\theta(x_t,\, t)\|^2 \tag{5}$$

where $\varepsilon_\theta(x_t,\, t)$ denotes the noise estimation model.

2.1.2 Score-based generative models

Score-based Generative Model (SGMs)[20, 21] is a method used to model and generate data distributions. The core idea is to use the gradient of the probability density function of the data, also known as the score function, to generate new samples.

For a given probability density function $p(x)$, its score function is defined as the gradient of its log probability density $\nabla_x \log p(x)$, and this gradient points to the fastest growing probability density.

In the process of SGMs implementation, Gaussian noise is gradually added to the original data x to form a series of noisy data $x_1, x_2, \ldots, x_t$. For each noise level, the model evaluates the score function $s_\theta(x_t,\, t)$ of the noisy data by training a deep neural network, known as the Noise Conditional Score Network (NCSN)[21], and the goal of this training process is to minimize the difference between the true score function and the model estimated score function, which is usually achieved by minimizing a loss function in the following:

$$\mathcal{L}(\theta) = E_{x_0,\, x_t}\left[\left\|\nabla_x \log p_t(x_t|x_0) - s_\theta(x_t,\, t)\right\|^2\right] \tag{6}$$

where $x_0 \sim p(x)$ and $x_t \sim p_t(x|x_0)$, which describe the evolution of the distribution of x over timestep t, given the initail stste x_0.

2.1.3 Stochastic differential equation

Stochastic Differential Equations (SDEs)[22] represent a significant development in the field of generative model, particularly for complex data. These models integrate forward backward stochastic differential equations and diffusion based models. This integration has resulted in the emergence of new variants of SDEs, such as sub-VP SDEs[23] and variance exploding SDEs[24].

For DDPMs and SGMs, it is essential to perturb the data distribution with multi-scale noise levels. As the noise intensity and timestep approach infinity, the perturbation and denoising processes become continuous-time stochastic processes that can be described by SDEs.

SDEs first perturbs the data using a diffusion process and decomposes it into random noise. The perturbation of the data by SDEs can be expressed as follows:

$$\mathrm{d}x = f(x,\, t)\mathrm{d}t + g(t)\mathrm{d}w \tag{7}$$

where $f(x,\, t)$ denotes drift coefficient, $g(t)$ denotes diffusion coefficient, t denotes timestep, and w denotes Brownian motion.

When the noise level becomes infinite, reverse SDEs can be utilized to invert the diffusion process. The reverse SDEs process is defined as follows:

$$\mathrm{d}x = \left[f(x,\, t) - g^2(t)\nabla_x \log p_t(x)\right]\mathrm{d}t + g(t)\mathrm{d}\overline{w} \tag{8}$$

where $\nabla_x \log p_t(x)$ represents the gradient of the log probability density $p_t(x)$ with respect to the data x at timestep t, and $\overline{w}$ represents the reverse time direction of the wiener process. To calculate the SDEs, the drift coefficients, diffusion coefficients, and scoring functions must be determined at each timestep in the

forward diffusion process. With this information, the reverse SDEs can be calculated to invert the diffusion process.

2.2 Sampling for diffusion models

Diffusion models generate data from random noise through a sampling process, thus effectively learning the distribution pattern of the target data and generating diverse and high-quality samples. In this paper, we introduce the guided and fast sampling methods for diffusion models.

2.2.1 Guided sampling methods

Guided sampling methods are used to generate new samples based on specific conditional data distributions in a diffusion model. These methods fall into two categories: classifier guidance and classifier-free guidance.

Classifier guidance. This method requires an additional classifier to identify or classify specific features or attributes that are subsequently used to guide the sampling process of the diffusion model to ensure that the generated samples match specific conditions or labels. The advantage of this method is that the classifier and the diffusion model can be trained independently, and if a diffusion model is already in place, an additional classifier is simply trained and used in combination with the diffusion model during sampling. Liu et al.[25] extended classifier guidance to semantic diffusion, enabling diffusion models to generate images based on image, text, and multimodal conditions. In recent research, Wallace et al.[26] provided a plug-and-play guiding method that does not require retraining or fine-tuning of existing models. The method calculates gradients on actual output and incorporates guidance in a semantically meaningful way, addressing the issues of gradient misalignment and inadequate control that been prevalent in previous classifier guidance methods.

Classifier-free guidance. This method works on the principle of speeding up the sampling algorithm by directly modifying the sampling algorithm for a given diffusion model without introducing an additional learning process. During training, the model learns to complete both conditional and unconditional generation tasks by randomly ignoring conditional information, allowing it to operate efficiently in both conditional and unconditional situations. The advantage of this method lies in the fact that it does not require a separate classifier, thus simplifying the model's training and deployment process. Furthermore, classifier-free guidance enables the model to directly learn how to adjust its generation process based on the provided conditional information, resulting in improved performance and higher sample quality.

Although classifier guidance and classifier-free guidance methods have made many advances in guided sampling research for diffusion models, these methods still require the support of labeled data and their application is limited to the use of conditional diffusion models. They also require extra training details and occasionally zero the category embeddings during the training stage, which adds complexity. To address these issues, Hong et al.[27] proposed novel unconditional and untrained strategies to improve the applicability of fine-grained information and intermediate sample structure through blurred guidance, This enables diffusion models to generate higher-quality samples with an appropriate scale of guidance. To further reduce the cost of sampling, Choi et al.[28] proposed a diffusion model without a prior guidance. This model uses the forward scores of the process probability distribution instead of the estimated scores of the hybrid unconditional model, effectively improving the sampling efficiency of the diffusion model.

2.2.2 Fast sampling methods

DDPMs require multiple iterations to generate high-quality samples. Each generated sample involves a Markov chain transformation, which leads to a slower iteration speed in the sampling process. This is due to the large number of steps that must be performed sequentially for each sample. This poses a challenge for DDPMs in practical applications, particularly when computational resources are limited and low latency is critical. Therefore, it is necessary to use both learning-free and learning-based methods to improve the sampling efficiency.

Learning-free sampling. This method achieves acceleration by directly modifying the sampling algorithm for a given diffusion model, without requiring the introduction of an additional learning process. A typical method is the Denoising Diffusion Implicit Model (DDIM)[29], which emerged as one of the important advances in fast sampling of diffusion models. Unlike the DDPM's inverse Markov diffusion process, DDIM uses a non-Markov process for sampling. However the training objective of both models is the same, which allows for a faster sampling

process by using DDIM's method on top of the DDPMs trained model. In addition, DDIM proposes a skip-step approach, in which the forward noise addition and the reverse denoising process are performed on only a subset of the original time points. Subsequently, Zhang et al.[30] improved the fractional network parameterization of DDIM and extended the application to a wider range of general diffusion models. Liu et al.[31] utilized pseudo numerical methods for diffusion model to treat the diffusion model as solving the problems of differential equations and accelerated the inference process by changing the classical numerical methods.

Learning-based sampling. This method requires an additional learning process to improve the sampling efficiency of the diffusion models and thus achieve faster sampling. Previous research has been centered around the discrete approach, truncated diffusion, and knowledge distillation. The discrete approach speeds up sampling by optimizing the diffusion model's discretization process. Zhang and Chen[32] used exponential integrators to discretize ordinary differential equations and the semilinear structure of the diffusion process to reduce the discretization error, thereby reducing the number of steps required to generate high-quality samples. Truncated diffusion methods reduce the sampling steps by truncating the diffusion process at a certain point. For example, instead of transforming the data completely into noise, the Truncated Diffusion Probabilistic Model (TDPM)[33] truncates the process at an earlier stage to achieve the effect of a hidden distribution of the noisy data and reduces the reverse diffusion steps required to generate the data. In addition, related studies have attempted to utilize knowledge distillation to enhance the sampling effect of diffusion models, resulting in various diffusion distillation strategies[34].

2.3 Architecture for diffusion models

2.3.1 U-Net architecture

U-Net is a neural network architecture consisting of an encoder-decoder, which is widely used in medical image computation, computer vision and other fields[35]. Its uniqueness lies in the combination of three core components: downsampling, upsampling, and skip-connection to realize the capture of high-level semantic information and recover accurate spatial details in images[36]. Figure 3 shows the architecture of U-Net.

The score function of the diffusion models has the same dimension as the input data because it is the derivative of the approximation to the latter. Similarly, the neural network predicts the noise with the same dimension as the input data because a separate Gaussian noise is added to the input data in each dimension, which satisfies the U-Net architecture's requirement for the resolution of inputs and outputs. From Fig. 3, it can be seen that the workflow of the U-Net architecture for diffusion models consists of several main stages: At first, the convolutional layer processes a batch of noisy images to compute the positional embedding of the noise level. Then, the

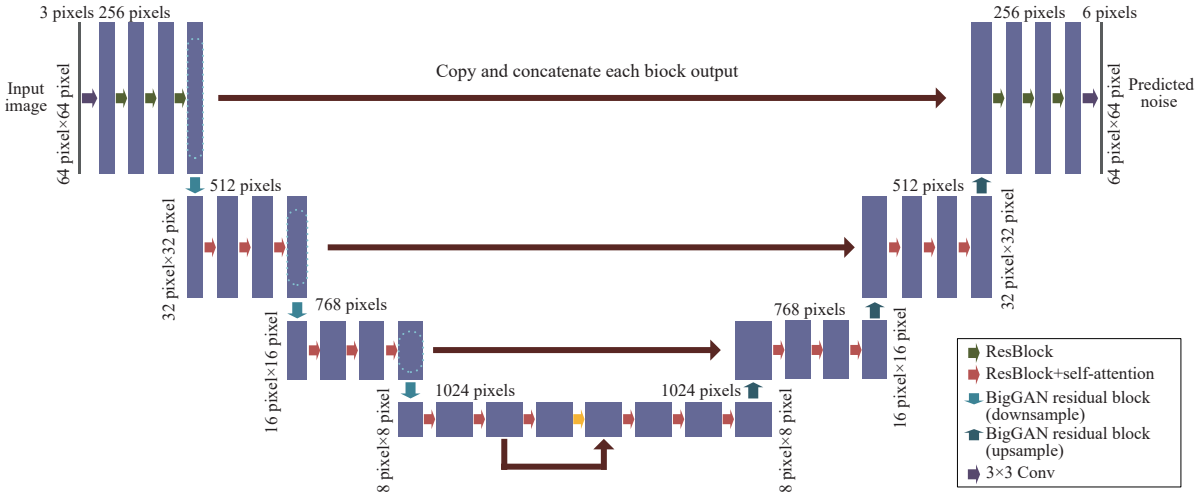


Fig. 3 Overview of U-Net architecture. The left side shows the downsampling stage of U-Net and the right side shows the upsampling stage. The downsampling stage reduces the image resolution to capture higher level features, while the upsampling stage restores the resolution and reconstructs the image details[36].

model goes through a series of downsampling and upsampling phases, each of which includes resnet blocks, group normalization, an attention mechanism, and residual concatenation. The downsampling phase reduces the image resolution to capture higher level features, while the upsampling phase restores the resolution and reconstructs the image details. Finally, to further refine and optimize the image, U-Net applies an additional resnet block and convolutional layer. This efficiently removes the noise from the noisy image while preserving or reconstructing the key features and details of the image, ultimately generating a high-quality clear image.

2.3.2 Transformer architecture

Recently, related researchers found that the use of transformer in diffusion models can also achieve good results[37], whose architecture is shown in Fig. 4. The unique feature of the transformer architecture is its self-attention mechanism, which allows the model to process each element of a sequence comprehensively, taking into account all other elements in the sequence[38]. This greatly enhances the ability to capture complex relationships within the sequence. Meanwhile, transformer abandons the traditional structure of Convolutional Neural Networks (CNNs)[39] and Recurrent Ueural Networks (RNNs)[40], which allows for more efficient parallel processing of data and improves training efficiency. In addition, the transformer model typically comprises multiple cascading encoders and decoders, which further strengthens its processing capabilities.

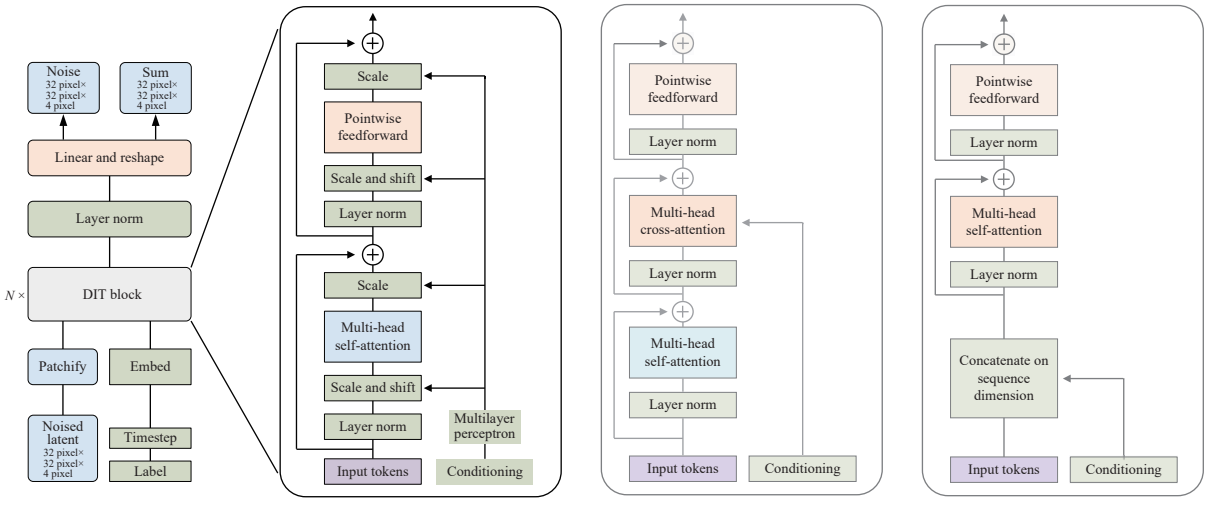
In the diffusion model, noisy data at each step can be encoded using the transformer architecture, subsequently, these encoded results are used to predict the expectation and variance of the transition kernel for the next step[41]. In the implementation, the transformer architecture first splits an image into patches, and converts them into serialized tokens that can be processed by the transformer blocks. Subsequently, each transformer block processes the input image tokens sequentially, while the final outputs are generated by a linear decoder, including noise prediction and covariance prediction. In this process, the self-attention mechanism allows the model to consider the overall context of the entire image, thereby synthesizing the before and after information at each step of the image generation[42].

3 Application

3.1 Image generation

In the field of MIC, image generation mainly includes image synthesis and image reconstruction[43]. The diffusion models can be used to generate synthetic and reconstructed samples that are consistent with the original image distribution. Table 1 provides information on the application of diffusion models in medical image generation tasks.

Clinical importance. By generating high-quality, high-resolution medical images, diffusion models can



(a) Latent diffusion transformer (b) DIT block with adaLN-Zero (c) DIT block with cross-attention (d) DIT block with in-context conditioning

Fig. 4 Overview of diffusion transformer (namely DIT) architecture. DIT first processes the input content using patch embedding to obtain several tokens. Then, a vision transformer based positional embedding is applied to the input tokens, followed by fast processing of the tokens using multiple transformers[37].

Table 1 Detailed information on comparison methods for medical image generation tasks.

Reference	Year	Algorithm	Dataset	Conference/Journal
[44]	2023	DPM	COVID-19, CGMH Pelvis	International Joint Conferences on Artificial Intelligence (IJCAI)
[45]	2023	DDIM	Chest Xray, Lung CT, IDRID	*Multimedia Systems*
[46]	2023	DDPM	ChestXR, VinDr-CXR	IEEE International Symposium on Biomedical Imaging (ISBI)
[47]	2022	DDPM	ICTS	Medical Imaging with Deep Learning (MIDL)
[48]	2022	LDM	UK Biobank	International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)
[49]	2023	LDM	ADNI, LIDC, DUKE, MRNet	*Scientific Reports*
[51]	2022	DDPM	ACDC	MICCAI
[52]	2021	SGM	LIDC, LDCT, BraTS	International Conference on Learning Representations (ICLR)
[53]	2022	SGM	FastMRI	*Medical Image Analysis*
[54]	2022	DM	FastMRI, SKM-TEA	MICCAI
[55]	2022	DM	IXI, FastMRI	*Medical Image Analysis*
[56]	2023	SGM	AAPM2016, BraTS	IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)
[57]	2023	DM	SARS-CoV-2 CT	International Conference on Information Networking (ICOIN)

help physicians more accurately identify and evaluate disease features, especially in complex or hard-to-diagnose cases. For patients who require regular radiological examinations, such as those with certain cancers and chronic diseases, diffusion models can generate alternative images to reduce the number of actual scans, thereby reducing the risk of long-term radiation exposure. In addition, in medical education and professional training, synthetic medical images can provide a rich resource for learning and practice, especially in diagnostic training for rare diseases or special cases.

In research on image synthesis tasks. Shao et al.[44] proposed a four-stage model, called DiffuseExpand, which is based on the Diffusion Probabilistic Model (DPM). This model aims to expand medical image datasets by employing a series of stages and techniques. This model first synthesizes a segmentation mask from Gaussian noise, and then uses the segmentation mask as a conditional prior to generate the corresponding image, thus realizing image-mask sample pair generation. Zhang et al.[45] proposed a Generalized Hybrid Denoising Diffusion Model (GH-DDM) for medical image generation. This model combines the global modeling capability of transformer with the texture modeling capability of CNN. Additionally, it includes a cross-attention module to the skip connection of the U-Net structure, which has demonstrated strong generative capability on multiple medical datasets. Huy and Quan[46] proposed a multi-branch Denoising Diffusion Medical Model (DDMM), which consists of two independent DDPMs sharing noise latent codes. This model can improve the quality of downstream tasks by generating synthetic X-rays and labels in an unsupervised manner.

MRI and CT images often contain 3D data with rich diagnostic information. Dorjsembe et al.[47] proposed 3D-DDPM for the first time to generate 3D medical image generation. In the experiments, the images generated from this model are judged as real by experts more frequently than actual images, and the model output shows better visual accuracy. Pinaya et al.[48] generated a 3D adult brain MRI based on Latent Diffusion Model (LDM), a model can effectively control image generation by using conditional variables, such as age and gender, and all metrics outperform the GAN-based method. Khader et al.[49] used LDM-based DPM for the latent representation compressed by VQ-GAN[50]. The model can generate realistic 3D synthetic images even with a small dataset. For parts such as the human heart that are constantly changing, 4D medical images containing time-series information are usually required for analysis. Kim and Ye[51] proposed a Diffusion Deformation Model (DDM) based on DDPM to generate 4D time-series images. By learning the latent space encoding between

the source image and the target image, it can continuously output any intermediate time point image.

In research on image reconstruction tasks. Song et al.[52] proposed an unsupervised approach to learn the prior distribution of medical images based on a generative model of scores. This approach enables the generation of image samples that adhere to both the prior distribution and the measured images, providing a valuable tool for CT and MRI reconstruction tasks. The denoising score-matching method was employed by Chung et al.[53] to train a continuous-time autoregressive score model. This model achieves MRI reconstruction through iteration between solving the backward SDEs and enforcing data consistency steps during the inference stage utilizing a Variational Euler Stochastic Differential Equation (VE-SDE) solver. The use of conditional prior-based diffusion models has become a hot topic for CT and MRI reconstruction, as it allows the incorporation of prior knowledge and the generation of high-quality image reconstructions. In their work, Peng et al.[54] introduced a diffusion model-based MRI reconstruction method called DiffuseRecon. This method involves guiding the reverse process of the diffusion model by observing k-space signals to achieve MRI reconstruction. The reverse process employs a cosine noise scheme and a U-Net architecture with multi-head attention, achieving excellent performance in generating reconstructed images from the original acquisition signals. In addition to using static image priors, Güngör et al.[55] introduced the initial adaptive diffusion prior model, known as AdaDiff, to generate MRI reconstruction images. It employs adversarial learning to implement reverse diffusion, accelerating the generation of high-quality reconstructed samples with fewer steps.

For 3D medical image reconstruction. Chung et al.[56] introduced DiffusionMBIR that utilizes a pre-trained 2D diffusion model for 3D medical image reconstruction. The forward diffusion step adopts a Model-Based Iterative Reconstruction (MBIR) optimization strategy. The reverse step applies the 2D diffusion model on different slices of the 3D image, aggregating slices and combining an Adam optimization framework to achieve data consistency. This approach is capable of reconstructing out-of-distribution data with significant variations.

Discussion. Diffusion models have made significant advances in the field of medical image generation. However, the training of most diffusion models relies on specific datasets, which may limit the diversity and representativeness of the images generated by the models. Particularly with respect to imaging data for rare diseases or specific populations (e.g., specific age groups or ethnicities), existing datasets may not be sufficient to train models that can be widely applicable. In addition, targeting high-resolution and 3D models often requires significant computational resources and time. This limits their application in resource-limited environments. Recently, Nguyen et al.[57] proposed a lightweight diffusion model to generate medical images without the need to train on large amounts of data, but the difference between the synthesized images and the original images is not significant, and research on lightweight models is yet to be conducted.

3.2 Modality conversion

Medical image modality conversion refers to the conversion of one type of medical image into another type. Table 2 shows information on the application of diffusion models in medical image modality conversion tasks.

Clinical importance. Doctors often need various modalities of imaging information to aid in disease diagnosis. Diffusion models can provide alternative diagnostic means when specific types of medical scans are not applicable due to equipment limitations or patient conditions. For instance, a patient with sensitivity to specific radiation types could receive necessary diagnostic data through the conversion of images from alternate modalities, thereby diminishing

Table 2 Detailed information on the comparison methods for the task of medical imaging modality conversion.

Reference	Year	Algorithm	Dataset	Journal/Preprint server
[58]	2023	DDPM	IXI, BraTS Pelvic MRI-CT	*IEEE Transactions on Medical Imaging*
[59]	2022	DDPMs SDEs	Gold Atlas	arXiv
[60]	2023	DDPMs	Gold Atlas, BRATS2018	arXiv
[61]	2023	DM	Head and Neck Dataset, Lung Dataset	arXiv
[62]	2023	DDPMs	–	arXiv
[63]	2023	DDM	CAMUS	*IEEE Access*

the potential health risk for the patient. Furthermore, the integration of information from various imaging modalities can be achieved through diffusion models. This approach provides a more comprehensive understanding of the pathology, such as the ability to accurately locate tumors and monitor their response to treatment by combining data from PET and MRI.

Özbey et al.[58] proposed the first unsupervised medical imaging modality conversion method SynDiff based on the adversarial diffusion model. Compared with ordinary diffusion models, SynDiff uses a larger diffusion step size, and adopts adversarial mapping for the reverse process. Efficient and high-fidelity modality conversion can be achieved by using the conditional diffusion process to gradually generate the target image guided by the source image. Lyu and Wang[59] carried out a conversion from MRI to CT using DDPMs and SDEs, based on T2-weighted MRI. Their method produces much better results than other models. Li et al.[60] proposed the Denoising Diffusion Model for Medical image Synthesis (DDMM-Synth). This model combines the anatomical information from MRI and the relevant information from sparsely sampled CT to achieve the synthesis of high-quality CT images. The synthesized CT maintains the data consistency with sparsely sampled CT while retaining the anatomical information in MRI. Li et al.[61] proposed a Frequency-Guided Diffusion Model (FGDM), which uses only target-domain samples for training and can be directly applied to source-to-target medical image modality conversion. It is the first model that performs the task of medical image modality conversion through zero-shot learning at the anatomical level, outperforming other state-of-the-art methods in the task of cone-beam CT-to-CT mode conversion.

For 3D medical imaging modality conversion, Pan et al.[62] proposed MC-DDPM based on DDPM for MRI to CT conversion. This model represents a significant advancement in the field, offering a novel approach to address the task of synthesizing CT images from MRI data in a three-dimensional context. The reverse process of the model uses the optimized and trained Swin-Vnet structure. Noisy CT is denoised and then guided by MRI to generate a synthetic CT that matches the MRI anatomy. In the study of the heart's motion, echocardiograms that include time information are also often used. Tiago et al.[63] trained an adversarial denoising diffusion model combined with GAN to synthesize echocardiograms and realize image translation between different domains. This model adopts GAN to learn the denoising process. It preserves relevant anatomical structures under the guidance of prior images and utilizes a larger step size to generate a variety of image samples with less sampling time.

Discussion. Diffusion models have demonstrated potential for modality conversion tasks in medical images. However, they face challenges, such as high computational resource requirements, difficulties in processing large volumes of 3D data, and balancing between accuracy and speed. Bieder et al.[64] proposed a patch-based training method and a coordinate encoding strategy to improve the efficiency and performance of the models, but these methods still need to be further optimized for practical applications. Particularly, the balance between accuracy and processing speed is still an important research direction.

3.3 Image segmentation

Image segmentation involves separating the area of interest in the medical image from the background, so that physicians can perform more accurate analysis and diagnosis. Table 3 presents information for the application of diffusion models in medical image segmentation tasks.

Clinical importance. Diffusion models maintain high accuracy even with poor image quality or tiny target structures, which is valuable for early disease diagnosis and precise treatment, especially in the fields of oncology, neurology, and cardiovascular diseases. Furthermore, diffusion models exhibit exceptional efficiency and reliability in automated image segmentation. They can reduce the workload of radiologists and image specialists, accelerate the diagnostic process by segmenting images quickly and accurately, and improve the overall workflow efficiency.

Wu et al.[65] introduced MedSegDiff, which stands as the pioneering DPM-based general medical image segmentation model. This model extends DPM by incorporating dynamic conditional encoding and FF-Parser. It achieves adaptive calibration of the segmentation masks at the current step by integrating them with image priors at multiple scales. It also suppresses high-frequency noise contained in the feature maps using the Fourier transform. Compared to other methods, the model produces more accurate

Table 3 Detailed information on the comparison methods for medical image segmentation tasks.

Reference	Year	Algorithm	Dataset	Conference/Preprint server
[65]	2023	DDPMs	REFUGE-2, BraTS2021, DDTI	MIDL
[66]	2023	DDPMs	AMOS, BraTS2021, REFUGE-2, DDTI	AAAI Conference on Artificial Intelligence (AAAI)
[67]	2022	DDPMs	CheXpert, BraTS2020	MIDL
[68]	2023	DDPMs	AMOS, Prostate MR	MICCAI
[69]	2023	DDM	MSD Liver, BTCV, BraTS2020	arXiv
[70]	2023	DDPMs	WMH	arXiv
[71]	2023	DM	LIDC-IDRI, MSMRI	CVPR
[72]	2023	DM	LIDC-IDRI, BraTS2021	arXiv
[73]	2023	DPM	QUBIQ	arXiv

segmentation results, particularly in blurry regions. Subsequently, Wu et al.[66] introduced the MedSegDiff-V2 model, which combines a transformer-based U-Net framework with the diffusion model. It adds anchor conditions and SS-Former to address the limitations of direct integration, achieving more stable and accurate results compared to the MedSegDiff model. Wolleb et al.[67] proposed a new semantic segmentation model based on the deep diffusion probability model. The original MRI serves as a conditional prior to enhance anatomical information in the generated image. This model generates five different segmentation masks for the same MRI through random sampling. By implicitly integrating these segmentation masks, the integrated image further improves the segmentation performance of the model. Guo et al.[68] presented the PD-DDPM model, a pre-segmentation diffusion sampling model designed to accelerate medical image segmentation. This model can generate segmentation results with fewer inverse steps by combining noise prediction with pre-segmentation results.

For 3D medical image segmentation, Fu et al.[69] proposed a 3D medical image multi-class segmentation model based on DDPM. The model predicts segmentation masks and directly optimizes them using the dice loss. The previous step's mask is used to generate noise-disturbed masks to reduce information leakage and decrease diffusion steps, thereby improving the efficiency and performance of the model. Xing et al.[70] proposed Diff-Unet, an end-to-end 3D medical image segmentation model based on the diffusion model, to solve the problem of high-dimensional medical image segmentation. This model takes volumetric images and noisy segmentation maps as inputs. During the inference process, a fusion module based on step uncertainty is introduced to combine the model output at each step. It outperforms other state-of-the-art methods in multiple segmentation tasks.

By applying multiple random perturbations to the input image, diffusion models can generate a series of diverse image samples. Rahman et al.[71] proposed the Collective Intelligent Medical Diffusion model (CIMD) that is based on a single diffusion model. It employs the random sampling procedure of the diffusion model to capture the ambiguity of medical images. Obtaining multiple segmentation masks by generating from a single input image, the proposed approach surpasses existing blurry segmentation networks and exhibits superior performance in the task at hand. Chen et al.[72] presented the conditional Bernoulli Diffusion model (BerDiff) as a novel technique for medical image segmentation. This model uses bernoulli noise instead of Gaussian noise, resulting in the generation of more precise segmentation masks. Leveraging the stochasticity of the diffusion process, the model performs multiple sampling to generate different segmentation masks. This approach highlights the regions of interest and provides valuable references for medical professionals. The annotation information may be uncertain due to potential variations in segmentation annotations provided by different doctors for the same image. Due to different doctors possibly providing varying segmentation annotations for the same image, the annotation information possesses uncertainty. Amit et al.[73] proposed a multi-annotator segmentation method that combines different segmentation annotations using the diffusion model. It generates a unified segmentation map representing a consensus among different physicians, improving the efficiency and quality of medical image segmentation.

Discussion. While diffusion models have potential in medical image segmentation tasks, they still have

limitations. Their generalization abilities are limited, and their performance may be unstable, especially when dealing with different types of data from the training set. In addition, these models require huge computational resources when processing large-scale or high-resolution 3D medical images. During the pursuit of high-precision segmentation, processing speed may decrease, which can impact clinical applications. Noise and artifacts in medical images may also affect the segmentation accuracy, especially when the image quality is poor.

3.4 Image denoising

Image denoising is the process of reconstructing noisy images into high-quality images under the premise of retaining important information about medical images. Table 4 shows information on the application of diffusion models in medical image denoising tasks.

Clinical importance. Diffusion models can effectively remove noise from medical images while maintaining key anatomical structures and pathological features. This is particularly important for early disease diagnosis, especially in areas such as tumor identification and cardiovascular disease assessment. By improving image clarity and contrast, diffusion models help radiologists and other medical professionals more accurately identify and evaluate diseased areas. In addition, automated denoising through diffusion models reduces the need for manual editing, speeds up the diagnostic process, and enables medical teams to process image data more efficiently. This is especially important in emergency and high intensity medical environments.

Abirami et al.[74] used a finite difference approximation scheme to create a spatiotemporal variable order fractional diffusion equation for medical image denoising. The proposed model is superior to fractional and integer order diffusion models in maintaining details and edge information, which can better perform image denoising. Hu et al.[75] employed unsupervised DPM for denoising retinal OCT denoising. By adjusting the number of steps in the reverse process, the model produces different degrees of denoising results, providing controllable denoising capabilities. Chung et al.[76] proposed the image denoising method based on regularized reverse diffusion uses a score-based diffusion model for image denoising. After training on knee joint MRI data, the model can be used for liver MRI denoising. Liu et al.[77] proposed an unsupervised model based on DDPM for low-dose CT denoising. This model first uses normal dose CT to train the unconditional DDPM and then integrates the trained unconditional DDPM into the denoising framework to solve the problem of denoising medical images. In each iteration, low-dose CT is used as the conditional prior of the denoising process to generate high-quality images corresponding to low-dose CT to achieve low-dose CT denoising. Xia et al.[78] applied conditional DDPM for low-dose CT denoising and used a fast Ordinary Differential Equation (ODE) solver to improve sampling efficiency. Using the ODE solver is 20 times faster than using the original DDPM denoising without compromising the denoising effect.

Discussion. Most existing studies rely on limited, specific datasets, which can lead to inadequate generalization capabilities of the models. The lack of diverse and large-scale medical image datasets restricts the application of diffusion models across different equipment, patient groups, and types of diseases. When faced with various types of noise and different qualities of images, the robustness of diffusion models is a challenge. Current research has not fully addressed the issue of performance stability under extreme or abnormal conditions.

3.5 Anomaly detection

Detecting anomalies in medical images using diffusion models is usually done by weakly supervised or unsupervised learning based on image reconstruction. The image with lesions is reconstructed into a lesion-free image, and the difference between the diseased

Table 4 Detailed information on the comparison methods for medical image denoising tasks.

Reference	Year	Algorithm	Dataset	Conference/Journal/Preprint server
[74]	2021	SGMs	–	*Mathematical Problems in Engineering*
[75]	2022	DPM	–	SPIE Medical Imaging
[76]	2022	SGMs	FastMRI	*IEEE Transactions on Medical Imaging*
[77]	2023	DDPMs	LDCT-PD	arXiv
[78]	2022	DDPMs	NIH-AAPM-Mayo 2016	arXiv

image and the reconstructed image forms a pixel-level abnormal image that can be used to display the location of the lesion. Table 5 shows information on the application of diffusion models in medical image anomaly detection tasks.

Clinical importance. Diffusion models show a high degree of flexibility and accuracy when dealing with complex and high-dimensional medical imaging data. They can learn from many normal and abnormal images, effectively identifying potential abnormal features and maintaining a high detection rate even for subtle or atypical lesions. In addition, diffusion models provide a powerful tool for early detection and monitoring of disease progression by automatically detecting abnormal changes. In oncology, for example, this technique can be used to monitor tumor growth or response to treatment, thereby effectively guiding treatment decisions.

Sanchez et al.[79] trained a diffusion probabilistic model using healthy and diseased images. By employing implicit guidance and attention modulation, the generation process is controlled to generate healthy images corresponding to the input images. Wolleb et al.[80] proposed an iterative method that combines DDIM with noise injection and denoising processes. They used DDIM's reverse sampling scheme to encode anatomical information and the deterministic sampling scheme for denoising. With the guidance of a classifier, the diseased images are transformed into healthy images. Wyatt et al.[81] introduced a partially diffused unsupervised anomaly detection model called AnoDDPM based on DDPM. In this model, a partial Markov chain is used diffusion to accelerate the training and inference processes. Simplex noise is used for larger region anomaly detection instead of Gaussian noise. The model is trained on healthy images and maps abnormal data to a normal distribution.

Iqbal et al.[82] described an unsupervised anomaly detection model called mDPPM that reconstructs the generation task of diffusion models in DDPM by introducing a mask-based regularization. The model is trained on healthy images to eliminate abnormal regions in diseased images and then generate healthy images. Pinaya et al.[83] employed VQ-VAE for image dimensionality reduction to accelerate the processing of diffusion models. They first trained VQ-VAE and DDPM on healthy images and then utilized DDPM to process the compressed latent representations to eliminate abnormal regions and generate healthy images. Behrendt et al.[84] proposed a patch-based unsupervised anomaly detection model called p-DDPM. It only adds noise and performs denoising on image patches of the input image while incorporating global image information and eventually reconstructs the entire image by integrating the denoised patches. This method allows for a better reconstruction of the brain MRI and generates corresponding healthy images.

Discussion. Currently, diffusion models have made some progress in medical image anomaly detection tasks, but the running time of this method is relatively long, mainly due to the long Markov chain sequence required, which leads to an increase in sampling time and affects the scalability and practicality of the model. Meanwhile, the Gaussian diffusion model is a probability-based generative model that reconstructs or generates images by gradually adding noise and then gradually removing it. This approach tends to smooth and remove extreme values or large deviations when processing an image, which can make it difficult to accurately capture larger anomalous regions, especially if these regions are not common in the overall dataset.

4 Experimental and Result Analysis

In this section, we conduct fine-tuning experiments on Stable Diffusion, as well as comparative experiments based on OpenAI's open source projects: Improved DDPM (IDDPM)[85] and Guided DDPM (GDDPM)[86].

Stable Diffusion is a deep learning text-to-image generation model based on LDM[13]. When given a text prompt, the model generates an image that matches the

Table 5 Detailed information on the comparison methods for medical imaging anomaly detection tasks.

Reference	Year	Algorithm	Dataset	Conference/Preprint server
[79]	2022	DDPMs	BraTS2021	MICCAI
[80]	2022	DDIMs	BraTS2020, CheXpert	MICCAI
[81]	2022	DDPMs	NFBS	CVPR
[82]	2023	DDPMs	IXI, MSLUB, BraTS2021	arXiv
[83]	2022	DDPMs	MedNIST, UKB, BraTS, WM, MSLUB	MICCAI
[84]	2023	DDPMs	IXI, BraTS2021, MSLUB	MIDL

given prompt. Compared to other models, Stable Diffusion is an open source model, which makes it easier to realize its application potential in different scenarios. However, in practice, Stable Diffusion may have uncontrollable output that require domain fine-tuning to meet the requirements of a particular scenario.

IDDPM reduces the number of forward passes required in the generation process by optimising the algorithm[85]. GDDPM balances diversity and fidelity in conditional image synthesis by introducing classifier guidance[86]. Through literature research, we find that there are many models related to diffusion models that are constructed based on IDDPM and GDDPM[54, 65, 66, 70, 73, 80]. To further explore their effects in different tasks, we apply the diffusion model to image generation, modality conversion, image segmentation, image denoising, and anomaly detection tasks by adjusting the architecture and parameters of IDDPM and GDDPM, and compare them with other generation models to analyze the performance and improvement strategies of IDDPM and GDDPM in different tasks.

4.1 Fine-tuning experiments

The fine-tuning methods used in the experiments include LoRA[87], DreamBooth[88], HyperNetwork[89], and Embedding[90]. LoRA is an efficient fine-tuning method that optimises the model output by fine-tuning the weights of the U-Net cross-attention layer. Meanwhile, DreamBooth fully tunes all the weights of the model, paying special attention to preserving the integrity of the original generated topics during the training process, while HyperNetwork influences the final performance of the model by adding a small auxiliary network in front of the cross-attention layer in the noisy predictor of U-Net. In addition, Embedding, also known as textual inversion, greatly improves the model's ability and efficiency in processing image data by transforming the input data into a vector representation. The application of these methods enables the Stable Diffusion to gain a wider range of adaptability and application while maintaining its original performance.

To measure the impact of dataset size on the experimental results, 100, 500, and 1000 images are selected for the experiments. Meanwhile, Peak Signal-to Noise Ratio (PSNR) and Structural SIMilarity (SSIM) metrics are used to objectively evaluate the quality of the generated results generated by the model. Figure 5 shows the objective evaluation scores for the generated results. In addition, to further validate the applicability of the generated images in clinical diagnosis, we invite 10 physicians with professional knowledge and clinical experience to subjectively evaluate the accuracy and clinical applicability of the generated results. Table 6 shows the subjective evaluation scores for the generated results.

In Fig. 5, higher scores of PSNR and SSIM indicate better performance of the fine-tuning method. As shown in Fig. 5d, the PSNR scores of the Embedding method demonstrate relative stability under different dataset sizes and different prompt conditions, Prompts 1−3 denote "photo of a lung X-ray", "photo of a lung X-ray with cardiomegaly", and "photo of a long X-ray with visible pleural effusion", respectively. However, as seen in Fig. 5h, the Embedding method generally scores lower than other methods in terms of SSIM value. Intuitively, the LoRA method exhibits some advantages in PSNR and SSIM, as evidenced in Figs. 5a and 5e. In Table 6, higher accuracy and suitability scores indicate better fine-tuning. It is observed that the LoRA method achieves the best average scores in both accuracy and suitability with a dataset size of Image100.

The generated results of different fine-tuning methods are shown in Fig. 6. After inputting Prompts 1−3, the model generates the images corresponding to the descriptions. As shown in Fig. 6, the outputs image of Stable Diffusion without fine-tuning is more abstract, which is far from the expectation. After fine-tuning, the subjective effect of the model's output image is significantly improved, where the effects generated after fine-tuning using the LoRA method and the HyperNetwork method are visually closer to the real CT X-ray. However, the Embedding method appeares to have incomplete thoracic structures and produces more artifacts.

The fine-tuning experiments on Stable Diffusion show its capability in medical image generation tasks. Specifically, the LoRA method outperforms others in both objective and subjective evaluations, effectively generating synthetic CT X-ray samples that closely match the prompts and have high clinical applicabilities. Although the Embedding method achieves high PSNR values, it falls short in single channel grayscale medical image contexts, especially in subjective assessments and generation quality.

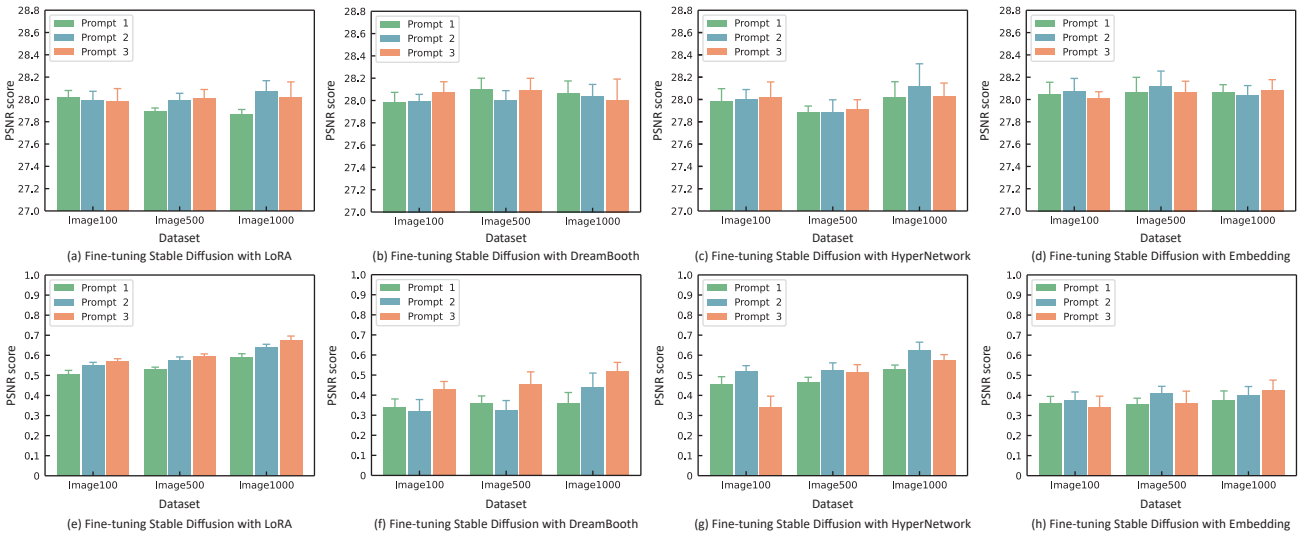


Fig. 5 Average and standard deviations of objective evaluation of different fine-tuning methods. Evaluation results of four different fine-tuning methods (LoRA, Dreambooth, HyperNetwork, and Embedding) on Stable Diffusion model for three different sizes of datasets (Image100, Image500, and Image1000), and three different prompts. (a)–(d) Show the evaluation results based on PSNR, and (e)–(h) show the evaluation results based on SSIM score.

Table 6 Subjective scores for different fine-tuning methods. The results are evaluated by 10 physicians, who score them based on the accuracy of the generated outcomes and the applicability to clinical diagnosis. Each evaluation index has a scoring range from 0 to 10, and the bold value in each column is the optimal value.

Strategy	Dataset	Accuracy score				Suitability score			
		Prompt 1	Prompt 2	Prompt 3	Average	Prompt 1	Prompt 2	Prompt 3	Average
LoRA	Image100	**6.333**	**9.167**	**6.306**	3.417	**7.300**	**8.800**	7.100	**7.733**
	Image500	6.083	4.167	3.500	4.583	6.300	4.500	3.100	4.633
	Image1000	4.750	7.167	4.338	5.418	5.400	8.700	**8.400**	7.500
DreamBooth	Image100	5.000	5.000	4.500	4.833	4.300	4.200	4.400	4.300
	Image500	3.417	4.000	4.167	3.861	3.400	3.700	4.400	3.833
	Image1000	4.422	5.083	4.838	4.781	5.500	4.700	5.300	5.167
HyperNetwork	Image100	4.917	6.333	5.083	**5.444**	6.200	6.100	7.000	6.433
	Image500	4.167	3.833	4.667	4.222	5.600	7.000	7.000	6.533
	Image1000	3.000	7.167	4.088	4.752	3.000	7.100	6.300	5.467
Embedding	Image100	4.808	5.250	5.088	4.809	4.200	4.600	4.700	4.500
	Image500	3.177	6.177	4.760	4.704	4.100	6.100	5.500	5.233
	Image1000	3.088	5.265	5.250	4.534	3.900	5.700	5.000	4.867

Moreover, as shown in Fig. 5 and Table 6, a larger dataset size dose not necessarily correlate with improved experimental outcomes. This might have been due to increased complexity and potential compromises in image quality with larger datasets, highlighting the importance of high-quality medical images for effectively fine-tuning Stable Diffusion.

4.2 Comparative experiments

The comparison experimental environment is the Ubuntu20.04 Linux operating system, and it utilizes a single RTX 3090 Ti GPU. The software framework is PyTorch. Table 7 shows an introduction to each generative model. Table 8 shows a description of the datasets used in the comparison experiments.

Tables 9–13 show the quantitative evaluation results on the tasks of image generation, modality conversion, image segmentation, image denoising, and anomaly detection. Each task is evaluated by three indicators. Some tasks, such as image generation, image segmentation, image denoising, and anomaly detection, are performed on three different datasets. In modality

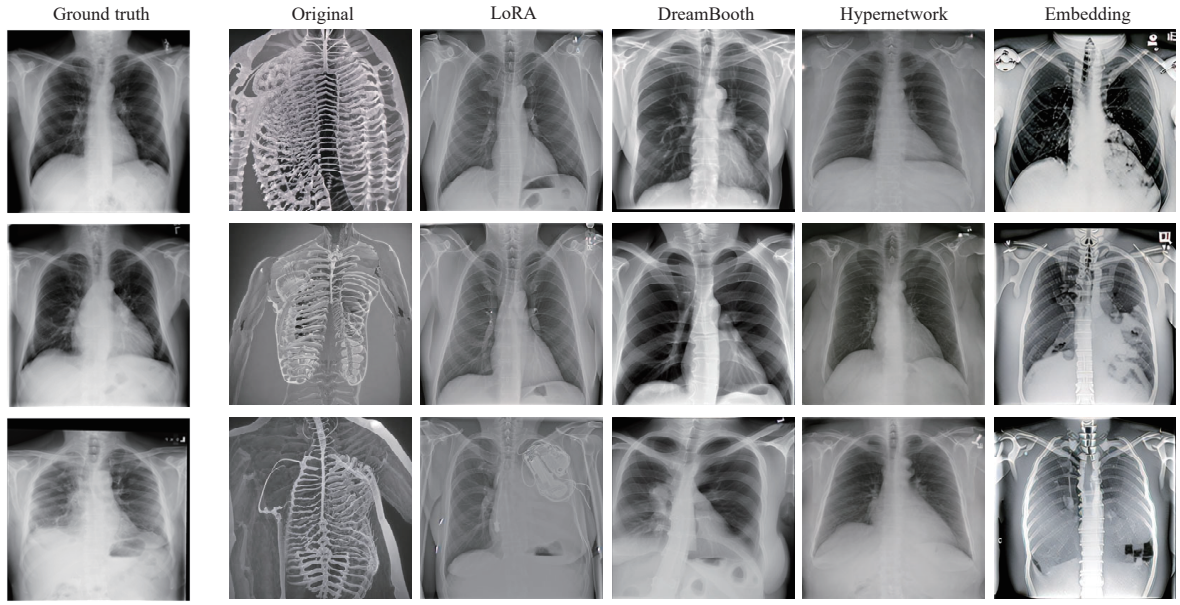


Fig. 6 Generation effects of different fine-tuning methods. In (b), the original output of Stable Diffusion and the outputs after fine-tuning using different methods are shown with the inputs of Prompts 1–3 in top row, middle row, and bottom row, respectively.

conversion task, three different conversion directions on the BRATS2020 dataset are selected.

4.2.1 Performance of diffusion models on image generation tasks

Table 9 shows a comparison between the diffusion models and the traditional generative models for image generation on the ISIC2018, BRATS2018, and LiTS2017 datasets. The evaluation uses Normalized Mean Absolute Error (NMAE), PSNR, and SSIM metrics, where lower values of NMAE indicate better quality, while higher values of PSNR and SSIM correspond to improved results. Figure 7 shows the effect of generation of different models. From Table 9, it can be seen that on the three datasets, IDDPM has the lowest NMAE values, indicating the smallest errors in image generation tasks. At the same time, IDDPM also scores the highest values in PSNR and SSIM, showing the best image quality and structural similarity. GDDPM is close to RegGAN in NMAE, but performes better in PSNR and SSIM, especially on the BRATS2018 and LiTS2017 datasets. CycleGAN and Pix2pix show average performance in these metrics, while RegGAN performs better in some aspects, but overall still lower than IDDPM and GDDPM. The reason for these results might be that, compared to some traditional generative models, IDDPM offers a deeper understanding of the probability distribution of generated images. It is able to produce diverse outcomes while maintaining consistency with the real data distribution.

Table 7 Introduction to generative models.

Year	Model	Infrastructure	Reference
2016	ResNet	ResNet	[91]
2015	VGG	VGG	[92]
2017	DNCNN-B	CNN	[93]
2017	SegNet	SegNet	[94]
2017	CycleGAN	GAN	[95]
2017	Pix2Pix	GAN	[96]
2019	A-Unet	U-Net	[97]
2021	RegGAN	GAN	[98]
2019	nnUnet	U-Net	[99]
2022	EDCNN	CNN	[100]
2020	DETR	Transformer	[101]
2022	VT-Unet	U-Net	[102]

4.2.2 Performance of diffusion models on modality conversion tasks

Table 10 shows a comparison between the diffusion models and the traditional generative models for modality conversion on the BRATS2018 dataset. The evaluation employs NMAE, PSNR, and SSIM metrics. Figure 8 shows the effect of modality conversion for different models. From Table 10, it can be seen that IDDPM performs the best overall performance in all

Table 8 Information of the dataset used for the comparison experiment.

Task	Dataset	Description	Reference
Image generation	ISIC2018	The training set contains 2,074 skin lesion JPEG images and the test set contains 520 skin lesion JPEG images.	[103]
	BRATS2018	The training set consists of MRI data from 285 patients, saved in NIfTI format.	[104–106]
	LiTS2017	The training set contains 130 CT images and the test set contains 70 CT images.	[107]
Modality conversion	BRATS2020	The dataset comprises a collection of 112 120 chest X-ray films obtained from 30 805 individual patients.	[104–106]
Image segmentation	MSD	The dataset comprises 10 different datasets, including a cardiac dataset that consists of 30 MRI images dedicated to left atrium segmentation.	[108]
	BRATS2021	The training set and validation set contain MRI of 1251 and 219 patients respectively.	[104, 105, 109]
	ROSE	It is divided into ROSE-1 and ROSE-2, containing 117 and 112 OCTA images, respectively.	[110]
Image denoising	BrainWeb	The dataset includes simulated brain MRI data based on two anatomical models: normal and multiple sclerosis. It provides three modalities of MRI data.	[111]
	ISBI2015	The dataset consists of 400 skull measurement X-ray images with a resolution of 2400×1935 pixels.	[112]
	NIH AAPM-Mayo Clinic	The training set consists of paired full-dose CT and LDCT images from 10 patients.	[113]
Anomaly detection	HKU-SZH X-ray Set	The dataset consists of 326 normal X-ray images and 336 abnormal X-ray images, all of which are saved in JPEG format.	[114]
	PALM	The training set consists of 800 color fundus photographs with image resolutions of either 1444 pixel × 1444 pixel or 2124 pixel × 2056 pixel.	[115]
	Digital Knee X-ray	This dataset contains 1650 digital X-ray images of the knee from hospitals and diagnostic centers.	[116]

Table 9 Performance comparison between diffusion models and traditional generative models on image generation tasks. The bold value in each column is the optimal value.

Model	ISIC2018			BRATS2018			LiTS2017		
	NMAE	PSNR	SSIM	NMAE	PSNR	SSIM	NMAE	PSNR	SSIM
CycleGAN	0.092	22.8	0.83	0.089	23.6	0.83	0.080	23.6	0.84
RegGAN	0.079	24.9	**0.87**	0.076	25.4	0.85	0.074	25.7	0.87
Pix2pix	0.085	23.8	0.82	0.080	24.9	0.85	0.081	24.9	0.86
IDDPM	**0.071**	**26.8**	**0.87**	**0.068**	**26.8**	**0.87**	**0.064**	**27.1**	**0.88**
GDDPM	0.079	25.1	0.85	0.071	25.9	0.86	0.070	26.1	**0.88**

tasks. Pix2pix and GDDPM have similar performances, especially in the T1, FLAIR → T2 and T2, FLAIR → T1 conversions. CycleGAN performs relatively poorly on these tasks. These performance differences may be attributed to the way that each model processes and learns from the imaging data. Traditional models typically require two generators and two discriminators for modality conversion. Each generator is responsible for one direction of the conversion. The discriminators work to distinguish real images from those generated by the generators. In contrast, IDDPM enhances this process by adding a guiding mechanism. This mechanism directs the diffusion process using the image's middle layer feature conditions. There have been studies based on IDDPM and GDDPM for image conversion tasks[117, 118]. However, the experimental results in this paper show that a simple structural change in IDDPM and GDDPM can yield good results in modality conversion. While there is still room for optimization, these findings inspire the application of diffusion models in modality conversion tasks.

4.2.3 Performance of diffusion models on image segmentation tasks

Table 11 shows a comparison between the diffusion models and the traditional generative models for image segmentation on the MSD, BRATS2021, and ROSE

Table 10 Performance comparison between diffusion models and traditional generative models on modality conversion tasks. The bold value in each column is the optimal value.

Model	T1, T2 → FLAIR			T1, FLAIR → T2			T2, FLAIR → T1		
	NMAE	PSNR	SSIM	NMAE	PSNR	SSIM	NMAE	PSNR	SSIM
A-Unet	0.075	25.27	0.85	0.068	**27.64**	0.92	0.081	24.98	0.91
CycleGAN	0.080	24.75	0.84	0.084	23.48	0.88	0.081	23.69	0.91
Pix2pix	0.075	25.13	0.86	0.072	26.88	0.92	0.069	26.01	0.92
IDDPM	**0.070**	**26.69**	**0.89**	**0.067**	27.24	**0.93**	**0.066**	**27.14**	**0.93**
GDDPM	0.072	26.01	0.88	0.068	26.96	0.92	0.069	26.85	0.92

Table 11 Performance comparison between diffusion models and traditional generative models on image segmentation tasks. The bold value in each column is the optimal value.

Model	MSD			BRATS2021			ROSE		
	DICE	95HD	IOU	DICE	95HD	IOU	DICE	95HD	IOU
nnUnet	**0.91**	**4.73**	0.84	**0.88**	**13.46**	**0.79**	**0.96**	**3.91**	**0.92**
VT-Unet	**0.91**	6.29	**0.85**	0.87	13.87	0.77	0.92	3.98	0.85
SegNet	0.81	9.20	0.68	0.83	16.74	0.71	0.88	7.62	0.79
IDDPM	0.90	6.31	0.81	0.83	16.79	0.71	0.90	5.59	0.81
GDDPM	0.86	5.79	0.75	0.82	17.21	0.69	0.86	6.47	0.76

Table 12 Performance comparison between diffusion models and traditional generative models on image denoising tasks. The bold value in each column is the optimal value.

Model	BrainWeb			ISBI2015			NIH AAPM-Mayo-Clinic		
	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM
Reg-GAN	0.0199	30.99	0.80	0.0159	33.75	0.88	0.0095	40.96	**0.98**
DNCNN-B	0.0206	30.81	0.75	0.0182	32.08	0.81	0.0117	39.79	0.91
EDCNN	0.0213	30.89	0.78	0.0161	33.16	0.85	0.0079	41.59	0.97
IDDPM	**0.0185**	**31.41**	0.81	**0.0147**	**34.71**	**0.90**	**0.0071**	**43.24**	**0.98**
GDDPM	0.0189	31.19	**0.83**	0.0158	33.94	0.88	0.0076	41.82	**0.98**

Table 13 Performance comparison between diffusion models and traditional generative models on anomaly detection tasks. The bold value in each column is the optimal value.

Model	HKU-SZH X-ray Set			PALM			Digital Knee X-ray		
	Top-1 error	AUC	Top-5 error	Top-1 error	AUC	Top-5 error	Top-1 error	AUC	Top-5 error
DETR	**23.64**	**84.63**	**6.57**	**22.99**	**84.68**	**6.21**	**23.35**	**82.67**	**6.31**
ResNet	25.69	82.79	7.23	24.87	83.79	7.23	26.58	82.52	7.92
VGG	28.87	82.58	8.89	28.74	81.28	8.60	29.38	80.01	9.57
IDDPM	37.85	70.87	14.14	35.91	63.62	12.76	36.16	62.69	13.12
GDDPM	38.96	70.09	15.38	37.69	60.89	13.98	35.99	60.13	12.81

datasets. The evaluation uses dice coefficient (DICE), 95% Hausdorff Distance (95HD), and Intersection Over Union (IOU) metrics, where lower 95HD values indicate better quality, while higher values of DICE and IOU correspond to improved results. From Table 11, it can be seen that nnUnet performs the best on these three datasets, especially on the ROSE dataset. While IDDPM and GDDPM show good performance in some aspects, they are overall still inferior to nnUnet and VT-Unet. The reason for this could be that the segmentation task focuses on predicting the label of each pixel, while the diffusion model generates new pixels during the diffusion process. Thus, the final segmentation effect is affected by the pixel value. The convolution and deconvolution operations of U-Net and its variants enable the model to obtain context information and spatial information of each scale, thereby accurately performing pixel-level

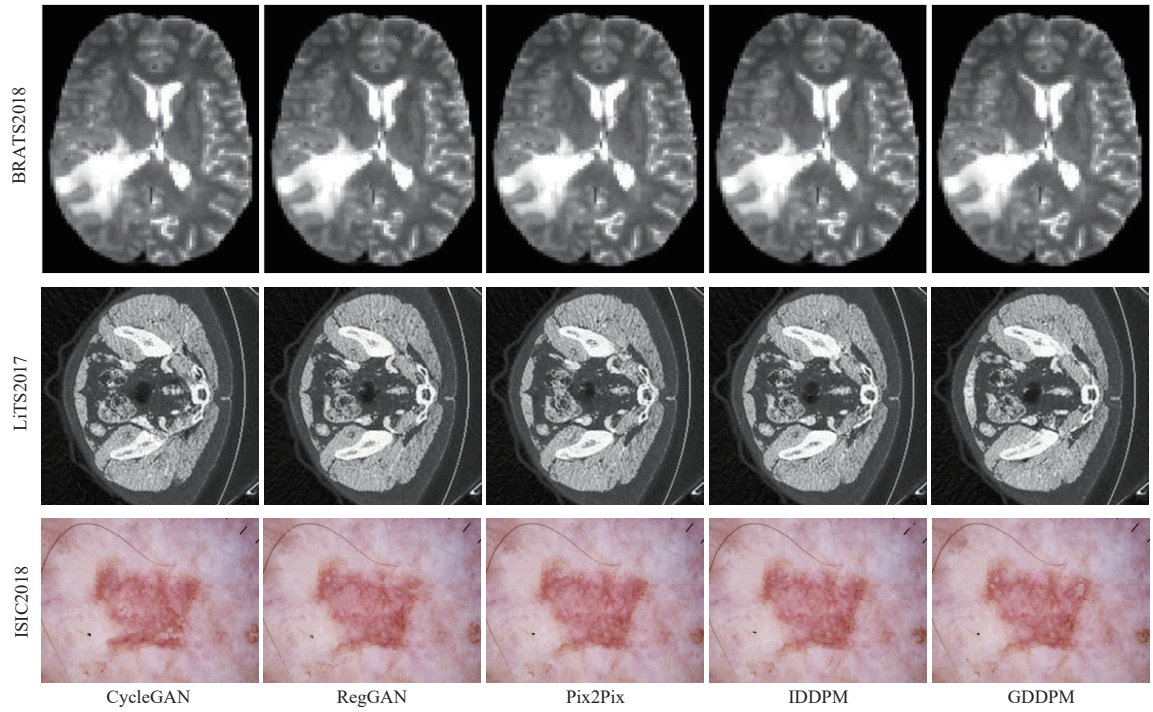


Fig. 7 Comparison of the performance between the diffusion models and traditional generative models in image generation tasks. The generation performances of different models on BRATS2018, LiTS2017, and ISIC2018 datasets are shown in the top, middle, and bottom rows, respectively.

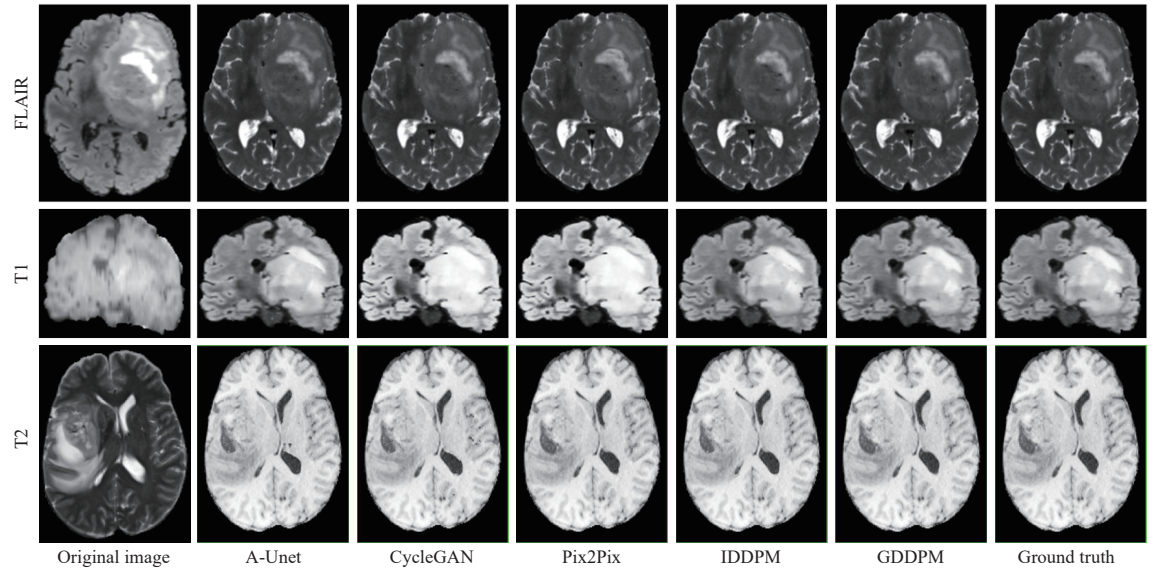


Fig. 8 Performance comparison between the diffusion models and conventional generative models in modality conversion task. The conversion from the FLAIR modality to T2 modality, from T1 modality to the FLAIR modality, and from T2 modality to T1 modality are shown in the top, middle, and bottom rows, respectively.

segmentation. Wu et al.[65] constructed FF-Parser to eliminate the negative impact of high-frequency noise in the diffusion process, while Hu et al.[119] used image-level annotations to obtain the predicted mask of the target object, which does not require the guidance of an external classifier and avoids the influence of noise in the diffusion process.

4.2.4 Performance of diffusion models on image denoising tasks

Table 12 shows a comparison between the diffusion

models and the traditional generative models for image denoising on the BrainWeb, ISBI2015, and NIH AAPM-Mayo-Clinic datasets. The evaluation uses RMSE, PSNR, and SSIM metrics, where lower RMSE values indicate better quality. Figure 9 shows the effect of denoising for different models. From Table 12, it can be seen that IDDPM consistently shows the best performance across all datasets, particularly excelling in reducing RMSE and achieving high PSNR and SSIM scores, indicating superior denoising ability and image quality retention. GDDPM also performs well, especially in maintaining structural integrity as indicated by its SSIM scores. Reg-GAN and EDCNN-B show strong performances in specific datasets, while DNCNN-B generally lags behind the others in these tasks. This could be explained by the generative process of IDDPM and GDDPM, which typically starts with noise diffusion and ends with target distribution. This makes it conducive to image denoising. When there is explicit conditional information, such as a textual description or label, the diffusion model can be guided to generate images with specific properties.

4.2.5 Performance of diffusion models on anomaly detection tasks

Table 13 shows a comparison between the diffusion models and the traditional generative models for anomaly detection on the HKU-SZH X-ray Set, PALM, and Digital Knee X-ray datasets. The evaluation uses Top-1 error, Area Under Curve (AUC), and Top-5 error metrics, where lower Top-1 error and Top-5 error values indicate better quality, while higher AUC values correspond to improved results. From Table 13, it can be seen that DETR and ResNet perform better than VGG, IDDPM, and GDDPM in all datasets, with DETR showing the best overall performance. This model has the highest AUC and the smallest Top-1 error and Top-5 error, indicating its strong capability for anomaly detection. However, IDDPM and GDDPM show weaker anomaly detection performance in these specific datasets. Traditional models can locate and identify abnormal regions in the image in the form of bounding boxes, while IDDPM and GDDPM, which are reconstruction-based methods for anomaly detection, need to compare the pixels and characteristics of the input image and the reconstructed image to locate the anomaly. Therefore, the diffusion model is easily affected by the quality of image reconstruction when performing anomaly detection tasks. Fontanella et al.[120] presented a weakly supervised technique that integrates DDPM and DDIM at each stage of the sampling process. This method guarantees the reconstruction quality of images and greatly improves the effectiveness of anomaly detection.

5 Challenge and Prospect

5.1 Current challenges

As an advanced generative model, diffusion models have great potential in MIC. However, they are faced with several challenges:

Data acquisition and annotation. Training and fine-

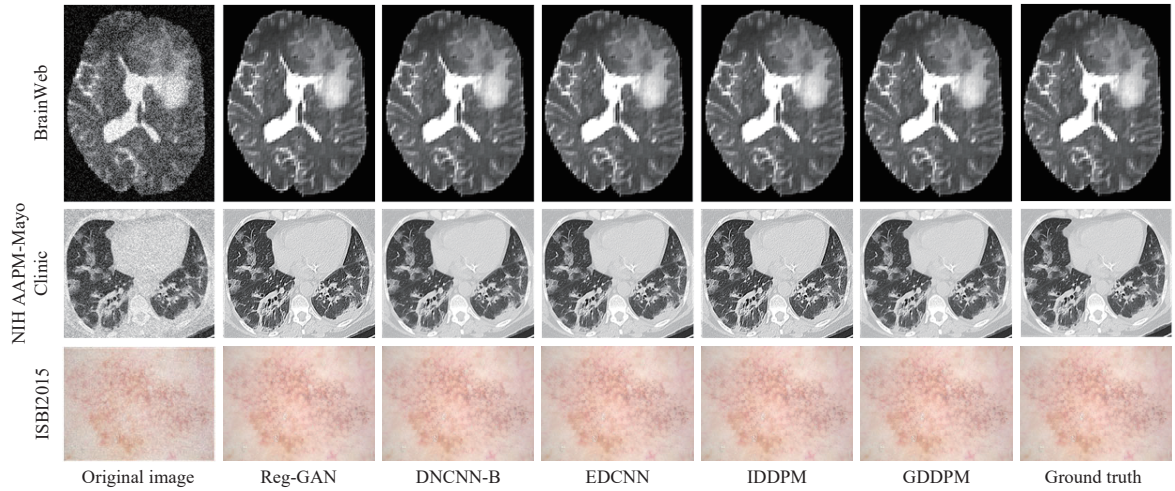


Fig. 9 Performance comparison between the diffusion models and conventional generative models in image denoising task. The denoising performances of different models on BrainWeb, NIH AAPM-Mayo Clinic, and ISBI2015 datasets are shown in the top, middle, and bottom rows, respectively.

tuning diffusion models require a large amount of medical image data, and acquiring and annotating such data is costly and time-consuming. Obtaining medical imaging data entails collaborating with medical institutions and raises concerns regarding data privacy. Accurate annotation of medical image data requires specialized medical knowledge and expertise.

Model complexity and computational resources. The design and training of diffusion models are typically more complex than those of traditional generative models, requiring more computational resources and time. This can limit many researchers and healthcare institutions, particularly when the models are applied to large-scale datasets or real-time diagnostic systems.

Model generalizability and interpretability. Current diffusion models may perform well on specific datasets. However, their ability to generalize across different tasks is still an issue, and their robustness and adaptability to different environments need further improvement. In addition, diffusion models are often regarded as black-box models, which presents challenges in interpreting the internal decision-making and reasoning processes of the model. This is an important issue in MIC, where physicians need to understand the model's decision-making basis to make accurate judgments and decisions about diagnostic results.

5.2 Future prospects

The application of diffusion models in MIC is still in its early stages. It requires further improvements in both theoretical development and empirical investigation. The following areas of research may be possible future directions:

Combining diffusion models with large models. From an algorithmic perspective, both diffusion models and large models are generative pre-training methods. The inclusion of human-annotated feedback and reinforcement learning in the training process of diffusion models, like how ChatGPT is fine-tuned based on human feedback, is worth exploring. Additionally, efficient fine-tuning of Stable Diffusion for MIC tasks is also worth further investigation.

Theoretical explanations of diffusion models. Diffusion models are powerful models, particularly in applications where they can rival GANs without the need for adversarial strategies. Therefore, it is crucial to understand why diffusion models are more efficient than other models in performing specific tasks. The theoretical interpretation of diffusion models represents an important research direction. Furthermore, exploring various modeling approaches within the framework of diffusion model theory presents a promising avenue.

Personalized medicine and precision diagnosis. Diffusion models can be used to analyze patient-specific medical image data and reveal disease characteristics and pathological changes. This provides customized diagnostic information for each patient by learning patterns and associations in the data. The integration of deep learning and artificial intelligence technologies will further improve the accuracy and adaptability of diffusion models in processing complex medical imaging data, and combining them with other health data can provide physicians with evidence-based personalized treatment recommendations.

6 Conclusion

In this study, we review the core theoretical framework of the diffusion models and their recent applications in five MIC tasks. We not only summarize the progress of the diffusion models in MIC, but also explore the performance of the diffusion models in different MIC tasks through fine-tuning experiments and a series of comparison experiments. Through these experiments, we show the unique advantages of the diffusion models in handling different MIC tasks, and also explore potential strategies for its performance improvement. We also recognize that diffusion models still face challenges in practical applications, such as computational efficiency and model generalization ability. Future research could aim to address these issues and develop more efficient and accurate model variants to further expand their applications in areas such as medical diagnosis, disease monitoring, and treatment planning.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos. 62366050, 61966033, and 61866035)

References

[1] X. Chen, X. Wang, K. Zhang, K. M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, Recent advances and clinical applications of deep learning in medical image analysis, *Med. Image Anal.*, vol. 79, p. 102444, 2022.

[2] G. Varoquaux and V. Cheplygina, Machine learning for medical imaging: Methodological failures and recommendations for the future, *NPJ Digit. Med.*, vol. 5, no. 1, p. 48, 2022.

[3] Y. Zhao, X. Wang, T. Che, G. Bao, and S. Li, Multi-task deep learning for medical image computing and analysis: A review, *Comput. Biol. Med.*, vol. 153, p. 106496, 2023.

[4] S. Wang, G. Cao, Y. Wang, S. Liao, Q. Wang, J. Shi, C. Li, and D. Shen, Review and prospect: Artificial intelligence in advanced medical imaging, *Front. Radiol.*, vol. 1, p. 781868, 2021.

[5] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, Deep learning-enabled medical computer vision, *NPJ Digit. Med.*, vol. 4, no. 1, p. 5, 2021.

[6] J. M. B. Haslbeck, L. F. Bringmann, and L. J. Waldorp, A tutorial on estimating time-varying vector autoregressive models, *Multivar. Behav. Res.*, vol. 56, no. 1, pp. 120–149, 2021.

[7] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, Generative adversarial networks: An overview, *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, 2018.

[8] D. P. Kingma and M. Welling, An introduction to variational autoencoders, *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019.

[9] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, Diffusion models in medical imaging: A comprehensive survey, *Med. Image Anal.*, vol. 88, p. 102846, 2023.

[10] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M. H. Yang, Diffusion models: A comprehensive survey of methods and applications, *ACM Comput. Surv.*, vol. 56, no. 4, p. 105, 2024.

[11] J. Ho, A. Jain, and P. Abbeel, Denoising diffusion probabilistic models, in *Proc. 34th Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2020, p. 574.

[12] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, Hierarchical text-conditional image generation with CLIP latents, arXiv preprint arXiv: 2204.06125, 2022.

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-resolution image synthesis with latent diffusion models, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, pp. 10684–10695.

[14] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, S. K. S. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al., Photorealistic text-to-image diffusion models with deep language understanding, in *Proc. 36th Int. Conf. Neural Information Processing Systems*, New Orleans, LA, USA, 2022, p. 2643.

[15] F. A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, Diffusion models in vision: A survey, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10850–10869, 2023.

[16] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, DiffuSeq: Sequence to sequence text generation with diffusion models, in *Proc. 11th Int. Conf. Learning Representations*, Kigali, Rwanda, https://doi.org/10.48550/arXiv.2210.08933, 2023.

[17] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. B. Hashimoto, Diffusion-LM improves controllable text generation, in *Proc. 36th Int. Conf. Neural Information Processing Systems*, New Orleans, LA, USA, 2022, p. 313.

[18] Z. Zhu, H. Zhao, H. He, Y. Zhong, S. Zhang, H. Guo, T. Chen, and W. Zhang, Diffusion models for reinforcement learning: A survey, arXiv preprint arXiv: 2311.01223, 2023.

[19] Y. Fan, H. Liao, S. Huang, Y. Luo, H. Fu, and H. Qi, A survey of emerging applications of diffusion probabilistic models in MRI, arXiv preprint arXiv: 2311.11383, 2023.

[20] Y. Song and S. Ermon, Improved techniques for training score-based generative models, in *Proc. 34th Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2020, p. 1043.

[21] Y. Song and S. Ermon, Generative modeling by estimating gradients of the data distribution, in *Proc. 33rd Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2019, p. 1067.

[22] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, Score-based generative modeling through stochastic differential equations, in *Proc. 9th Int. Conf. on Learning Representations* (*ICLR*), https://doi.org/10.48550/arXiv.2011.13456, 2021.

[23] W. Du, H. Zhang, T. Yang, and Y. Du, A flexible diffusion model, in *Proc. 40th Int. Conf. Machine Learning*, Honolulu, HI, USA, 2023, p. 347.

[24] C. Yu, Y. Guan, Z. Ke, K. Lei, D. Liang, and Q. Liu, Universal generative modeling in dual domains for dynamic MRI, *NMR Biomed.*, vol. 36, no. 12, p. e5011, 2023.

[25] X. Liu, D. H. Park, S. Azadi, G. Zhang, A. Chopikyan, Y. Hu, H. Shi, A. Rohrbach, and T. Darrell, More control for free! image synthesis with semantic diffusion guidance, in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision* (*WACV*), Waikoloa, HI, USA, 2023, pp. 289–299.

[26] B. Wallace, A. Gokul, S. Ermon, and N. Naik, End-to-end diffusion latent optimization improves classifier guidance, in *Proc. IEEE/CVF Int. Conf. Computer Vision*, Paris, France, 2023, pp. 7246–7256.

[27] S. Hong, G. Lee, W. Jang, and S. Kim, Improving sample quality of diffusion models using self-attention guidance, in *Proc. IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Paris, France, 2023, pp. 7428–7437.

[28] W. G. Choi, S. J. Kim, T. Kim, and J. H. Chang, Prior-free Guided TTS: An improved and efficient diffusion-based text-guided speech synthesis, in *Proc. INTERSPEECH 2023*, Dublin, Ireland, 2023, pp. 4289–4293.

[29] J. Song, C. Meng, and S. Ermon, Denoising diffusion implicit models, in *Proc. 9th Int. Conf.on Learning*

Representations (*ICLR*), https://doi.org/10.48550/arXiv.2010.02502, 2021.

[30] Q. Zhang, M. Tao, and Y. Chen, gDDIM: Generalized denoising diffusion implicit models, in *Proc. 11th Int. Conf. on. Learning Representations* (*ICLR*), Kigali, Rwanda, https://doi.org/10.48550/arXiv.2206.05564, 2023.

[31] L. Liu, Y. Ren, Z. Lin, and Z. Zhao, Pseudo numerical methods for diffusion models on manifolds, in *Proc 10th Int. Conf. on. Learning Representations* (ICLR), https://doi.org/10.48550/arXiv.2202.09778, 2022.

[32] Q. Zhang and Y. Chen, Fast sampling of diffusion models with exponential integrator, in *Proc. 11th Int. Conf. on Learning Representation*s, Kigali, Rwanda, https://doi.org/10.48550/arXiv.2204.13902, 2023.

[33] H. Zheng, P. He, W. Chen, and M. Zhou, Truncated diffusion probabilistic models and diffusion-based adversarial auto-encoders, in *Proc. 11th Int. Conf. on Learning Representations*, Kigali, Rwanda, https://doi.org/10.48550/arXiv.2202.09671, 2023.

[34] W. Luo, A comprehensive survey on knowledge distillation of diffusion models, arXiv preprint arXiv: 2304.04262, 2023.

[35] G. Du, X. Cao, J. Liang, X. Chen, and Y. Zhan, Medical image segmentation based on U-Net: A review, *J. Imaging Sci. Technol.*, vol. 64, no. 2, p. 020508, 2020.

[36] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *Proc. 18th Int. Conf. Medical Image Computing and Computer-Assisted Intervention* (*MICCAI*), Munich, Germany, 2015, p. 234–241.

[37] W. Peebles and S. Xie, Scalable diffusion models with transformers, in *Proc. IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Paris, France, 2023, pp. 4172–4182.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[39] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, A survey of convolutional neural networks: Analysis, applications, and prospects, *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, 2022.

[40] L. R. Medsker and L. C. Jain, *Recurrent Neural Networks*. Boca Raton, FL, USA: CRC Press, 2001, p. 392.

[41] A. Hatamizadeh, J. Song, G. Liu, J. Kautz, and A. Vahdat, DiffiT: Diffusion vision transformers for image generation, arXiv preprint arXiv: 2312.02139, 2023.

[42] G. J. Chowdary and Z. Yin, Diffusion transformer U-Net for medical image segmentation, in *Proc. 26th Int. Conf. Medical Image Computing and Computer-Assisted Intervention* (*MICCAI*), Vancouver, Canada, 2023, pp. 622–631.

[43] H. A. Bedel and T. Çukur, DreaMR: Diffusion-driven counterfactual explanation for functional MRI, arXiv preprint arXiv: 2307.09547, 2023.

[44] S. Shao, X. Yuan, Z. Huang, Z. Qiu, S. Wang, and K. Zhou, DiffuseExpand: Expanding dataset for 2D medical image segmentation using diffusion models, arXiv preprint arXiv: 2304.13416, 2023.

[45] S. Zhang, J. Liu, B. Hu, and Z. Mao, GH-DDM: The generalized hybrid denoising diffusion model for medical image generation, *Multimed. Syst*., vol. 29, no. 3, pp. 1335–1345, 2023.

[46] P. N. Huy and T. M. Quan, Denoising diffusion medical models, in *Proc. 20th Int. Symp. Biomedical Imaging* (*ISBI*), Cartagena, Colombia, 2023, pp. 1–5.

[47] Z. Dorjsembe, S. Odonchimed, and F. Xiao, Three-dimensional medical image synthesis with denoising diffusion probabilistic models. in *Proc. Int. Conf. Medical Imaging with Deep Learning*, Zurich, Switzerland, https://openreview.net/pdf?id=Oz7lKWVh45H, 2022.

[48] W. H. L. Pinaya, P. D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, Brain imaging generation with latent diffusion models, in *Proc. 2nd MICCAI Workshop on Deep Generative Models*, Singapore, 2022, pp. 117–126.

[49] F. Khader, G. Müller-Franzes, S. T. Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baeßler, S. Foersch, et al., Denoising diffusion probabilistic models for 3D medical image generation, *Sci. Rep.*, vol. 13, no. 1, p. 7303, 2023.

[50] P. Esser, R. Rombach, and B. Ommer, Taming transformers for high-resolution image synthesis, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 12868–12878.

[51] B. Kim and J. C. Ye, Diffusion deformable model for 4D temporal medical image generation, in *Proc. 25th Int. Conf. Medical Image Computing and Computer Assisted Intervention* (*MICCAI*), Singapore, 2022, pp. 539–548.

[52] Y. Song, L. Shen, L. Xing, and S. Ermon, Solving inverse problems in medical imaging with score-based generative models, in *Proc. 10th Int. Conf. on Learning Representations*, https://doi.org/10.48550/arXiv.2111.08005, 2022.

[53] H. Chung and J. C. Ye, Score-based diffusion models for accelerated MRI, *Med. Image Anal.*, vol. 80, p. 102479, 2022.

[54] C. Peng, P. Guo, S. K. Zhou, V. M. Patel, and R. Chellappa, Towards performant and reliable undersampled MR reconstruction via diffusion model sampling, in *Proc. 25th Int. Conf. Medical Image Computing and Computer Assisted Intervention* (*MICCAI*), Singapore, 2022, pp. 623–633.

[55] A. Güngör, S. U. H. Dar, Ş. Öztuük, Y. Korkmaz, H. A. Bedel, G. Elmas, M. Ozbey, and T. Çukur, Adaptive diffusion priors for accelerated MRI reconstruction, *Med. Image Anal.*, vol. 88, p. 102872, 2023.

[56] H. Chung, D. Ryu, M. T. McCann, M. L. Klasky, and J. C. Ye, Solving 3D inverse problems using pre-trained 2D diffusion models, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 22542–22551.

[57] L. X. Nguyen, P. S. Aung, H. Q. Le, S. B. Park, and C. S.

Hong, A new chapter for medical image generation: The stable diffusion method, in *Proc. Int. Conf. Information Networking* (*ICOIN*), Bangkok, Thailand, 2023, pp. 483–486.

[58] M. Özbey, O. Dalmaz, S. U. H. Dar, H. A. Bedel, Ş, Öztürk, A. Güngör, and T. Çukur, Unsupervised medical image translation with adversarial diffusion models, *IEEE Trans. Med. Imaging*, vol. 42, no. 12, pp. 3524–3539, 2023.

[59] Q. Lyu and G. Wang, Conversion between CT and MRI images using diffusion and score-matching models, arXiv preprint arXiv: 2209.12104, 2022.

[60] X. Li, K. Shang, G. Wang, and M. D. Butala, DDMM-Synth: A denoising diffusion model for cross-modal medical image synthesis with sparse-view measurement embedding, arXiv preprint arXiv: 2303.15770, 2023.

[61] Y. Li, H. C. Shao, X. Liang, L. Chen, R. Li, S. Jiang, J. Wang, and Y. Zhang, Zero-shot medical image translation via frequency-guided diffusion models, *IEEE Trans. Med. Imaging*, vol. 43, no. 3, pp. 980–993, 2024.

[62] S. Pan, E. Abouei, J. Wynne, T. Wang, R. L. J. Qiu, Y. Li, C. W. Chang, J. Peng, J. Roper, P. Patel, et al., Synthetic CT generation from MRI using 3D transformer-based denoising diffusion model, arXiv preprint arXiv: 2305.19467, 2023.

[63] C. Tiago, S. R. Snare, J. Šprem, and K. McLeod, A domain translation framework with an adversarial denoising diffusion model to generate synthetic datasets of echocardiography images, *IEEE Access*, vol. 11, pp. 17594–17602, 2023.

[64] F. Bieder, J. Wolleb, A. Durrer, R. Sandkühler, and P. C. Cattin, Denoising diffusion models for memory-efficient processing of 3D medical images, in *Proc. Medical Imaging with Deep Learning*, Nashville, TN, USA, 2023, pp. 552–567.

[65] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, and Y. Xu, MedSegDiff: Medical image segmentation with diffusion probabilistic model, in *Proc. Medical Imaging with Deep Learning*, Nashville, TN, USA, 2023, pp. 1623–1639.

[66] J. Wu, W. Ji, H. Fu, M. Xu, Y. M. Jin, and Y. Xu, MedSegDiff-V2: Diffusion-based medical image segmentation with transformer, in *Proc. AAAI Conf. Artificial Intelligence*, Vancouver, Canada, 2024, pp. 6030–6038.

[67] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin, Diffusion models for implicit image segmentation ensembles, in *Proc. Int. Conf. Medical Imaging with Deep Learning*, Zurich, Switzerland, 2022, pp. 1336–1348.

[68] X. Guo, Y. Yang, C. Ye, S. Lu, B. Peng, H. Huang, Y. Xiang, and T. Ma, Accelerating diffusion models via pre-segmentation diffusion sampling for medical image segmentation, in *Proc. 20th Int. Symp. Biomedical Imaging* (*ISBI*), Cartagena, Colombia, 2023, pp. 1–5.

[69] Y. Fu, Y. Li, S. U. Saeed, M. J. Clarkson, and Y. Hu, Importance of aligning training strategy with evaluation for diffusion models in 3D multiclass segmentation, in *Proc. 3rd MICCAI Workshop on Deep Generative Models*, Vancouver, Canada, 2024, pp. 86–59.

[70] Z. Xing, L. Wan, H. Fu, G. Yang, and L. Zhu, Diff-UNet: A diffusion embedded network for volumetric segmentation, arXiv preprint arXiv: 2303.10326, 2023.

[71] A. Rahman, J. M. J. Valanarasu, I. Hacihaliloglu, and V. M. Patel, Ambiguous medical image segmentation using diffusion models, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 11536–11546.

[72] T. Chen, C. Wang, and H. Shan, BerDiff: Conditional Bernoulli diffusion model for medical image segmentation, in *Proc. 26th Int. Conf. Medical Image Computing and Computer Assisted Intervention*, Vancouver, Canada, 2023, pp. 491–501.

[73] T. Amit, S. Shichrur, T. Shaharabany, and L. Wolf, Annotator consensus prediction for medical image segmentation with diffusion models, in *Proc. 26th Int. Conf. Medical Image Computing and Computer Assisted Intervention*, Vancouver, Canada, 2023, pp. 544–554.

[74] A. Abirami, P. Prakash, and Y. K. Ma, Variable-order fractional diffusion model-based medical image denoising, *Math. Probl. Eng.*, vol. 2021, p. 8050017, 2021.

[75] D. Hu, Y. K. Tao, and I. Oguz, Unsupervised denoising of retinal OCT with diffusion probabilistic model, in *Proc. SPIE 12032, Medical Imaging 2022*: *Image Processing*, San Diego, CA, USA, 2022, p. 1203206.

[76] H. Chung, E. S. Lee, and J. C. Ye, MR image denoising and super-resolution using regularized reverse diffusion, *IEEE Trans. Med. Imaging*, vol. 42, no. 4, pp. 922–934, 2023.

[77] X. Liu, Y. Xie, S. Diao, S. Tan, and X. Liang, A diffusion probabilistic prior for low-dose CT image denoising, arXiv preprint arXiv: 2305.15887, 2023.

[78] W. Xia, Q. Lyu, and G. Wang, Low-dose CT using denoising diffusion probabilistic model for 20× speedup, arXiv preprint arXiv: 2209.15136, 2022.

[79] P. Sanchez, A. Kascenas, X. Liu, A. Q. O'Neil, and S. A. Tsaftaris, What is healthy? Generative counterfactual diffusion for lesion localization, in *Proc. 2nd MICCAI Workshop on Deep Generative Models*, Singapore, 2022, pp. 34–44.

[80] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, Diffusion models for medical anomaly detection, in *Proc. 25th Int. Conf. Medical Image Computing and Computer Assisted Intervention* (*MICCAI*), Singapore, 2022, pp. 35–45.

[81] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks, AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops* (*CVPRW*), New Orleans, LA, USA, 2022, pp. 649–655.

[82] H. Iqbal, U. Khalid, C. Chen, and J. Hua, Unsupervised anomaly detection in medical images using masked diffusion model, in *Proc. 14th Int. Workshop on Machine Learning in Medical Imaging*, Vancouver, Canada, 2023,

pp. 372–381.
[83] W. H. L. Pinaya, M. S. Graham, R. Gray, P. F. D. Costa, P. D. Tudosiu, P. Wright, Y. H. Mah, A. D. MacKinnon, J. T. Teo, R. Jager, et al., Fast unsupervised brain anomaly detection and segmentation with diffusion models, in *Proc. 25th Int. Conf. Medical Image Computing and Computer Assisted Intervention* (*MICCAI*), Singapore, 2022, pp. 705–714.
[84] F. Behrendt, D. Bhattacharya, J. Krüger, R. Opfer, and A. Schlaefer, Patched diffusion models for unsupervised anomaly detection in brain MRI, in *Proc. Medical Imaging with Deep Learning*, Nashville, TN, USA, 2023, pp. 1019–1032.
[85] A. Q. Nichol and P. Dhariwal, Improved denoising diffusion probabilistic models, in *Proc. 38th Int. Conf. Machine Learning*, Vienna, Austria, 2021, pp. 8162–8171.
[86] P. Dhariwal and A. Q. Nichol, Diffusion models beat GANs on image synthesis, in *Proc. 34th Advances in Neural Information Processing Systems*, https://doi.org/10.48550/arXiv.2105.05233, 2021.
[87] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning, arXiv preprint arXiv: 2307.04725, 2023.
[88] D. Li, J. Li, and S. C. H. Hoi, Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing, in *Proc. 37th Advances in Neural Information Processing Systems*, https://doi.org/10.48550/arXiv.2305.14720, 2023.
[89] N. Ruiz, Y. Li, V. Jampani, W. Wei, T. Hou, Y. Pritch, N. Wadhwa, M. Rubinstein, and K. Aberman, HyperDreamBooth: HyperNetworks for fast personalization of text-to-image models, arXiv preprint arXiv: 2307.06949, 2023.
[90] N. Deckers, J. Peters, and M. Potthast, Manipulating embeddings of stable diffusion prompts, arXiv preprint arXiv: 2308.12059, 2023.
[91] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 2016, pp. 770–778.
[92] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, in *Proc. 3rd Int. Conf. on Learning Representations* (*ICLR*), https://doi.org/10.48550/arXiv.1409.1556, 2015.
[93] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising, *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, 2017.
[94] V. Badrinarayanan, A. Kendall, and R. Cipolla, SegNet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.
[95] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in *Proc. IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 2242–2251.
[96] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, Image-to-image translation with conditional adversarial networks, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 5967–5976.
[97] Y. Li, X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, and X. Wang, Attention-guided unified network for panoptic segmentation, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 7019–7028.
[98] L. Kong, C. Lian, D. Huang, Z. Li, Y. Hu, and Q. Zhou, Breaking the dilemma of medical image-to-image translation, in *Proc. 34th Advances in Neural Information Processing Systems*, https://doi.org/10.48550/arXiv.2110.06465, 2021.
[99] F. Isensee, P. F. Jäger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, Automated design of deep learning methods for biomedical image segmentation, arXiv preprint arXiv: 1904.08128, 2019.
[100] Y. Wang, Y. Yang, Z. Ma, K. C. Wong, and X. Li, EDCNN: Identification of genome-wide RNA-binding proteins using evolutionary deep convolutional neural network, *Bioinformatics*, vol. 38, no. 3, pp. 678–686, 2022.
[101] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, End-to-end object detection with transformers, in *Proc. 16th European Conf. Computer Vision* (*ECCV*), Glasgow, UK, 2020, pp. 213–229.
[102] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, A robust volumetric transformer for accurate 3D tumor segmentation, in *Proc. 25th Int. Conf. Medical Image Computing and Computer Assisted Intervention* (*MICCAI*), Singapore, 2022, pp. 162–172.
[103] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al., Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC), arXiv preprint arXiv: 1902.03368, 2019.
[104] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (BRATS), *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.
[105] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features, *Sci. Data*, vol. 4, no. 1, p. 170117, 2017.
[106] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge, arXiv preprint arXiv: 1811.02629, 2018.
[107] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen,

G. Kaissis, A. Szeskin, C. Jacobs, G. E. H. Mamani, G. Chartrand, et al., The liver tumor segmentation benchmark (LiTS), *Med. Image Anal.*, vol. 84, p. 102680, 2023.

[108] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, et al., A large annotated medical image dataset for the development and evaluation of segmentation algorithms, arXiv preprint arXiv: 1902.09063, 2019.

[109] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, et al., The RSNA-ASNR-MICCAI BRATS 2021 benchmark on brain tumor segmentation and radiogenomic classification, arXiv preprint arXiv: 2107.02314, 2021.

[110] Y. Ma, H. Hao, J. Xie, H. Fu, J. Zhang, J. Yang, Z. Wang, J. Liu, Y. Zheng, and Y. Zhao, ROSE: A retinal OCT-angiography vessel segmentation dataset and new model, *IEEE Trans. Med. Imaging*, vol. 40, no. 3, pp. 928–939, 2021.

[111] R. K. S. Kwan, A. C. Evans, and G. B. Pike, MRI simulation-based evaluation of image-processing and classification methods, *IEEE Trans. Med. Imaging*, vol. 18, no. 11, pp. 1085–1097, 1999.

[112] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre, et al., Longitudinal multiple sclerosis lesion segmentation: Resource and challenge, *NeuroImage*, vol. 148, pp. 77–102, 2017.

[113] C. McCollough, TU-FG-207A-04: Overview of the low dose CT grand challenge, *Med. Phys.*, vol. 43, no. 6Part35, pp. 3759–3760, 2016.

[114] S. Jaeger, S. Candemir, S. Antani, Y. X. J. Wáng, P. X. Lu, and G. Thoma, Two public chest X-ray datasets for computer-aided screening of pulmonary diseases, *Quant. Imaging Med. Surg.*, vol. 4, no. 6, pp. 475–477, 2014.

[115] H. Fu, F. Li, J. I. Orlando, H. Bogunović, X. Sun, J. Liao, Y. Xu, S. Zhang, and X. Zhang, Palm: Pathologic myopia challenge, *IEEE Dataport*, doi: 10.21227/55pk-8z03.

[116] S. Gornale and P. Patravali, Digital knee X-ray images, *Mendeley Data*, doi: 10.17632/t9ndx37v5h.1.

[117] J. O. Cross-Zamirski, P. Anand, G. Williams, E. Mouchet, Y. Wang, and C. B. Schönlieb, Class-guided image-to-image diffusion: Cell painting from brightfield images with class labels, in *Proc. IEEE/CVF Int. Conf. Computer Vision Workshops*, Paris, France, 2023, pp. 3802–3811.

[118] S. Pan, C. W. Chang, J. Peng, J. Zhang, R. L. J. Qiu, T. Wang, J. Roper, T.Liu, H. Mao, and X. Yang, Cycle-guided denoising diffusion probability model for 3D cross-modality mri synthesis, arXiv 2305.00042, 2023.

[119] X. Hu, Y. J. Chen, T. Y. Ho, and Y. Shi, Conditional diffusion models for weakly supervised medical image segmentation, in *Proc. 26th Int. Conf. Medical Image Computing and Computer Assisted Intervention*, Vancouver, Canada, 2023, pp. 756–765.

[120] A. Fontanella, G. Mair, J. Wardlaw, E. Trucco, and A. Storkey, Diffusion models for counterfactual generation and anomaly detection in brain images, arXiv preprint arXiv: 2308.02062, 2023.

Yaqing Shi received the BS degree from Hebei University, China in 2021. He is currently a master student in management science and engineering at Xinjiang University of Finance and Economics, China. His research interests include generative AI and medical intelligence.

Abudukelimu Abulizi received the PhD degree from Tsinghua University, China in 2018. He is currently an associate professor at School of Information Management, Xinjiang University of Finance and Economics, China. His research interests include business intelligence, artificial intelligence, and medical intelligence.

Hao Wang is currently an undergraduate student at School of Information Management, Xinjiang University of Finance and Economics, China. His research interests include generative AI and medical intelligence.

Ke Feng received the BS degree from Qufu Normal University, China in 2021. She is currently a master student in management science and engineering at XinJiang University of Finance and Economics, China. Her research interests include generative AI and medical intelligence.

Nihemaiti Abudukelimu received the bachelor degree of medicine from Xinjiang Medical University, China in 2005. He is currently an associate chief physician at Department of Orthopaedic Surgery, Yili Friendship Hospital, China. His research interests include clinical medicine and medical intelligence.

Youli Su received the PhD degree from Tokyo University of Agriculture and Technology, Japan in 2010. She is currently an associate professor at School of Information Management, XinJiang University of Finance and Economics, China. Her research interests include artificial intelligence and data analysis.

Abudukelimu Halidanmu received the PhD degree from Tsinghua University, China in 2018. She is currently an associate professor at School of Information Management, Xinjiang University of Finance and Economics, China. Her research interests include natural language processing and medical intelligence.