# Image Processing Based on Improved Denoising Diffusion Probability Model and its Application in Medical Imaging

**MARCH 2024**

**DOCTOR OF ENGINEERING**

**JINCHENG PENG**

**AKITA PREFECTURAL UNIVERSITY**

# Contents

# Abstract

Image resolution, as one of the fundamental properties of an image, affects its clarity and quality. In the real world, various factors such as hardware limitations and environmental noise contribute to the inevitable degradation of image quality, resulting in the loss of high-frequency details and texture information. To effectively address such issues, research on image super-resolution reconstruction has become an important topic in the fields of computer vision and image processing. Traditional super-resolution techniques include image interpolation, edge information statistics, and frequency domain reconstruction. In recent years, with the breakthrough progress of deep learning methods in computer vision tasks, deep learning-based super-resolution methods have gradually become mainstream. These methods include approaches that use convolutional neural networks to construct mappings and methods that employ generative adversarial networks (GANs) to learn generated images.

Denoising Diffusion Probabilistic Models (DDPM), as a novel image generation method, has shown promising results in computer graphics tasks. Recent studies indicate that DDPM has surpassed traditional generative adversarial networks in various image processing domains. Super-resolution reconstruction is the inverse process of image degradation. This paper first investigates several degradation factors of images and employs a strategy of randomly shuffling degradation factors to construct a low-resolution (LR) image training set that reflects real-world conditions. Building upon the research on the DDPM generation process and DDPM super-resolution reconstruction model, the paper proposes a super-resolution reconstruction image processing method based on an improved denoising diffusion probability model. This method enables high-definition super-resolution image reconstruction. The improvement primarily focuses on noise addition methods and the introduction of Variational Autoencoder (VAE) and Vector Quantized VAE (VQVAE) autoencoder structures. Due to the large GPU memory consumption of the original DDPM model, processing images with resolutions above 256x256 becomes challenging. The main improvement in this paper involves compressing the image using the encoding module of VAE/VQVAE to obtain a smaller resolution latent feature variable. The compressed latent feature variable is then used for DDPM synthesis, and finally, the decoding

module of VAE/VQVAE is employed to upscale the latent feature variable to generate a high-resolution image. This method addresses the difficulty of super-resolution reconstruction for 256x256 and 512x512 resolution high-definition images, enhancing the realism and intricacy of the generated images. Through experimental validation, our improved DDPM model demonstrates significant effectiveness in generating high-definition, high-quality images. We showcase the application of this improvement on different datasets of degraded low-resolution facial images. Furthermore, we apply this technology to the field of medical imaging for the super-resolution processing of medical pathology images, enriching high-quality, high-resolution medical images and making a positive contribution to the development of image generation, image super-resolution, image processing, and image colorization technologies.

# Acknowledgements

First, I would like to express my deep gratitude to Professor Guoyue Chen, my advisor, for his specific guidance and fatherly support during my doctoral studies at Akita Prefectural University. Throughout my three years in Japan, Professor Chen has been caring for every aspect of my life. Without his strong and continuous assistance and inspiration, completing my thesis and even my academic journey in Japan would have been impossible.

In addition to my advisor, I would like to thank the other members of the thesis defense committee: Professor Masayuki Nishiguchi (Akita Prefectural University), Professor Yanwei Chen (Ritsumeikan University), Professor Kazuki Saruta (Akita Prefectural University), and Associate Yuki Terata (Akita Prefectural University). I appreciate their enthusiasm, willingness to read the thesis, pose instructive questions, and provide insightful comments, despite their busy schedules.

I also want to express gratitude to my graduate advisors, Professor Guoyong and Professor Xiaoling Zhong(Chengdu University of Technology), for supporting my overseas studies in Japan and writing letters of recommendation for my doctoral program.

Furthermore, I extend my thanks to two good friends, Dr. Lingi Kong (University of Chinese Academy of Sciences) and HeDi Qu (Shenzhen Smartchip Company, graduated from the University of Electronic Science and Technology of China), for their assistance with paper submissions and English proofreading.

I would like to acknowledge the Japanese language teachers at Akita Prefectural University for their kind assistance, which greatly contributed to improving my Japanese proficiency.

Special thanks go to my parents, without whose support I could not have pursued studies in Japan. Although there have been long periods without seeing them due to being overseas, I always carry their thoughts with me.

Lastly, I want to express my gratitude to Akita Prefectural University for providing a comfortable living and learning environment, as well as a platform that supports my innovative research.

# Chapter 1
# Introduction

## 1.1 Background of research

Image, as a crucial carrier of digitized information in the modern era, is one of the key sources for humans to understand and comprehend the world. Image resolution, defined as the number of pixels per unit area on an image, is a fundamental property that serves as a vital indicator in evaluating image quality. Generally, higher image resolution results in richer information transfer, leading to better visual perception. However, various factors such as imaging environment, bandwidth limitations, and characteristics of light-sensitive devices can lead to natural degradation of image quality. This often results in acquired images having lower resolutions than the original ones, impacting the accuracy of computer vision analysis. Thus, effective enhancement of spatial resolution is crucial for intelligent image processing.

Super-resolution reconstruction technology, a fundamental task in image processing, plays a significant role in the field of computer vision. This technology involves transforming a sequence of low-resolution images with pixel values into a higher-resolution image through specific image processing algorithms. Essentially, it enlarges a low-resolution image sequence with LMxLN pixels related or complementary to the same scene by a factor of L, creating a high-resolution image with LMxLN resolution, this process retains the original image structure while enriching details, increasing image size, and supplementing high-frequency components, resulting in an overall clearer image, Super-resolution reconstruction technology has proven useful in various fields such as security surveillance, facial recognition, and medical imaging.

In the domain of security surveillance[1], law enforcement often relies on monitoring images to identify suspect's facial features, identities, or vehicle license plates. However, challenging conditions like adverse weather, crowded environments, and high human traffic can lead to unclear images, affecting visual effectiveness. Super-resolution technology effectively aids law enforcement agencies in improving efficiency under such circumstances.

In facial recognition, issues arise when facial movement or incorrect lens focusing causes image blurring, rendering facial recognition algorithms ineffective. This situation hinders operations such as access control, mobile payments, and facial unlocking,

High-resolution images are essential to enhance the accuracy of the verification system[2].

In the medical field[3], super-resolution reconstruction technology assists doctors in treatment by reducing misdiagnosis caused by blurry images. This, in turn, improves the accuracy of disease diagnosis, aiding in the formulation of effective treatment plans and holding significant value in medical research.

However, current research on image super-resolution reconstruction technology faces three challenges:

**1. Image Reconstruction Model and Low-Resolution Data Set:** Image reconstruction involves restoring the original image by reversing the image degradation model. Often, due to a relatively scarce low-resolution dataset, researchers attempt to use prior information obtained during the image degradation process to restore the original image. However, many studies overlook external factors affecting the image degradation model, resulting in low-resolution image data with varying degradation quality. Some research relies on simple bicubic interpolation degradation types, which are only effective for specific bicubic degradation images[4]. To address this, reliable prior information about the image degradation model and the quality of low-resolution images needs to be studied to standardize the acquisition of optimal reconstruction results.

**2. Increasing Complexity with Magnification:** As magnification increases, the complexity of the super-resolution problem also rises, at higher magnification levels, restoring lost scene details becomes more intricate, often leading to the recovery of erroneous texture information[5]. This instability makes the reconstruction of images exceptionally challenging, especially when high-frequency noise exists in the original image, many existing generative models struggle to recover lost high-frequency details, resulting in significant pixel variations and a lack of image continuity.

**3. Application to Medical Imaging:** While some studies on image super-resolution have shown promising results in natural image datasets, the transferability of reconstruction models to the medical field remains a challenge. Research on medical images is relatively scarce due to the difficulty in obtaining specialized medical datasets, unlike standardized datasets for natural images, medical institutions hold sensitive information in medical images, making it impractical to release them publicly. Additionally, most natural image super-resolution algorithms use color images, while common medical images are predominantly black and white, human visual perception has a strong color recognition ability, making the identification of details in black-and-white images challenging for doctors. Therefore, employing efficient image su-

per-resolution reconstruction algorithms for high-definition medical imaging becomes crucial.

In conclusion, while image super-resolution reconstruction technology has shown success in various applications, addressing these challenges will contribute to its further advancement and application in diverse fields.

## 1.2 Existing Research

Current research on super-resolution reconstruction can be categorized into traditional super-resolution methods and deep learning-based super-resolution reconstruction, traditional super-resolution methods generally exhibit comparatively lower performance, but they have the advantage of faster processing speeds when it comes to magnification. On the other hand, approaches based on deep learning for super-resolution reconstruction present a novel perspective and represent the primary focus of researchers in the current stage[6]. In our study, we particularly emphasize deep learning-based super-resolution reconstruction.

The effectiveness of reconstruction in this category surpasses that of the former, although it requires the extraction of rich image feature maps, consequently, the training time for these networks is typically longer.

### 1.2.1 Traditional Super-Resolution Reconstruction Research

Traditional super-resolution reconstruction research is generally divided into interpolation methods, edge information statistical methods, and frequency domain methods. Image interpolation methods mainly involve calculating the value of an interpolated pixel by considering the values of surrounding pixels in its neighborhood to enhance the image resolution, the three most commonly used interpolation methods are Nearest-neighbor interpolation, Bilinear interpolation, and Bicubic interpolation. Nearest-neighbor interpolation is a simple method that assigns the value of the nearest pixel in the original image to the interpolated pixel, while computationally simple and fast, this method can lead to significant distortion, blurring, and blocky artifacts in certain cases, resulting in suboptimal interpolation. Bilinear interpolation utilizes linear interpolation in both horizontal and vertical directions among the four nearest pixels to obtain the value of the interpolated pixel, this method produces smoother images compared to nearest-neighbor interpolation but still exhibits some artifacts. Bicubic interpolation, more complex than the previous two methods, uses the values of 16 surrounding pixels to perform cubic interpolation. Although computationally more de-

manding, bicubic interpolation yields better results and is more widely applicable. Li and Orchard[13] proposed a directional interpolation algorithm tailored to the edges of natural images, emphasizing the adaptive edge orientation property based on covariance, allowing adjustment of interpolation coefficients to match any direction of step edges. In literature [8], various interpolation techniques were employed to reconstruct high-resolution MR brain images from low-resolution MR brain images acquired in axial, sagittal, and coronal directions.

Edge information statistical methods: Example-based methods involve learning the mapping relationship from low resolution to high resolution by using pairs of high and low-resolution images in a training dataset, once the learning process is complete, similar mappings can be applied to new low-resolution images. Stevenson et al[8]. proposed using this algorithm for image super-resolution reconstruction in 1996. This method transformed super-resolution reconstruction into a parameter estimation problem, analyzing spatial and temporal information in short image sequences for reconstruction, optimizing noise and improving image quality.

Frequency domain-based reconstruction methods involve transforming low-resolution images into the frequency domain using Fourier transform, adding missing high-frequency information during this process to enhance resolution. Finally, the image is restored to the spatial domain through inverse Fourier transform. Huang[9] and others in 1980 initially proposed a frequency domain method for motion images, the principle involves performing Fourier transform on the original image, obtaining the HR reconstruction image spectrum by blending the spectra of multiple LR original images, and finally reconstructing the HR image by inverse discrete Fourier transform. Subsequently, in 1984[10], a frequency domain condition-based super-resolution technique was introduced, combining relationships between multiple low-resolution images and utilizing the properties of Fourier transform to derive a complex mapping relationship between two images, ultimately reconstructing the image.

### 1.2.2 Research on Deep Learning Super-Resolution Reconstruction

The traditional methods mentioned above model noise and blur using mathematical formulas, solving the inverse process of these formulas to obtain clear images. In recent years, breakthroughs in super-resolution reconstruction technology have been achieved in computer vision through the use of deep learning techniques, such as Generative Adversarial Networks (GANs) [11] and Variational Autoencoders (VAEs)

[12], owing to their powerful representation capabilities and nonlinear modeling abilities.

Deep neural networks, known for their effective extraction of local image features and efficient processing of high-dimensional images, have found widespread applications in computer vision and image processing. Convolutional neural networks (CNNs) can effectively extract high-frequency features from low-resolution images, enabling end-to-end mapping from low to high resolution, addressing challenges faced by traditional methods in accurately extracting and mapping features, deep learning super-resolution techniques are mainly divided into two research directions: constructing mappings and learning image features through convolutional neural networks, and generating transformations based on image conditions[14], the super-resolution reconstruction technique studied in this paper falls into the latter category.

Dong et al[15]. introduced an innovative approach for deep convolutional super-resolution image reconstruction in their research, the method utilizes a three-layer convolutional neural network designed for synthesizing super-resolution images through a feedforward Super-Resolution Convolutional Neural Network (SRCNN), the input low-resolution image is upsampled to the target image size using bicubic interpolation and subsequently passes through three convolutional layers: feature extraction layer, nonlinear mapping layer, and reconstruction layer. Training the three-layer convolutional network allows it to learn complex mapping relationships between images, producing the final reconstructed image as output. While this method uses SRCNN to extract crucial features from low-resolution images and enhances resolution through bicubic interpolation, the authors acknowledge its high computational complexity and slow convergence speed. Subsequent literature[16] improved upon this by incorporating deconvolution for upsampling operations, allowing the network to directly input low-resolution images without the need for interpolation enlargement. This modification uses smaller convolutional kernels to expedite network speed and introduces a local feedback mechanism to enhance model performance, aiding in better learning of image details. Kim et al[17]. constructed a deeper convolutional neural network for feature extraction, incorporating an interpolation method in the preprocessing stage, residual learning was also employed to mitigate the slow convergence and gradient vanishing issues. However, the increase in the number of network layers results in a higher parameter count and model complexity, leading to problems such as overfitting during training. For medical CT images, Styner et al[18]. proposed a sparse coding super-resolution method, significantly improving the visual quality and objective evaluation metrics of CT images using sparse representation techniques, ef-

fectively recovering high-resolution images from low-dimensional image representations.

In recent years, with the rapid development of Generative Adversarial Networks (GANs) composed of generators and discriminators, some researchers have approached super-resolution reconstruction as an image-to-image translation problem, this perspective has led to new developments in image super-resolution as a branch of image restoration technology, resulting in many classic works based on Generative Adversarial Networks (GANs).

Ledig et al[19]. first proposed a Super-Resolution Generative Adversarial Network (SRGAN), marking a groundbreaking application of GANs in SR technology, this method effectively handles the phenomenon of image edge smoothing under larger magnification factors, resulting in richer high-frequency details. Although the numerical performance of this method in terms of objective evaluation metrics, such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM), may not be satisfactory, the visually improved quality of the reconstructed high-resolution images is evident. Wang et al[20]. (ESRGAN) enhanced the Super-Resolution Generative Adversarial Network (SRGAN) by introducing a more advanced architecture, they replaced the residual blocks in the original generator with residual dense blocks after each convolutional layer, and further improved the quality of the reconstructed images by using a relative discriminator. However, GAN-based super-resolution image reconstruction methods have stringent Nash equilibrium conditions, requiring precise adjustment of parameters and learning rates during image synthesis[21]. They also suffer from problems such as gradient explosions, mode collapse, and the generation of artifacts. While generating networks can smooth out generated images, they may fail to preserve fine details, occasionally resulting in artifacts. Based on the current research status, we explored the latest denoising diffusion generative models.

## 1.3 Research Objectives

The primary focus of this paper is on image processing based on the improved denoising diffusion probability model, with a particular emphasis on its application in super-resolution reconstruction techniques. Drawing upon a wealth of literature on generative model algorithms and super-resolution reconstruction algorithms, which are well-established for natural images, several limitations of existing research:

1. Previous studies primarily target specific types of image degradation, whereas real-world scenarios involve multiple degradation processes, this research delves into the inverse process of super-resolution, studying the degradation and deterioration of im-

ages caused by blur, downsampling, and noise. Unlike previous studies that often rely on simplistic degradation types obtained through bicubic interpolation, this study adopts a more realistic approach by considering a variety of degradation factors and employs a strategy of randomly mixing degradation to acquire the required low-resolution dataset.

2. Recent advancements in denoising diffusion probability networks (DDPM) have shown promising capabilities in generating realistic texture features in images, while DDPM has outperformed traditional generative adversarial networks in various synthesis tasks such as image synthesis, translation, restoration, and colorization.

3. The original DDPM model has limitations in handling high-resolution images due to its large model parameter size, high GPU memory consumption, and long inference times. Moreover, its application in medical image research remains underdeveloped.

Based on the research on diffusion probability generative models and diffusion probability super-resolution models, this paper proposes an improved denoising diffusion super-resolution reconstruction model by introducing latent variables as an intermediate bridge during the reconstruction process. During the forward diffusion process, similar to the original DDPM model, the input high-resolution image is transformed into latent variables Z via a feature decoder. These latent variables Z undergo T iterations of Gaussian noise injection, resulting in a Gaussian noise distribution. Due to the necessity of constraining the solution space of high-resolution (HR) images in super-resolution reconstruction, during the forward process, the low-resolution image (LR) and the noise image of the current HR latent variables are stacked together for conditional sampling. Subsequently, the powerful parameter fitting capability of the UNet within the diffusion denoising probability model is utilized to fit the conditional feature update model. In the backward derivation process, the model combines the low-resolution image as a guiding condition with random Gaussian noise, refining the noise through iterative inverse processes. The random Gaussian noise is gradually transformed into a distribution similar to that of the high-resolution image latent variable data distribution. Finally, the latent feature variables are elevated to a distribution similar to that of the original generated image using the decoding module of VAE/VQVAE, achieving the reconstruction of high-resolution images, this results in more natural-looking generated images.

By improving the noise schedule timetable and introducing cosine time step control during the iteration process, the noise distribution becomes more uniform, ensuring more stable diffusion. The improved latent variable autoencoder DDPM, compared to the original model, features a reduced parameter count, saving GPU memory and reducing inference time. It addresses issues such as the single degradation of low-

resolution dataset images and problems encountered in traditional generative adversarial networks (GAN), such as gradient explosion and mode collapse during synthesis. Additionally, the generated images from the improved model exhibit greater naturalness, and the high-resolution images reconstructed by VQVAE-DDPM are clearer. Furthermore, the technology is applied to the medical imaging domain for studying multi-class texture pathology images of colorectal cancer, enriching the availability of high-quality, high-resolution medical images, it addresses common issues in medical image datasets, such as small sample sizes, unclear textures, and inconsistent sizes. Finally, the model is applied to other tasks, such as image colorization, demonstrating its advanced performance in image coloring tasks through comparative experiments. Therefore, the enhanced denoising diffusion probability model holds significant academic value and research significance in fields such as facial recognition, medical image super-resolution reconstruction, dataset augmentation, and image colorization.

# Chapter 2
# Prior research

## 2.1 Image Quality Degradation

In the process of single-image super-resolution reconstruction, obtaining the true corresponding low-resolution data is challenging due to the uncertainty in its collection. Therefore, in deep learning, to ensure consistency in learning, it is necessary to establish a one-to-one correspondence between low-resolution image data and its high-resolution counterparts. The traditional approach often involves downsampling the high-resolution images to low resolution, defining this process as an image quality degradation process. Subsequently, by utilizing the mapping relationship between high-resolution and low-resolution images, the corresponding high-resolution images are reconstructed. Hence, the quality of image degradation directly affects the quality of the low-resolution dataset. Based on literature review, we believe that in the actual image sampling process, blur, downsampling, and noise are three key factors leading to real image degradation. Therefore, from a mathematical perspective, we define the image quality degradation model as the process of downsampling a high-resolution image to a low-resolution image, expressed as follows:

$$y = (x \otimes k) \downarrow_s + n \tag{2.1}$$

From the above formula, it can be inferred that the low-resolution (LR) image is obtained by convolving the high-resolution (HR) image with an isotropic/anisotropic Gaussian kernel (or point spread function) $k$ to produce a blurred image $x \otimes k$. Subsequently, downsampling operation $\downarrow$s is applied with a scaling factor s, and white Gaussian noise $N$ with a standard deviation of $\sigma$ is added.

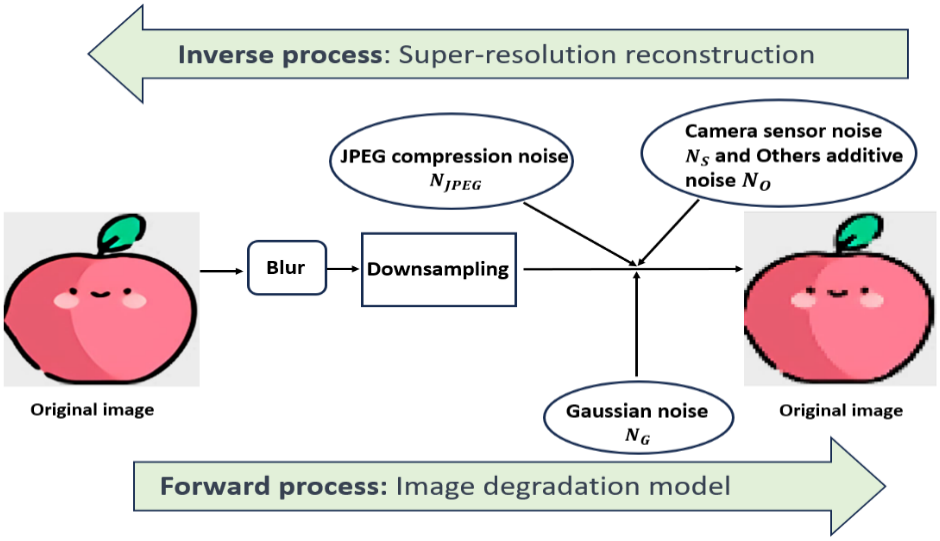


Figure 2.1: Image degradation and super-resolution reconstruction process

This allows us to approximate the high-resolution image after the aforementioned real degradation as a low-resolution image. **Figure 2.1** illustrates a schematic diagram of the image degradation and reconstruction process. From the diagram, it can be observed that the image degradation and reconstruction processes are mutually inverse, the objective of super-resolution reconstruction is to determine the inverse process $f^{-1}$ of the degradation[22].

Specifically, blur is achieved through two convolutions with isotropic and anisotropic Gaussian kernels, as illustrated in **Figure 2.2** common downsampling methods include bilinear and bicubic interpolation, as shown in **Figure 2.3** There are various types of noise, mainly categorized into Gaussian white noise $N_G$ at different noise levels and JPEG compression noise $N_{JPEG}$, as shown in **Figure 2.4**. Additionally, there are additive noises such as camera sensor noise $N_S$, (This camera sensor noise is simulated by the reverse-forward image signal processing (ISP) [23]pipeline model and RAW image noise model, because there are too many factors influencing image quality degradation in the real world, it is not feasible to exhaustively enumerate all possible combinations. Additionally, this paper primarily focuses on super-resolution models rather than image degradation models. Therefore, we simplify our study by investigating only three degradation factors and adopting the degradation combination strategy from the referenced paper[24], this research employs a strategy of randomly shuffling the aforementioned degradation factors, aiming to cover the degradation space of real images for synthesizing low-resolution (LR) images, for experimental

simplicity, we omit additive noises such as camera sensor noise from the original degradation model.

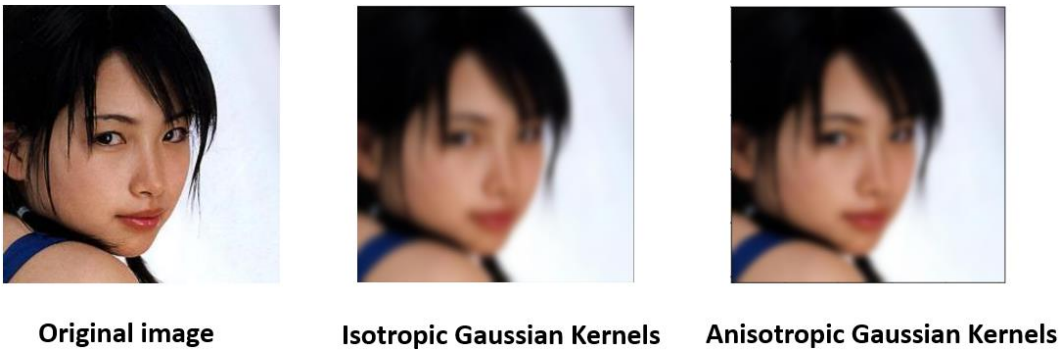


Figure 2.2: The results of blurring with isotropic and anisotropic gaussian kernels


Figure 2.3: The results of different downsampling methods


Figure 2.4: The results of Gaussian white noise $N_G$ at different noise levels and JPEG compression noise $N_{JPEG}$.

## 2.2 Variable auto-encoder

### 2.2.1 Introduction to Variational Autoencoder Model

The autoencoder (AE) [25] is a type of neural network model designed to learn or encode low-dimensional representations of high-dimensional data, such as images or text, autoencoder models find widespread applications in areas like dimensionality reduction and feature extraction. The structure of an autoencoder includes an encoder and a decoder: the encoder takes the input image, learns its latent features, and the decoder reconstructs the image from those features, by constraining the output image to be consistent with the input image, the autoencoder model achieves compression of information, with the dimensionality of the latent space being smaller than that of the input image. However, this model does not model the distribution of variables in the latent space and cannot generate new samples by sampling in the latent space, making it not a generative model. In 2013, Kingma and Welling extended the autoencoder and introduced the Variational Autoencoder (VAE) model[12], the introduction of the VAE model propelled the development of generative models, the VAE uses variational inference, training an autoencoder with a regularized latent variable space to compress high-dimensional data into a latent variable space representation vector.

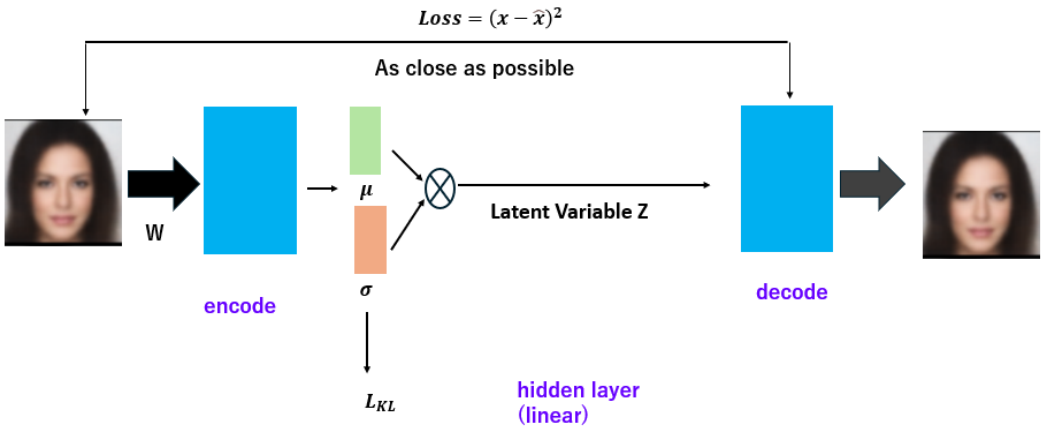


Figure 2.5: Variational autoencoder encoder-decoder framework

The VAE model is structurally similar to the autoencoder, as shown in **Figure 2.5**, and primarily consists of an encoder generative network $p_\theta(x|z)$, and an approximate inference network or decoder $q_\emptyset(x|z)$. The encoder transforms images into the latent space representation, and the decoder transforms the representation in the latent space back into the image space. During training, the latent space is subjected to KL

divergence constraints to make the latent space variables follow a Gaussian distribution. In the generation phase, new images can be generated by first sampling from the Gaussian distribution in the latent space and then inputting the sampled latent variables into the decoder.

The variational autoencoder has demonstrated its effectiveness in generating various complex data, including handwritten digits, facial images, house numbers, CIFAR images, and physical model segmentation of scenes. However, the network also has limitations, due to the limited expressive capacity of the inference model, the noise introduced during sampling, and the use of a flawed pixel-level loss function (such as the Mean Squared Error (MSE) loss function), the generated images tend to be relatively blurry.

### 2.2.2 Variational Inference in the Variational Autoencoder Model

The Variational Autoencoder introduces the variational inference algorithm, where the main idea is to transform inference into an optimization problem, the specific approach involves sampling from an approximating distribution $q^*(z)$ derived from a tractable Gaussian distribution that can often be probabilistically decomposed, this is achieved by maximizing the variational lower bound of the log-likelihood function, aiming to approximate the true posterior distribution $p(z|x)$, this is mathematically expressed as:

$$q^*(z) = \underbrace{argmin}_{q(z)\epsilon Q} KL(q(z)\|p(z|x)) \tag{2.2}$$

where $KL$ denotes the Kullback-Leibler divergence (or divergence), defined as:

$$KL\big(q(z)\|p(z|x)\big) = \int q(z)\log\frac{q(z)}{p(z|x)}dz \tag{2.3}$$

Here, $q(z)$ represents the distribution of the latent variable z, and $p(z|x)$ follows the normal distribution $N(0,1)$. In information theory, the KL divergence function is used to measure the information difference contained in two distributions, the objective is to minimize the $KL$ distance between the approximating distribution and the true distribution $p(z|x)$. Therefore, we define the distribution generated by the VAE model as $L_{VAE}\big(\phi,\theta;x^i\big)$ according to (2.2).

$$L_{VAE}(\phi,\theta;x^i)=\underbrace{argmin}_{q(z)\epsilon Q}\left(E_{z\sim q_x}(logq_x(z)-E_{z\sim q_x}(log\frac{p(x|z)p(z)}{p(x)}\right)$$

$$=\underbrace{argmin}_{q(z)\epsilon Q}\big(E_{z\sim q_x}(logq_x(z)-E_{z\sim q_x}(logp(z))-E_{z\sim q_x}(logp(x|z))+$$
$$E_{z\sim q_x}(logp(x))\big) \quad (2.4)$$

The term $E_{z\sim q_x}(logp(x))$ is a constant.

$$L_{VAE}(\phi,\theta;x^i)=\underbrace{argmin}_{q(z)\epsilon Q}\big(E_{z\sim q_x}(logp(x|z))-D_{KL}(q_x(z|x^i)\|p_\theta(z|x^i))\big)$$

(2.5)

On the other hand, the variational autoencoder belongs to the category of typical latent variable generative models $p(x|z)$, Its objective function is equivalent to maximizing the Evidence Lower Bound (ELBO) of the log-likelihood of the data, given the prior $P(\omega)$, the data distribution $P(x)$ can be obtained through integration:

$$P(x)=\int p(x|z)P(\omega)d\omega \quad (2.6)$$

During inference, the posterior distribution $\log p(x)$ is obtained through Bayes' rule:

$$p(z|x)=\frac{p(X|Z)P(\omega)}{\int p(X|Z)P(\omega)d\omega} \quad (2.7)$$

Using maximum log-likelihood estimation involves optimizing the log-likelihood function with respect to the dataset $X$. Thus, the log-likelihood function for the entire dataset is obtained by summing up the log-likelihood functions for each corresponding sample:

$$\log p_\theta\,(x^1,x^2,\cdots x^n)=\sum_{i=1}^{N}\log p_\theta\,(x^{(i)}) \quad (2.8)$$

For an individual sample data:

$$\log p_\theta\,(x^i)=L(\phi,\theta;x^i)+D_{KL}(q_x(z|x^i)\|p_\theta(z|x^i)) \quad (2.9)$$

Where the first term on the right side is the Evidence Lower Bound (ELBO) [26] of the log-likelihood function, and the second term is the Kullback-Leibler (KL) [27] divergence between the approximate posterior probability and the true posterior probability, this term is non-negative:

$$\log p_\theta\left(x^i\right) \geq L\left(\phi,\theta;x^i\right) = E_{z\sim q_x}\left(logp(x|z)\right) - E_{z\sim q_x}\left(logq(x|z)\right) \quad (2.10)$$

$$L\left(\phi,\theta;x^i\right) = E_{z\sim q_x}\left(logp(x|z)\right) - D_{KL}\left(q_x\left(z|x^i\right)\|p_\theta\left(z|x^i\right)\right) \quad (2.11)$$

The final objective formula for VAE using standard gradient optimization and reparameterization techniques is:

$$L_{VAE}\left(\phi,\theta;x^i\right) = arg\min_{\theta,\phi}\left[|\mu_\theta(z)-x|^2 + 0.5\sum_{j=1}^{\dim(z)}{\sigma_\phi}^2(x)_j + {\mu_\phi}^2(x)_j - log{\sigma_\phi}^2(x)_j\right] \quad (2.12)$$

## 2.3 Vector quantization auto-encoder

Vector Quantization (VQ) is a method for signal compression, the basic idea is to organize several scalar data into a vector and then quantize the entire vector space as a whole, this approach enables data compression without losing important information. In practice, VQ exhibits high compression rates and good visual quality in image processing. While Variational Autoencoders (VAE) are constrained by the assumption of Gaussian distribution modeling, real-world data distributions may be more complex. Therefore, inspired by Vector Quantization, DeepMind[28] introduced a novel discrete variational autoencoder generative model called VQ-VAE, building upon the foundation of VAE. VQ-VAE introduces the concept of vector quantization by discretizing the continuous vectors in the latent space into discrete encoding vectors, this not only shares similarities with continuous latent variable VAE models but also offers the flexibility of a discrete distribution.

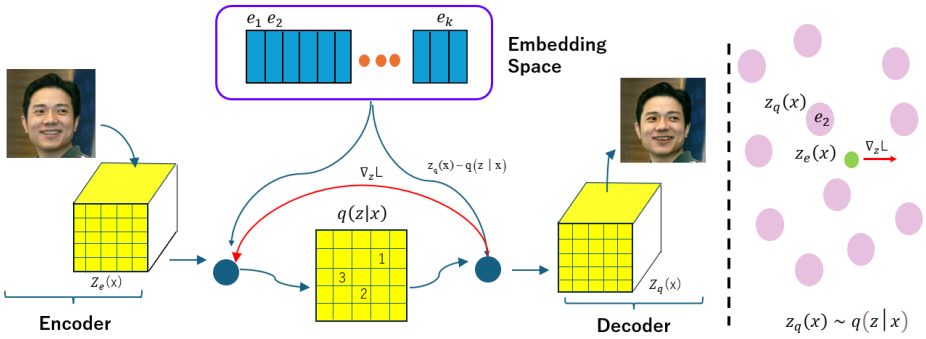


Figure 2.6: Vector quantization variational autoencoder encoder-decoder framework

The structure of VQ-VAE is still based on the encoder-decoder framework, as shown in **Figure 2.6**, the original image $X$ passes through the convolutional layers in the encoder to obtain a continuous encoding vector $Z_e(X)$ with a size of $L \times W \times D$, between the encoder and the decoder, there is a component called the VQ layer, which is responsible for discretizing the latent variables, the VQ layer includes an embedding dictionary module, where the embedding dictionary re-encodes the continuous latent variables output by the encoder. It quantizes the continuous vector in the latent space of the encoder's output by mapping it to the nearest embedding in the embedding dictionary. These encoding vectors are often learned atomic vectors (e.g.cluster centers) from the training data, each denoted as $e_i$ which is a vector of size D. Subsequently, VQ-VAE performs a nearest-neighbor search to map $Z_e$ to one of these $X$ vectors:

$$Z \to e_x,\ k = \arg min \left\| Z - e_j \right\|_2 \tag{2.13}$$

The embedding corresponding to $Z_e$ is denoted as $e_{emb}$ (final encoding result), and these embeddings $Z_q$ are then fed into the decoder network for decoding, By combining information from VAE with discretized latent variables, VQ-VAE demonstrates powerful potential in neural network learning. The advantage of VQ-VAE lies in the fact that while the encoder's latent variables in VAE are continuous. The latter outputs discrete variables, by enforcing the discretization of the encoding in the latent space, it becomes a representative vector from a finite set, achieving vector quantization, this helps reduce the dimensionality of the latent space, improve training efficiency, and prevent posterior collapse.

## 2.4 Generative Adversarial Network (GAN) Model

### 2.4.1 Introduction to Generative Adversarial Network (GAN) Model

In the realm of image generation models, a classic work is the Generative Adversarial Network (GAN), first introduced by Ian Goodfellow in 2014, this deep learning generative model immediately drew widespread attention in the academic community. GANs are capable of generating data in the form of one-dimensional signals, two-dimensional matrices, or three-dimensional images, with a focus on image data in this example.

A Generative Adversarial Network is composed of a generative model and a discriminative model, as shown in **Figure 2.7**. the generative model learns the distribu-

tion of real data, while the discriminative model is a binary classifier responsible for distinguishing whether the input is real or generated data, Both real and generated data are fed into the discriminative model $D$, which outputs the corresponding classification. In the original GAN framework, the generator network employs a U-Net architecture consisting of an encoder and a decoder. This structure comprises downsampling convolutional modules, a bridging module, and upsampling convolutional modules. The $D$ is discriminative model, being a binary classifier, is trained using binary cross-entropy loss to control the content generated by the network, for real images, the given label is 1, while for generated images, the given label is 0. The generative model $G$ attempts to synthesize images in a way that the discriminative model $D$ is inclined to classify them as genuine. Assuming $x$ represents real data following the distribution $P_r(x)$, $P_g$ denotes the data distribution of generated images $G(z)$, and $P_z$ represents the prior distribution of random noise vector $z$ with $N(0,1)$. The generator network and the discriminator network are denoted as $G$ and $D$, respectively, where $D$ can be regarded as a binary classifier. Using the cross-entropy loss function, the optimization objective of the Generative Adversarial Network (GAN) can be expressed as follows:

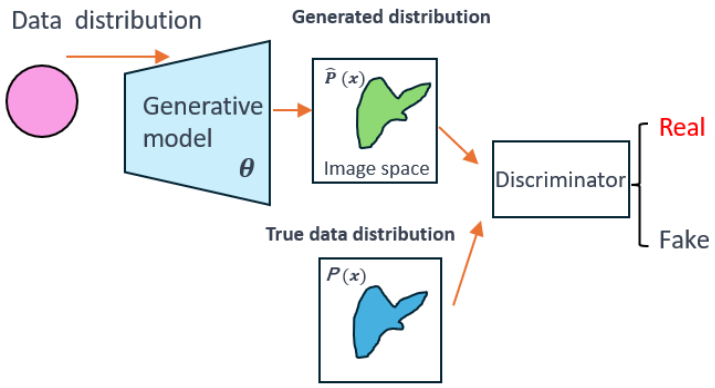


Figure 2.7: Generative adversarial network framework

$$\min_G \max_D = E_{x\sim P_r}[logD(x)] + E_{x\sim P_z}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad (2.14)$$

The Generative Adversarial Network operates through a mechanism akin to a Max-Min game, alternately optimizing the generator network $G$ and the discriminator network $D$ until they reach a Nash equilibrium point. As the alternating optimization proceeds, the discriminator network $D$ gradually approaches the optimal discriminator.

When this approximation reaches a certain level, the optimization objective of the Generative Adversarial Network is approximately equivalent to minimizing the Jensen-Shannon Divergence (JS divergence) [29] between the data distributions of real and generated images. In brief, it is equivalent to optimizing the distribution distance between real and generated data.

### 2.4.2 Some drawbacks of Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) can produce realistic high-quality images, but they may face issues such as mode collapse and convergence difficulties. This is because the natural data distribution is highly complex and multimodal, meaning there are many "peaks" or "modes" in the data distribution, each mode represents similar data samples distinct from other modes, mode collapse occurs when the generated samples lack diversity, and the generator believes it can deceive the discriminator by focusing on a single mode. In other words, the generator only produces samples from that specific mode, the discriminator eventually identifies the samples from this mode as fake, prompting the generator to switch to another mode, this cycle repeats, fundamentally limiting the diversity of samples generated by the GAN.

Moreover, GANs face challenges in training and slow convergence, GAN training is relatively difficult because the balance between the two networks is a continuous competition in a high-dimensional parameter space to achieve a Nash equilibrium. Finding a Nash equilibrium in this continuous high-dimensional space is challenging, Gradient descent algorithms are commonly used in current research to optimize the GAN's objective function, aiming to minimize the loss function rather than engaging in the competitive game of finding a true Nash equilibrium in simulated space. The objective function is non-convex, and achieving a true Nash equilibrium in the continuous high-dimensional parameter space is challenging, in practical GAN training, the limited fitting capacity of the generator and discriminator, coupled with the inability to guarantee optimal optimization during iterative training, prevents GANs from reaching a true Nash equilibrium state, this can lead to problems such as vanishing or exploding gradients and oscillations, resulting in slow or non-convergent model training.

Furthermore, GANs lack explicit inference capabilities and cannot directly extract the probability density of the data from the model, this might necessitate the use of other generative models, such as Variational Autoencoders (VAEs), to achieve more intuitive probability density estimation.

## 2.5 Denoising diffusion probability model

### 2.5.1 Introduction to Denoising diffusion probability model

In 2015, Jascha Sohl-Dickstein, Eric, and others proposed the Probability Diffusion Model (DPM)[30], abbreviated as Diffusion Model (DM), the inspiration for this model came from non-equilibrium statistical physics. Similar to other generative models, the diffusion model synthesizes images by learning the distribution of images in the training dataset, the forward process involves gradually adding noise to a normal image, analogous to dropping ink into a cup of water. Over time, the ink diffuses throughout, eventually uniformly distributing in the solution, resulting in a murky water cup, this diffusion process is one reason why the model is called a diffusion model.

If the positions, movement speeds, and directional properties of pigment molecules are recorded during the diffusion process, it becomes possible to infer the dropping positions of pigments in a cup of dissolved water. In 2020, Jonathan Ho and others [31] proposed improvements to the original diffusion probability model's mathematical calculation methods and introduced the concept of Denoising Diffusion Probability Model (DDPM), this model demonstrated improved performance in image synthesis by providing a complete strategy for adding noise and denoising, and it was applied in the field of image synthesis.

DDPM comprises a forward noisy diffusion Markov process and a backward denoising Markov process. By learning the Markov chain, the model ensures that real images, after $T$ steps of noise addition, conform to a Gaussian noise distribution. Conversely, the backward denoising process optimizes the distance between generated data and the real data distribution through the learning of a U-net network. By iteratively denoising for $T$ steps, useful samples resembling the real data distribution can be generated.

The detailed derivation process of the denoising diffusion probability model is explained in Chapter 3 of this paper. Subsequent research based on DDPM has propelled its development, for instance, Guided Diffusion introduced a classifier to guide the DDPM sampling process, achieving good results in unconditional image generation and category-based image generation. Classifier-Free Diffusion ingeniously proposed how to generate high-quality diverse images without utilizing a classifier. Further improvements, such as the application of denoising diffusion models in image super-resolution reconstruction models, have been explored, the fourth chapter of this paper builds on this research, presenting an improved denoising super-resolution diffusion

model that outperforms previous simple denoising diffusion models in terms of performance and applicability.

Research[32] indicates that DDPM has surpassed traditional Generative Adversarial Networks (GANs) in the field of synthesis, numerous studies[33-35] demonstrate the outstanding performance of denoising diffusion probability models in unconditional and conditional image generation, serving as a promising approach in tasks such as image synthesis, translation, restoration, coloring, composition, and speech synthesis. Therefore, as a novel image generation method, DDPM has shown remarkable success in the field of computer vision.

### 2.5.2 Introduction to Some Application Areas of Denoising Diffusion Probability Model (DDPM)

**Figure 2.8** shows an award-winning artwork named "Théâtre D'opéra Spatial" (French for "Spatial Opera Theater"), created by the generative AI platform Midjourney (based on the denoising diffusion probability model) under the guidance of Jason Michael Allen[36], this image won the 2022 Annual Art Competition at the Colorado State Fair, using DDPM. The artwork breaks the dimensions of traditional image construction, describing a Renaissance-era space opera theater by adding keywords like "grand" and "luxurious", the painting is a combination of steampunk and surrealism.


Figure 2.8: Based on the denoising diffusion probability model, the award-winning work synthesizes a space opera theater.

**Figure 2.9** represents an experimental application case conducted beyond the scope of this paper, it involves using the denoising diffusion probability network to generate

cartoon images of real faces based on prompts. The model employs techniques such as connecting text and images and conditional diffusion, the left side shows a real person's image, and the right side displays the final generated cartoon image. Due to the complexity of the network, which involves text embedding, detailed explanations are not provided here.

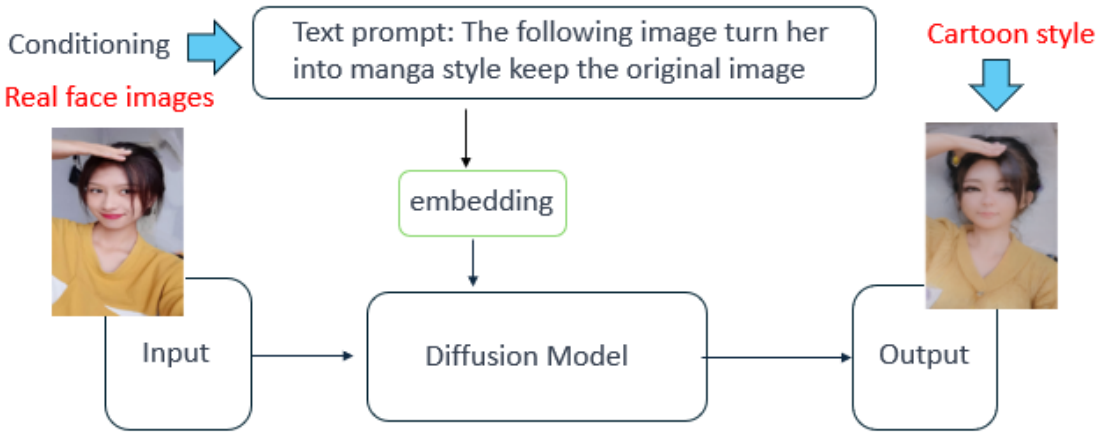


Figure 2.9: Connecting text and images" and "diffusion technology" are applied to generate cartoon images based on prompt words

# Chapter 3

# DDPM Generation Process and DDPM Super-Resolution Principles

## 3.1 DDPM Generation Process

DDPM (Denosing Diffusion Probabilistic Models) generation process consists of a forward process and a backward process, both of which can be viewed as parameterized Markov chains, the entire model training process is illustrated in **Figure 3.1**. In the forward process of the DDPM image, the initial image is first subjected to noise using a pre-defined noise scheduling timetable, according to the formula, the image after noise at a certain moment is determined, and the time T of that moment is recorded, the gradient of the U-net network is updated by calculating the loss function of the noise-added image and the noise. Subsequently, a series of noise iterations are performed to completely destroy the image, once the model training is complete, the estimated model $\varepsilon_\theta(x_t, \mathrm{t})$ is obtained.

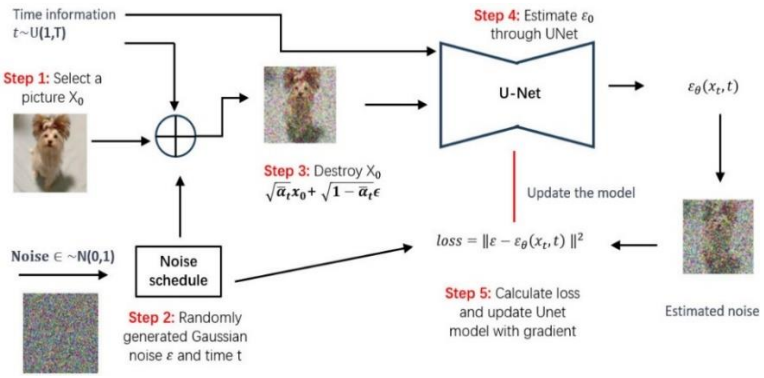


Figure 3.1: DDPM generation process

Then, utilizing the reverse denoising formula derived in Section 3.2, the noise vector $X_T$ is gradually denoised and recovered until a high-quality output image $X_0$ is obtained, DDPM transforms the standard normal distribution into an empirical data distribution (similar to Langevin dynamics) through a series of refining steps. This al-

lows for simple neural network parameterization regularization, reducing model collapse, and retaining more details in the image.

### 3.1.1 DDPM Forward Process

The forward process of the denoising diffusion network, also known as the diffusion process, refers to the continuous addition of Gaussian-distributed noise to image data, transforming it into pure Gaussian noise. The purpose is to use a Markov chain to convert the complex distribution $q_{complex}$ formed by the original target data variable $X_0$ into a simple normal prior distribution $p_{prior}$ (Gaussian noise).

As shown in **Figure 3.3**, at each step, the added noise is derived from the noise added at the previous time step (this process is a Markov process). As T increases, $X_T$ gradually transforms into a Gaussian noise data distribution, and in T time steps, the image $X_0$ is transformed into Gaussian white noise $X_T$～N(0,1).

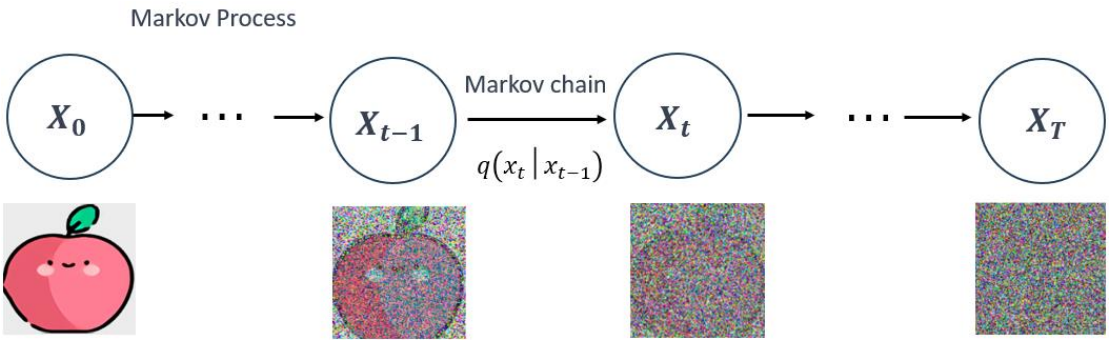


Figure 3.2: DDPM forward process

The variance of $q(x_t|x_{t-1})$ is defined to be independent of $x_{t-1}$ and is expressed as $\beta_t I$, the distribution $q(x_t|x_{t-1})$ is given by:

$$q(x_t|x_{t-1})=N(x_t;\sqrt{1-\beta_t}x_{t-1},\beta_t I) \tag{3.1}$$

Here, $N$ represents the normal distribution with a mean of $\sqrt{1-\beta_t}x_{t-1}$ and a variance of $\beta_t I$. Consequently, by sampling from the standard normal distribution $\epsilon_t$～$N(0,1)$, the relation is:

$$x_t=\sqrt{a_t}x_{t-1}+\sqrt{1-a_t}\epsilon_t \tag{3.2}$$

Based on the previous analysis, the entire diffusion process is a Markov process, the posterior probability distribution from input $X_0$ to $X_t$ can be expressed as:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{3.3}$$

Therefore, in the forward process, it is necessary to calculate each time step's $x_t$, similar to the VAE model, a recursive loop needs to be implemented, progressing from $x_0$ to $x_T$ step by step using Formula 3.3, the process requires repeated sampling T times, making the efficiency of the recursive loop calculation low. The original paper did not adopt a step-by-step forward noise addition calculation but utilized a computational trick to directly compute any $x_t$ from $x_0$, the specific derivation is as follows:

$$x_t = \sqrt{a_t}x_{t-1} + \sqrt{1-a_t}\epsilon_t \tag{3.4}$$

$$= \sqrt{a_t a_{t-1}}x_{t-2} + \sqrt{a_t - a_t a_{t-1}}\epsilon_{t-1}) + \sqrt{1-a_t}\epsilon_t$$

Adding two independent Gaussian distributions with zero mean

$$\sqrt{a_t a_{t-1}}x_{t-2} + \sqrt{\sqrt{a_t - a_t a_{t-1}}^2 + \sqrt{1-a_t}^2}\,\epsilon \tag{3.5}$$

Adding variances and replacing with a new Gaussian distribution:

$$\sqrt{a_t a_{t-1}}x_{t-2} + \sqrt{1 - a_t a_{t-1}}\epsilon$$

$$= \sqrt{\prod_{i=1}^{t} a_i}\, x_0 + \sqrt{1 - \prod_{i=1}^{t} a_i}\,\epsilon$$

$$= \sqrt{\bar{a}_t}x_0 + \sqrt{1-\bar{a}_t}\epsilon \ , \ \bar{a} = \prod_{i=1}^{t} a_i, \epsilon \sim N(0,1) \tag{3.6}$$

Therefore, the formula derived in the original paper is as follows:

$$x_t = \sqrt{\bar{a}_t}x_0 + \sqrt{1-\bar{a}_t}\epsilon \tag{3.7}$$

Where $a_t = 1 - \beta_t$, and as $\beta_t$ continuously increases, with the parameter set to be between 0.0001 and 0.002 in the original paper, the weight of noise influence becomes increasingly significant as the forward time steps progress. As $t$ approaches positive infinity, $x_t$ becomes equivalent to a Gaussian white noise distribution.

Through this formula, $x_t$ with added noise at any time can be calculated, in fact, due to the assumption that $q(x_{t-1}|x_t)$ is a linear Gaussian, it is possible to parallelize the computation of all $x_t$.

### 3.1.2 DDPM Reverse Process

The forward diffusion process is a noise generation process, and accordingly, the reverse process is a denoising process. As the denoising diffusion model is a standard variational inference generative model, the reverse process is also referred to as the inference process, in simple terms, it involves iteratively inferring and gradually restoring meaningful data, resembling the original data, from Gaussian noise. If we have the true data distribution $q(x_{t-1}|x_t)$ at each denoising step, the reverse iteration involves continuously sampling, denoising, and progressively reconstructing the Gaussian noise until a complete image $q(x_0)$ is obtained.

However, obtaining the distribution $q(x_{t-1}|x_t)$ directly is challenging because it requires the entire training dataset. Therefore, the denoising diffusion model employs the construction of a neural network parameterized by $\theta$ to approximate this distribution. Assuming a distribution $p_\theta(x_{t-1}|x_t)$ represents the distribution for the reverse generation process, and this distribution follows a Gaussian distribution, with its mean $\mu_\theta$ and variance $\sigma_\theta$ as parameters depending on $x_t$ and $t$, we have:

$$p_\theta(x_{t-1}|x_t) := N(x_t; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)) \tag{3.9}$$

In order to reduce the training complexity of the neural network and facilitate computation, during the training process, the variance $\sigma_\theta$ is typically set as a constant $\beta_t$ that does not require the involvement of the neural network and is time-dependent, only the neural network is used to train the mean $\mu_\theta$. Given the values of $x_t$ and $x_0$ at time $t$, the posterior probability $q(x_{t-1}|x_t)$ can be calculated, then, utilizing Bayes' theorem, the posterior distribution $p_\theta(x|z)$ is given by:

$$p_\theta(z|x) = \frac{p_\theta(x|z)p(z)}{p_\theta(x)} \tag{3.10}$$

Typically, we employ variational inference, as introduced in the previous section on variational autoencoders, to solve for the posterior distribution $p_\theta(z|x)$, combining the formula 3.2 from section 3.1.1 with Bayes' theorem[37], we get:

$$q(x_{t-1}|x_t, x_0) = q(x_t|x_{t-1}, x_0)\frac{q(x_{t-1}|x_0)}{q(x_t|, x_0)}$$

$$= N(x_{t-1}; \tilde{\mu_t}\ (x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\bar{z}_t), \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t I) \quad (3.11)$$

From the above equation, we can deduce that:

$$x_{t-1} = \frac{1}{\sqrt{a_t}}x_t - \frac{1-a_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t, t) + \sigma_t z \quad (3.12)$$

This formula represents the reverse inference process of DDPM, and it allows us to deduce $x_t$ from $x_{t-1}$. Here, Loss $= \|\varepsilon - \varepsilon_\theta(x_t, t)\|^2$, $\varepsilon_\theta(x_t, t)$ is the noise model estimated during the training of the DDPM model based on $x_t$ and $t$, $\theta$ represents the model training parameters, and $\sigma_t$ is Gaussian noise following a normal distribution $N(0,1)$, used to represent the error between the actual and predicted values. Finally, the complete image can be generated through a step-by-step reverse iteration process.

## 3.2 DDPM U-net model structure

The denoising diffusion probability model (DDPM) employs a U-Net network structure[38], as shown in **Figure 3.3**. Taking advantage of the "image-to-image" transformation capabilities of the encoder-decoder architecture, the DDPM diffusion model utilizes a U-shaped encoder and decoder structure to predict the noise model.

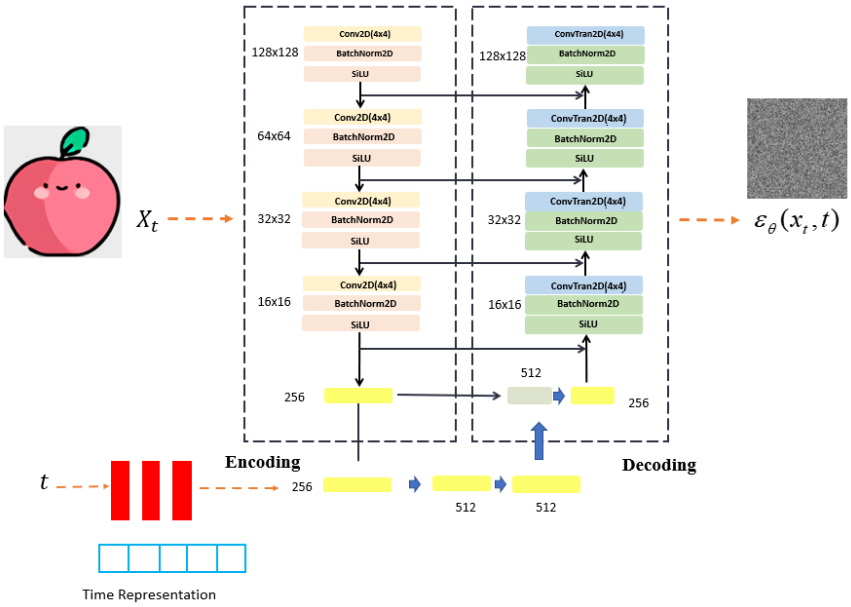


Figure 3.3: Denoising diffusion probabilistic model U-net model network structure

On the right side is the encoder, consisting of 4 downsampling residual blocks, each residual block includes two normalization layers, a SiLU activation function (SiLU is a symmetric activation function compared to ReLU, preserving information in the negative range), a 4x4 convolutional layer with a stride of 2, and an embedding layer, the input $X_t$ is a single-channel tensor of size 128×128, embedding is used to encode the time $t$, and the feature map is added to the time $t$ processed by the embedding layer, forming a residual block structure, the output is the predicted value $\varphi_z$ of the noise $z$, with the same number of output channels and size as the input. The decoder is similar to the encoder, consisting of 4 upsampling residual blocks, the decoder uses nearest-neighbor interpolation for upsampling, restoring the resolution of the feature map to the original image size. Skip connections are established between the middle parts of the encoder and decoder, concatenating corresponding feature maps to enrich features and enhance the details of image synthesis.

The U-Net network is regularized using the smooth L1 loss function to constrain latent information such as texture, shape, and style between feature variables and noise. In the denoising diffusion probability model, the primary role of the U-Net network is to combine image features and map them to the intermediate layers of the U-Net, through the structure of the encoder and decoder, it can learn more features of the original image after noise addition, simultaneously updating the learning parameters of the entire model.

## 3.3 DDPM Super-Resolution Reconstruction Model Principles

From the previous discussion, it's evident that DDPM has shown promising results in computer synthesis tasks. However, DDPM is based on unconditional or simple conditional model inputs. Therefore, there is a need to enhance the model to make it suitable for super-resolution reconstruction tasks. According to references from previous literature, the super-resolution reconstruction model can be simplistically viewed as a conditional denoising diffusion probability generative model [39], as illustrated in **Figure 3.4**. In simpler terms, it leverages the conditional Markov chain of conditional DDPM to transform latent variables from a Gaussian distribution based on conditions into a conditionally complex distribution, this is employed for the task of super-resolution image reconstruction.

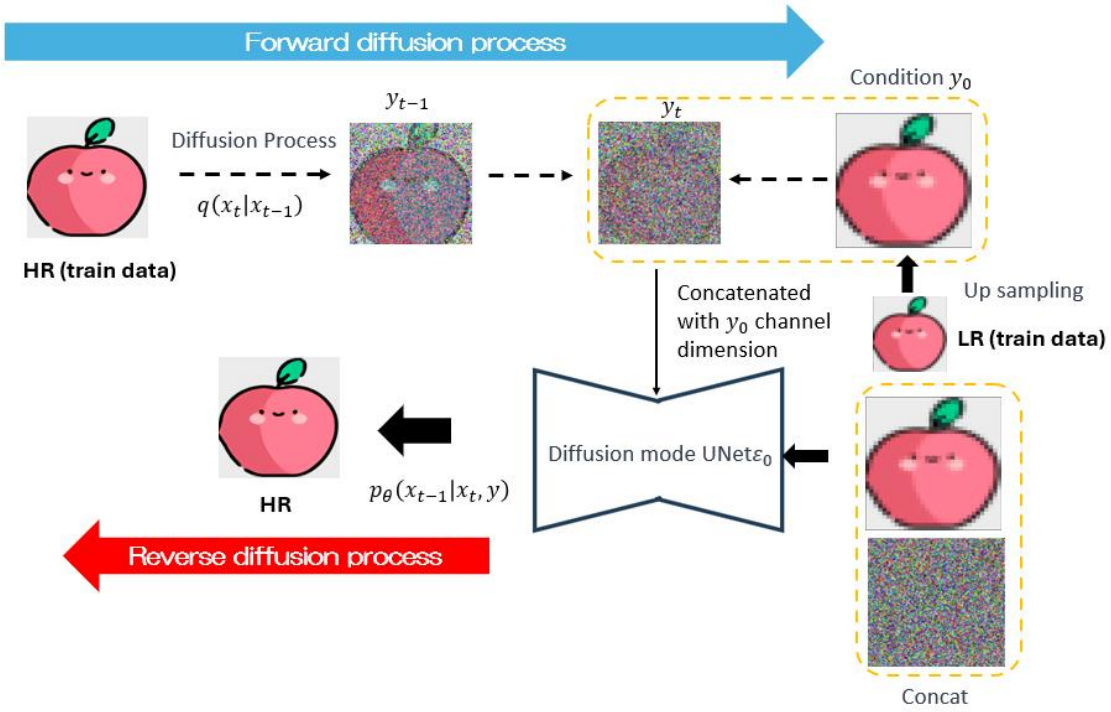


Figure 3.4: Denoising diffusion probability super-resolution reconstruction model network structure

In the denoising diffusion probability super-resolution reconstruction model, the condition for generation is set as a low-resolution image. Therefore, it is necessary to use the LR (Low-Resolution) image as the conditional input to constrain the solution space for HR (High-Resolution) images, the LR image is utilized as a conditional input to the function $\epsilon_\theta(\cdot)$ to control the synthesis of the HR image. In the forward process, the LR image is treated as a conditional dependency and is stacked and merged with the high-resolution image after adding noise at the current time step along the channel $y_0$[40], this combined input is then fed into the U-Net network model to predict the loss between the noise distribution.

During the forward process, conditional sampling is performed, and the U-Net network model learns based on the contextual details and semantic information related to the LR image, according to the assumptions of the reverse process model from the previous section, the conditional distribution can be expressed as $p_\theta(x_{t-1}|x_t, y)$, through the inverse diffusion process $p_\theta(x_{t-1}|x_t, y)$, the conditional distribution is learned without modifying the forward diffusion process $q(x_{1:T}|x_{t-1})$, this approach ensures that the sampled $x$ with $y$ as a condition has high fidelity, the process can be represented as:

$$p_\theta(x_{0:T}|y) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, y) \tag{3.13}$$

To learn the conditional distribution, the stacked images $y$ are used to constrain the neural network at all time steps $t$, the training objective is defined as follows:

$$L_{simple} = E_{t,x_0,e}[\epsilon - \epsilon_\theta(x_t, y, t)^2] \quad (3.14)$$

Once the conditional distribution is obtained, the conditional model can be used for super-resolution reconstruction inference. In the reverse inference process of super-resolution reconstruction, starting from random Gaussian noise, given a low-resolution image LR condition $y$, the model combines the low-resolution image as a guiding condition with random Gaussian noise, through an iteratively refined inverse process, the random Gaussian noise is gradually transformed into a distribution similar to the data distribution of high-resolution images. According to the derivation in Section 3.2, the formula for calculating the conditional distribution $x_{t-1}$ of $p_\theta(x_{t-1}|x_t, y)$ is given by:

$$x_{t-1} = \frac{1}{\sqrt{a_t}} x_t - \frac{1-a_t}{\sqrt{1-\bar{a}_t}} \varepsilon_\theta(x_t,\ y,t) + \sigma_t z \quad (3.15)$$

Finally, based on $x_{t-1}$ after $T$ steps of the denoising sampling of the Markov chain, the high-resolution image HR is inferred.

However, it is essential to note that the super-resolution reconstruction task of DDPM is different from the original DDPM generation task. In the super-resolution reconstruction task, DDPM's reverse diffusion process requires complex probability distributions to model the denoising distribution. Therefore, DDPM in the forward diffusion process requires thousands of evaluation steps to sample a feature. If DDPM uses a small number of sampling steps, it can result in issues such as low-quality generated high-resolution images.

# Chapter 4
# Improving the DDPM Super-Resolution Model

## 4.1 Improvement of the Noise Schedule Timetable in DDPM

According to the research in Chapter 3 on the forward process, it is known that as T increases, $X_T$ gradually becomes Gaussian noise data. In $T$ time steps, the image $X_0$ is transformed into Gaussian white noise $X_T \sim$N(0,1). The definition of the variance of the forward process $q(x_t|x_{t-1})$ is independent of $x_{t-1}$, with a mean of $\sqrt{1-\beta_t}x_{t-1}$ and a variance of $\beta_t I$ in a normal distribution, the addition of noise at each step should maintain a consistent noise diffusion amplitude as much as possible. In the early stages of the image distribution, adding some noise can alter the original distribution, however, as time progresses, more noise needs to be added, accelerating the diffusion. Therefore, it is essential to ensure a consistent noise diffusion amplitude, $\beta_t$ is a hyperparameter for the DDPM noise scheduling timetable, and $\beta_t$ needs to increase. Hence, a scheduling timetable is required to control the variation of $\beta_t$ from 0.001 to 0.02, as shown in **Figure 4.1**, which illustrates several noise scheduling timetables.

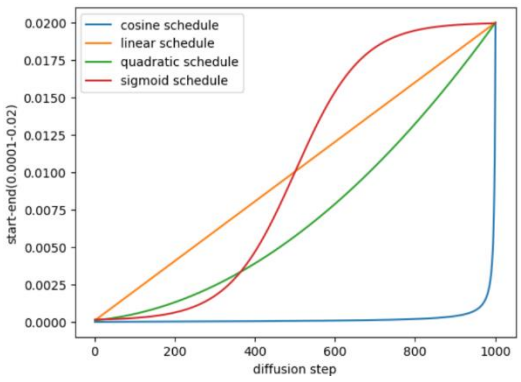


Figure 4.1 Different noise schedule timetables in DDPM

The linear noise schedule is a simple noise scheduling timetable that adds noise to the image in a linear manner, the increase in noise is uniform and linear, without considering specific structures or content in the image. The cosine noise schedule adds

noise to the image in the form of a cosine function, the goal of this strategy is to effectively reduce noise while preserving the basic semantic information in the image as much as possible, the shape of the cosine function helps smooth the noise addition process, reducing damage to the overall features of the image. The quadratic noise schedule is a noise scheduling method based on a quadratic function, by adding noise to the image along the curve of a quadratic function, finer control over the noise addition process is achieved, this method may adjust the intensity of noise based on different parts of the image. The sigmoid noise schedule adds noise to the image in the shape of a sigmoid function, the sigmoid function is an S-shaped curve, with the characteristic that the output approaches zero or one when the input is small or large. Applying this function introduces some non-linear changes to the noise increase in the image, creating a smooth transition from low to high values, this helps maintain certain subtle features of the image while adding noise.

The original DDPM model used a linear noise schedule, which led to excessive noise in the early stages, causing rapid and abrupt data diffusion, making the reverse restoration difficult. Additionally, since the data itself is close to random noise in the later stages, adding insufficient noise, equivalent to small changes, slows down diffusion, requiring more events for the chain length, this results in wasted steps in the diffusion or reverse diffusion process. Through the study of the DDPM, it is found that the model needs to add noise more slowly in the early stages and faster in the later stages. Therefore, adjusting the time step frequency information in the noise scheduling timetable is necessary to obtain a more efficient noise addition strategy. In this work, after repeated studies, the hyperparameter setting for the noise scheduling timetable will be changed to a cosine noise schedule, this simple modification helps to make the process of noise addition smoother, reducing the disruption to the overall features of the image. Consequently, noise diffuses better during the forward process, aiming to retain the fundamental semantic information in the image as much as possible, thereby improving the accuracy of synthesis.

## 4.2 Latent Variable Model

The modeling of deep generative models mainly relies on Latent Variable Models[41], where latent variables are unobserved but crucial variables in the model. A core problem in statistics and machine learning is to learn a complex probability distribution $p_\theta(x)$ with given observable high-dimensional sample points x, where $\theta$ represents the distribution parameters. However, directly modeling complex high-dimensional distributions is a challenging task. In addressing this issue, latent variable

models do not directly model $p_\theta(x)$, but introduce an unobservable or unmeasurable latent variable z and define a conditional distribution $p_\theta(x|z)$ for the data, often referred to as the likelihood, the latent variable z itself can be interpreted as a continuous random variable, the purpose of introducing these latent variables is to capture latent structures, patterns, or causal relationships in the data, enabling the model to have a deeper understanding of the data generation process.

For example, if one wants to learn the probability distribution of images of apples, it is necessary to define a distribution that can model the complex correlations between all pixels constituting each image, the latent variable z may include latent features such as the type, color, or shape of the apple. Furthermore, a prior distribution $p(z)$ can be introduced for the latent variable z, representing the model's understanding of the latent variables before learning observable data, based on the prior $p(z)$ and the likelihood function $p_\theta(x|z)$, the joint distribution of observable variable x and latent variable z can be defined:

$$p_\theta(x,z) = p_\theta(x|z)p(z) \tag{4.1}$$

According to the definition, a mathematical model that includes latent variables is called a latent variable model. Typically, the dimensionality of latent variables is much lower than that of the original data vectors, forming a compressed representation of the data. Therefore, they can be considered as a dimensionality reduction representation of the original data. In deep generative models, latent variables play a crucial role in understanding the data generation process and controlling the behavior of the model. Specifically, by jointly modeling latent variables with observable real data, one can not only understand the data generation process from latent variables to samples but also infer latent variables based on given observed data for downstream tasks such as classification, clustering, and regression.

Hence, research on latent variable models is essential for studying learning and inference methods in super-resolution reconstruction tasks, the super-resolution reconstruction method studied in this paper is based on image conditional generation transformation. This reconstruction method requires taking a given input image as the reconstruction target, minimizing the differences between the generated image and the target image in both overall and detail aspects. In this context, latent variable models can serve as useful features, acting as control valves in the reconstruction process. By more precisely controlling specific attributes during the reconstruction process, such as changing the skin color, texture, shape, facial details, etc. Through the rightmost latent variable feature map in image generation as shown in **Figures 4.2** and **Figures**

**4.3**, latent variable models contribute to learning effective representations of the conditional low-resolution data distribution.

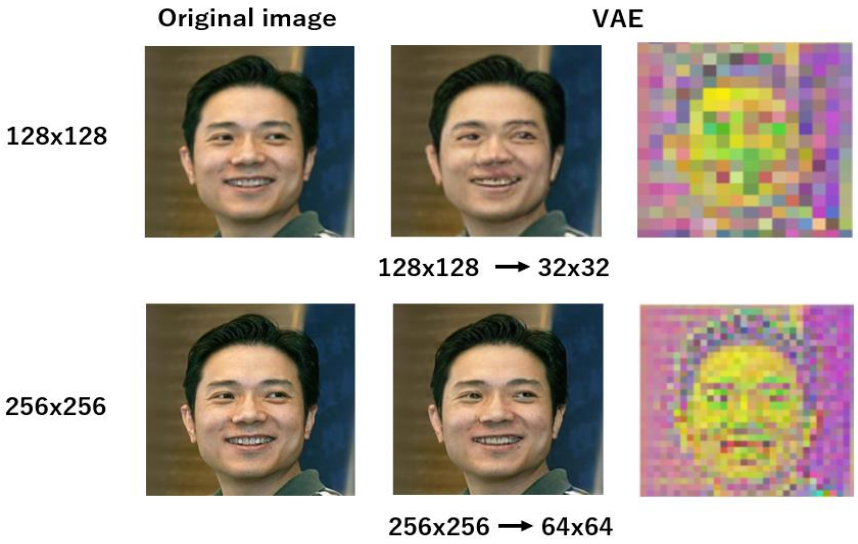


Figure 4.2 Latent variable feature map obtained by the variational autoencoder (VAE) model

The concept of latent variable models is applied in various machine learning tasks, including clustering, dimensionality reduction, generative models, semi-supervised learning, etc. Common examples include principal components in Principal Component Analysis (PCA), factors in Factor Analysis, topics in Latent Dirichlet Allocation (LDA), Gaussian Mixture Models (GMMs), etc. In the field of deep learning, latent variables frequently appear in models such as autoencoders and generative adversarial networks (GANs) to learn the distribution of data and latent variables.

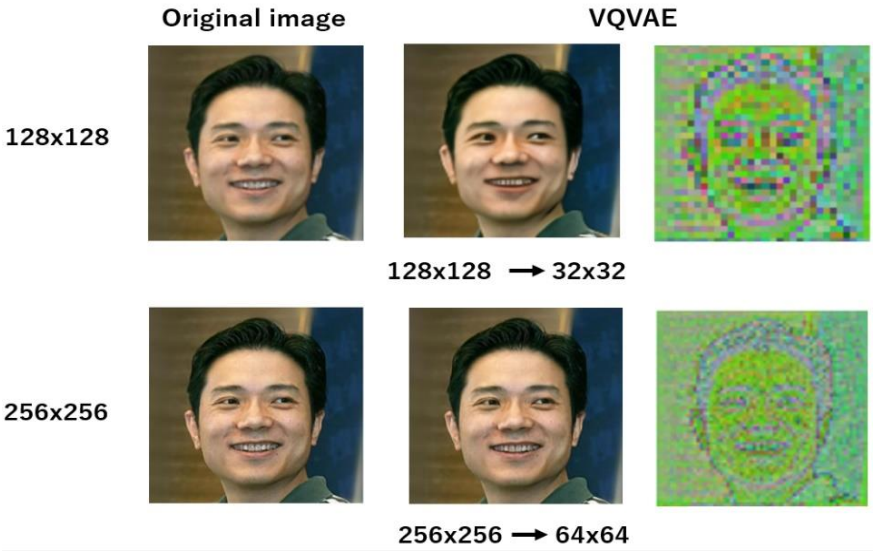


Figure 4.3 Latent variable feature map obtained by the vector quantization variational autoencoder (VQ-VAE) model

## 4.3 Latent Variable Denoising Diffusion Super-Resolution Reconstruction Model

### 4.3.1 Latent Variable Denoising Diffusion Probabilistic Model Autoencoder

Deep learning for generation aims to find a reasonable mapping relationship such that parameters in a low-dimensional space can be sampled onto a corresponding sample point in the embedded high-dimensional space distribution, the parameters sampled from the low-dimensional space are the latent variables, and the sampled low-dimensional space corresponds to the latent variable space. To address challenges in synthesizing images at resolutions of 256x256 and 512x512, we employ an encoder in the latent variable denoising diffusion probabilistic model to encode and map the distribution of training image data to a low-dimensional latent space, obtaining the latent variable code z for that image. The denoising diffusion probabilistic model is responsible for generation based on the low-dimensional latent space, and the decoder is responsible for the inverse operation of decoding, mapping the latent variable z from the low-dimensional compressed latent space to the high-dimensional image data space.

To avoid arbitrary scaling in the latent space, we introduce a regularization loss term $L_{reg}$, which normalizes the latent variable z to be zero-centered and have a small variance, we investigate two different latent variable regularization methods: (i) VAE (Variational Autoencoder), a standard variational autoencoder that constrains the true distribution and the generated distribution by introducing a low-weighted KL divergence between $q_E(z|x)$=$N(z; E_\mu, E_{\delta^2})$ and the standard normal distribution $N(z; 0,1)$, (ii) VQ-VAE, which regularizes the latent space using a vector quantization layer by learning an embedding dictionary with $|e|$ embeddings [42].

For high-fidelity reconstruction, we use very small regularization in both VAE and VQ-VAE. Through autoencoders, latent variables are constructed, where high-frequency, imperceptible details are abstracted, achieving access to an efficient, low-dimensional latent variable space. Compared to the high-dimensional pixel space, this space can (i) focus on attention mechanisms and semantic information in the data, and (ii) allow the model to be trained in a lower-dimensional, computationally efficient space.

Furthermore, VAE/VQVAE pull the latent variable distribution computed by the decoder to a normal distribution and train the output data distribution to closely match the distribution of real training data. Therefore, latent variables sampled from the

normal distribution can be effectively mapped to the target image space through the decoder, reasonable reconstruction of graphical information can be achieved based on the encoded latent variables. Additionally, by applying the reparameterization trick to the autoencoder and diffusion model, a balance between generation speed and image quality can be effectively maintained.

### 4.3.2 Generation Process of Latent Variable Denoising Diffusion Super-Resolution Reconstruction Model

The generation process of the latent variable denoising diffusion super-resolution reconstruction model primarily utilizes the latent variable features generated by the VAE/VQVAE, autoencoders to approximate the latent variable Z corresponding to the high-resolution image and the conditionally low-resolution LR image that has been degraded, the goal is to generate an approximately high-resolution image constrained by the condition of the degraded low-resolution image.

As shown in **Figure 4.3**, In the forward diffusion process of the latent variable denoising diffusion super-resolution reconstruction model, first, as discussed in the previous section, the VAE/VQVAE encoder is used to fit the distribution of the latent space, the high-resolution image is compressed through automatic encoding to replace the original image data distribution with the latent variable z. We obtain a latent variable z corresponding to the high-resolution image, which we refer to as the latent code. Next, in the DDPM forward process, the latent code is perturbed by noise, here, $X_i$ includes the HR (High-Resolution) image $X$ and, at each step, Gaussian noise is added to $X_{i-1}$, with $T$ being the total diffusion steps. In DDPM, the latent code gradually incorporates Gaussian noise to generate a noisy latent code.

As described in the earlier sections on super-resolution reconstruction models, LR image features are needed as conditions to constrain the generation space of HR images, the low-resolution LR image, processed after degradation according to the image degradation strategy studied in Chapter 2, serves as a conditional dependency. This LR image, along with the latent code obtained by adding noise to the high-resolution image, is merged and input into the U-Net network module to predict noise, perform conditional sampling, and combine image features, mapping them to the intermediate layers of U-Net, this process guides the U-Net network to learn more latent variable features from LR images and transfers the conditional features to the latent space.

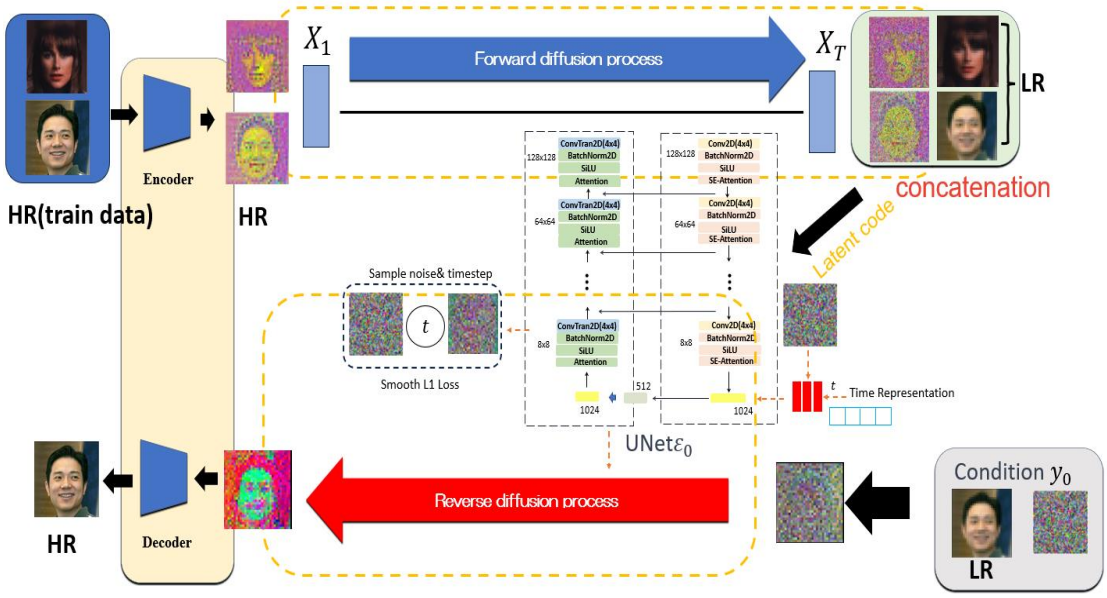


Figure 4.3 Network structure of latent variable denoising diffusion super-resolution reconstruction model

In the reverse inference process, the conditional probability is mapped to the output of the U-Net encoder. Convolutional layers are used to learn the mean $F_\mu$ and variance $F_\sigma$ of the feature map $F_x$ output by the U-Net encoder, the conditional probability mapping feature is then used to predict noise by utilizing convolutional layers in the U-Net, the model reduces the Kullback-Leibler (KL) divergence of the noise probability distribution between the denoising model and the real model, the low-resolution image is taken as a guiding condition and combined with random Gaussian noise. Through an iteratively refined inverse process, the combination of random Gaussian noise and guiding conditions is gradually transformed into a distribution similar to the latent variable data distribution of the high-resolution image. Finally, the decoding module of VAE/VQVAE is used to elevate the latent feature variables to a distribution similar to the originally generated image, achieving the reconstruction of high-resolution images. In the improvement made during the reverse inference process, the original model doesn't directly predict the result of each denoising step but predicts the noise directly for the noiseless image and then obtains the result of each denoising step through reparameterization. The quantized diffusion model utilizing the reparameterization trick achieves faster generation (about 15 times faster than DDPM models), while obtaining better image quality.

Through the latent variable Markov chain, using randomly sampled Gaussian distribution as the input latent variable Z for the feature decoder, the traditional denoising diffusion super-resolution reconstruction model experiences a quadratic growth in image generation speed with increasing resolution. However, the latent variable denoising diffusion super-resolution reconstruction does not depend on the previous generation results when generating each discrete code, making the generation speed independent of the image resolution. VAE/VQVAE not only effectively fills in the missing information due to LR image enlargement but also constrains the solution space for reconstructing HR images, this reduction in computational overhead makes it easier for the model to learn the current moment's noise and mitigates the negative impact of model collapse on HR image reconstruction during rapid sampling. It enables the rapid production of high-quality HR images with stable style and content consistency. While predicting the probability distribution of HR images is challenging, the proposed latent variable denoising diffusion super-resolution reconstruction method mitigates the impact of the inherent randomness in maximizing the variational lower bound in DDPM, this results in a stable training process and the generation of images that are consistent with the original LR image in terms of both style and content, producing more natural-looking results.

# Chapter 5
# Experiment and Evaluation

## 5.1 Experimental Datasets

### 5.1.1 CelebFaces Attribute Dataset

The CelebFaces Attribute dataset is a widely used large-scale dataset in the fields of face synthesis, attribute analysis, and editing. As shown in **Figure 5.1**, it is primarily employed for training and evaluating deep learning models related to faces. The dataset comprises thousands of face images of celebrities, accompanied by rich attribute annotations, these annotations include information such as age, gender, ethnicity, facial expression, hair color, and image background, researchers can utilize these annotations to train models for predicting various attributes of individuals in the images. The images are sourced from publicly available pictures on the internet, The dataset's large scale makes it a powerful resource and benchmark dataset for researching and developing face recognition algorithms. Additionally, some artificial intelligence competitions, such as face attribute prediction competitions, use the CelebFaces Attribute dataset as a publicly available competition dataset, This dataset is highly valuable for training algorithms related to face synthesis, editing, and other relevant fields.

Figure 5.1 Different individuals from the CelebFaces Attribute dataset

### 5.1.2 MRI Dataset

We validated our network model using the MRI Brain Tumor Classification dataset downloaded from the Kaggle website [45], the MRI Brain Tumor Classification dataset is a publicly available training set for brain MRI classification, the dataset includes over 3000 brain tumor images, comprising pituitary tumors, meningiomas, and gliomas, captured in 2D MRI cross-sectional slices in grayscale format, as shown in **Figure 5.2**. The original size of these images is 256×256, but due to constraints in GPU memory during experiments, all image sizes were cropped to 128×128.

Figure 5.2 Scanned images from the MRI Brain Tumor Classification dataset

### 5.1.3 Kather Colorectal Cancer Histology Multi-class Texture Analysis dataset

This image dataset is sourced from the Heidelberg University and Mannheim University Medical Center Pathology Institute, provided on the Kaggle website, as shown in **Figure 5.3.** It consists of completely anonymous histopathological images of human colorectal cancer tissue, totaling 4,000 images covering eight different types of tissue.

Figure5.3 The pathological images are sourced from a multi-class texture analysis dataset of colorectal cancer.

For training and validation, we selected 3,000 images with clear key details such as cell nuclei, cytoplasm, and glandular structures as the training set, and 500 images as

the validation set. Each pathological image has a size of 256×256 pixels, stored in TIF format, and a pixel size of 0.495 micrometers, these images were digitized using Aperio ScanScope (Aperio/Leica Biosystems) at a magnification of 20x.

## 5.2 Experimental Conditions and Hyperparameter Settings

This article conducted experiments using the pytorch framework, running on both Google Colab and a local laboratory server environment, all model training was performed using GPU acceleration. The server utilized an Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz processor, with two NVIDIA GeForce GTX 1080 Ti GPUs, each having a 12GB memory, totaling 24GB. The Google Colab platform used NVIDIA Tesla T4 and A100 GPUs. The programming language used in this work is Python, and the deep learning architecture is implemented using pytorch 1.7.1 with CUDA 11.1 support, the initial learning rate was set to lr=5e-4, and the loss function employed was the Smooth L1 loss function. During the experiments, Set the training batch size based on the model size and GPU capacity, the initial learning rate was 0.0002, and the optimization algorithm used was Adam, the noise schedule $\beta 1$ for the diffusion model employed a cosine schedule, the total steps for the generative model were set to 1000, while for the super-resolution model, it was set to 2000, the training loop was configured for 300 epochs.

## 5.3 Quality Evaluation Standards for Super-Resolution Reconstruction

### 5.3.1 Objective Evaluation Criteria

For the quality evaluation of images, this paper primarily utilizes two commonly used image assessment metrics: Peak Signal-to-Noise Ratio (PSNR) [44] and Structural Similarity Index (SSIM) [46] to judge the reconstruction image quality. PSNR is one of the most widely used quality assessment metrics in lossy transformation tasks, such as image denoising, compression, and deblurring. It represents the ratio of the maximum possible power of a signal to the power of noise that affects its accuracy, expressed in dB. For post-image super-resolution reconstruction, reconstruction error can be considered as noise, typically represented by Mean Squared Error (MSE). Given a ground truth HR image $I_{HR}$ and the reconstructed SR image $I_{SR}$, MSE is defined as follows:

$$MSE = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(I_{HR}(i,j) - I_{SR}(i,j)\right)^2 \quad (5.1)$$

where M and N represent the width and height of the image. Therefore, the PSNR of the reconstructed image $I_{SR}$ is defined as:

$$PSNR = 10 \times log_{10}\left[\frac{L^2}{MSE}\right] \quad (5.2)$$

PSNR, to some extent, reflects the approximation of the reconstructed image to the original image, it objectively evaluates the noise level in the image. However, because PSNR is only related to per-pixel MSE and focuses solely on the differences between corresponding pixels without considering visual perception, it may not be consistent with human perception in real-world scenarios. Despite this, PSNR remains widely used in image super-resolution reconstruction tasks due to its prevalence in previous literature and the absence of completely accurate perceptual metrics. A higher PSNR value indicates better image quality. Structural Similarity Index (SSIM) compares the similarity and differences between the original and reconstructed images in terms of brightness, structure, and contrast, thereby judging the quality of the image after reconstruction. A larger SSIM value indicates less difference between the reconstructed image and the original reference image, the formula for calculating SSIM is as follows:

$$SSIM(X,Y) = \frac{(2\mu_X\mu_Y + c_1)(2\sigma_{XY} + c_2)}{(\mu_X^2 + \mu_Y^2 + c_1)(\sigma_X^2 + \sigma_Y^2 + c_2)} \quad (5.3)$$

where $x$ and $y$ represent the two images being compared, $\mu_X$ is the average pixel value of image $x$, $\mu_Y$ is the average pixel value of image $y$, $\mu_X{}^2$and $\mu_Y{}^2$ are the variances of images $x$ and $y$, and $\sigma_{XY}$ is their covariance. Constants $c_1 = (k_1L)^2$ and $c_2 = (k_2L)^2$ are added to avoid division by zero, where $L$ represents the dynamic range of pixel values (usually 255). Typically, $k_1$ is set to 0.01, $k_2$ to 0.03, and $L$ to 255. The SSIM value ranges from $-1$ to 1, with higher values indicating greater similarity, SSIM is generally more suitable for evaluating structural similarity between local regions of images.

### 5.3.2 Subjective Evaluation Criteria

In recent years, subjective assessment of image super-resolution quality has become increasingly important. Subjective assessment relies on the observation of an image by the human eye, involving a subjective judgment of whether the image conforms to visual characteristics and an evaluation of the perceived quality of the reconstructed image. Since the human eye is sensitive to information such as edge contours, textures, color, and brightness in an image, it can quickly capture differences between two images, particularly in regions with distinct edge contours, this direct reflection of human visual perception results in subjective evaluation.

To measure the quality of super-resolution reconstructed images, we focused on aspects such as clarity and texture details and compared them with the information from the original HR images. We designed a subjective evaluation criterion to assess image quality, we organized a group of 20 friends and family members as testers to score images reconstructed by different algorithms based on their personal visual perceptions and some standard criteria, the images were rated on different levels of quality, ranging from high to low, divided into four categories: Excellent, Good, Fair, and Poor. Each category corresponds to different scoring criteria, for images with good texture details and clarity, and high color contrast, a score of 4 points was given. For images with acceptable texture details and clarity, a score of 3 points was assigned, images with poor texture details and clarity received a score of 2 points, while images with completely blurred texture details were given a score of 1 point. Testers judged the image effects based on their subjective feelings and the evaluation table for image quality.

To simplify the recording of scoring evaluation results, we introduced user preference (up) as a qualitative evaluation metric with high subjectivity. By categorizing 4 points and 3 points as user satisfaction and 2 points and 1 point as user dissatisfaction, we performed statistical analysis and presented the results in percentage form. Subjective evaluation is highly subjective and influenced by various factors such as the subjective perception and visual acuity of the testers, evaluation criteria, and other uncertainties, this introduces a significant degree of variability in the results, making it challenging to precisely assess the performance of the algorithm.

## 5.4 Experimental Results of Latent Variable DDPM

### 5.4.1Experimental Results of Latent Variable DDPM Generation

We first use our improved latent variable autoencoder DDPM generative model to synthesize natural images and medical pathological images with completely random Gaussian noise, as shown in **Figure 5.4**.

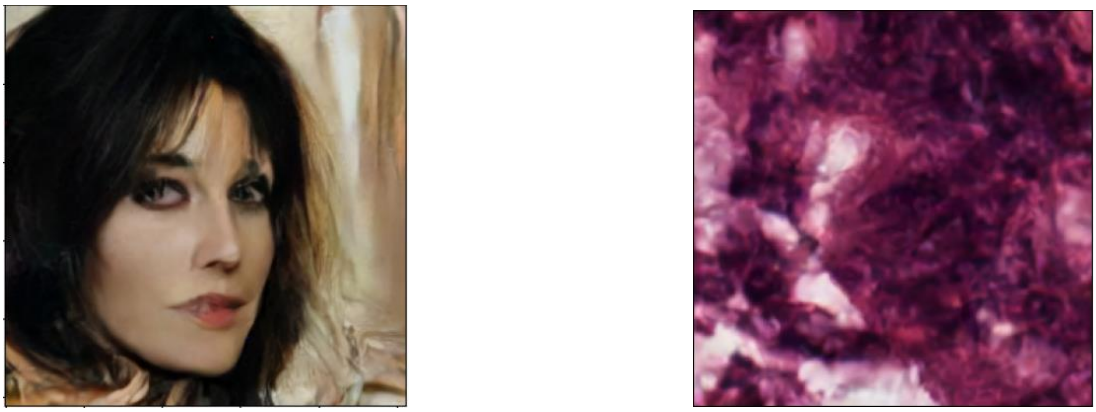

Figure 5.4: The two images above are generated using our improved DDPM model from Gaussian noise, including natural images and pathological images

To validate the results of image generation by VAE/VQVAE latent variable autoencoder DDPM, we compared different generative network models, all network models were obtained from publicly available code, as shown in **Figure 5.5**. From the figures, it is clear that the GAN-generated adversarial network synthesizes MRI with blurred boundaries and artifacts, while the anatomical effects are poor, and the gray and white matter regions in the brain are unclear. In contrast, DDPM, VAE-DDPM, and VQVAE-DDPM use a noise addition and denoising image synthesis method, avoiding the issues of GAN structures not meeting KL divergence Nash equilibrium and "mode collapse." The image quality difference between DDPM, our improved VAE-DDPM, and VQ-DDPM is not very noticeable. Compared to other methods, DDPM, our improved VAE-DDPM, and VQ-DDPM can sample high-quality 2D MRI slice images with clear details and realistic textures, with higher image restoration quality than other generative adversarial network methods, the parameter count of VAE-DDPM and VQ-DDPM denoising diffusion probability models is less than that of the denoising diffusion probability model, occupying less GPU memory and achieving higher efficiency in backward inference synthesis.

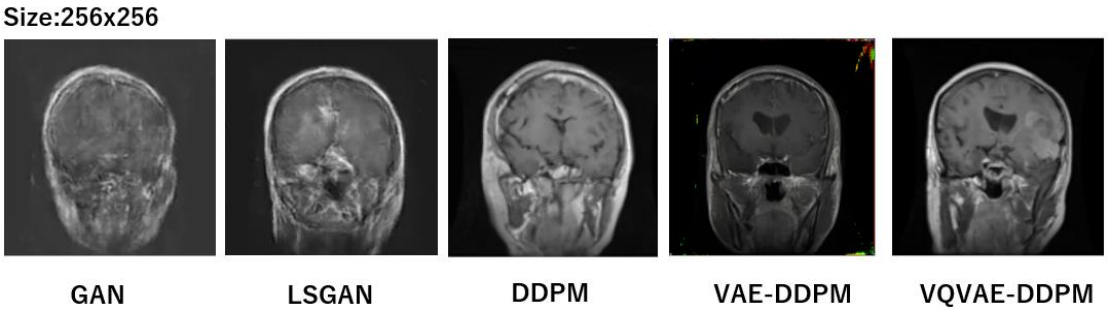


Figure 5.5: Results of MRI generated by different generation network models

### 5.4.2 Experimental results of latent variable DDPM super-resolution image reconstruction

As shown in **Figure 5.5**, we compared various super-resolution reconstruction models. We observed that SRCNN and FSRCNN, both based on convolutional neural networks for mapping in super-resolution reconstruction, overall exhibit poor reconstruction effects, although the reconstructed images are slightly clearer than the original LR images, the texture information cannot be restored clearly, this method only considers the correlation between pixels in the neighborhood, and due to degradation issues, there are often problems such as edge aliasing and blurring of edge and texture information within the neighborhood. It fails to restore the high-frequency information of the image and address the loss of edge information, resulting in unsatisfactory reconstruction effects.


Figure 5.5: Facial super-resolution reconstruction results using different super-resolution models

The last four methods are based on the approach of generating and reconstructing transformations through conditioned image generation. From the images, it can be observed that these methods, which learn the conditional image features for image transformation, demonstrate better generative results. This approach effectively protects information such as image edges and textures, particularly addressing false phenomena in edge information that may occur in high-noise images. Not only does it preserve fine details, but it also enhances the overall quality of image reconstruction, resulting in clearer and more detailed images. We improved the VAE-DDPM VQVAE-DDPM network, and compared it with DDPM, highlights the advantage of adding VAE/VQVAE autoencoder latent variables. To demonstrate that latent variables can better extract features and semantics from images, We compared using SSIM heatmaps, as shown in **Figure 5.6**, The heatmap indicates that the VQVAE-DDPM produces images with the least red regions, proving that the implicitly extracted features by its autoencoder contribute to preserving image details. Therefore, the reconstructed super-resolution images exhibit the best results.

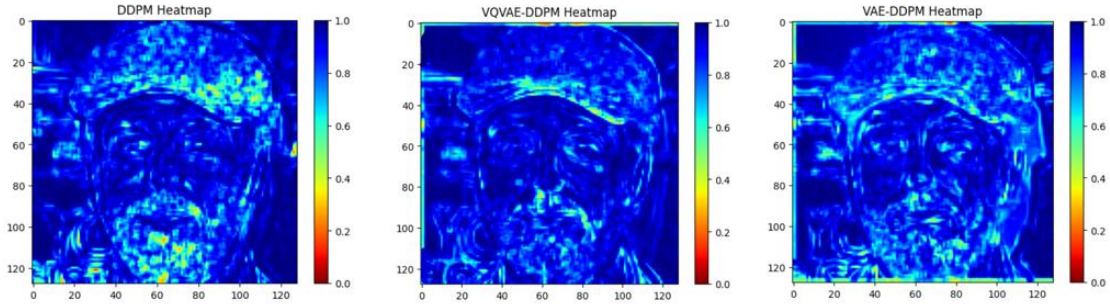


Figure 5.6: Comparison of SSIM heatmap effects for DDPM, VQVAE-DDPM, and VAE-DDPM. Red indicates a small SSIM value, signifying large differences between images, while blue areas represent small differences between the two images

As the generation of results relies on learning conditioned image features, the quality of the generated results determines the effectiveness of the reconstruction, this approach is dependent on the construction of the conditional dataset, and controlling latent feature information in the dataset distribution is challenging. Although VAE-DDPM performs well in recovering high-frequency features such as edges and textures of features like beards and eyes in the image, as seen in the SSIM heatmap, it demonstrates better results than DDPM. However, the overall color contrast in the reconstructed super-resolution images by VAE-DDPM deviates from the original high-resolution images, this deviation leads to suboptimal reconstruction results, future work could address this issue by studying specific adjustments to color parameters.

From the loss graph in the left image of **Figure 5.7**, we can observe that the improved model, VQVAE-DDPM, exhibits smaller losses compared to the original DDPM. It converges faster, reaches stability more quickly, and shows a smoother progression, this indicates that VQVAE-DDPM has an enhanced accuracy in reconstructing super-resolution images.

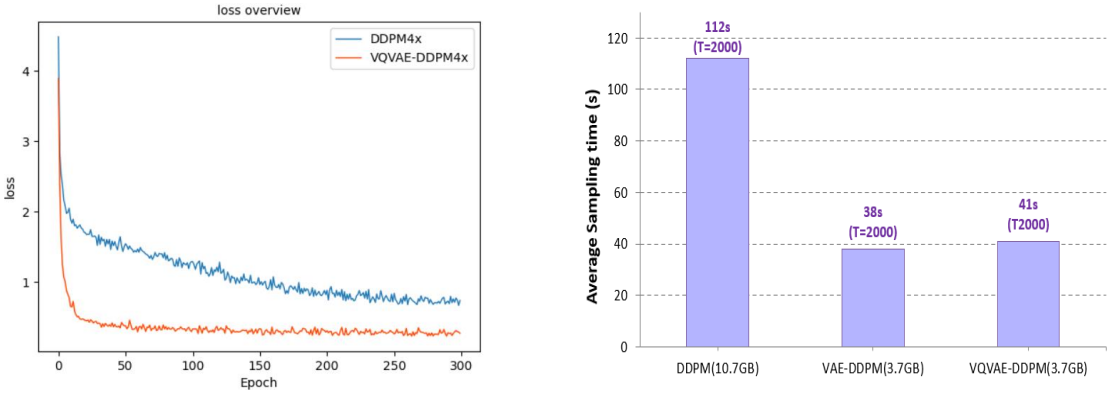


Figure 5.7: Results of the loss overview for DDPM and VQVAE-DDPM, the average sampling time and memory consumption for DDPM, VAE-DDPM, and VQVAE-DDPM

The right graph in **Figure 5.7** and **Figure 5.8** illustrates that methods for generating transformation reconstructions through image-conditioned approaches have a large number of parameters, especially in the case of DDPM, VAE-DDPM, and VQVAE-DDPM. One of the advantages of a large number of model parameters is the ability to linearly fit different details. However, excessively large model parameters can lead to issues such as high GPU memory usage and extended training times.

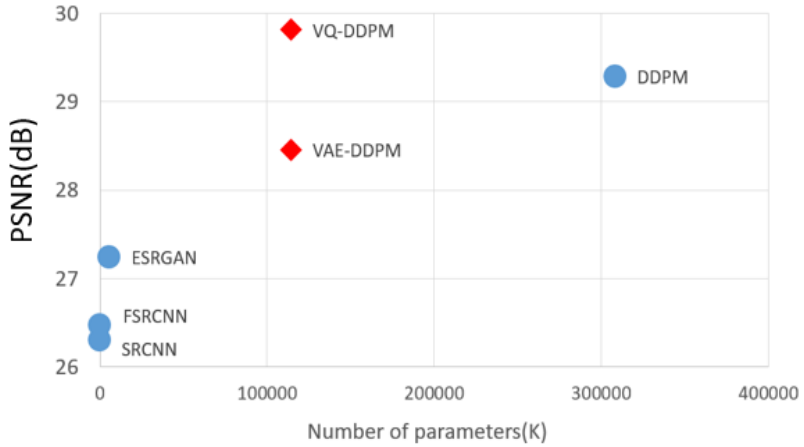


Figure 5.8: The figure presents a comparison of parameter count and performance metrics for various super-resolution models, ours model is represented by a different color or symbol for clarity

In our improved models, VAE-DDPM and VQVAE-DDPM, the number of model parameters is reduced compared to the original DDPM model, yet their performance remains unaffected. As seen in the right graph of **Figure 5.7**, the GPU memory usage for VAE-DDPM and VQVAE-DDPM is noticeably lower than that for DDPM. Therefore, by introducing VAE/VQVAE autoencoders to model the distribution of latent space, we have reduced the model's parameter count, alleviated GPU memory usage, and simultaneously decreased training times and the duration of reverse inference. Importantly, these improvements do not compromise the clarity and quality of the synthesized images.

We compared various super-resolution reconstruction models based on PSNR and SSIM metrics, as shown in **Table 5.1**, along with an evaluation of model parameters. SRCNN, being an early method relying on mean square error for super-resolution reconstruction, exhibits relatively low PSNR and SSIM values. While FSRCNN and SRCNN show improvements in PSNR and SSIM, there is still a noticeable gap in super-resolution effectiveness compared to real facial images. DDPMSR outperforms previous works in terms of both PSNR and SSIM evaluation metrics, demonstrating its ability to generate high-quality super-resolution images. However, VAE-DDPM has some color parameter issues, leading to slight deviations in the reconstruction, resulting in lower PSNR and SSIM values. On the other hand, VQVAE-DDPM achieves the best super-resolution reconstruction results among the compared models.

Table 5.1: Objective evaluation table for the quality of super-resolution reconstructed images

| Methods | Timesteps | PSNR | SSIM | Parameters(K) |
|---|---|---|---|---|
| SRCNN | / | 26.30 | 0.723 | 69 |
| FSRCNN | / | 26.47 | 0.748 | 25 |
| ESRGAN | / | 27.24 | 0.838 | 5949 |
| DDPM | 2000 | **29.28** | **0.864** | **308569** |
| **VAE-DDPM** | 2000 | **28.45** | **0.856** | **114446** |
| **VQVAE-DDPM** | 2000 | **29.81** | **0.878** | **114446** |

We also conducted a subjective evaluation of super-resolution reconstruction results, and the user satisfaction (up) scores are presented in **Table 5.2**. From the table, it can be seen that VQVAE-DDPM achieves higher subjective satisfaction scores. The satisfaction scores for DDPM-based methods are higher than those for convolutional neural network-based mapping and generative adversarial network-based reconstruction methods. It's important to note that subjective evaluations are based on individual

opinions, and when the visual differences between the reconstructed and original images are minimal, subjective evaluation methods may not provide highly accurate judgments. Therefore, subjective evaluation is used as a complementary metric to objectively assess our super-resolution results.

Table 5.2 Subjective satisfaction (up) evaluation table for the quality of super-resolution reconstructed images

| Methods | CNN Methods | SRCNN | FSRCNN | ESRGAN |
|---|---|---|---|---|
| UP | / | 10% | 11% | 40% |
| Methods | Diffusion Methods | DDPM | VAE-DDPM | VQVAE-DDPM |
| UP | / | 70% | 62% | 75% |

We conducted experiments and comparisons on two synthetic network models, VAE-DDPM, and VQVAE-DDPM, with training batches at 30, 60, 120, 240, and 300 epochs, displaying their synthetic effects as shown in **Figure 5.9**. In the first column from the left, it can be observed that at 30 epochs of training, due to the incomplete training of the denoising model, it can only make simple predictions on randomly generated noise distributions, the facial image details recovered from random noise are somewhat blurry, and the predicted synthetic distribution results are low.

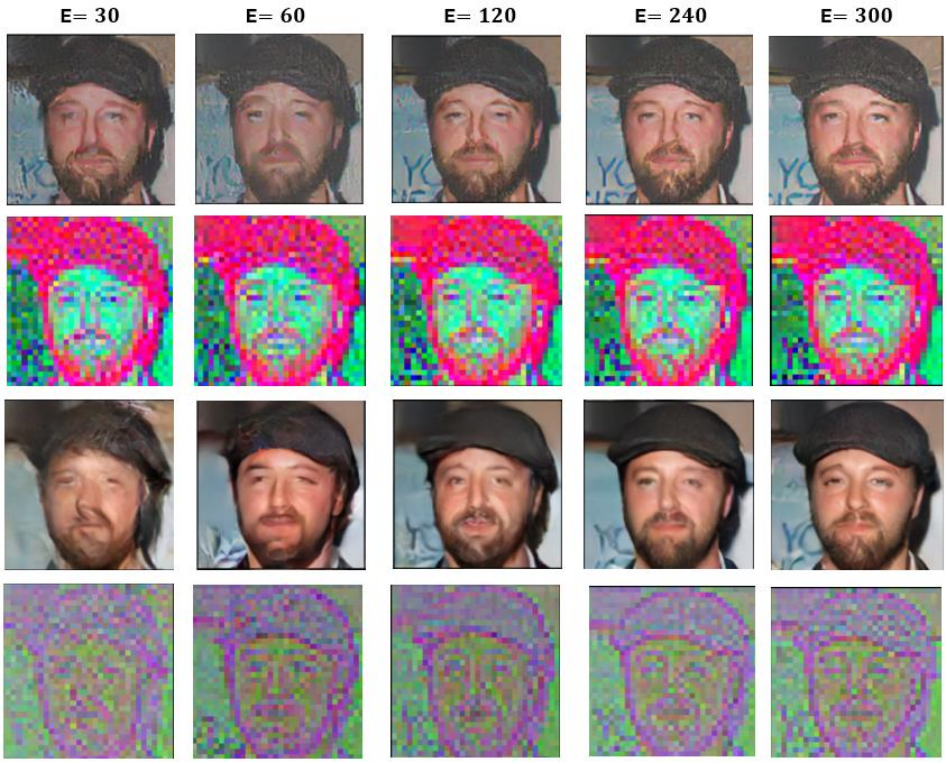


Figure 5.9: Reconstruction results and latent feature comparison results for two synthetic network models (VAE-DDPM, VQVAE-DDPM) at training batches 30, 60, 120, 240, and 300

The generated images exhibit obvious traces of artificial synthesis, by comparing with the second image, we can see that VAE-DDPM has slightly poorer synthetic image quality than VQVAE-DDPM. When the training batch is above 120 epochs, VAE-DDPM stabilizes in terms of its discrete fitting ability and latent feature information, resulting in stable reconstruction quality.

In our final study, we investigated different image quality degradation strategies for degraded super-resolution reconstruction, we experimented with randomly mixing bicubic interpolation and degradation factors as strategies for low-quality reconstruction of low-resolution images. As shown in **Figure 5.10**, simple bicubic interpolation for image degradation results in a relatively blurry low-resolution image, making it closer to true low resolution due to the degradation factors. On the other hand, our adopted random mixing degradation strategy for low-quality reconstruction produces low-resolution images with texture details that are more akin to real image degradation processes. This degradation strategy effectively preserves a significant amount of image semantics and high-frequency details, therefore, using a random mixing degradation strategy enhances the clarity of the reconstructed images, making the data generation process more natural.


Figure 5.10: Comparison of low-resolution images reconstructed using different image degradation methods

To validate the effectiveness of the latent variable DDPM for super-resolution image reconstruction and its transfer application to medical pathology images, we conducted comparative experiments on the Kather colorectal cancer histology multi-class texture dataset, the results of 4x resolution reconstruction are shown in **Figure 5.11**. From the images, it is evident that SRGAN, DDPMSR, VAE-DDPM, and VQVAE-DDPM produce results very close to real high-resolution images. However, in the fourth SRGAN reconstruction image at the bottom of **Figure 5.11**, there are slight pale purple artifacts in the lower white region, whereas DDPMSR and VQVAE-DDPM closely resemble the real image with an all-white appearance. This difference is attributed to the fact that DDPMSR and VQVAE-DDPM do not require additional modules, such as discriminators. During training, they do not need to balance Nash equilibrium conditions, avoiding issues like gradient explosions and mode collapse. VQVAE-DDPM shows excellent super-resolution results on colorectal cancer multi-class texture pathology images, and compared to generative adversarial networks, it avoids artifacts, resulting in more natural-looking reconstructions and reconstructs clearer textures, maintaining a strong consistency with HR images. However, VAE-DDPM may have color deviation issues due to its color settings, making it less suitable for medical image transfer and not achieving the desired super-resolution effect.

Nevertheless, these experimental results demonstrate that our latent variable denoising diffusion super-resolution reconstruction model can be successfully applied to the field of medical image super-resolution processing, indicating its high practical value.

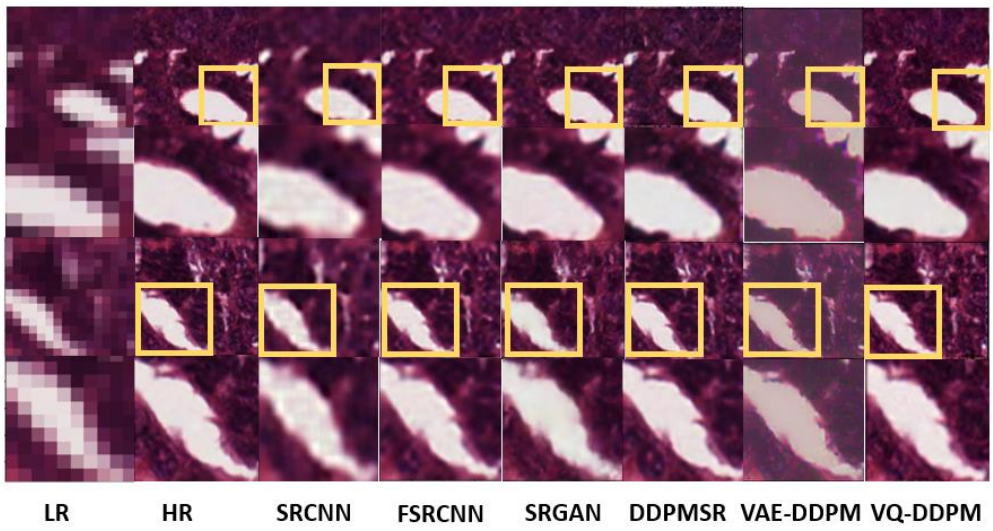


Figure 5.11: Results of a different super-resolution methods 4x reconstruction (32×32→ 128×128)

Additionally, we performed super-resolution reconstruction on medical pathology images with an 8x enlargement factor. In the standard 8x super-resolution reconstruction experiment, we randomly selected three 16x16 resolution images from the test set and reconstructed them for comparison, we then compared the reconstructed images after 8x enlargement from different networks. Due to the larger enlargement factor, there are inherent differences between the reconstructed images and the reference images, the results are shown in **Figure 5.12**.

From the images, it can be observed that the images reconstructed by SRCNN are relatively blurry, with poor texture details. On the other hand, both SRGAN and DDPMSR networks achieve better visual reconstruction results, clear organizational structural textures and richer details are visible in these images, closely resembling real HR images. However, the images reconstructed by SRGAN exhibit blurriness and artifacts at the edges, such as ripples and aliasing, resulting in a lack of edge details and a less natural appearance compared to the original HR images.

Our improved model shows strong recovery of image texture details, although the color restoration may not be perfect. In contrast, DDPMSR's reconstructed images present a more natural appearance in terms of edge details, indicating that the DDPMSR model can finely preserve the complete information of the images even at larger enlargement factors.

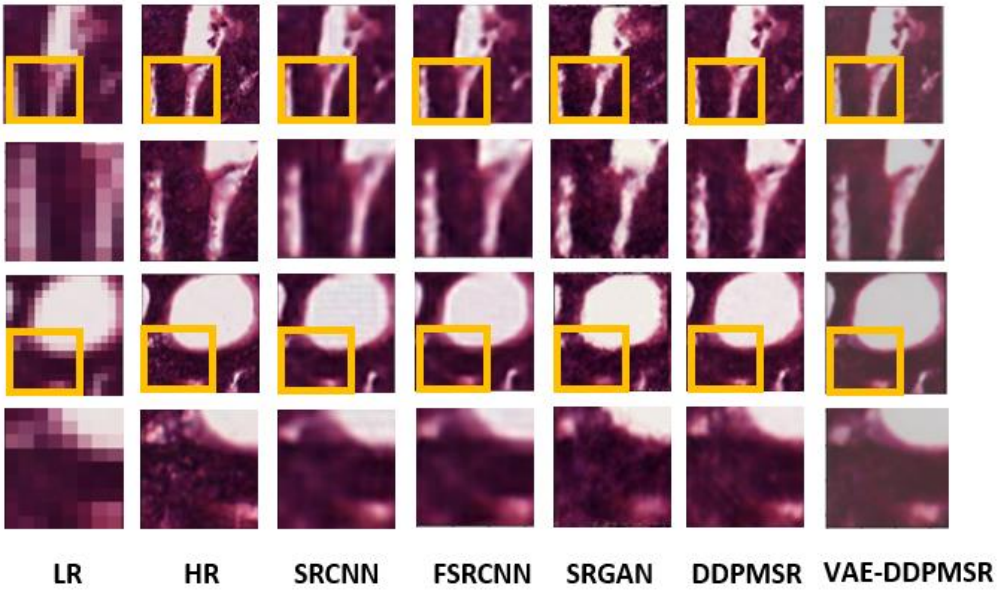


Figure 5.12: Results of a different super-resolution methods 8x reconstruction (16×16 → 128×128)

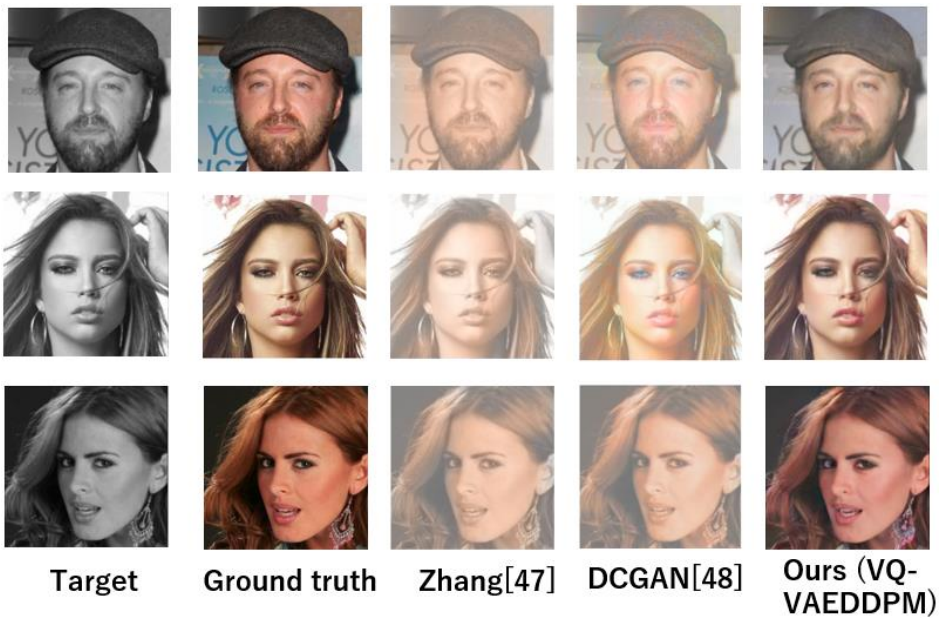


Figure 5.13: Different colorization method celebfaces attribute dataset coloring effect comparison

We also extended our improved latent variable denoising diffusion model to other domain tasks, such as image colorization. Initially, we utilized the CelebFaces Attribute dataset for colorization, a comparison of colorization results among various methods is presented in **Figure 5.13**.


Figure 5.14: Different colorization method animes coloring effect comparison

Contrasting with real images, it can be observed that previous methods like Zhang's[47] and DCGAN's[48] colorization outcomes were subpar, with most areas of the images appearing yellowish and lacking semantic fidelity in background color restoration. However, in our VQVAE-DDPM method, the color fidelity is higher, and the colorized images closely resemble real ones. Furthermore, we validated our improved latent variable denoising diffusion model by training it on the Rem (a character from anime) dataset for anime colorization experiments. As depicted in the **Figure 5.14**, the colorized images generated by our VQVAE-DDPM model exhibit rich and vibrant colors, with strong semantic information, better preserving Rem's original colors. Thus, this validates the advancement of our improved denoising diffusion model in the field of image colorization.

# Chapter 6
# Conclusions

This paper primarily investigates four aspects of image processing based on an improved denoising diffusion probabilistic model: including image super-resolution and degradation principles and strategies, denoising diffusion probability models, improvement schemes for latent variable denoising diffusion super-resolution models, and the application of improved models to medical image processing and other tasks in computer vision image processing such as image colorization.

There are various processes of image quality degradation in the real world. In contrast to some current image reconstruction research that only targets specific degradation types, this study first investigates the process of image degradation and degradation, which includes three degradation factors that lead to real image degradation: blur, downsampling, and noise. We adopt a degradation strategy that randomly mixes degradation factors as much as possible to train, and the results generated by this mixed degradation strategy are found to be better in face super-resolution experiments.

Simultaneously, we improve the original denoising diffusion probability network and propose a latent variable denoising diffusion super-resolution reconstruction model for synthesizing clear images and image super-resolution reconstruction, this improvement introduces intermediate latent variables into the model for transition. In the forward diffusion process, similar to the original DDPM model's diffusion process, the high-resolution image input is transformed into latent variables Z through feature decoders, after T times of adding Gaussian noise to latent variable Z, it is transformed into a Gaussian noise distribution. Since the image super-resolution reconstruction model needs to constrain the solution space of HR images, the latent variables Z of the low-resolution image LR and the high-resolution image HR after noise addition at the current moment are stacked together for conditional sampling. Then, the powerful parameter fitting ability of the diffusion denoising network is used to learn the distribution of latent variables Z. In the reverse inference process, the model combines the low-resolution image as a guiding condition with random Gaussian noise distribution. Through iteratively refined inverse processes, the randomly combined Gaussian noise is gradually transformed into latent variable $Z_{\varphi}$ of the high-resolution image. Finally, the latent feature variable $Z_{\varphi}$ is promoted to a distribution similar to the original generated image distribution using the decoding module of VAE/VQVAE, thus achieving the reconstruction of high-resolution images. Experimental results show that

VAE/VQVAE-DDPM produces clearer and more natural images after reconstruction compared to other methods.

The improved super-resolution reconstruction network model solves the problems of single degradation in low-resolution dataset images and the gradient explosion of traditional generative adversarial network (GAN) images, as well as the problem of pattern collapse synthesis producing artifacts. Finally, we conducted comparative experiments on both a facial dataset and the Kather colorectal histology multi-class texture dataset. The results indicate significant achievements in super-resolution reconstruction of facial images, outperforming the DDPM method in terms of peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), as well as traditional methods such as ESRGAN, SRCNN, and FSRCNN. Meanwhile, applying this method to the study of colorectal tissue unit pathology images, the experimental results show that the method surpasses traditional generative adversarial network super-resolution methods in high-resolution reconstruction, demonstrating significant advantages, it not only preserves fine details but also improves the overall image reconstruction quality. Finally, we conducted research on the improvement of the denoising diffusion model in image colorization tasks, we validated on face datasets and anime datasets, and comparative experimental results show that the improved denoising diffusion model also has advanced performance in image colorization fields.

This research has made significant contributions to face recognition, pathology diagnosis and treatment, color restoration of old photos, and automatic colorization of animations, providing a powerful computer image processing tool. However, there are still some issues in our research. For example, although our improved denoising diffusion probability super-resolution model optimizes the reverse inference time, the overall inference time is still long. Additionally, how to control the integration of latent feature information, such as image edges and textures, may affect the colors and semantic details of high-resolution pathological images generated by DDPM super-resolution models. Therefore, improving inference methods to shorten inference time, integrating multiscale resolutions into DDPM super-resolution models, and addressing improvements in video super-resolution and video colorization will be the focus of future research.

# Chapter 7
# References

[1] Farooq M, Dailey M N, Mahmood A, et al. Human face super-resolution on poor quality surveillance video footage[J]. Neural Computing and Applications, 2021, 33: 13505-13523.

[2] Nan F, Jing W, Tian F, et al. Feature super-resolution based Facial Expression Recognition for multi-scale low-resolution images[J]. Knowledge-Based Systems, 2022, 236: 107678.

[3] Chen L , Yang X , Jeon G , et al. A Trusted Medical Image Super-Resolution Method based on Feedback Adaptive Weighted Dense Network[J]. Artificial Intelligence in Medicine, 2020, 106(2):101857

[4] Xiao J, Yong H, Zhang L. Degradation model learning for real-world single image super-resolution[C]//Proceedings of the Asian Conference on Computer Vision. 2020.

[5] Baker S, Kanade T. Limits on super-resolution and how to break them[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(9): 1167-1183.

[6] Ha V K, Ren J, Xu X, et al. Deep learning based single image super-resolution: A survey[C]//Advances in Brain Inspired Cognitive Systems: 9th International Conference, BICS 2018, Xi'an, China, July 7-8, 2018, Proceedings 9. Springer International Publishing, 2018: 106-119.

[7] Mahmoudzadeh A P, Kashou N H. Interpolation-based super-resolution reconstruction: effects of slice thickness[J]. Journal of Medical Imaging, 2014, 1(3): 034007-034007.

[8] Schultz R R, Stevenson R L. Extraction of high-resolution frames from video sequences[J]. IEEE transactions on image processing, 1996, 5(6): 996-1011.

[9] Tsai R Y, Huang T S. Multiframe image restoration and registration[J]. Multiframe image restoration and registration, 1984, 1: 317-339.

[10] Bevensee R M. Fundamental limitations in antenna resolution by maximum entropy methods[R]. Lawrence Livermore National Lab., CA (USA), 1984.

[11] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

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

[13] Li X, Orchard M T. New edge-directed interpolation[J]. IEEE transactions on image processing, 2001, 10(10): 1521-1527.

[14] Qiao J, Song H, Zhang K, et al. Image super‐resolution using conditional generative adversarial network[J]. IET Image Processing, 2019, 13(14): 2673-2679.

[15] Dong C, Loy C C, He K, et al. Image super-resolution using deep convolutional networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(2): 295-307.

[16] Dong C, Loy C C, Tang X. Accelerating the super-resolution convolutional neural network[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer International Publishing, 2016: 391-407.

[17] Kim J, Lee J K, Lee K M. Accurate image super-resolution using very deep convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1646-1654.

[18] Ota J, Umehara K, Ishimaru N, et al. Evaluation of the sparse coding super-resolution method for improving image quality of up-sampled images in computed tomography[C]//Medical Imaging 2017: Image Processing. SPIE, 2017, 10133: 509-517.

[19] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4681-4690.

[20] Wang X, Yu K, Wu S, et al. Esrgan: Enhanced super-resolution generative adversarial networks[C]//Proceedings of the European conference on computer vision (ECCV) workshops. 2018: 0-0.

[21] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C]//International conference on machine learning. PMLR, 2017: 214-223.

[22] Cheng X, Fu Z, Yang J. Zero-shot image super-resolution with depth guided internal degradation learning[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. Springer International Publishing, 2020: 265-280.

[23] Sutcu Y, Bayram S, Sencar H T, et al. Improvements on sensor noise based source camera identification[C]//2007 IEEE International Conference on Multimedia and Expo. IEEE, 2007: 24-27.

[24] Zhang K, Liang J, Van Gool L, et al. Designing a practical degradation model for deep blind image super-resolution[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 4791-4800.

[25] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786): 504-507.

[26] Yang X. Understanding the variational lower bound[J]. variational lower bound, ELBO, hard attention, 2017, 22: 1-4.

[27] Van Erven T, Harremos P. Rényi divergence and Kullback-Leibler divergence[J]. IEEE Transactions on Information Theory, 2014, 60(7): 3797-3820.

[28] Van Den Oord A, Vinyals O. Neural discrete representation learning[J]. Advances in neural information processing systems, 2017, 30.

[29] Menéndez M L, Pardo J A, Pardo L, et al. The jensen-shannon divergence[J]. Journal of the Franklin Institute, 1997, 334(2): 307-318.

[30] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//International conference on machine learning. PMLR, 2015: 2256-2265.

[31] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[32] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis[J]. Advances in neural information processing systems, 2021, 34: 8780-8794.

[33] Batzolis G, Stanczuk J, Schönlieb C B, et al. Conditional image generation with score-based diffusion models[J]. arXiv preprint arXiv:2111.13606, 2021.

[34] Hong S, Lee G, Jang W, et al. Improving sample quality of diffusion models using self-attention guidance[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 7462-7471.

[35] Chung H, Sim B, Ryu D, et al. Improving diffusion models for inverse problems using manifold constraints[J]. Advances in Neural Information Processing Systems, 2022, 35: 25683-25696.

[36] Théâtre D'opéra Spatial online:
https://en.wikipedia.org/wiki/Th%C3%A9%C3%A2tre_D%27op%C3%A9ra_Spatial

[37] Joyce J. Bayes' theorem[J]. 2003.

[38] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, et al. (editors). Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. Springer, Cham; 2015. pp. 234–241.

[39] Saharia C, Ho J, Chan W, et al. Image super-resolution via iterative refinement[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(4): 4713-4726.

[40] Wang Y , Li X , Nan F , et al. Image super-resolution reconstruction based on generative adversarial network model with feedback and attention mechanisms[J]. Multimedia Tools and Applications, 2022, 81(5):6633-6652

[41] Louizos C, Shalit U, Mooij J M, et al. Causal effect inference with deep latent-variable models[J]. Advances in neural information processing systems, 2017, 30.

[42] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695.

[43] Brain tumor classification (MRI). Available online: https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri

[44]Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. In: Proceedings of the 6th International Conference on Learning Representations (ICLR 2018); 30 April–3 May 2018; Vancouver, BC, Canada. pp. 1–26.

[45] Kather Colorectal Cancer Histology Multi-class Texture Analysis dataset Available online: https://www.kaggle.com/datasets/kmader/colorectal-histology-mnist.

[46]Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 2004; 13(4): 600–612. doi: 10.1109/TIP.2003.819861.

[47] Zhang R, Isola P, Efros A A. Colorful image colorization[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. Springer International Publishing, 2016: 649-666.

[48] Nazeri, Kamyar, Eric Ng, and Mehran Ebrahimi. "Image colorization using generative adversarial networks." Articulated Motion and Deformable Objects: 10th International Conference, AMDO 2018, Palma de Mallorca, Spain, July 12-13, 2018, Proceedings 10. Springer International Publishing, 2018.