Medical CT Image Generation Based on Latent Space Compression

Jinxing Zhu[#]
College of Big Data and Internet
Shenzhen Technology University
Shenzhen, China
ronkoc@outlook.com
(Equal contribution)

Jiayu Wu[#]
College of Big Data and Internet
Shenzhen Technology University
Shenzhen, China
2210413001@email.szu.edu.cn
(Equal contribution)

Tao Wang
College of Big Data and Internet
Shenzhen Technology University
Shenzhen, China
wangtao@sztu.edu.cn

Bingding Huang[*]
College of Big Data and Internet
Shenzhen Technology University
Shenzhen, China
huangbingding@sztu.edu.cn
*Corresponding author

2025 IEEE 3rd International Conference on Control, Electronics and Computer Technology (ICCECT) | 979-8-3315-1855-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCECT64621.2025.11339847

***Abstract*—In recent years, with deep learning development, research based on medical image data has received widespread attention. However, deep learning algorithms require model training on a large number of annotated datasets. Access to medical image data is strictly regulated due to the high cost of acquiring and legal, ethical, and privacy considerations. All these reasons lead to the fact that it is often difficult to obtain large-scale medical image datasets. How to expand the existing data by generating simulated medical images through deep learning methods has become an urgent issue. Here, we propose a generation model for 3D medical CT images based on VQGAN and DDPM. Our algorithm employs DDPM to fit the distribution of latent space data of VQGAN, which realizes the generation of CT images at the 3D level and effectively preserves the consistent coherence among CT slices. We conducted experimental validation using the MICCAI FLARE 2022 dataset and comparative experiments for the parameters used in the VQGAN training process. Our proposed model's 3D average slice FID coefficient reaches 20.79, and the FVD coefficient reaches 252.97 in 128×128×64 dimensions. These results demonstrated that our model can effectively generate medical CT images with sound generation quality and stable performance.**

Keywords—Image Generation, Medical Images, Adversarial Learning, Diffusion Model

I. INTRODUCTION

Medical imaging data is vital in artificial intelligence and machine learning for automated image analysis and computer-aided diagnosis[1-4]. However, since medical imaging data collection methods are often accompanied by ionizing radiation that harms human health, medical images usually require specialized equipment, which increases the cost of data acquisition[5]. Regarding ethical considerations and privacy, medical image datasets contain sensitive patient information and require strict privacy protection. For various reasons, obtaining large-scale medical image datasets is often difficult, resulting in a scarcity of data for training deep models[6-9]. How to generate simulated medical images through deep learning methods to expand existing data has become an urgent issue that needs to be solved.

Image generation focuses on developing algorithms and models to generate realistic images or modify existing ones while preserving their semantic content[10]. Medical image generation is also emerging as an essential research area as it has the potential to address various challenges in medical imaging[11-14]; for example, medical image generation can mitigate privacy concerns by generating images that closely resemble accurate patient data, enabling researchers to share and collaboratively process datasets without compromising patient privacy, and thus accelerating medical image analysis and computer-aided diagnostic research. Acquiring medical images is expensive, time-consuming, and sometimes risky for patients; therefore, medical image generation also plays a crucial role in various clinical applications such as disease diagnosis, treatment planning, and surgical simulation[15, 16].

With the development of computer vision, some influential generative models have emerged, such as VAE(variational autoencoders)[17], PixelCNN[18], Glow (Generative Latent Optimization)[19], GANs (Generative Adversarial Networks)[20], DM(Diffusion Model)[21], etc. Among them, GAN was proposed by Goodfellow et al. in 2014 and received widespread attention. GAN contains two main components: generator and discriminator. These two parts confront and cooperate to generate realistic data through confrontation training. Esser et al.[22] combined the effectiveness of the inductive bias of CNN with the expressiveness of the transformer and proposed VQGAN (Vector Quantized Generative Adversarial Network), which can model and thereby synthesize high-resolution images. Dong et al.[23] used adversarial generation strategies for generation from MR images to CT images. However, GANs also have problems with unstable training and mode collapse. In recent years, diffusion models have become a promising image generation technology that can generate realistic and high-quality images and have been widely used in general image generation[24, 25]. Ho et al.[26] proposed DDPMs (Denoising Diffusion Probabilistic Models), which can provide high-quality image synthesis results. On the unconditional CIFAR10 data set, this method's Inception index and FID index reached 9.46 and 3.17, respectively. The diffusion models' mechanisms allow them to effectively capture changes in anatomy, pathology, and imaging modalities. Deep generative models have shown great potential in solving the problem of medical image scarcity.

However, applying deep generative models to medical images has not been extensively and systematically evaluated. This is mainly because medical images are entirely different from natural images. Medical images, such as CT images, are three-dimensional voxel data requiring higher computational costs to train deep models. In addition, the modality, signal-to-noise ratio, difficulty of data acquisition, and other characteristics of medical images are also entirely different

from natural images. The above problems pose considerable challenges to deep generative models applied to medical images.

Moreover, in terms of structured information, generic images usually lack inherent structured information and exhibit high complexity, including different object classes, backgrounds, lighting conditions, and viewpoints. Generative models for generic images must capture this complexity and generate visually realistic samples. Medical images, on the other hand, have specific structured information related to anatomical structures, pathologies, and imaging modalities[27, 28]. These structured features are often similar and homogeneous. Generative models for medical image design need to focus on capturing and reproducing these structured features, such as organ shapes, tissue textures, or disease-specific features. Understanding and utilizing these features allows us to generate simulated images that resemble authentic medical images.

Previous studies have shown that VQGAN[22] has good image reconstruction and feature extraction capabilities. The diffusion model performs strongly in capturing data features and is the best-performing method among current generative models. Therefore, we combine VQGAN and DDPM in this work to achieve a better medical CT image generation model.

In this work, we propose a network architecture that combines an adversarial autoencoder and diffusion model to generate CT images, first by training the VQVAE with a CT image dataset, which improves the model's ability to reconstruct the CT image through adversarial strategies, then compressing the data onto potential spatial representations by the trained VQVAE, and using it to train the diffusion model to learn the distribution of the data from these potential spatial representations. In the inference process, the latent space representations reflecting the CT images are generated by the trained diffusion model, which is then mapped into CT images by the decoder of the VQVAE for image generation. Our model leverages DDPM to fit the latent space data distribution of VQGAN, resulting in the impressive generation of high-quality three-dimensional CT images. It effectively maintains consistency between CT slices, ensuring a reliable and accurate representation of the original data.

II. Method

The generative model proposed in this work combines VQGAN and DDPM for CT image reconstruction, latent space compression, and feature generation, respectively. The main design of VQGAN in this work is based on VQVAE, including an encoder and a decoder, in which the two modules' design are symmetrical. The encoder is responsible for converting the input image into a low-dimensional latent space representation, which mainly consists of a series of 3D convolutional layers to extract image features and gradually reduce the spatial dimension of the data. The 3D convolutional layers are adapted to the dimensions of the 3D voxel data and are used for data feature extraction and transformation. The downsampling layer is used to reduce the spatial dimension of the input data, which can reduce the number of parameters and computational complexity of the model and improve the robustness and generalization ability. In this work, we replace the pooling layer by downsampling through a convolutional layer. The downsampling effect can be achieved by setting a more significant step size in the convolutional layer. A larger step size will reduce the spatial size of the output feature map, thus achieving the downsampling effect. A residual module stack is connected at the end of the encoder, which consists of an attentional residual module, a batch normalization layer, and a ReLU activation function. The attentional residual module is constructed using a 3D convolution operation and a multi-head attentional module in three dimensions. The attentional residual module is a module that combines the attentional mechanism and residual connectivity, which is commonly used to enhance the representational and generalization capabilities of neural networks. It combines residual connectivity with the attention mechanism, allowing the network to pay more attention to essential features and better propagate the gradient, thus improving the model's performance. With normalization, the distribution of inputs in each network layer can be made more stable, hence speeding up the network's training process. Due to the more stable gradients, a more significant learning rate can be used, and convergence to a better solution is faster. The latent representation output by the encoder is quantized in the latent space to reduce the complexity.

The structure of the decoder is symmetric to the encoder, with the convolutional layer replaced by a deconvolutional layer. The decoder is responsible for reconstructing the representation of the latent space into a high-resolution image. At the same time, the discriminator is used to distinguish whether the input image is from a real dataset or a fake image generated by the generator. The generator and discriminator compete against each other through adversarial training, with the generator trying to create increasingly realistic images and the discriminator trying to better distinguish between these natural and fake images.

VQGAN is a variant based on Vector Quantised Variational AutoEncoder (VQVAE), which utilizes discrete bottlenecks as the latent space for reconstruction. It utilizes GAN loss to improve reconstruction quality and the bottleneck compression rate. We follow the VQGAN architecture for 3D CT generation, replacing its 2D convolution operations with 3D convolutions. The main ideas are introduced as follows: Given a three-dimensional image $x \in \mathbb{R}^{T\times H\times W\times 3}$, VQVAE consists of an encoder $f_{\mathcal{E}}$ and a decoder $f_{\mathcal{G}}$. The discrete latent space representation $z = q\left(f_{\mathcal{E}}(x)\right) \in \mathbb{Z}^{t\times h\times w}$ is calculated using the quantization operation q, which applies a nearest-neighbor search algorithm to the mapping, where the trainable codebook $c_z \in \mathbb{R}^{t\times h\times w\times c}$ is used as a reference. The token embedding is then fed into the decoder to reconstruct the input $\hat{x} = f_{\mathcal{G}}(c_z)$. The training loss of VQVAE is as (1).

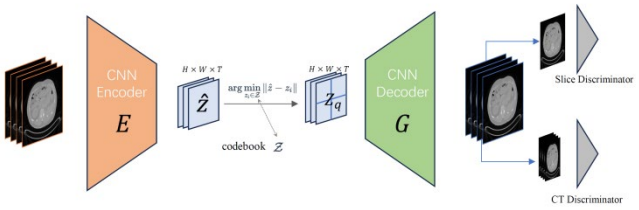

Fig. 1. Our proposed CT image reconstruction model framework

$$\mathcal{L}_{\text{vqvae}} = \underbrace{\| x - \hat{x}\|_1}_{\mathcal{L}_{\text{rec}}} + \underbrace{\| \text{sg}[f_{\mathcal{E}}(x)] - c_z\|_2^2}_{\mathcal{L}_{\text{codebook}}} + \beta \underbrace{\| \text{sg}[c_z] - f_{\mathcal{E}}(x)\|_2^2}_{\mathcal{L}_{\text{commit}}} \quad (1)$$

where sg represents the stop gradient operation. This article follows the VQGAN method and utilizes β=0.25. At the same time, it uses EMA update to optimize script cap L sub codebook and a straight-through gradient estimator to avoid the non-differentiable quantization step q.

Inspired by the GAN-based video generation model proposed by Wang et al.[29], we design two types of discriminators in our CT image reconstruction model (Fig. 1). The spatial discriminator $f_{\mathcal{D}_s}$ accepts randomly reconstructed slices $\hat{x}_i \in \mathbb{R}^{H\times W}$ to encourage the quality of slice generation, while the temporal discriminator $f_{\mathcal{D}_t}$ accepts the entire reconstructed 3D image $\hat{x} \in \mathbb{R}^{H\times W\times T}$ to penalize abnormal movements:

$$\mathcal{L}_{disc} = log\, f_{\mathcal{D}_{s/t}}(x) + log\,(1 - f_{\mathcal{D}_{s/t}}(\hat{x})) \quad (2)$$

Based on this, the training objectives of our overall reconstruction model are as follows:

$$\begin{aligned}\mathcal{C} = & \min_{f\mathcal{E},f\mathcal{G},\mathcal{C}} (\lambda_{\text{rec}}\mathcal{C}_{\text{rec}} + \mathcal{C}_{\text{codebook}} + \beta\mathcal{C}_{\text{commit}}) \\ & + \min_{f\mathcal{E},f\mathcal{G},\mathcal{C}} (\max_{f\mathcal{D}_s,f\mathcal{D}_t} \lambda_{\text{disc}}\mathcal{C}_{\text{disc}}))\end{aligned} \quad (3)$$

In addition to the VQGAN reconstruction model, the sample generation of the latent space is completed by DDPM. After the reconstruction model VQGAN is trained, the three-dimensional CT image is compressed into a latent space representation $z = q(f_{\mathcal{E}}(x))$. Afterward, a diffusion model is applied to perform regression modeling on the data distribution of this representation.

Our model is mainly based on the denoising diffusion probability model DDPM. The following is a detailed review of the theoretical derivation. First, we define the data distribution $x_0 \sim q(x_0)$ and the Markov noise process q, which gradually adds Gaussian noise to the data to produce noisy samples x_1 to x_T. Each step of the Markov noise process q adds Gaussian noise according to the variance table based on β_t:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \boldsymbol{I}) \quad (4)$$

where t represents the total step size of process q. Furthermore, we do not need to repeatedly sample from $x_0 \sim q(x_t|x_0)$. Due to the step-by-step recursive nature of the formula(4), $q(x_t|x_0)$ can be expressed as:

$$\begin{aligned} q(x_t|x_0) &= \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\, I) \\ &= \sqrt{\bar{\alpha}_t}x_0 + \epsilon\sqrt{1-\bar{\alpha}_t}, \epsilon \sim \mathcal{N}(0, I)\end{aligned} \quad (5)$$

Among them, $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$. where $1 - \alpha_t$ represents the noise variance at any time step, and we can equivalently use it to define noise scheduling instead of β_t. Using Bayes' theorem, Ho et al.[26] found that the posterior $q(x_{t-1} \mid x_t, x_0)$ is a Gaussian distribution with a mean of $\tilde{\mu}_t(x_t, x_0)$, the variance is $\tilde{\beta}_t$, defined as follows:

$$\begin{aligned} \tilde{\mu}_t(x_t, x_0) &:= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \\ \tilde{\beta}_t &:= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \\ q(x_{t-1}|x_t, x_0) &= \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \boldsymbol{I}) \end{aligned} \quad (6)$$

If we want to sample from the data distribution $q(x_0)$, we can first sample from $q(x_T)$ and then reversely sample $q(x_{t-1} \mid x_t)$ until we reach x_0. When β_t and T are reasonably set, the distribution $q(x_T)$ is almost consistent with the isotropic Gaussian distribution, so there is no need to sample x_T. Further, a neural network is used to approximate $q(x_{t-1} \mid x_t)$, because it cannot be accurately calculated by formulas when the data distribution is unknown. Therefore, Sohl-Dickstein et al.[21] demonstrated that $q(x_{t-1} \mid x_t)$ is approximately a diagonal Gaussian distribution at $T \to \infty$ and the corresponding $\beta_t \to \infty$. It is sufficient to train a neural network to predict the mean of μ_θ and the diagonal covariance matrix Σ_θ:

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \boldsymbol{\Sigma_\theta}(x_t, t)) \quad (7)$$

In order to let $p(x_0)$ learn the real data distribution $q(x_0)$, we can optimize the following variational lower bound L_{vlb} for $p_\theta(x_0)$:

$$\begin{aligned} L_{vlb} &:= L_0 + L_1 + \ldots + L_{T-1} + L_T \\ L_0 &:= -\log p_\theta(x_0|x_1) \\ L_{t-1} &:= D_{KL}(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)) \\ L_T &:= D_{KL}(q(x_T|x_0) \parallel p(x_T)) \end{aligned} \quad (8)$$

Although the above goals are reasonable, different variational bounds goals will produce better samples in practice. Therefore, we directly parameterize $\mu_\theta(x_t, t)$ as a neural network and train a model $\epsilon_\theta(x_t, t)$ to predict ϵ in formula (5). This simplified variational boundary objective is defined as follows:

$$L_{\text{simple}} := E_{t\sim[1,T], x_0\sim q(x_0), \epsilon\sim\mathcal{N}(0,I)}[||\epsilon - \epsilon_\theta(x_t, t)||^2] \quad (9)$$

When sampling, we use the substitution method to deduce $\mu_\theta(x_t, t)$ from $\epsilon_\theta(x_t, t)$:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) \quad (10)$$

L_{simple} does not provide any learning signal for $\Sigma_\theta(x_t, t)$. Instead of learning $\Sigma_\theta(x_t, t)$, we can fix it to a constant, choosing $\beta_t I$ or $\widetilde{\beta_t} I$. These values correspond to upper and lower bounds on the true inverse process variance.

For this simplified goal, we just need to find a way to calculate $\epsilon - \epsilon_\theta(x_t, t)$. We choose Mean Squared Error (MSE) as the loss function of DDPM:

$$MSELoss = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i) \quad (11)$$

Among them, y_i and $\hat{y}_i$ represent the real noise value and the model prediction value, respectively.

To evaluate the quality of the generated 3D CT images, we need to evaluate the fidelity and diversity of CT slices, as well as the consistency between CT slices. The FID[30] index is a metric used to evaluate the effectiveness of generative models. FVD[31] is an indicator used to evaluate the quality of generated videos. Three-dimensional CT images can be regarded as continuous videos with CT slices as frames. Thus, we use FVD to assess the generation quality of three-dimensional CT images.

Tables I and II show some network parameters of VQGAN and DDPM in our model.

TABLE I. PARAMETERS USED IN THE VQGAN NETWORK.

Parameter	Value
Image size	256×256×128
The number of basic channels in the hidden layer	16
Downsampling scale	[2,2,2]
The number of vectors in the codebook	16384
Dimensions of vectors in the codebook	8
Attention mechanism type	axial
Number of attention heads	2

TABLE II. PARAMETERS USED IN THE DDPM NETWORK.

Parameter	Value
Image size	64×64×32
The number of basic channels in the hidden layer	256
Number of residual modules per layer	2
Number of attention heads	4
Whether the model learns the variance of the noise	true

III. RESULTS

A. Datasets and Preprocessing

The MICCAI FLARE 2022 challenge[32] provides a large-scale data set for abdominal CT. The challenge is a multi-organ medical image segmentation competition. The anatomical organs include the liver, spleen, pancreas, right kidney, left kidney, stomach, gallbladder, esophagus, aorta, inferior vena cava, right adrenal gland, left adrenal gland, and duodenum. It contains 2,300 three-dimensional CT image data from more than 20 centers, including 50 labeled images and 2,000 unlabeled images. A total of 2,050 images are publicly available and can be used to train our generative model.

Since most images have inconsistent resolutions, all CT images are resampled at intervals of (0.72mm, 0.72mm, 2.5mm) in the x, y, and z directions. This sampling scale is determined by the average resolution obtained from the entire data set. The abdominal area is then cropped out at 320×320×160 and then resized to 128×128×64. Additionally, the range of intensity value in CT images is too large, making it difficult to input it directly into the model for effective training. Thus, all images are normalized to between -1 and 1. Before normalization, we removed intensity values less than 0.2% and greater than 99.8, to reduce the impact of outliers. Furthermore, we use a random 50% probability image inversion method for data enhancement.

B. Reconstruction Model Loss Function

The reconstruction of the VQGAN model is very important for the entire CT image generation framework. The design of the loss function plays a major role in improving reconstruction accuracy. $\mathcal{L}_{\mathrm{rec}}, \mathcal{L}_{\mathrm{cb}}, \mathcal{L}_{\mathrm{cm}}$ and $\mathcal{L}_{\mathrm{disc}}$ represent reconstruction loss, codebook consistency loss, encoder commitment loss and discriminator loss, respectively. Their impact on generative model training is shown in Table III.

TABLE III. THE IMPACT OF DIFFERENT LOSS FUNCTIONS ON THE GENERATION EFFECT.

loss function	$\mathrm{FID_{xy}}$	$\mathrm{FID_{xz}}$	$\mathrm{FID_{yz}}$	FVD
$\mathcal{L}_{\mathrm{rec}}$	38.73	40.42	42.43	473.43
$\mathcal{L}_{\mathrm{rec}}+\mathcal{L}_{\mathrm{disc}}$	23.83	25.91	24.68	368.36
$\mathcal{L}_{\mathrm{rec}}+\mathcal{L}_{\mathrm{cb}}+\mathcal{L}_{\mathrm{cm}}$	31.35	29.96	32.53	293.34
$\mathcal{L}_{\mathrm{rec}}+\mathcal{L}_{\mathrm{cb}}+\mathcal{L}_{\mathrm{cm}}+\mathcal{L}_{\mathrm{disc}}$	20.45	21.57	20.34	252.97

$\mathrm{FID_{xy}}$, $\mathrm{FID_{xz}}$ and $\mathrm{FID_{yz}}$ are used to evaluate the FID indicators in the three plane dimensions of XY, XZ, YZ respectively, where$\mathrm{FID_{xy}}$ is used to evaluate the generation effect of CT slices and $\mathrm{FID_{xz}}$ and $\mathrm{FID_{yz}}$ are used to evaluate the overall effect between slices to a certain extent.

A network trained only based on reconstruction loss can be regarded as a VQVAE network. Table III shows that the pure VQVAE network performs poorly in generating effect evaluation indicators. Compared with the pure reconstruction loss, after introducing the discriminator loss of the GAN network mechanism, the $\mathrm{FID_{xy}}$, $\mathrm{FID_{xz}}$ and $\mathrm{FID_{yz}}$ indices decreased from 38.73, 40.42, and 42.43 to 23.83, 25.91, and 24.68, respectively. The FVD index decreased from 473.43 to 368.36, demonstrating the importance of the adversarial learning mechanism. Furthermore, compared to pure reconstruction loss, codebook consistency loss and encoder commitment loss are mechanistically responsible for consistent translation alignment between latent space representations and codebooks. After applying the two loss functions, the $\mathrm{FID_{xy}}$, $\mathrm{FID_{xz}}$, $\mathrm{FID_{yz}}$ and FVD metric scores decrease from 38.73, 40.42, 42.43, and 473.43 to 31.35, 28.96, 35.53, and 293.34, respectively, which also demonstrates that alignment consistency between the latent spatial representations and the codebook is necessary. After combining the above two mechanisms, the $\mathrm{FID_{xy}}$, $\mathrm{FID_{xz}}$, $\mathrm{FID_{yz}}$ and FVD indicators are further decreased to 20.45, 21.57, 20.34, and 252.97, which are also the optimal results of the method in this paper. FVD is used to evaluate the generation effect of three-dimensional volume images directly. The numerical range is much higher than the FID indicator of two-dimensional images, but the performance trend is broadly consistent with the appeal results. When the optimal loss function design is applied, the generative model can generate CT images with better quality. The model in this article can generate diverse and high-quality three-dimensional images, as shown in Fig. 2. The first row is an example of a slice of the generated image in the XY plane. Our model generates vivid details and organ features. The second row is an example of slices on the XZ and YZ planes. Our model not only generates vivid images on the XY plane but also shows good coherence in the z-axis direction.

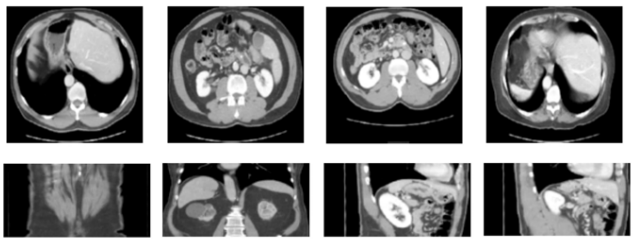

Fig. 2. CT images examples generated by our model.

C. The Impact of Latent Space Dimensions on Generation Effects

The dimensionality of the latent space determines the image features and complexity that the model can represent. Higher-dimensional latent spaces usually have greater expressive power and can capture richer and more complex image features, and the generated images may be more detailed and diverse. We found that when each spatial dimension is compressed by a factor of 16, that is, the potential dimensionality of a 128×128×64 image is 8×8×4, the image generation quality is seriously affected. When each spatial dimension is compressed by a factor of 8, that is, the potential dimensionality of a 128×128×64 image is 64×64×32, relevant image features are lost. When training the VQ-GAN autoencoder, the compression factor is four times, that is, the potential dimension of the 128×128×64 image is 32×32×16, and the generation of image features is more accurate.

TABLE IV. THE IMPACT OF DIFFERENT COMPRESSION COEFFICIENTS OF THE ENCODER ON THE GENERATED INDICATORS

compression coefficients	FID_{xy}	FID_{xz}	FID_{yz}	FVD
16	52.75	49.38	56.82	373.54
8	27.37	27.82	29.76	289.74
4	20.45	21.57	20.34	252.97

The dimensionality of the latent space determines the image features and complexity that the model can represent. Higher-dimensional latent spaces usually have greater expressive power and can capture richer and more complex image features, and the generated images may be more detailed and diverse.

IV. CONCLUSIONS

Medical image generation plays a vital role in all aspects of healthcare, computer-aided diagnosis, medical research, and education. Its importance stems from its ability to replicate real-world medical imaging scenarios and help develop and validate medical imaging techniques and computer-aided diagnosis algorithms. Simulation data can generate large datasets containing a variety of pathologies and imaging variations, thus improving the training and testing of deep learning models based on medical images, which is essential for robust model training.

Here we propose a latent space-based image generation method that has the potential to revolutionize medical image analysis. Our method effectively combines VQGAN and DDPM, leveraging the strengths of their respective modules. By fitting the distribution of the VQGAN latent space representation using DDPM, we maintain the consistency between CT slices while achieving high-fidelity CT image generation. This innovative approach could mark a significant leap forward in medical imaging.

In this work, the FLARE 22 dataset is used as the main dataset for training validation. We meticulously analyze the effectiveness of the generative model under different parameter settings and verify the superiority of our proposed model on the image generation task through quantitative evaluation and comparative analysis. Our comparative experiments for the loss functions used in the VQGAN training process further demonstrate the thoroughness of our research. The size of the dimension of the latent space determines the image features and complexity that the model can represent. To analyze the effect of the latent space dimension on the quality of image generation, we trained the VQ-GAN self-encoder with two different compression coefficients.

We demonstrated that the 3D average FID coefficient reaches 20.79, and the FVD coefficient reaches 252.97 under 128×128×64 size. This work is of great significance in improving the accuracy and efficiency of medical image generation and segmentation and is expected to play an important role in medical research.

ACKNOWLEDGMENT

This study was supported by the Project of the Educational Commission of Guangdong Province of China (No. 2022ZDJS113) and the Shenzhen Science and Technology Program (No. KJZD20240903095605007).

REFERENCES

[1] ROSSI A, VANNUCCINI G, ANDREINI P, et al. Analysis of brain NMR images for age estimation with deep learning [J]. Procedia Computer Science, 2019, 159: 981-989.

[2] BONECHI S, BIANCHINI M, BONGINI P, et al. Fusion of Visual and Anamnestic Data for the Classification of Skin Lesions with Deep Learning; proceedings of the International Conference on Image Analysis and Processing, F, 2019 [C].

[3] TOGNETTI L, BONECHI S, ANDREINI P, et al. A new deep learning approach integrated with clinical data for the dermoscopic differentiation of early melanomas from atypical nevi [J]. Journal of Dermatological Science, 2020.

[4] BONECHI S, ANDREINI P, MECOCCI A, et al. Segmentation of Aorta 3D CT Images Based on 2D Convolutional Neural Networks [J]. Electronics, 2021.

[5] WOLTERINK J M, DINKLA A M, SAVENIJE M H F, et al. Deep MR to CT Synthesis Using Unpaired Data; proceedings of the Simulation and Synthesis in Medical Imaging, Cham, F 2017//, 2017 [C]. Springer International Publishing.

[6] COLLINS D L, ZIJDENBOS A P. Design and construction of a realistic digital brain phantom [J]. IEEE Transactions on Medical Imaging, 1998, 17(3): P.463-468.

[7] ANDREINI P, BONECHI S, BIANCHINI M, et al. A Deep Learning Approach to Bacterial Colony Segmentation: 27th International Conference on Artificial Neural Networks, Proceedings, Part III, F, 2018 [C].

[8] ANDREINI P, BONECHI S, BIANCHINI M, et al. Image generation by GAN and style transfer for agar plate image segmentation - ScienceDirect [J]. Computer Methods and Programs in Biomedicine, 2020, 184.

[9] BONECHI S, BIANCHINI M, MECOCCI A, et al. Segmentation of Petri Plate Images for Automatic Reporting of Urine Culture Tests [J]. 2022.

[10] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [J]. Computer Science, 2013.

[11] KUGELMAN J, ALONSO-CANEIRO D, READ S A, et al. Data augmentation for patch-based OCT chorio-retinal segmentation using generative adversarial networks [J]. Neural Computing and Applications, 2021, 33(4).

[12] WAHEED A, GOYAL M, GUPTA D, et al. CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection [J]. in IEEE Access, 2021.

[13] BO H, YE T, I-CHAO C E, et al. Unsupervised Learning for Cell-level Visual Representation in Histopathology Images with Generative Adversarial Networks [J]. IEEE Journal of Biomedical and Health Informatics, 2017

[14] SHIN H C, TENENHOLTZ N A, ROGERS J K, et al. Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks [J]. Springer, Cham, 2018.

[15] MADANI A, MORADI M, KARARGYRIS A, et al. Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation, F, 2018 [C].

[16] SRIVASTAV D, BAJPAI A, SRIVASTAVA P. Improved Classification for Pneumonia Detection using Transfer Learning with GAN based Synthetic Image Augmentation; proceedings of the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), F, 2021 [C].

[17] NASH C, WILLIAMS C K I. The shape variational autoencoder: A deep generative model of part-segmented 3D objects [J]. Computer Graphics Forum, 2017, 36(5): 1-12.

[18] OORD A V D, KALCHBRENNER N, KAVUKCUOGLU K. Pixel Recurrent Neural Networks [J]. 2016.

[19] KINGMA D P, DHARIWAL P. Glow: Generative Flow with Invertible 1x1 Convolutions [J]. 2018.

[20] GOODFELLOWIAN, POUGET-ABADIEJEAN, MIRZAMEHDI, et al. Generative adversarial networks [J]. Communications of the ACM, 2020.

[21] SOHL-DICKSTEIN J, WEISS E A, MAHESWARANATHAN N, et al. Deep Unsupervised Learning using Nonequilibrium Thermodynamics [J]. JMLRorg, 2015.

[22] ESSER P, ROMBACH R, OMMER B. Taming Transformers for High-Resolution Image Synthesis [J]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 12868-12878.

[23] NIE D, TRULLO R, LIAN J, et al. Medical Image Synthesis with Context-Aware Generative Adversarial Networks [J]. 2017.

[24] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-Resolution Image Synthesis with Latent Diffusion Models [J]. 2021.

[25] DHARIWAL P, NICHOL A. Diffusion Models Beat GANs on Image Synthesis [J]. 2021.

[26] HO J, JAIN A, ABBEEL P. Denoising Diffusion Probabilistic Models [J]. ArXiv, 2020, abs/2006.11239.

[27] FRID-ADAR M, KLANG E, AMITAI M, et al. Synthetic data augmentation using GAN for improved liver lesion classification; proceedings of the International Symposium on Biomedical Imaging, F, 2018 [C].

[28] CHUQUICUSMA M J M, HUSSEIN S, BURT J, et al. How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis; proceedings of the International Symposium on Biomedical Imaging, F, 2018 [C].

[29] WANG Y, BILINSKI P, BREMOND F, et al. ImaGINator: Conditional Spatio-Temporal GAN for Video Generation; proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), F, 2020 [C].

[30] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium [J]. 2017.

[31] UNTERTHINER T, STEENKISTE S V, KURACH K, et al. Towards Accurate Generative Models of Video: A New Metric & Challenges [Z]. 2018

[32] MA J, ZHANG Y J, GU S, et al. Unleashing the Strengths of Unlabeled Data in Pan-cancer Abdominal Organ Quantification: the FLARE22 Challenge [J]. ArXiv, 2023, abs/2308.05862.