Manuscript Click here to view linked References

FQCDM: Feature Quantization-Based Cardiac Image Diffusion Synthesis Model

Jiahui Shi[a], Bo Chen[a], Jiacheng Liu[a], Weigang Lu[c,d], Shugang Zhang[d], Weimin Ma[e,*], Fei Yang [a,b,*] and Dong Li[a,b]

[a]School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China
[b]Shandong Key Laboratory of Intelligent Electronic Packaging Testing and Application, Shandong University, Weihai, China
[c]Department of Educational Technology, Ocean University of China, Qingdao, China
[d]Department of Computer Science and Technology, Ocean University of China, Qingdao, China
[e]Department of Intensive Care Unit, Limin Hospital of Weihai High District, Weihai, China

ARTICLE INFO

Keywords:
Cardiac Image Synthesis
Generative Model
Denoising Diffusion Probabilistic Model
Feature Quantization
Dual-branch Discriminator

ABSTRACT

Background and Objective: Deep learning-based cardiac image segmentation algorithms, including those using CNNs and Transformers, have demonstrated superior accuracy compared to traditional methods. However, segmentation performance is often hindered by the limited availability of high-quality cardiac datasets, which are constrained in size, suffer from class imbalance, and are prone to artifacts. To address these limitations, this study proposes the Feature Quantization-based Cardiac Diffusion Model (FQCDM), aiming to enhance synthetic cardiac data generation and improve segmentation tasks.
Methods: The proposed FQCDM integrates feature quantization into the Denoising Diffusion Probabilistic Model (DDPM) to synthesize high-quality cardiac images. Real image labels guide the reverse denoising process to improve semantic focus, while a dual-branch discriminator evaluates both global features and edge details for better quality control. The study explores three segmentation training strategies using synthetic data: (1) mixed training with real and synthetic images, (2) self-supervised pretraining with synthetic data, and (3) integration of synthetic images with traditional data augmentation.
Results: Comparative experiments demonstrated that FQCDM-generated synthetic cardiac images excel in quality, diversity, and distribution similarity compared to baseline methods. Corresponding ablation studies confirm that, when used with appropriate strategies, synthetic data can enhance the segmentation performance of cardiac models.
Conclusions: The FQCDM framework effectively addresses challenges in cardiac image segmentation by generating high-quality synthetic data, thereby mitigating dataset limitations and improving segmentation accuracy. The findings highlight the potential of synthetic cardiac data in advancing medical imaging tasks, offering valuable insights for clinical applications and theoretical research.

1. Introduction

Cardiac image segmentation aims to automatically delineate and label various semantic structures within cardiac images, facilitating more accurate diagnoses by physicians when combined with medical expertise. However, the inherent complexities of cardiac structures and noise interference [1], among other factors, expose traditional image segmentation methods based on prior rules and feature engineering to several challenges, including poor adaptability to different imaging modalities, high sensitivity to noise and artifacts, and strong coupling with prior knowledge. In contrast, deep learning algorithms excel in feature learning and representation, enabling effective automatic extraction of complex latent features and patterns from large datasets. Compared to traditional methods, deep learning-based cardiac image segmentation algorithms exhibit greater robustness, enhanced accuracy, and reduced dependence on prior knowledge.

Despite the superior performance of deep learning-based cardiac image segmentation algorithms across various cardiac datasets, several challenges persist. Firstly, annotating cardiac image datasets requires significant professional resources, resulting in a scarcity of labeled data. Secondly, existing datasets often suffer from class imbalance and the presence of artifacts. Furthermore, the acquisition and use of medical data must comply with relevant laws and regulations, necessitating strict protection of sensitive information within the images. These factors can significantly impact the generalization ability and segmentation accuracy of cardiac segmentation models. Synthetic images hold substantial promise for data augmentation, enhancing rare cases, and ensuring privacy protection, rendering synthetic medical data an important area of research in contemporary medical image processing tasks. The exploration of high-quality cardiac image synthesis methods to enhance segmentation performance or to apply in other tasks is an urgent need in the field of cardiac medical imaging, potentially aiding physicians in obtaining more medically defined rare cardiac

*Corresponding authors: Fei Yang (feiyang@sdu.edu.cn) and Weimin Ma (maweimin4562@163.com) are jointly responsible for correspondence regarding this work.
202417561@mail.sdu.edu.cn (Jiahui Shi); markcbo2021@gmail.com (Bo Chen); 202200800027@mail.sdu.edu.cn (Jiacheng Liu); luweigang@ouc.edu.cn (Weigang Lu); zsg@ouc.edu.cn (Shugang Zhang); maweimin4562@163.com (Weimin Ma); feiyang@sdu.edu.cn (F.); dongli@sdu.edu.cn (Dong Li)
ORCID(s): 0000-0003-0342-133X (Jiahui Shi)

examples, which is crucial for clinical guidance and theoretical research.

However, early efforts in synthesizing medical data displayed significant discrepancies from real data in terms of physiological shape, intensity, size, and texture, rendering synthetic data unsuitable for model training. In recent years, innovations in natural image generation methods have contributed to improved quality and diversity of medical synthetic data. This paper leverages the strengths of DDPM for high-quality image generation and VQGAN for quantizing discrete features, producing synthetic cardiac data that excels in image quality, structural texture, diversity, and similarity. This approach addresses the issues of scarce and costly labeled cardiac image datasets, semantic class imbalance, and noise interference, ultimately enhancing the performance of existing segmentation methods. The innovations presented in this study are as follows:

(1) This paper proposes the Feature Quantized Cardiac Diffusion Model (FQCDM), a cardiac image synthetic framework based on DDPM and feature quantization. FQCDM utilizes a denoising probabilistic model guided by real label maps to generate compressed latent features, which are then quantized to produce high-quality synthetic cardiac images.
(2) A dual-branch GAN discriminator (DBD) is designed to simultaneously evaluate both the global features and edge information of synthetic images, thereby further enhancing the quality and diversity of the generated cardiac images.
(3) The paper outlines three strategies for applying synthetic cardiac images in cardiac segmentation tasks. We implement mixed training by combining real and synthetic cardiac images in varying proportions. Additionally, we propose a pre-training strategy utilizing synthetic cardiac data within self-supervised tasks. Lastly, we design a strategy that integrates synthetic images with data augmentation for model training. Experimental results demonstrate that the proposed data-driven strategies effectively enhance the segmentation performance of existing cardiac segmentation networks.

The rest of the paper is organized as follows. Section 2 reviews relevant work and recent advancements in the field of medical image segmentation, medical image generation, and their applications. Section 3 provides a detailed explanation of our proposed method. Section 4 describes the datasets used in the experiments and presents a comprehensive explanation of the experimental procedure, along with an extensive evaluation of the proposed method. Section 5 presents three data-driven training strategies, while Section 6 discusses the experimental results and the performance of our approach. Finally, Section 7 offers concluding remarks on the findings of the paper.

2. Related work

2.1. Deep learning-based medical image segmentation

Before the mainstream adoption of deep learning, traditional medical image segmentation relied on techniques such as (1) threshold-based, (2) texture and morphology-based, (3) anatomy-based, and (4) level set model-based methods, which laid the groundwork for subsequent advances.

Since 2016, deep learning breakthroughs have continually improved medical image segmentation. Ronneberger et al. introduced U-net[2], based on fully convolutional networks[3]. U-net's encoder-decoder architecture with skip connections set a standard for subsequent models in segmentation tasks. However, simple concatenation in skip connections limits feature fusion, potentially leading to information loss. Zhou et al.[4] enhanced U-net with dense feature reorganization in U-net++ to improve multi-scale feature utilization, achieving high segmentation accuracy. Ibrahim et al.[5] successfully employed U-net++ for myocardium segmentation. To further enlarge the receptive field, Sander et al.[6] proposed DCNN, incorporating dilated convolutions.

To address the issue that medical image segmentation requires more focus on the region of interest(ROI), Oktay et al. proposed Attention U-net[7], which uses an attention gate model to focus on targets of varying shapes and sizes while masking irrelevant areas of the input image, thereby highlighting salient features important for specific tasks. This led to the development of models that balance increasing the receptive field for multi-scale feature extraction with attention mechanisms.

With the introduction of Vision Transformers (ViT)[8], Transformer-based models emerged as powerful alternatives, enabling better long-range dependencies. As a result, many researchers began exploring the combination of ViT and U-net for medical image segmentation. Chen et al.[9] proposed TransUnet, which encodes CNN feature maps as input sequences to extract global context while using a CNN decoder to cascade upsample the features. This approach demonstrated the strong potential of Transformers as encoders in cardiac image segmentation. Cao et al.[10] introduced the hierarchical Swin Transformer[11], which replaced the ViT-based encoder and CNN-based decoder in TransUnet, creating a pure Transformer architecture known as Swin-UNet. Subsequent works combining Vision Transformers and CNNs in cardiac segmentation have emerged, including models that continuously improved segmentation accuracy on the ACDC dataset, such as MISSFormer[12], CASCADE[13], nnFormer[14], G-CASCADE[15], MERIT[16], MIST[17], SwinUNETR[18], and FCT[19].

Unlike existing research that focuses on improving model architectures, our work proposes leveraging synthetic medical data to address challenges such as the scarcity of cardiac image datasets, class imbalance, and privacy protection. By generating high-quality synthetic cardiac images, we

effectively enhance the model's generalization capability while mitigating the challenges posed by high annotation costs and data scarcity, all while ensuring patient privacy.

2.2. Medical image generation and applications of generated images

Methods for image generation can be classified into four main types: (1) Variational Autoencoder (VAE)-based methods; (2) Generative Adversarial Network (GAN)-based methods; (3) Diffusion Model-based methods; and (4) Flow-based methods. Each of these approaches has its advantages in learning latent data distributions, generating diverse samples, producing high-resolution and high-quality samples, and reversibly controlling the image generation process. In terms of diversity and image generation quality, most medical image generation work focuses on GANs and Diffusion Models.

Before the rise of DDPM[20], GAN-based methods dominated medical organ image synthesis tasks. Kwon et al.[21] successfully generated various types and modalities of 3D MRI brain images from limited training data using GAN networks. Dragan et al.[22] introduced Deep Convolutional GAN (DCGAN) to generate diverse diabetic retinopathy (DR) images, addressing the overfitting issue caused by imbalanced datasets in DR classification. Saad et al.[23] proposed an attention-guided multi-scale gradient GAN (MSG-SAGAN), generating diverse X-ray images by simulating the long-range dependencies of biomedical image features.

Although GANs ensure image diversity, they suffer from mode collapse, unstable training, and convergence issues. Additionally, they cannot fully address the inherent noise and artifacts present in medical images. Due to the excellent properties of diffusion models in generating high-quality samples through denoising, they have gained more attention in recent years. To balance high-quality generation with computational resource consumption, various approaches[24, 25, 26, 27, 28, 29, 30, 31, 32, 33] based on DDPM or latent diffusion models[34] have combined cascaded operations, text prompts, conditional guidance, wavelet transforms, and other modules to generate cross-modal chest X-rays, 3D brain MRI images, and 3D cardiac MRI images. These methods have been widely applied to downstream tasks such as image denoising, reconstruction, object detection, image pairing, and registration.

In terms of applying synthetic images to medical image segmentation, Khader et al.[35] proposed the Medical Diffusion network, modifying the 2D-Unet denoising diffusion network in DDPM to a 3D-Unet network, successfully synthesizing CT and MRI data for various organs. Khader et al. demonstrated that synthetic images used for self-supervised pre-training improved chest image segmentation model performance when real datasets were scarce. Saragih et al.[36] used DDPM to generate polyp image data, resulting in more realistic synthetic samples. Applying synthetic data to the polyp segmentation task proved that using synthetic data to train segmentation models can enhance their performance. Xu et al.[37] proposed the TDASD method based on stable diffusion to address the challenges of limited spread through air spaces data in lung cancer diagnosis. The method they proposed not only safeguards patient privacy but also enhances the diversity of medical images under limited data conditions. Du et al.[38] proposed an adaptive semantic refinement diffusion model (ArSDM) to generate colonoscopy images favorable for polyp segmentation and detection tasks, and experiments showed that the data generated by ArSDM could significantly improve the performance of baseline segmentation and detection methods. Khosravi et al[39]. demonstrated the effectiveness of using denoising diffusion probabilistic models (DDPMs) as feature extractors to enhance medical image segmentation in a few-shot setting. Current work on cardiac data synthesis mainly focuses on cardiac image pairing and reconstruction tasks, with applications of DDPM-based cardiac image segmentation awaiting further exploration. In this regard, we propose a cardiac image generation framework based on DDPM and feature quantization. This framework employs a DDPM guided by real label maps to generate compressed latent features, which are then quantized to produce high-quality synthetic cardiac images. The network effectively generates synthetic cardiac data excelling in image quality, structural texture, diversity, and similarity.

3. Methods

3.1. Overview

Fig.1 illustrates Feature Quantized Cardiac Diffusion Model (FQCDM) framework proposed in this paper for cardiac image synthesis. The model primarily consists of three modules: the Feature Diffusion Generation Block (FDGB), the Feature Quantized Vector Block(FQB), and the DBD. The FQB module includes not only the standard encoder and decoder but also a Feature Vector Quantization Block (FQB). The Double Branch Discriminator is made up of two components: PatchGAN, responsible for global feature discrimination, and EdgeGAN, which focuses on edge feature detection.

Let I represent the input cardiac image, and I_{syn} represent the synthetic cardiac image output by the framework. The FQCDM generation process can be expressed as in (1):

$$I_{syn} = H_{FQCDM}(I), \tag{1}$$

where $H_{FQCDM}()$ represents the FQCDM network architecture. Let E denote the encoder and D denote the decoder, the feature extraction process F by the encoder can be expressed as in (2):

$$F = E(I), \tag{2}$$

where $E()$ denotes the feature extraction process. The process of generating new features F' from the features F using the Feature Diffusion Generation Block (FDGB) can be expressed as in (3):

$$F' = H_{FDGB}(F), \tag{3}$$

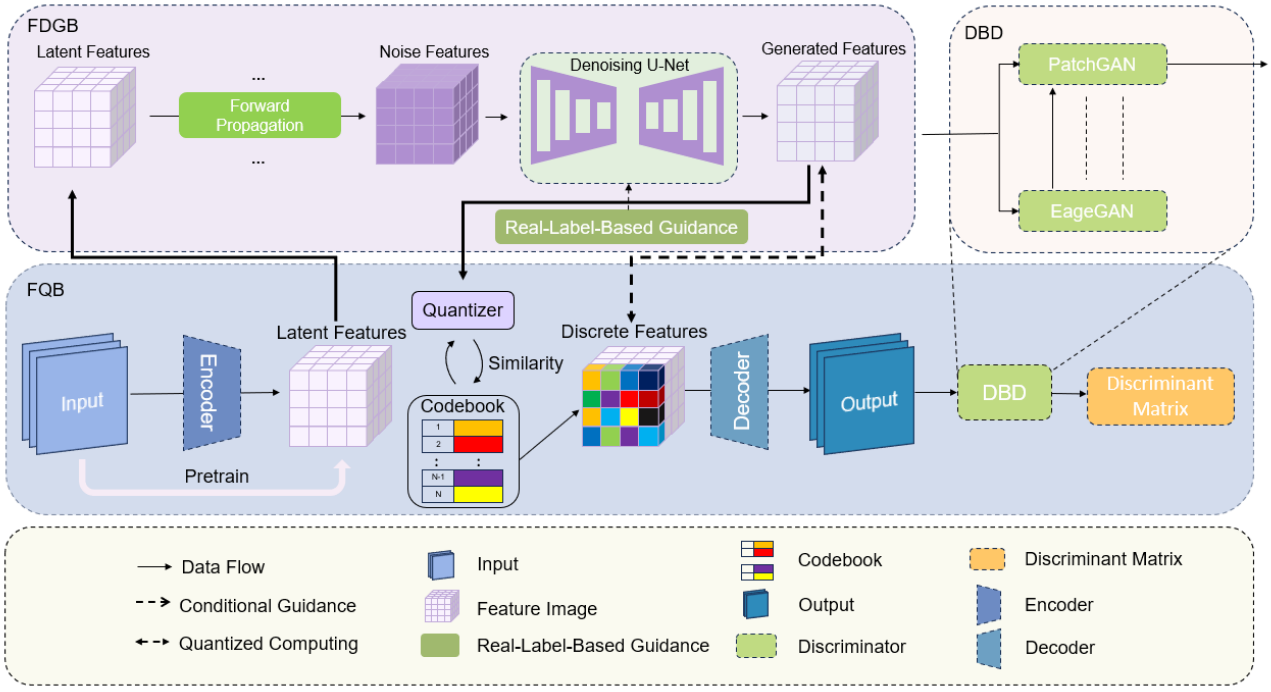

Figure 1: Overview of our proposed FQCDM framework

where $H_{FDGB}()$ represents the FDGB module. The new features $F^{'}$ are input to the codebook and, after vector quantization, the discrete features F_q are input to the decoder. This process can be expressed as in (4):

$$F_q = H_{FQB}(F^{'}). \tag{4}$$

The process of decoding the discrete features output by the codebook into an image using the decoder D can be expressed as follows:

$$I_{syn} = D(F_q), \tag{5}$$

where $D()$ represents the feature decoding process. The process of generating the discrimination matrix G by the discriminator can be expressed by the equation:

$$G = H_{DBD}(H_{PatchGAN}(I_{syn}, I)||H_{EageGAN}(I_{syn,1})), \tag{6}$$

where $H_{DBD}()$ represents the DBD module, $H_{PatchGAN}()$ represents the PatchGAN module, and $H_{EageGAN}()$ represents the EageGAN module.

3.2. Feauture quantization

Feauture Quantization maps latent feature vectors to the codebook for quantization and further improves image reconstruction quality by applying GAN discriminator loss on the output image. In this block the image $I \in R^{H\times W\times C}$ is input into the encoder to obtain the compressed latent feature $F \in R^{(H/s)\times(W/s)\times m}$, where H is the image height, W is the image width, and C is the number of channels. m represents the number of latent feature maps, and s represents the compression scale. In the feature quantization (FQ) step, each latent feature vector is replaced by the closest corresponding codebook vector, and then the quantized feature vectors are input into the decoder G for reconstruction. The loss function of Feature Quantization Blcok is a joint function of reconstruction loss, codebook quantization error, and GAN discriminator error, expressed as in (7). Here, reconstruction loss uses perceptual loss, codebook quantization error uses mean squared error, and the GAN discriminator uses PatchGAN[40]. λ is the weight that adaptively controls the error.

$$L_{FQB} = L_{reconstruct} + L_{codebook} + \lambda L_{GAN}. \tag{7}$$

The codebook is based on a Transformer architecture for image synthesis. FQB scans each pixel of the image in a top-left to bottom-right order, predicting the reconstruction of the i-th pixel S_i based on the sequence of preceding pixels $S_{<i}$. This sequential prediction enables the model to effectively capture contextual dependencies within the image.

3.3. Conditioned label-guided DDPM continuous latent feature generation

Unlike unconditional image synthesis methods, this study aims for the segmentation results of the generated synthetic cardiac images to closely align with the ground truth annotations of the input images, effectively "replacing" the original data to some extent. To achieve this objective, we designed the FDGB.

Initially, we normalize the latent features $F_0 \in R^{h\times w\times c}$ obtained from the encoder, then input them into the DDPM

to generate the features $F' \in R^{h\times w\times c}$. During the forward diffusion process, a series of noisy images F_1,F_2,…,F_T are generated by adding Gaussian noise to F_0 at continuous time steps $t = 1, 2, \ldots, T$, as expressed in the following equation:

$$F_t = \sqrt{\bar{\alpha}_t}F_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{8}$$

where the sampling noise $\epsilon \sim N(0, I)$,$\alpha_t = 1 - \beta_t$, where β_t is a predefined hyperparameter that gradually decays over time, with $\bar{\alpha}_t = \prod_{i=1}^{T}\alpha_i$. Subsequently, F_T and time step t are input into the reverse denoising network ϵ_θ, which in this study is implemented as a 2D U-net, through its encoder E. The true label $L_0 \in R^{H\times W}$ is then resized to dimensions $l \in R^{h\times w}$ and embedded into the decoder D. The reverse denoising process is described as in (9):

$$F' = \epsilon_\theta(F_t, t, l) = D(E(F_t, t), l). \tag{9}$$

The loss function for the denoising training process, L_{FDGB}, is given by the following equation:

$$L_{FDGB} = \mathbb{E}_{t,x_0,\epsilon\sim N(0,I)}\left[\left\|\epsilon - \epsilon_\theta\left(F_t, t, l\right)\right\|^2\right]. \tag{10}$$

The pseudocode algorithm for the FDGB training process is shown in Algorithm 1.

Algorithm 1 FDGB Module Process.

Input: Time step $t \in \{1, T\}$, sampling noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, real labels L_0, input features F_0 .

Output: Denoised model $\bar{\epsilon}$, generated features F'.

$F_t = \sqrt{\bar{\alpha}_t}F_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ ▷ Forward Diffusion Process

for $i = T, T-1, \ldots, 1$ **do** ▷ Reverse Denoising Process

  $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ **if** $i > 1$ **else** $z = 0$

  $\bar{F}_{i-1} = \frac{1}{\sqrt{\bar{\alpha}_i}}\left(\bar{F}_i - \frac{1-\alpha_i}{\sqrt{1-\bar{\alpha}_i}}\epsilon_\theta(\bar{F}_i, i, L_0)\right) + \sigma_i z$

end for

$F' = R(\bar{F}_0)$

return F'

3.4. Dual-branch cardiac image discriminator based on GANs

In the original GAN architecture, the discriminator outputs a single evaluation value that indicates whether the synthetic image is real or synthetic. This value serves as a judgment for the entire image produced by the generator. In contrast, PatchGAN utilizes convolutional operations to map the input to an $N \times N$ matrix, replacing the single evaluation value found in traditional GANs. Each point in this $N \times N$ matrix (true/false) corresponds to the evaluation of a small region, or patch, within the original image. By employing this matrix approach, PatchGAN can focus on a broader area of the image. However, this method of "increasing the receptive field" does not adequately address the intricate details of edge texture features. To overcome this limitation, we designed the DBD for cardiac images, which includes the EageGAN module as described by the following equation:

$$0 = H_{EageGAN}(I_{syn}, I_{real}), \tag{11}$$

where $H_{EageGAN}()$ denotes the EageGAN module, I_{syn} is the input synthetic image, I_{real} is the input real image, and O is the output. First, the EageGAN branch fuses the edge results of the real image obtained via a Canny edge detector with the synthetic image to obtain the fused image I_{fuse}, as shown in (12). Following this, multiple convolutional blocks (comprising 3×3 convolution, 1×1 convolution, and ReLU functions) are cascaded to extract features from the fused image, resulting in the fused feature output O. This process are shown in (13) and (14):

$$I_{fuse} = I_{syn}||Canny(I_{real}), \tag{12}$$

$$I_{fuse_i} = \delta\Big(Conv_{3\times3}\big(I_{fuse_{(i-1)}}\big) \parallel Conv_{1\times1}\big(I_{fuse_{(i-1)}}\big)\Big), \quad i = 1, 2, \ldots, n \tag{13}$$

$$O = Reshape\Big(Conv_{3\times3}\big(I_{fuse_n}\big)\Big), \tag{14}$$

where $Canny()$ denotes the Canny edge detector, $Conv_{3\times3}()$ denotes the 3×3 convolution operation, $Conv_{1\times1}()$ denotes the 1×1 convolution operation, and δ denotes the ReLU function. The fused image O is input to the PatchGAN network as auxiliary supervision information to ensure that PatchGAN can provide an accurate evaluation matrix focusing on global features and edge texture information. The PatchGAN module first inputs the synthetic image I_{syn} and the real image I_{real}, and processes them through 4×4 convolution and LeakyReLU functions to obtain fused features, as shown in (15). Then, multiple feature extraction blocks composed of 4×4 convolution, batch normalization (BN), and LeakyReLU activation functions are cascaded to obtain deep features, as shown in (16). Finally, the deep features are fused with the fused features output by the EageGAN to obtain the edge-fused features. These features are then processed through a 4×4 convolution to produce the discriminator matrix G, as expressed in (17):

$$I_0 = \sigma Conv_{4\times4}(I_{syn(real)}), \tag{15}$$

$$I_i = \sigma BN\Big(Conv_{4\times4}\big(I_{i-1}\big)\Big), i = 1, 2, ..., n \tag{16}$$

$$G = Conv_{4\times4}(I_n||O), \tag{17}$$

where σ denotes the LeakyReLU function, $Conv_{4\times4}()$ denotes the 4×4 convolution, and $BN()$ denotes batch normalization. The structure of the double-branch discriminator DBD is shown in Figure 2.

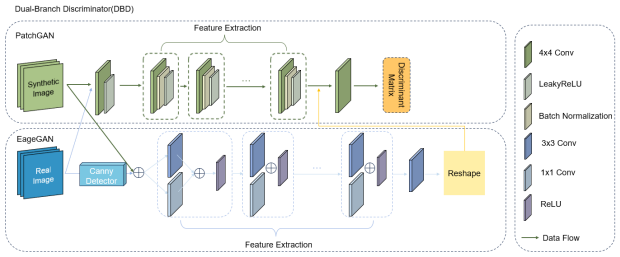

Figure 2: Double-Branch Cardiac Image Discriminator

4. Experiments and results

4.1. Dataset description

This study utilizes the ACDC[41] and MMWHS-2017 datasets for the synthesis tasks. The MMWHS-2017 cardiac segmentation challenge dataset[42] comprises 120 multi-modal whole-heart images from multiple institutions, including 60 cardiac CT/CTA scans and 60 3D cardiac MRI scans that encompass the entire cardiac substructure. The data covers the full heart region from the upper abdomen to the aortic arch. Specifically, the training set includes 20 CT and 20 MRI images, while the test set consists of 40 CT and 40 MRI images.

The ACDC dataset consists of multi-slice 2D cardiac magnetic resonance images (cine MRI) from 100 patients, encompassing five categories of clinical diagnoses: Normal (NOR), Dilated Cardiomyopathy (DCM), Hypertrophic Cardiomyopathy (HCM), Myocardial Infarction with Heart Failure (MINF), and Arrhythmogenic Right Ventricular Cardiomyopathy (ARVC). Each case in the dataset includes a series of short-axis cardiac MRI images covering the entire cardiac cycle, with segmentation labels provided for the left ventricle (LV), right ventricle (RV), and myocardium (Myo) at two time points: end-diastole (ED) and end-systole (ES), serving as benchmarks for evaluation.

The slice images were divided into training, validation, and test sets in a ratio of 7:2:1. Various data augmentation strategies were applied to the new training set, including random flipping, random rotation, distortion, random cropping, and random horizontal/vertical translation and flipping.

4.2. Training details and evaluation metrics

The experiments were conducted using the PyTorch 1.10 framework with Python 3.8 on Ubuntu 20.04, utilizing a GeForce RTX 3090 graphics card. The loss function is $L = L_{FGB} + L_{reconstruct} + L_{codebook} + \lambda L_{GAN}$. In the Feature Vector Quantization Block, the compression scale S was set to (4,4), and the total number of codebooks N was set to 1024. Within the FDGB module, DDPM operated with a time step T of 100. The initial learning rate was set to 0.0001 with a learning rate decay strategy using ReduceLROnPlateau. The optimizer employed was AdamW with a weight decay of 0.05. The batch size was set to 1. The number of epochs was 800, and beyond epoch 400, the improvements in PSNR and MS-SSIM, and the decrease in FID values, notably slowed down, indicating convergence.

For evaluation metrics, the study employed three metrics to assess the quality, multi-scale structural similarity, and distribution similarity of synthesized cardiac images: Peak Signal-to-Noise Ratio (PSNR), Multi-Scale Structural Similarity Index (MS-SSIM), and Fréchet Inception Distance (FID).

PSNR is a metric used to measure image quality, typically assessing the similarity between an image and its original. A higher PSNR value indicates greater similarity and better quality between the two images. The formula for calculating PSNR is given as in (18).

$$PSNR = 10 \times log_{10}\left(\frac{MAX^2}{MSE}\right), \tag{18}$$

where MAX represents the maximum possible pixel value of the image, and MSE stands for Mean Squared Error, aiming to compute differences between two images pixel by pixel, then squaring and averaging them.

SSIM, based on the hypothesis that the human eye extracts structured information from images, measures image similarity through brightness, contrast, and structural aspects. MS-SSIM extends SSIM by examining image details at different scales. The synthetic image s and the real image t are input into the model and downsampled N times. At each downsampling scale, SSIM calculates contrast measure $C_i(s,t)$ and structural measure $S_i(s,t)$, and at the final scale $scale_N$, it computes brightness measure $L_N(s,t)$. By integrating results from different scales, MS-SSIM evaluates the synthesized image comprehensively. The methods for calculating brightness, contrast, and structural measures at each scale are represented by (19), (20), and (21), respectively. The formula for computing MS-SSIM is shown in (22).

$$L(s,t) = \frac{2\mu_s\mu_t + c_1}{\mu_s^2 + \mu_t^2 + c_1}, \tag{19}$$

$$C(s,t) = \frac{2\sigma_s\sigma_t + c_2}{\sigma_s^2 + \sigma_t^2 + c_2}, \tag{20}$$

$$S(s,t) = \frac{\Sigma_{st} + c_3}{\sigma_s\sigma_t + c_3}, \tag{21}$$

$$MS-SSIM(s,t) = [L(s,t)]^{\alpha_N} \times \sum_{i=1}^{N} [c(s,t)]^{\beta_i} \cdot [s(s,t)]^{\gamma_i}, \tag{22}$$

where μ represents the mean of the image, σ represents the standard deviation of the image, and Σ represents the covariance between images. Constants c_1, c_2, and c_3 are used to prevent division by zero. Exponents α_N, β_i, and γ_i are parameters that adjust the importance of the three measures.

FID calculates the Fréchet distance based on the feature vector space between two image distributions. Formula (23) represents the calculation formula for FID:

$$FID(S,T) = \left\|\mu_T - \mu_S\right\|^2 + Tr(\Sigma_S + \Sigma_T - 2 * \sqrt{\Sigma_S\Sigma_T}),$$

(23)

where T represents the set of feature vectors for the real image distribution, typically represented using outputs from intermediate layers of an Inception network. S represents the set of feature vectors for the synthetic image distribution, similarly represented in the same manner. μ_T and μ_S are the means of the feature vector sets T and S, respectively. Σ_T and Σ_S are the covariance matrices of the feature vector sets T and S, respectively. $Tr(\Sigma_S + \Sigma_T - 2 * \sqrt{\Sigma_S \Sigma_T})$ represents the square root of the trace of the covariance matrix. A lower FID value indicates a greater similarity between the distributions of synthetic and real images, correlating with higher image quality in the synthetic images.

4.3. Analysis of experimental results

In this paper, experiments were conducted on the MMWHS-2017 and ACDC datasets to compare the proposed FQCDM method with classical medical image synthesis techniques across multiple dimensions using evaluation metrics such as PSNR, MS-SSIM, and FID. These synthesis algorithms include VQVAE[43], DCGAN[44], CycleGAN, VQGAN, and DDPM.

4.3.1. Analysis of synthesized image quality evaluation

The experiments used the PSNR metric to evaluate the quality of synthesized images, with higher PSNR values indicating better image synthesis quality. The experimental results are shown in Table 1.

According to Table 1, it is observed that diffusion models generate higher-quality images compared to GAN and VAE methods. This is because GAN networks are prone to mode collapse relative to DDPM, whereas diffusion models focus more on the probability distribution of image samples, providing better stability and controllability. VQVAE performs poorer compared to GAN networks and DDPM because VAE uses mean squared error for training in reconstruction tasks, which forces pixel values to be as close as possible but does not guarantee perceptual similarity, resulting in relatively blurry outcomes. The use of perceptual loss in VQGAN and FQCDM has shown better results.

From Table 1, it is also evident that VQGAN, with its codebook quantization constraint, generates image patterns more effectively compared to other classical GAN image generation networks. Additionally, both VQGAN and FQCDM quantize features at compression scale sizes, whereas DDPM directly applies diffusion models to original images, resulting in longer training times per epoch. FQCDM demonstrates comparable image generation quality results to DDPM but in a more efficient manner.

4.3.2. Multi-scale similarity assessment of synthetic images

In this paper, we evaluate the diversity of synthetic images using the MS-SSIM metric. A higher MS-SSIM value indicates greater similarity in terms of brightness, contrast, and structure across multiple scales. The experimental results are presented in Table 2.

Table 1
Quantitative Comparison of PSNR Values on MMWHS-2017 and ACDC Datasets

Method	MMWHS-MRI	MMWHS-CT	ACDC
VQVAE	21.59	25.60	23.82
DCGAN	26.70	32.84	29.46
CycleGAN	29.78	35.17	34.10
VQGAN	30.41	37.18	35.92
DDPM	**33.76**	38.06	37.61
FQCDM (ours)	32.69	**38.42**	**37.89**

Note: Bold numbers in each column indicate the highest score for the corresponding dataset.

Table 2
Quantitative Comparison of MS-SSIM Values with Other Methods on MMWHS-2017 and ACDC Datasets

Method	MMWHS-MRI	MMWHS-CT	ACDC
VQVAE	73.77	72.16	70.91
DCGAN	78.14	74.89	73.45
CycleGAN	82.08	77.62	76.74
VQGAN	84.52	80.30	78.92
DDPM	85.66	82.72	81.76
FQCDM(ours)	**86.15**	**85.77**	**87.58**

Note: Bold numbers in each column indicate the highest score for the corresponding dataset.

As demonstrated in Table 2, the proposed FQCDM outperforms other baseline methods and commonly used classical generative models. This indicates that FQCDM implicitly incorporates the advantages of both diffusion models and GANs. Whether for CT or MRI images, FQCDM consistently ensures superior generation quality of synthetic images across multiple scales.

4.3.3. Evaluation of distribution similarity of synthetic images

Furthermore, we evaluate the distribution similarity of the synthetic images using the FID metric. A lower FID value indicates that the distribution of the synthetic dataset is more similar to the real data distribution. The experimental results are presented in Table 3.

Table 3 shows that the FID values for synthetic heart images are generally higher compared to natural images. This higher FID value can be attributed to the inherent challenges in medical images, such as blurring, noise, and artifacts, which make it difficult to accurately model their distribution. Consequently, diffusion models, which focus on distributional aspects, demonstrate superior performance in this regard compared to GANs and VAEs. Additionally, the proposed model performs better on the ACDC dataset than on the MMWHS-2017 dataset. This is likely because the ACDC dataset involves only three semantic classes, while

Table 3
Quantitative Comparison of FID Values with Other Methods on MMWHS-2017 and ACDC Datasets

Method/dataset	MMWHS-MRI	MMWHS-CT	ACDC
VQVAE	74.82	70.01	68.23
DCGAN	67.44	58.34	60.24
CycleGAN	60.81	53.27	58.83
VQGAN	59.74	50.94	48.08
DDPM	53.59	43.60	44.73
FQCDM(ours)	**49.61**	**41.97**	**38.60**

Note: Bold numbers in each column indicate the highest score for the corresponding dataset.

the MMWHS-2017 dataset involves seven, adding to its complexity. Finally, the FQCDM proposed in this paper is capable of generating synthetic heart datasets with lower FID values.

4.4. Qualitative results

From the visual perspective presented in Fig.3, it is evident that among the synthetic cardiac images, those generated by VQVAE exhibit the poorest clarity, least realistic contrast, and minimal perceptual semantic diversity compared to other generative results. In comparisons with GAN networks, DCGAN demonstrates the lowest performance, while VQGAN and CycleGAN perform similarly in generating MMWHS semantic classes, with VQGAN showing slightly higher clarity. CycleGAN, primarily used for style transfer in cross-domain medical imaging, does not match VQGAN's advantage in generating high-resolution images.

Moreover, the diffusion models (DDPM and FQCDM) consistently produce the highest quality synthetic images across all comparative experiments, excelling in overall brightness, contrast, and semantic diversity. Notably, FQCDM proposed in this paper is capable of generating prominent semantic regions while also capturing relatively rare semantic classes found in the real dataset distribution (highlighted by red/green boxes). This demonstrates its superiority in both structural similarity and distributional fidelity compared to the other models.

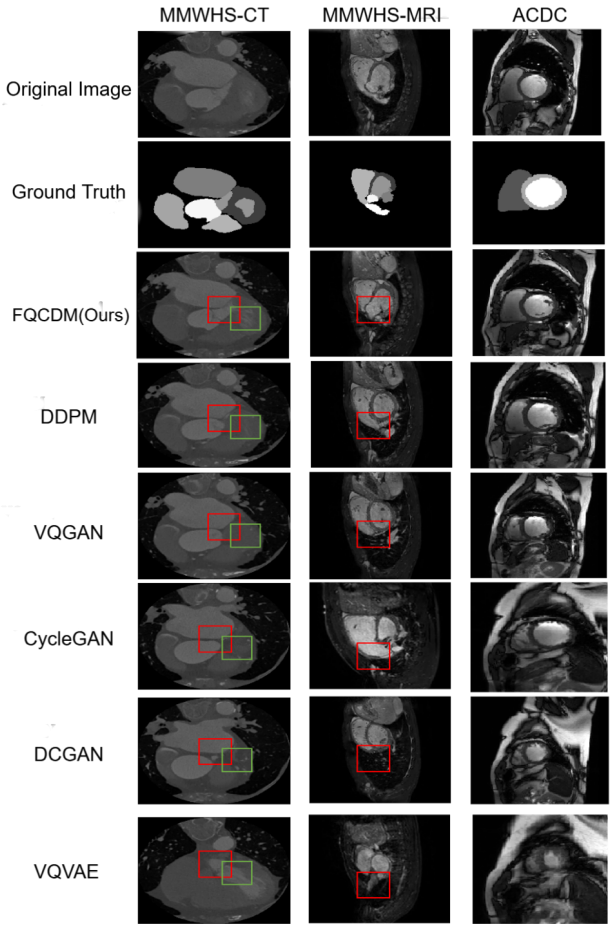

Figure 3: Qualitative comparison of the results from different synthesis methods. Each row presents one typical method from top to bottom: (a) raw cardiac images (b) ground truth labels (c) synthetic images from FQCDM (d) synthetic images from DDPM (e) synthetic images from VQGAN (f) synthetic images from CycleGAN (g) synthetic images from DCGAN (h) synthetic images from VQVAE. Each column represents results from different datasets, including MMWHS-CT, MMWHS-MRI, and ACDC. In each case, green and red boxes are used to highlight regions of interest in the segmentation outputs.

5. Synthetic image-driven cardiac segmentation strategies

Current work on synthetic data, such as Medical Diffusion and ArSDM, has shown that synthetic data can effectively support segmentation and detection tasks for organs like the brain, chest, and abdomen, achieving promising results. Inspired by these findings, we designed several training strategies to investigate whether the synthetic cardiac data generated by the proposed FQCDM method can enhance the segmentation performance of classical cardiac segmentation networks like Swin Unet and U-net. This section primarily outlines three data-driven training strategies: mixing synthetic and real data, self-supervised pretraining, and a combination of synthetic data with traditional data augmentation.

5.1. Mixed training strategy with synthetic and real data

Given the privacy and ethical concerns associated with medical images, many research organizations prefer to use high-quality synthetic images as a proxy rather than sharing real data directly when collaborating with external entities[45]. To address this need, we have designed two strategies for utilizing both synthetic and real data.

To evaluate whether adding relatively low-cost synthetic data can improve segmentation performance on a test set, we examines various ratios of synthetic to real data as supplementary training data. The amount of real data used is 100 images. This study references the advanced work on Synthetically Enhance[46] to define the ratio of synthetic

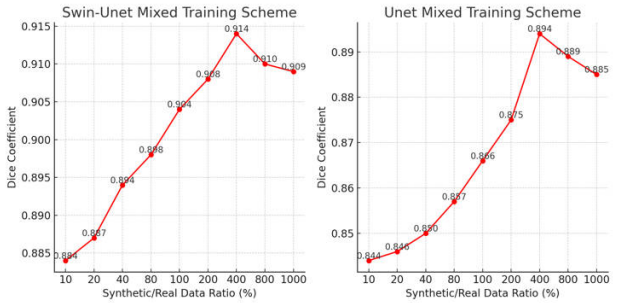

Figure 4: Results of mixed training with synthetic and real data. The figure consists of two subplots illustrating the Dice Coefficient for Swin-Unet and Unet mixed training schemes: (a) Left subplot presents the Dice Coefficient for the Swin-Unet scheme as the ratio of synthetic to real data changes. (b) Right subplot shows the Dice Coefficient for the Unet scheme with the same varying data ratios.

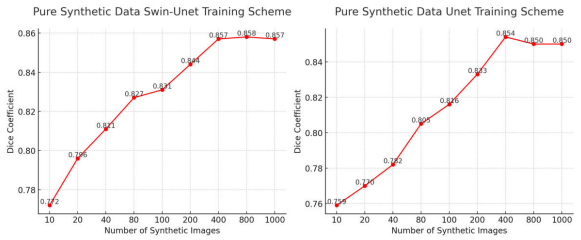

Figure 5: Results of training with synthetic data only. Each subplot displays the results for a different training scheme: (a) Left: Swin-Unet results. (b) Right: Unet results.

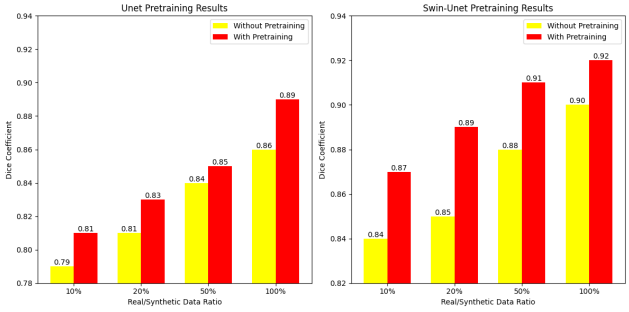

Figure 6: Results with and without Self-Supervised pretraining. (a) Left subplot: Unet Pretraining Results, showing the Dice Coefficient at different Real/Synthetic Data Ratios. (b) Right subplot: Swin-Unet Pretraining Results, also displaying the Dice Coefficient at varying Real/Synthetic Data Ratios.

to real data. We created mixed ACDC datasets with synthetic/real data ratios of [0.1, 0.2, 0.4, 0.8, 1.0, 2.0, 4.0, 8.0, 10.0] and conducted training and evaluation using two common cardiac segmentation networks, U-net and Swin U-net. The experimental results are illustrated in Fig.4.

From Fig.4, it is evident that when the synthetic-to-real data ratio is below 400%, meaning the number of synthetic images does not exceed four times that of real images, both segmentation networks show continuous improvement in segmentation performance as the number of synthetic images increases. However, when the synthetic data exceeds four times the amount of real data, segmentation performance deteriorates. This decline might be due to quality and annotation issues with the synthetic data; beyond a fourfold increase, the diversity of the synthetic dataset may reach saturation. Excessive synthetic data could introduce low-quality images or overly repetitive semantic classes, leading the model to overfit to frequently occurring classes and ignore rarer ones, thus reducing performance. Therefore, using FQCDM synthetic images with a synthetic-to-real data ratio of 400% is recommended.

To simulate a scenario where only synthetic images are used, this section presents a training strategy that employs solely synthetic data as the training set. This approach allows us to investigate whether using synthetic images in place of real images would lead to a noticeable reduction in segmentation accuracy. The amounts of synthetic data and segmentation networks follow the same strategy as (1), with the synthetic data quantities set at 10, 20, 40, 80, 100, 200, 400, 800, and 1000 images. The experimental results are shown in Fig.5.

Fig.5 indicates that the best segmentation performance and Dice score are achieved with 400 synthetic images, consistent with the findings from the mixed training strategy. This suggests that using 400 synthetic images is a good choice for training with FQCDM synthetic data. However, compared to the mixed training strategy, pure synthetic data training yields less satisfactory results. This highlights that synthetic data cannot fully replace real data, as well-annotated real data remains highly valuable and plays a crucial role in model training and evaluation.

5.2. Self-supervised pretraining for cardiac segmentation networks

Self-supervised learning is a method of supervised learning that does not require manually annotated labels. Instead, it leverages the intrinsic properties of the data for training. In the medical imaging field, where acquiring large amounts of finely annotated data is challenging, self-supervised pretraining can address data scarcity issues by using unlabeled data, thereby enhancing model generalization and performance. Additionally, it provides a robust initialization for segmentation networks, which helps accelerate convergence and improve segmentation outcomes.

In this section, we introduce a self-supervised pretraining strategy using a context encoder[47] auxiliary task with synthetic data, applied to the Swin-Unet and U-net segmentation models. For this approach, 500 synthetic images were used for self-supervised training. Fine-tuning was conducted with the ACDC real dataset, with real image quantities set to [50, 100, 250, 500] and real-to-synthetic data ratios of [0.1, 0.2, 0.5, 1]. The results on the ACDC test set are illustrated in Fig.6.

From Fig.6, it is evident that the pre-training mode, which combines synthetic data with real data fine-tuning, effectively enhances the segmentation performance of both

Table 4
Performance Comparison of Segmentation Networks Using Synthetic Data and Data Augmentation

Method	Model	Average Dice↑
Synthetic Images	Swin U-net	85.37
	U-net	81.63
Real Images	Swin U-net	88.43
	U-net	83.46
Synthetic Images+Data Augmentation	Swin U-net	87.49
	U-net	83.04
Real Images+Data Augmentation	Swin U-net	90.01
	U-net	85.69
Synthetic Images+Real Images	Swin U-net	88.06
	U-net	83.38
Synthetic Images+Real Images+Data Augmentation	Swin U-net	**90.87**
	U-net	**86.72**

networks. This improvement is particularly pronounced as the quantity of real data used for fine-tuning increases. When the fine-tuning ratio is at 20%, the segmentation enhancement is notable, while a 100% fine-tuning ratio achieves optimal performance. Thus, we can hypothesize two scenarios: in cases with limited real data, it is advisable to pretrain with a 20% fine-tuning ratio to achieve cost-effective segmentation results. Conversely, if a substantial amount of real data is available, a 100% ratio for self-supervised pretraining can yield advanced segmentation outcomes.

5.3. Combining training strategies with data augmentation

Data augmentation has long been a common method for increasing the dataset size during pre-processing, with extensive work demonstrating its effectiveness in enhancing segmentation accuracy. The synthetic images proposed in this paper also contribute to increasing data volume and thereby improving segmentation network performance, but they should not be conflated with data augmentation methods. Data augmentation generates “pseudo-original" images through geometric or color transformations that are reversible, whereas synthetic images are generated based on real scene simulations or entirely fictional scenarios, allowing customization of various attributes to achieve specific generation goals. The two methods serve different purposes in expanding data. Therefore, this study aims to explore whether combining data augmentation with synthetic data methods can further enhance segmentation network performance.

Our experiments utilized various data augmentation strategies, including random flipping, random rotation, distortion, random cropping, and random horizontal/vertical shifting, as default settings for data augmentation. The results of training using the combination of synthetic data methods and data augmentation are presented in Table 4.

From Table 4, it is evident that data augmentation is effective when applied to both synthetic and real data. For the Swin-Unet and U-net networks, the Dice scores improved by 2.12 and 1.41, respectively, for synthetic images with data augmentation. For real images with data augmentation, the Dice scores increased by 2.52 and 1.83. The magnitude of improvement is similar for both types of data, indicating that data augmentation does not preferentially benefit real images; synthetic images can also be enhanced effectively through data augmentation. When comparing models trained with real images versus those trained with synthetic images, the models trained on real images consistently achieved higher Dice scores, whether data augmentation was used or not. This suggests that relying solely on synthetic data cannot entirely replace real data for training segmentation models. However, the performance of segmentation networks trained with synthetic images did not show a substantial decrease, only about 1.8% lower in accuracy. Considering the cost of acquiring real versus synthetic data, this minor reduction in performance is acceptable. High-quality synthetic data can also be applied to cardiac segmentation tasks, such as participating in training strategies for self-supervised pretraining or serving as a warm-up for initializing model weights. Finally, as shown in the table, using both synthetic and real data combined with data augmentation achieves better results than training the segmentation networks separately with either type of data. Both segmentation networks reach their best Dice scores, indicating that synthetic data can effectively supplement real datasets to some extent.

6. Discussion

In this study, the Feature Quantization-based Cardiac Diffusion Model FQCDM was proposed to address the challenges in cardiac image synthesis and segmentation. Additionally, this work investigates three innovative training strategies to integrate synthetic cardiac data into segmentation tasks. These strategies demonstrate the feasibility and potential of using synthetic data to enhance segmentation performance. Compared to previous studies, our work makes two contributions, which are discussed in detail in Section 6.1 and 6.2.

6.1. Ablation study

Our proposed FQCDM comprises three key components: FDGB, FQB, and DBD. It integrates feature quantization into the generative model and leverages label-guided DDPM for high-quality cardiac image synthesis. To improve edge and texture feature extraction, we designed a dual-branch discriminator that enhances edge detection accuracy while preserving global feature discrimination.

To assess the effectiveness and scientific validity of each module within the FQCDM framework for cardiac image segmentation tasks, we conduct an ablation study on the FDGB, FQB, and DBD. The default network used for comparison is the original VQGAN, and the experiments are

Table 5
Ablation Study

FDGB	FQB	DBD	PSNR↑	MS-SSIM↑
/	/	/	35.92	78.92
✓	/	/	36.87	82.06
/	✓	/	36.46	81.89
/	/	✓	36.07	81.25
✓	✓	/	37.54	83.04
✓	/	✓	37.02	82.92
/	✓	✓	36.68	83.45
✓	✓	✓	**37.89**	**87.58**

Note: Bold numbers in each column indicate the highest score for the corresponding dataset.

performed on the ACDC dataset. The results are presented in Table 5.

Table 5 reveals that each individual module (FDGB, FQB, and DBD) enhances the image PSNR and MS-SSIM values to some extent when embedded in the original network, demonstrating the effectiveness of each module. Among these, the FDGB plays the most crucial role, providing the greatest contribution to image synthesis quality. When all three modules are combined, as in the FQCDM proposed in this paper, the method achieves the most competitive results.

6.2. Analysis of synthetic image-driven cardiac segmentation strategies

As shown in Figs. 4, 5, and 6, three data-driven training strategies explored in this study—mixed training, pretraining with synthetic data, and combining synthetic data with data augmentation—demonstrate improved performance in segmentation accuracy. The mixed training strategy shows optimal segmentation performance when the synthetic-to-real data ratio is up to 400%, while excessive synthetic data leads to diminishing returns. Pure synthetic training confirms the importance of quality and diversity, with 400 synthetic images achieving the best results, though real data remains indispensable for optimal accuracy. The self-supervised pretraining strategy demonstrates that synthetic data can effectively initialize models, particularly when real data is limited, with notable improvements even at low fine-tuning ratios. Combining synthetic data with data augmentation achieves the highest Dice scores, underscoring the complementary benefits of these approaches.

The results suggest that well-annotated real data remains critical for achieving superior segmentation performance, while synthetic data can complement segmentation tasks. However, the experimental results also reveal that one should not overly rely on synthetic data; well-annotated real data still plays a dominant role in achieving superior segmentation performance, while synthetic data can complement cardiac segmentation tasks from a data-driven perspective. Future work could focus on extending FQCDM to generate rare or anomalous structures and exploring its effectiveness across various cardiac imaging modalities.

7. Conclusion

In this work, we propose a novel cardiac image generation framework, VQCDM, which combines DDPM and VQGAN to address the challenges of high labeling costs and limited cardiac imaging datasets. By leveraging the VQGAN encoder to encode cardiac images into continuous features and introducing a label-guided DDPM in the DGB to generate high-quality latent features, the model quantizes these features via the VQGAN codebook before decoding them into synthetic cardiac images. To evaluate the quality and diversity of the generated images, a DBD was designed, enabling simultaneous discrimination of global features and edge information.

Extensive experiments on the MMWHS-2017 and ACDC datasets demonstrate that VQCDM is capable of synthesizing high-quality and diverse cardiac datasets, with evaluation distributions closer to those of real-world data. Moreover, this study explored the application of synthetic cardiac datasets in cardiac image segmentation tasks. By designing three strategies — mixed training with synthetic and real data, pretraining with self-supervision, and combining synthetic data with conventional augmentation methods — we examined the adaptability of synthetic data in enhancing segmentation performance. The results highlight the potential of synthetic data in improving segmentation accuracy. However, our findings also emphasize that excessive reliance on synthetic data may hinder segmentation performance, as fully annotated real-world data remains a dominant factor for achieving superior accuracy. Synthetic data, while beneficial, should complement rather than replace real data, offering a data-driven solution for cardiac segmentation tasks. These insights underscore the importance of balancing synthetic and real data in future research to optimize segmentation performance and address data scarcity challenges.

Ethical statement

The authors certify that this manuscript is original and has not been published and will not be submitted elsewhere for publication while being considered by Computer Methods and Programs in Biomedicine. And the study is not split up into several parts to increase the quantity of submissions and submitted to various journals or to one journal over time. No data have been fabricated or manipulated (including images) to support the conclusions.

The author ensure that all procedures were performed in compliance with relevant laws and institutional guidelines and have been approved by the appropriate institutional committee(s). If the authors have used the public dataset, that this has been appropriately cited or quoted and permission has been obtained where necessary.

The submission has been received explicitly from all co-authors. And authors whose names appear on the submission have contributed sufficiently to the scientific work and

therefore share collective responsibility and accountability for the results.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research was funded by the Natural Science Foundation of Shandong Province (No. ZR2021MF011) and the National Natural Science Foundation of China (No. 62306293). This research was also supported in part by the National Natural Science Foundation of China (No. 62376136 and No. 62076149), and Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515011935).

Acknowledgment

We gratefully acknowledge the reviewers and the support of NVIDIA Corporation with the donation of the Titan V used for this research.

References

[1] CUI Ke, TIAN Qichuan, et al. Review of medical image segmentation algorithms based on u-net variants. *Journal of Computer Engineering & Applications*, 60(11), 2024.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[3] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6810–6818, 2018.

[4] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.

[5] Mihaela Croitor Ibrahim, Nishant Ravikumar, Alistair Curd, Joanna Leng, Oliver Umney, and Michelle Peckham. Segmenting cardiac muscle z-disks with deep neural networks. In *Medical Imaging 2024: Digital and Computational Pathology*, volume 12933, pages 340–346. SPIE, 2024.

[6] Jörg Sander, Bob D de Vos, Jelmer M Wolterink, and Ivana Išgum. Towards increased trustworthiness of deep learning segmentation methods on cardiac mri. In *Medical imaging 2019: image Processing*, volume 10949, pages 324–330. SPIE, 2019.

[7] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[10] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[12] Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. Missformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*, 2021.

[13] Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6222–6231, 2023.

[14] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 2023.

[15] Md Mostafijur Rahman and Radu Marculescu. G-cascade: Efficient cascaded graph convolutional decoding for 2d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7728–7737, 2024.

[16] Md Mostafijur Rahman and Radu Marculescu. Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. In *Medical Imaging with Deep Learning*, pages 1526–1544. PMLR, 2024.

[17] Md Motiur Rahman, Shiva Shokouhmand, Smriti Bhatt, and Miad Faezipour. Mist: Medical image segmentation transformer with convolutional attention mixing (cam) decoder. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 404–413, 2024.

[18] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.

[19] Athanasios Tragakis, Chaitanya Kaul, Roderick Murray-Smith, and Dirk Husmeier. The fully convolutional transformer for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3660–3669, 2023.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[21] Gihyun Kwon, Chihye Han, and Dae-shik Kim. Generation of 3d brain mri using auto-encoding generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 118–126. Springer, 2019.

[22] Cristina-Madalina Dragan, Muhammad Muneeb Saad, Mubashir Husain Rehmani, and Ruairi O'Reilly. Evaluating the quality and diversity of dcgan-based generatively synthesized diabetic retinopathy imagery. In *Advances in Deep Generative Models for Medical Artificial Intelligence*, pages 83–109. Springer, 2023.

[23] Muhammad Muneeb Saad, Mubashir Husain Rehmani, and Ruairi O'Reilly. A self-attention guided multi-scale gradient gan for diversified x-ray image synthesis. In *Irish Conference on Artificial Intelligence and Cognitive Science*, pages 18–31. Springer, 2022.

[24] Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. Cascaded latent diffusion models for high-resolution chest x-ray synthesis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 180–191. Springer, 2023.

[25] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion

models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022.

[26] Lingting Zhu, Zeyue Xue, Zhenchao Jin, Xian Liu, Jingzhen He, Ziwei Liu, and Lequan Yu. Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 592–601. Springer, 2023.

[27] Jueqi Wang, Jacob Levman, Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, M Jorge Cardoso, and Razvan Marinescu. Inversesr: 3d brain mri super-resolution using a latent diffusion model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 438–447. Springer, 2023.

[28] Yongcheng Zong, Changhong Jing, Jonathan H Chan, and Shuqiang Wang. Brainnetdiff: Generative ai empowers brain network construction via multimodal diffusion. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[29] Finn Behrendt, Debayan Bhattacharya, Robin Mieling, Lennart Maack, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Guided reconstruction with conditioned diffusion models for unsupervised anomaly detection in brain mris. *arXiv preprint arXiv:2312.04215*, 2023.

[30] Shihan Qiu, Shaoyan Pan, Yikang Liu, Lin Zhao, Jian Xu, Qi Liu, Terrence Chen, Eric Z Chen, Xiao Chen, and Shanhui Sun. Spatiotemporal diffusion model with paired sampling for accelerated cardiac cine mri. *arXiv preprint arXiv:2403.08758*, 2024.

[31] Tianqi Xiang, Wenjun Yue, Yiqun Lin, Jiewen Yang, Zhenkun Wang, and Xiaomeng Li. Diffcmr: Fast cardiac mri reconstruction with diffusion probabilistic models. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 380–389. Springer, 2023.

[32] Xiaoxiao He, Chaowei Tan, Ligong Han, Bo Liu, Leon Axel, Kang Li, and Dimitris N Metaxas. Dmcvr: Morphology-guided diffusion model for 3d cardiac volume reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 132–142. Springer, 2023.

[33] Paul Friedrich, Julia Wolleb, Florentin Bieder, Alicia Durrer, and Philippe C Cattin. Wdm: 3d wavelet diffusion models for high-resolution medical image synthesis. *arXiv preprint arXiv:2402.19043*, 2024.

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[35] Firas Khader, Gustav Mueller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baessler, Sebastian Foersch, et al. Medical diffusion: denoising diffusion probabilistic models for 3d medical image generation. *arXiv preprint arXiv:2211.03364*, 2022.

[36] Daniel G Saragih, Atsuhiro Hibi, and Pascal N Tyrrell. Using diffusion models to generate synthetic labeled data for medical image segmentation. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–11, 2024.

[37] Yidan Xu, Jiaqing Liang, Yaoyao Zhuo, Lei Liu, Yanghua Xiao, and Lingxiao Zhou. Tdasd: Generating medically significant fine-grained lung adenocarcinoma nodule ct images based on stable diffusion models with limited sample size. *Computer Methods and Programs in Biomedicine*, 248:108103, 2024.

[38] Yuhao Du, Yuncheng Jiang, Shuangyi Tan, Xusheng Wu, Qi Dou, Zhen Li, Guanbin Li, and Xiang Wan. Arsdm: colonoscopy images synthesis with adaptive refinement semantic diffusion models. In *International conference on medical image computing and computer-assisted intervention*, pages 339–349. Springer, 2023.

[39] Bardia Khosravi, Pouria Rouzrokh, John P Mickley, Shahriar Faghani, Kellen Mulford, Linjun Yang, A Noelle Larson, Benjamin M Howe, Bradley J Erickson, Michael J Taunton, et al. Few-shot biomedical image segmentation using diffusion models: Beyond image generation. *Computer Methods and Programs in Biomedicine*, 242:107832, 2023.

[40] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[41] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[42] Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, et al. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical image analysis*, 58:101537, 2019.

[43] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[44] Alec Radford. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[45] Firas Khader, Gustav Mueller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baessler, Sebastian Foersch, et al. Medical diffusion: denoising diffusion probabilistic models for 3d medical image generation. *arXiv preprint arXiv:2211.03364*, 2022.

[46] Bardia Khosravi, Frank Li, Theo Dapamede, Pouria Rouzrokh, Cooper U Gamble, Hari M Trivedi, Cody C Wyles, Andrew B Sellergren, Saptarshi Purkayastha, Bradley J Erickson, et al. Synthetically enhanced: unveiling synthetic data's potential in medical imaging research. *EBioMedicine*, 104, 2024.

[47] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.