IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 February 2024, accepted 23 February 2024, date of publication 1 March 2024, date of current version 15 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3372863

RESEARCH ARTICLE

DiffusionDCI: A Novel Diffusion-Based Unified Framework for Dynamic Full-Field OCT Image Generation and Segmentation

BIN YANG[1], JIANQIANG LI[1], (Senior Member, IEEE), JINGYI WANG[1], RUIQI LI[1], KE GU[1], (Senior Member, IEEE), AND BO LIU[2], (Senior Member, IEEE)

[1]Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[2]School of Mathematical and Computational Sciences, Massey University, Auckland 0632, New Zealand

Corresponding author: Bo Liu (b.liu@massey.ac.nz)

This work was supported by the National Natural Science Foundation of China under Grant 62076015.

ABSTRACT Rapid and accurate identification of cancerous areas during surgery is crucial for guiding surgical procedures and reducing postoperative recurrence rates. Dynamic Cell Imaging (DCI) has emerged as a promising alternative to traditional frozen section pathology, offering high-resolution displays of tissue structures and cellular characteristics. However, challenges persist in segmenting DCI images using deep learning methods, such as color variation and artifacts between patches in whole slide DCI images, and the difficulty in obtaining precise annotated data. In this paper, we introduce a novel two-stage framework for DCI image generation and segmentation. Initially, the Dual Semantic Diffusion Model (DSDM) is specifically designed to generate high-quality and semantically relevant DCI images. These images not only serve as an effective means of data augmentation to assist downstream segmentation tasks but also help in reducing the reliance on expensive and hard-to-obtain large annotated medical image datasets. Furthermore, we reuse the pretrained DSDM to extract diffusion features, which are then infused into the segmentation network via a cross-attention alignment module. This approach enables our network to capture and utilize the characteristics of DCI images more effectively, thereby significantly enhancing segmentation results. Our method was validated on the DCI dataset and compared with other methods for image generation and segmentation. Experimental results demonstrate that our method achieves superior performance in both tasks, proving the effectiveness of the proposed model.

INDEX TERMS Semantic diffusion model, image synthesis, image segmentation, dynamic cell imaging.

I. INTRODUCTION

Accurate intraoperative diagnosis of breast cancer is crucial for guiding surgical procedures and enhancing patient outcomes. Traditional methods like intraoperative frozen tissue section diagnosis are hampered by challenges such as imprecise lesion localization, ambiguous margin assessment, and lengthy diagnostic processes, increasing the likelihood of additional surgeries [1], [2]. Dynamic Full Field Optical Coherence Tomography (D-FFOCT) [3], also known as DCI, is a breakthrough non-destructive imaging method that uses light scattering and interference measurements to reveal microscopic details within tissues. While D-FFOCT shows significant potential in replacing traditional pathological imaging methods by rapidly providing histology-like images, it also presents its own set of challenges. For instance, due to the unique characteristics of cellular activity imaging, the whole slide DCI image may exhibit obviously color variations and artifacts (as described in Section IV-B), complicating image analysis and accurate diagnosis [4]. Additionally, the sheer volume of images generated by these technology burdens manual interpretation with time-consuming and error-prone tasks, complicating intraoperative diagnoses. Leveraging automated medical image analysis based on deep learning provides an effective solution to these challenges. Advanced models such as Convolutional Neural Networks (CNNs) have significantly improved the speed and accuracy of medical image analysis, particularly in segmentation. However, these models largely depend on extensive annotated training data, which can be costly and limited by available

The associate editor coordinating the review of this manuscript and approving it for publication was Carmelo Militello.

medical resources and expertise. Therefore, new technologies are urgently needed to address these issues.

Denoising Diffusion Probabilistic Models (DDPMs) have shown impressive success in generating various medical images, such as MRIs [5], [6], CTs [7], ultrasounds [8], [9], and histopathology [10]. However, few studies have explored its application in semantically guided medical image synthesis, especially in scenarios involving limited datasets with semantic labels. In this paper, we introduce the Dual Semantic Diffusion Model (DSDM), specifically tailored to generate DCI images in high quality and semantic relevance. The model conditions on semantic masks and reference images to generate images conforming to specific semantic layouts. Semantic masks are integrated via Multi-layer Spatially-Adaptive Normalization (SPADE) [11], aligning the denoising process with semantic guidelines and yielding synthetic images that closely resemble real medical images both visually and semantically. Reference images, through an attention alignment module, introduce style information, ensuring feature alignment at corresponding spatial positions, thus enhancing global feature correlation and maintaining similarity with reference images. Our approach, as an effective data augmentation method, provides additional training data for downstream segmentation tasks, enhancing the model's learning and understanding of various image features, thereby improving segmentation accuracy and robustness. Moreover, it decreases reliance on large, costly, and hard-to-obtain annotated medical image datasets.

Diffusion models have recently demonstrated significant potential in various image segmentation tasks [12], [13], including medical images [14], [15], [16]. They mainly operate in three ways: one is using ground truth (semantic labels) as input, employing the diffusion model's random sampling process to generate implicit segmentation masks, with each step in training and sampling using images as priors. However, this method can generate high-frequency noise in the generated masks, affecting prediction accuracy. While post-processing methods [17] like averaging, median blurring, and Fully Dense CRM mitigate these issues, it is hard to avoid completely. An alternative strategy involves first pretraining a U-Net-based diffusion model on tasks like image generation, followed by fine-tuning to boost segmentation performance. Additionally, some methods [18], [19] extract fixed features of images from diffusion models, then generate segmentation maps using simple classifiers. To maximize the use of synthetic images and pretrained diffusion models for segmentation, we propose a unified segmentation framework. The first stage focuses on generating high-quality, semantically relevant DCI images, while the second stage is dedicated to enhancing image segmentation accuracy and efficiency. Specifically, we input noisy images (synthetic or real) into DSDM, extracting multi-level features from various noise levels across different blocks. These diffusion features are then infused into the segmentation network through a cross-attention alignment module (CAA). Notably, the segmentation model in the second stage reuses the pretrained diffusion U-Net backbone structure from the first stage and is initialized with pretrained weights.

Our extensive experiments on the DCI dataset demonstrate the framework's superiority. Both quantitative and qualitative results confirm that our generative model produces high-fidelity, diverse DCI images. The segmentation model identifies and segments cancerous areas more accurately, providing essential support for surgeons and enhancing surgical success rates and patient quality of life. In summary, the contributions of this paper are manifold:

1) A generation network, Dual Semantic Diffusion Model (DSDM), has been developed. This model uniquely combines semantic masks and reference images to create synthetic images that closely resemble real DCI images, both visually and semantically.
2) A segmentation model is also proposed, leveraging diffusion features derived from the pretrained diffusion model, thereby significantly enhancing the network's ability to segment images.
3) This work integrates image generation and segmentation into a cohesive framework, termed DiffusionDCI. In this framework, images generated by the diffusion model are directly utilized for training the segmentation network, enriching the quality and diversity of task data. An advanced attention alignment module has been incorporated, effectively transmitting knowledge of diffusion probability distributions from the image generation domain to the segmentation network.
4) Comprehensive experiments conducted on the DCI dataset demonstrate superior performance of this approach in both image generation and segmentation tasks.

II. RELATED WORK

A. MEDICAL IMAGE GENERATION

Previous studies have predominantly utilized Generative Adversarial Networks (GANs) [20] for creating realistic natural and medical images [21], [22], [23]. Calimeri et al. [23] introduced a versatile framework for image-to-image translation based on conditional GANs [24]. Wang et al. [25] employed a novel adversarial loss and a multi-scale architecture (Pix2pixHD) to synthesize high-resolution, lifelike images from semantic label maps. Isola et al. [26] developed DermGAN, based on the Pix2Pix architecture, to generate high-fidelity clinical skin images with pathology for specific medical conditions. Additionally, Wang et al. [27] utilized spatially conditioned normalization for photorealistic image synthesis from semantic layouts, emphasizing semantic coding and stylization. Zhang et al. [28] proposed a novel noise adaptation GAN, incorporating adversarial, style, and content losses for noise style transfer in OCT and Ultrasound images. However, GAN-based methods face significant challenges, including mode collapse, training non-convergence, and instability. In contrast, the Denoising Diffusion Probabilistic Model (DDPM) represents a new class of generative models that use a Markov chain process to

transform a simple distribution, such as Gaussian noise, into complex data distributions, thereby achieving high-quality image generation. Compared to GANs, diffusion models offer better interpretability, stability, and controllability and can produce more realistic image qualities [29]. They have also been explored for conditional generation tasks, like class-conditional, image-guided, and semantic-conditional generation [30]. Wolleb et al. [31] proposed classifier-guided diffusion models for image-to-image translation, focusing only on altering anomalous regions to healthy outcomes. Wang et al. [32] introduced a unified framework for Semantic Diffusion Model (SDM), allowing for either linguistic or image guidance, or both. However, semantic-guided synthesis in medical imaging, especially for DCI images, has been less explored. Parallelly, NASDM [33] leverages SDM to generate realistic histological images of colon tissue from semantic masks composed of various nuclei types. The main difference between these methods and our DSDM is that we have introduced not only semantic guidance but also new reference conditions, coupled with an attention alignment module to enhance image quality. Furthermore, based on a pretrained DSDM, we explore the use of diffusion features to assist segmentation tasks.

B. MEDICAL IMAGE SEGMENTATION

Medical image segmentation involves separating areas of interest (such as organs, tissues, cells, lesions) from the background or other regions in medical images, aiding doctors in disease diagnosis, treatment planning, and assessment [34]. Diffusion models have recently demonstrated significant potential in medical image segmentation. Segdiff [12] initially used a diffusion probabilistic approach to design an end-to-end medical image segmentation, injecting information into the denoising process to iteratively refine the segmentation map. Differently, they computed pixel-wise uncertainty maps of the segmentation, allowing an implicit ensemble of segmentations that enhance segmentation performance. Parallel work Medsegdiff [17] introduced dynamic conditional encoding and FF-Parser to improve DDPM-based segmentation models. The upgraded MedSegDiff-v2 [35] introduced a new transformer-based conditional U-Net framework and a novel Spectrum-Space Transformer (SSFormer), thereby enhancing the interaction between noise and semantic features and further enhancing segmentation performance. However, these methods do not explicitly apply diffusion features to segmentation. Inspired by [36] independently using diffusion features to aid segmentation, we infuse multi-scale diffusion features from the generative model into the segmentation network via CAA to enhance segmentation performance.

III. PRELIMINARIES: DENOISING DIFFUSION PROBABILISTIC MODELS

Denoising diffusion probabilistic model (DDPM), one of the generative models, is inspired by non-equilibrium thermodynamics and defines a Markov chain of diffusion steps. The purpose is to slowly add stochastic noise to the original data and then learn to reverse this process to construct desired samples from the noise. The core of DDPM involves a two-step process: forward and backward processes.

Given a data point sampled from a real data distribution $x_0 \sim q(x)$, let us define a forward diffusion process in which we introduce stochastic noises (usually parameterized with Gaussian kernel) with steps $t \in \{0, 1, 2, \ldots, T\}$, producing a sequence of noisy samples $x_1, \ldots, x_t$. The step sizes are controlled by a variance schedule $\{\beta_t \in (0, 1)\}_{t=1}^{T}$. Eventually x_t becomes an isotropic Gaussian distribution.

The forward process is described by the formulation:

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

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

where $I_{n\times n}$ is the identity matrix of size n. $T = 1000 \sim 5000$ is a typical choice for most works. The resulting distribution of x_t given x_0 is then expressed as:

$$q(x_t|x_0) = \mathcal{N}((x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I_{n\times n}), \tag{3}$$

The forward process has a nice property that it can sample x_t at an arbitrary time step t in a closed form via a reparameterization trick. Let $\bar{\alpha}_t = \prod_{i=1}^{T}(1-\beta_i)$, x_t can be sampled by:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{(1-\bar{\alpha}_t)}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I_{n\times n}). \tag{4}$$

If β_t is sufficiently small, the reverse process also follows a Gaussian distribution. Thus, the reverse (generative) process is also defined as a Markov chain parameterized by θ, gradually denoising an arbitrary Gaussian a noise $x_t \sim \mathcal{N}(0, I)$ to a clean data sample at each step, for $t \in \{T, T-1, \ldots, 0\}$. Unfortunately, since $q(x_{t-1}|x_t)$ depends on the entire dataset distribution, a neural network p_θ is trained to approximate the reverse denoising process. The reverse process can be represented as:

$$p_\theta(x_{0:T-1}|x_T) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{5}$$

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{6}$$

where $\Sigma_\theta(x_t, t)$ can be either a fixed covariance or a learned parameter as well which has been shown to improve the model quality [37]. In practice, rather than directly parameterizing $\mu_\theta(x_t, t)$ as a neural network, a noise predictor network $\epsilon_\theta(x_t, t)$ is trained to the noise component at the step t. Then, the mean $\mu_\theta(x_t, t)$ as a function of noise can be defined as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)), \tag{7}$$

The learning objective for the neural network to measures the distance between the real noise ϵ and the noise estimation $\epsilon_\theta(x_t, t)$ is derived by considering the variational lower bound (VLB) and simple by Ho et al. [38]:

$$L_{simple} = \mathbb{E}_{x_0,\epsilon,t}[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2], \tag{8}$$

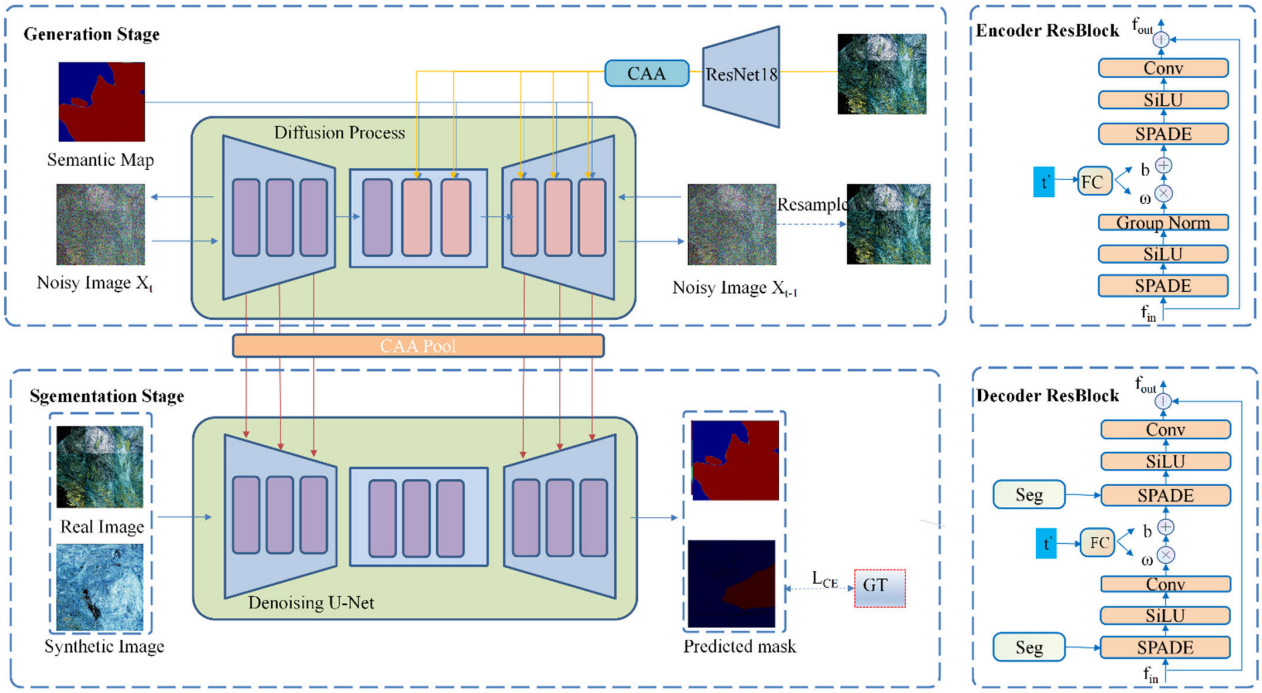


FIGURE 1. Overview of the proposed unified segmentation framework.

where x_0, ϵ_t and t are sampled from $q(x_0), \mathcal{N}(0, I_{n\times n})$ and the discrete uniform distribution $\mathcal{U}(0, T)$.

For inference, the reverse diffusion steps are performed starting from a random sample $x_t \sim \mathcal{N}(0, I)$. For each timestep $t \in \{T, T-1, \ldots, 1\}$, x_t is iteratively denoised to x_{t-1} based on reparametrize Eq.6:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) + \sigma_\theta z. \tag{9}$$

where σ_t denotes the variance scheme that can be learned by the model, as proposed in [37].

IV. METHOD

A. OVERVIEW

DSDM is implemented on a diffusion-based U-Net framework as depicted in Figure 1, which consists of two parts: semantic image synthesis and segmentation. In the first stage, two different conditioning manners for semantic guidance are applied to guide the U-Net-based diffusion network: reference image condition and semantic mask condition. During training, a compact encoder (Resnet 18) extracts coarse features F_1, including appearance and structure features, from a noise-free reference DCI image I. This conditioning imposes an initial high-level semantic reference on the diffusion model, aiding in the reduction of diffusion variances. The high-level features F_1 are then processed through the multi-scale CAA modules, transforming F_1 into various resolutions and progressively aligning with the middle image features in SPADE ResBlocks (SRES). Simultaneously, the semantic mask is injected into the SPADE ResBlocks (SRES) as a multi-layer spatially adaptive semantic condition. In the sampling process, given a semantic mask and a reference image, a semantically relevant and realistic synthetic DCI image is recovered from the initialized Gaussian distribution.

In the segmentation stage, the network's architecture, shown at the bottom of Figure 1, reuses the denoising U-Net's backbone, substituting the SPADE layer with group normalization. In order to preliminarily leverage the information learned in the image synthesis stage, the network is initialized with pre-trained weights from the first stage of the diffusion network. Given a noisy synthetic image X_t, it is passed through the first-stage diffusion network with frozen parameters to extract internal diffusion features F_1 from various ResBlocks. These features contain visual-linguistic semantic information and have proven effective in semantic segmentation tasks [39]. Then, these features are injected into the segmentation network via a cross-attention mechanism at the corresponding ResBlocks, in order to make use of the maximum semantic and diverse visual-linguistic information.

B. SYNTHETIC DCI IMAGE GENERATION BASED ON DUAL SEMANTIC GUIDANCE

1) COLOR NORMALIZATION

To construct a whole slide DCI image, axial (10 nanometers) and transverse (1.24 mm × 1.24 mm patch) scans are performed on each biopsy specimen using a Light-CT scanner, producing 512 regions of interest (ROIs). For each pixel, the power spectral density (PSD) is computed and subjected to saturation normalization, thereby generating an RGB image for an ROI patch. Subsequently, several ROI patches are combined along the horizontal and vertical axes

through continuous scanning to form a complete DCI image. As the imaging principle involves using the characteristic time period or frequency of the cell dynamic signal to attribute a color to each pixel, the shorter the imaging duration, the better the preservation of tissue freshness, and the better the quality of the generated DCI signal. However, the process of continuous scanning and generating multiple ROI patches may result in color discrepancies between the ROI patches of the complete DCI image, potentially reducing the effectiveness of training for the generation and segmentation models. Therefore, in this work, we employ a stain normalization method modified from Reinhard [26] to transform all ROI patches of a complete DCI image to match the average staining distribution of a whole slide DCI image.

2) SEMANTIC DIFFUSION MODEL

a: NETWORK

The DSDM backbone, a U-Net-based conditional diffusion model described in Wang et al. [27], maintains a structure that allows for injecting semantic mask via SPADE. The DSDM (shown in Figure 1) includes four components: a) Diffusion encoder that converts the noise samples to latent representations; b) Diffusion decoder that reconstructs or generates images conditioned on two types of guidance; c) Reference image encoder that encodes reference image into embeddings; d) Multi-scale Cross-Attention Alignment Module that aligns and measures the distance between conditional reference image features and middle image diffusion features. Diffusion encoder: we use residual blocks [40] with self-attentions [41] as encoder resblock which consists of convolution, GroupNorms, SiLU [42], and skip connections. Diffusion decoder: Unlike the encoder resblocks, the group normalization layer is replaced by the spatially-adaptive normalization (SPADE) [11] which injects the semantic mask into the diffusion network in the multi-layer spatially adaptive manner (as shown in Figure 1). Reference image encoder: Considering that the same cancer may present different types, using reference image guidance can generate images with the same or similar layout format. Therefore, we propose using a refined ResNet-18 for the reference image condition, which better captures DCI image features, such as texture and color. Specially, we modify the standard ResNet-18 architecture by stripping away the final classification layer and appending a fully connected layer which produces vectors of dimension 512. We use the contrastive learning strategy proposed in [43] to train our reference image encoder on DCI dataset. Multi-scale Cross-Attention Alignment Module (MCAAM): Rombach et al. [44] exploited the spatial transformer [41] as a flexible and powerful conditioning mechanism to combine the conditional features and the diffusion U-Net features. Inspired by this, we propose the MSCAA, consisting of an embedding transform layer and Cross-Attention Alignment (CAA) to inject the reference image condition. Embedding transform layer uses a fully connected layer Ψ_{FC} to project a lower-dimensional encoded image features into a higher-dimensional feature, reshaping it to align with the multi-scale resolutions of the middle features. Given the middle features $F_m \in \mathbb{R}^{b\times h\times w}$ and the reference image embedding $F_r \in \mathbb{R}^{b\times n}$, $K, V \in \mathbb{R}^{b\times h\times w}$ are the output of the embedding transform layer, CAA can be defined as:

$$Q = W_Q F_m,\ K = W_K F_r,\ V = W_V \Psi_{FC}(F_r) \tag{10}$$

$$MCAA(Q, K, V) = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{C}}\right)V \tag{11}$$

where b is the batchsize, h, w is the resolution of the middle features, n is the length of the reference image embedding. Furthermore, residual connection is applied between input and output of the CAA modules ensuring the integration of original features throughout the process.

b: LOSS

In the preliminaries section, we review the theory of unconditional DDPM. For convenience, we follow the definitions of the symbols in Section III. If the data is sampled from a conditional data distribution $x_0 \sim q(x_0|y)$, and y is the condition (e.g., semantic mask, image embedding), the sampling distribution becomes:

$$p_\theta(x_{0:T-1}|x_T, y) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, y), \tag{12}$$

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{13}$$

we can train DSDM using a hybrid loss function following previous work [37]:

$$L_{hybrid} = L_{sample} + \lambda L_{vlb}, \tag{14}$$

$$L_{simple} = \mathbb{E}_{x_0,\epsilon,t}[||\epsilon - \epsilon_\theta(x_t, y, t)||^2], \tag{15}$$

$$L_{vlb} = D_{KL}(p_\theta(x_{t-1}|x_t, y)||q(x_{t-1}|x_t, x_0), \tag{16}$$

where λ is the trade-off parameter to prevent L_{vlb} from overwhelming L_{simple}.

c: CLASSIFIER-FREE GUIDANCE

To enhance the correlation between the sampled images and semantic label maps, we adopt the Classifier-Free Guidance [45]. This approach eliminates the need for training a separate classifier model and has been widely applied to diffusion models conditioned on a variety of modalities, including text, images, and class labels [45], [46], [47]. During training, we replace the condition (reference embedding and semantic mask) with a null label at a certain probability, thus unifying the training of the conditional and unconditional models. During inference, a guidance scale s controls the contribution of the guidance condition to the noise estimation, enabling a balance between the quality and diversity of the sample. The process can be formulated as follows,

$$\hat{\epsilon}_\epsilon(x_t|y) = \epsilon_\theta(x_t|y) + s\cdot(\epsilon_\theta(x_t|y) - \epsilon_\theta(x_t|\emptyset)). \tag{17}$$

Since we empirically observe the best results for $s = 3$, we fix this choice over all generation experiments.

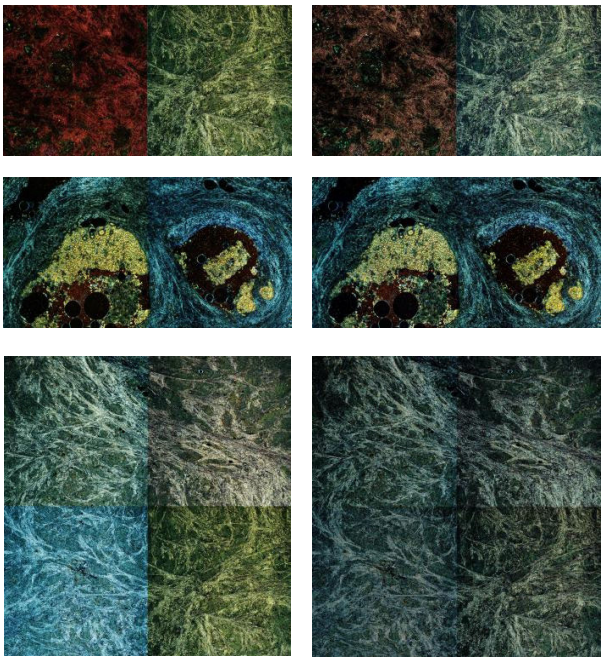


FIGURE 2. Comparison of color normalization result. (A) shows original images. (B) shows transformed images with the modified Reinhard method.

C. SEGMENTATION

1) EXTRACTING DIFFUSION REPRESENTATIONS

To identify the features contributing most significantly to the model's segmentation performance, we analyze the diffusion features produced by the different blocks within the U-Net of DSDM at every 100th diffusion timestep within the range of [100,1000]. A small subset of DCI image with masks from DCI dataset will be used for training and validation. Here, a straightforward decoder is trained on these features to predict the associated masks. Specifically, for a given image I at a diffusion timestep t, we initially extract internal features with the dual semantic condition input to the DSDM at that timestep. Subsequently, these features are used to train the corresponding decoder until convergence. The model's performance is then evaluated using the mean Intersection over Union (mIoU) metric on a select subset of the validation dataset.

Figure 3 illustrates the segmentation result utilizing intermediate features from various resblocks and timesteps. The features produced by the middle of the U-Net decoder {6,8,10} appear to be the most informative. Different from the findings of previous works [39] that the most useful information for segmentation appear only in a wider range of timesteps, from 200 to 800.

2) ARCHITECTURE

The structure of our segmentation model is shown in the bottom of Figure 1 and is the same as the first-stage denoising U-Net structure. Therefore, it can be initialized with the pretrained weights of the first-stage model. However, a domain gap exists between the diffusion embedding and the segmentation semantic embedding when predicting redundant noise from the noisy image. To address this issue, we reuse the Cross-Attention Alignment (CAA) (shown in Figure 4), which allows the model to learn the interaction between noise feature and semantic features, resulting in stronger feature representations. Following the semantic analysis of diffusion features in above section, we incorporate the intermediate features from blocks {6-10} of the DSDM into the corresponding blocks of segmentation model.

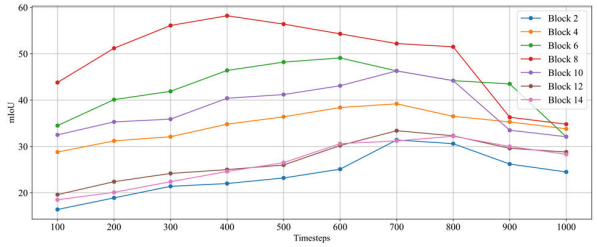


FIGURE 3. Performance of DSDM pixel-wise representations on the DCI dataset.

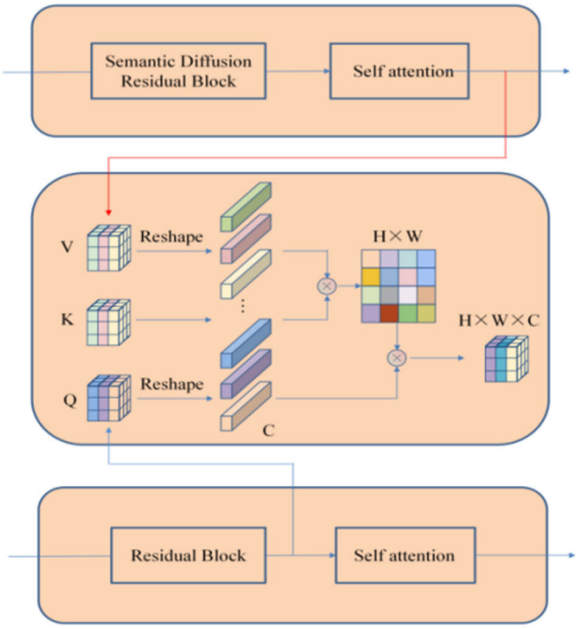


FIGURE 4. The cross-attention mechanism integrates the diffusion features obtained from DSDM into the corresponding feature space of the segmentation model.

V. EXPERIMENT

In this section, we aim to comprehensively detail the design and execution of our experiments. All the models are implemented with Pytorch framework and trained using a single NVIDIA RTX 3090.

A. EXPERIMENTAL SETUP

1) DATASET

DCI dataset contains 300 breast images from 141 patients in total, covering 235 malignant cases and 74 benign lesions or normal cases with their corresponding semantic masks and

the pathological diagnosis obtained through a standard H&E staining. Each DCI image is assigned a full segmentation annotation for types of malignant and non-malignant. In the annotation of the DCI dataset, three experienced pathologists utilize QuPath, a powerful open-source software for digital pathology. The process involves pathologists meticulously reviewing each slide and delineating regions of interest based on the corresponding histological diagnosis, thereby generating semantic masks. The semantic mask in the dataset consists of three class labels: 0 represents the background, 1 corresponds to the Benign tissue, and 2 indicates malignant tissue. The DCI dataset resolution ranged from 2592 × 2592 pixels to 21024 × 21024 pixels. Therefore, Regions of Interest (ROIs) of 3450 × 3450 pixels were generated on the original image using a sliding window algorithm with steps of 3,450 pixels (horizontal and vertical). We extract a total of 7000 including 5000 training patches and 2000 testing patches which are resized to 256 × 256 pixels for training and testing in the generation stage. In addition, the dataset used for the segmentation task contains the original DCI images and the synthetic images generated using the DSDM. We use total semantic maps in training dataset as input to the generative model and generate synthetic images with the corresponding semantic mask. Note that all samples displaying normal breast tissue and various cancerous changes are evenly distributed across both low-quality and high-quality image subsets. Consequently, two experienced clinicians have conducted a screening of the generated images for both quantitative and qualitative evaluations. In Section III, we explain how we apply synthetic images to image enhancement.

2) MODEL SETTINGS

In the generating experiment, the agriculture of our DSDM is described in the section V-B2. The denoising U-Net is training have all been trained from scratch while the image encoder ResNet-18 is pre-trained with contrastive learning strategy. Based on the generally effective performance of the learning rate in similar image generation tasks [33], we use Adam [48] optimizer with a learning rate of 1e−4 and a batch size of 10 to train the whole model for 50,000 iterations. Additionally, we choose a cosine noise schedule for all the diffusion models.

As for segmentation experiment, we optimize the parameters of segmentation network with the cross-entropy (CE) loss using the Adam optimizer with a learning rate of 0.001 as it is a widely accepted starting point for training segmentation models [35], while keeping the parameters of the pre-trained DSDM freezed. We train the model with a batch size of 12 for 200 epochs. The primary image augmentation techniques used include: horizontally flipping the image with a 50% probability; randomly adjusting the image's brightness, saturation, contrast, and hue with an 80% probability, and normalizing the tensor by calculating the mean and variance of the dataset and adjusting its range from (0,225) to (-1,1).

TABLE 1. Ablation study for different configurations of our model.

Agriculture	FID	LPIPS
Denoising U-Net (baseline)	31.92	**0.436**
Denoising U-Net + SM	26.25	0.386
Denoising U-Net + RI	29.84	0.395
Denoising U-Net + SM+RI	**21.36**	0.354

3) METRICS

To evaluate generation performance, we use two objective indicators Frechet Inception Distance (FID) [49] and Learned Perceptual Image Patch Similarity (LPIPS) [50]. FID measures the fidelity of the generated images. LPIPS measures the distance between the multimodal generated images, which reflects the diversity of the generated images. FID is evaluated on 2000 synthesis images against their respective 2000 images from the test dataset, while LPIPS which is computed on sets of 10 sample images for each of the 2000 test images and features.

To evaluate the segmentation performance of various neural network, we focus on four of the most widely used evaluation metric: mean Dice Coefficient (mDice) (a.k.a. F1 score), mean Intersection over Union (mIoU), precision and recall and accuracy.

B. GENERATING RESULT

1) ABLATION STUDY

Ablation study assesses the impact of different components on a U-Net-based denoising model for image generation. To fully evaluate the performance of the conditional module of the network, we retain the U-Net backbone of DSDM and replace the SPADE layer in the decoder with GroupNorm to obtain the initial unconditional denoising U-Net. We can train from scratch the baseline unconditional denoising U-Net.

In Table 1, we present both quantitative and qualitative results obtained by conditioning the model with either a reference image (RI), semantic masks (SM), or a combination of both. When SM and RI are used individually, we observe improvements in the FID scores and reductions in the LPIPS scores. Notably, the use of SM leads to a higher improvement compared to RI alone. This could be attributed to the explicit semantic constraints provided by SM, which more effectively guide the learning and understanding of data, thereby enhancing the quality of the generated images. When both SM and RI are utilized simultaneously, the FID scores achieve the best results, and the LPIPS scores show only a slight decrease, indicating a well-maintained balance between quality and diversity.

Furthermore, we also explore the influence of color normalization and classifier-free guidance strategy. As shown in Table 2, we observe that the introduction of color normalization does not significantly impact diversity but slightly improves FID. Comparing the third and fourth rows of the table, it is evident that the classifier-free guidance

TABLE 2. Ablation study for color normalization and the classifier-free guidance strategy on denoising U-Net baselines.

Setting		Meters	
Color normalization	Class-free guidance	FID	LPIPS
×	√	23.92	0.365
√	×	28.35	**0.412**
√	√	**21.36**	0.354

strategy significantly enhances the FID metric, albeit with a marginal increase in the LPIPS scores. This finding suggests that classifier-free guidance effectively boosts image quality, contributing to a notable improvement in the FID metric

2) COMPARED WITH OTHER GENERATING METHOD

We report FID and LPIPS results, compared to previous works. The specific comparative method chosen for this paper is as follows:

CycleGAN is a novel method for unpaired image-to-image translation, which means it can transform images from one domain to another without requiring corresponding images in the source and target domains. It is widely used in medical image generation.

AttributeGAN is designed to manipulate specific attributes of an image, such as facial features or expressions, allowing for controlled modifications while maintaining the identity and main features of the original image.

MorphDiffusion introduced prioritized morphology weighting and color normalization into the diffusion model to synthesize high-quality histopathological images of brain cancer.

NASDM is a nuclei-aware semantic tissue generation framework based on SDM, which realizes the perfect nuclear localization of pixels in the generated samples through the semantic instance mask of different kernel types.

To ensure a fair and comprehensive comparison, we evaluate all models under identical conditions using the same datasets and metrics. For the parameter settings of each model, we adhere as closely as possible to the common configurations specified in their original papers. By retraining and testing them on our DCI dataset, we ensure a direct comparison of each model's capabilities in a consistent environment.

As shown in Table 3, CycleGAN has the highest FID score, indicating the generated images are less similar to real images compared to other methods. AttributeGAN Shows improvement over CycleGAN with a lower FID score, suggesting better quality of generated images in terms of realism. The improved quality of generated images by NASDM primarily attributed to the incorporation of semantic label guidance into the diffusion process, effectively utilizing the information in semantic masks to direct the diffusion. Furthermore, our methods achieve superior sample quality to previous method that surpasses SDM-based method methods by 2.92 in FID. However, However, our method achieves a

TABLE 3. Quantitative comparison with different methods.

Method	Condition	FID	LPIPS
CycleGAN [51]		72.32	-
AttributeGAN [52]		58.45	-
MorphDiffusion [10]		38.32	0.337
NASDM [33]	SM	24.28	0.322
DSDM(Ours)	SM+RI	**21.36**	**0.354**

lower LPIPS than NASDM, mainly due to the introduction of RI guidance. This provides more constraints on the morphology and layout in the diffusion process, consequently reducing diversity.

3) EXPERT ASSESSMENT

The conventional norm for assessing the visual quality of generated medical images relies on evaluations by pathology experts [8], [22]. To assess the histological plausibility of our generated images (i.e., the realism of lesion appearance and their applicability in classification), we conducted a survey with five pathologists who have an average of five years of work experience. Their task involved evaluating 100 randomly selected image patches from both the original DCI dataset and the synthetic dataset, with a dual focus on identifying synthetic images and detecting the presence of cancerous tissues. Moreover, the five experts rated the images on a 5-point Likert scale (very poor, poor, Medium, good, very good) to quantify the similarity confidence between real and synthetic images.

The findings are summarized in Table 4. In classification tasks, the experts demonstrate comparable and balanced performance in identifying real images (average accuracy of 74.6%) and synthetic images (average accuracy of 75.8%). An interesting observation is that expert generally has a slightly higher accuracy rate for identifying synthetic images compared to real images, but the difference is not significant. These results indicate that there is no strong bias towards real and synthetic images and there exist the high visual similarity between real and synthetic images.

For the task of discerning image authenticity, pathologists relied on various microscopic and histological features. These included abnormalities in cellular nuclei and intercellular structures, overly uniform morphology of cancer cells, exaggerated tissue structures, unnatural color and texture distribution, and a lack of complexity in surrounding details, particularly evident in tumor regions.

As shown in Table 4, the classification accuracy of experts ranges from 62.5% to 71.5%, with an average accuracy of 67.3%. This indicates a certain degree of challenge in distinguishing between real and synthetic images. Confidence scores vary from 2.4 to 3.1, averaging 2.64 on the 5-point Likert scale. This implies that while accuracy rates were better than random guessing, experts exhibited relatively low confidence in their judgments. These findings underscore

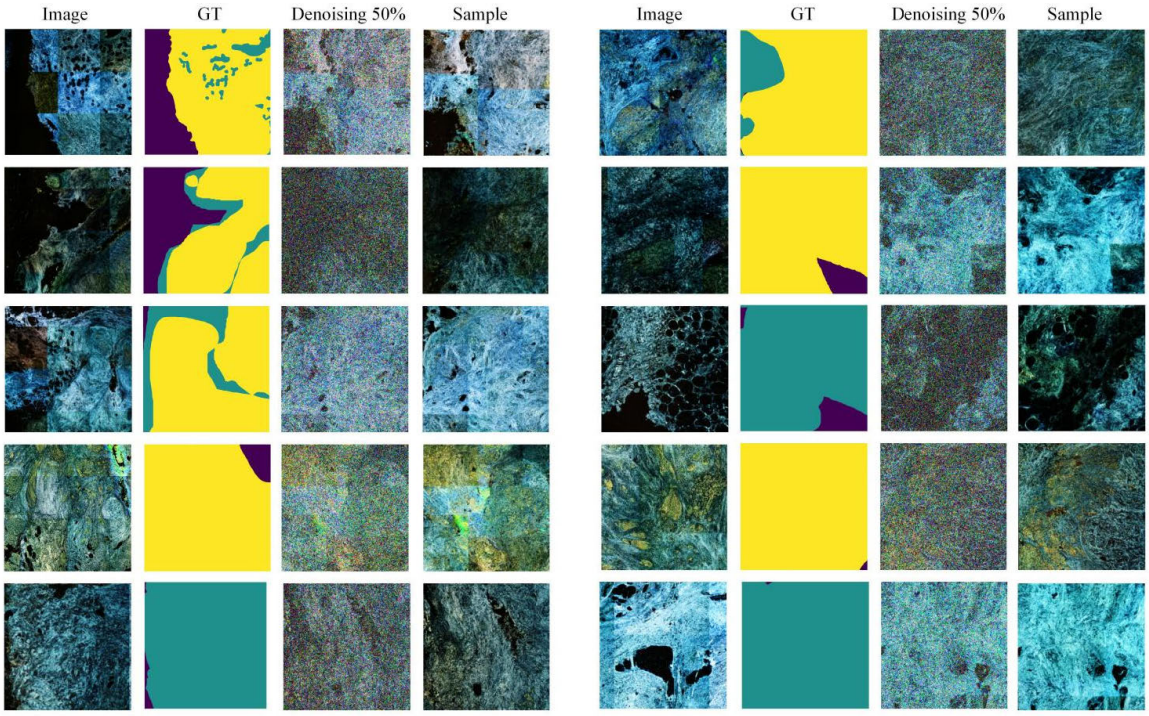


FIGURE 5. Comparison of sampling results generated based on dual semantic condition under different magnification. Samples start with random Gaussian noise and are denoised iteratively until a high-quality output is produced. Denoising 50% represents an intermediate state in the generation process, followed by a progressive enhancement of details or a gradual reduction of noise. Yellow is Malignant tissue, green is benign tissue, and blue is background.

TABLE 4. Results of the expert assessment.

	Classification Accuracy		Is Real image	
	Real	Synthetic	Accuracy	Confidence score
1	74	76	64.5	2.5
2	76	78	70.5	2.8
3	82	78	71.5	3.1
4	69	72	62.5	2.4
5	72	75	67.5	2.4
avg	74.6	75.8	67.3	2.64

the challenges experienced experts face in differentiating between real and synthetic medical images. The moderate accuracy rates, coupled with the less robust confidence levels, can be attributed to the high quality of synthetic imaging technology and its close resemblance to real images, further accentuating the complexity of discerning real from synthetic images in medical image analysis.

4) SAMPLING ANALYSIS

In this experiment, we train the two models respectively with images with ROI size of 1150 × 1150 and 3450 × 3450 and the sampling results. Some examples are shown in Figure 5. we can observe that both models can generate high-quality results. Specifically, the generated image on the right displays more details and has a better visual effect, but with lower semantic relevance. In contrast, the left samples may not exhibit as fine detail as those at right sample, but are likely to be smoother and more coherent overall. Meanwhile, the left samples can also capture and reproduces overall structure and main features more effectively. The result shows the diffusion model's limitations in handling higher resolutions, such as potential distortion of details or increased noise. Furthermore, while color normalization has been applied, achieving uniform color across various regions of the same image remains challenging, and this tends to be more noticeable in high-resolution images.

C. SEGMENTATION RESULT

1) ABLATION STUDY

We conduct an ablation study to validate the effectiveness of infusing DSDM's diffusion features into the segmentation model via CAA. Using the end-to-end supervised training results of the diffusion U-Net as our baseline. We use Dice scores and mIoU to assess performance on the segmentation task. As is shown in Table 5, the segmentation results of DCI images are significantly improved by using two diffusion features. Specially, the Skip Connection method improves the Dice score by approximately 14.42% and mIoU by about 21.88% over the baseline Denoising U-Net. CAA further enhances the Dice and mIoU scores by 17.47% and 26.90%, respectively. Moreover, comparing CAA to the Skip Connection method, we observe an improvement of around 2.68% in Dice and 4.12% in mIoU. This suggests that CAA's integration notably bolsters performance, likely attributed to

TABLE 5. Ablation study for different configurations of our model.

Method	Dice	mIoU
Denoising U-Net	0.7067	0.5620
DenoisingU-Net + skip connection	0.8085	0.6849
Denoising U-Net + CAA	**0.8302**	**0.7131**

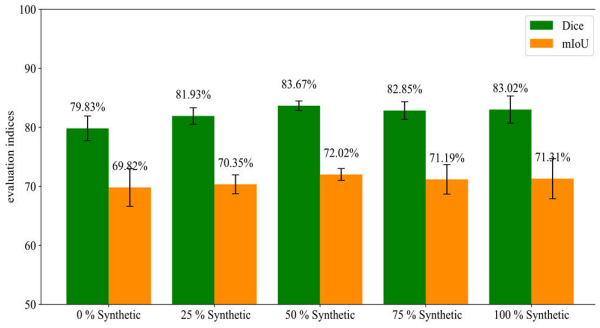


FIGURE 6. Dice coefficient and mIoU result under different ratios of synthetic DCI images.

its ability to learn the interplay between noise and semantic features and effectively leverage the context and correlations across the entire image.

2) SYNTHESIS DATASET

We have designed a set of experiment to evaluate the role of the synthetic data in the segmentation task. In all experiment, DSDM is used to generate images based on DCI training dataset. Guided by expert insights, we selected a dataset of 2,000 high-quality synthetic samples. We randomly sample in a certain proportion from the generated synthesis images which are merged into the original training dataset to finetune the segmentation model. Note that all experiment is evaluate in the real data. As shown in Figure 6, both Dice and mIoU scores increase significantly when a portion of synthetic data is added to the real data. The highest Dice score (0.8367) and mIoU score (0.7202) are achieved with 50% synthetic data, suggesting an optimal balance between real and synthetic data at this proportion. However, at 75% and 100% synthetic data, the standard deviations increase again, which might indicate variability in the quality or relevance of the synthetic data at higher proportions. The increase in indicators may benefit from increasing the diversity of the dataset without introducing too much noise. However, as the amount of synthetic data introduced increases, the model introduces irrelevant features that could arise from an over reliance on synthetic data. Therefore, the synthetic image can be effectively used as a means of image enhancement to improve the segmentation result.

3) COMPARE WITH OTHER METHOD

Table 6 presents a comprehensive comparison of our method against existing approaches across multiple metrics: Dice, mIoU, Precision, Recall, and Accuracy. Compared to non-diffusion-based methods such as Unet, PSP, and nnU-Net, both Medsegdiff and our method demonstrate substantial and comprehensive improvements. Particularly, our method, which incorporates diffusion features into the segmentation model, achieves state-of-the-art results across all metrics. When compared to the second-best result, Medsegdiff, our method shows improvements of 3.62% in Dice and 5.25% in mIoU, which increased by 3.62% and 5.25%, respectively. This significant enhancement in Dice and mIoU suggests that DSGM is more adept at accurately identifying cancerous regions in DCI images. The superior performance of DSGM could be attributed to advanced techniques in diffusion models and more effective integration of diffusion features.

Figure 7 shows the qualitative results of some examples in the DCI dataset with predicted semantic maps produced by all compared methods. From left to right, the columns represent the original images, ground truth (GT), and the segmentation result of various methods. In Figure 7, we observe that the U-Net manifests notable deficiencies, particularly in segmenting intricate regions, with numerous omissions indicating missed segmentation opportunities. For the five sub-image of U-Net, the malignant proportion (the bright yellow areas) is 97.9%, 89.2%, 75.1% and 49.4% respectively, compared to GT's 76.2%, 82.5%, 55.6%, 5.9% and 59.9%. This may suggest U-Net's notable deficiencies in segmenting intricate regions, with a tendency to over-mark malignant areas. In contrast, the PSP method, while showing marginal improvements, still leaves considerable areas unsegmented, especially the malignant areas of complex images. For instance, in the fourth and fifth sub-images of PSP, the Cyan proportions are 20.9% and 74.8% against GT's 45.7% and 30.0%, respectively, while its Dark Purple proportions are 56.1% and 9.2% compared to GT's 48.3% and 10.0%. The nnU-Net shows a significant enhancement that it successfully identified the malignant region of the first image, while the first two methods do not identify it. The Label-efficient DM's performance appears to regress relative to nnU-Net, especially in boundary delineation, shown in the red box of the figure. The Medsegdiff method excels in boundary definition and segmentation coherence, as evidenced by the color uniformity indicating high fidelity to the ground truth. This demonstrates a closer alignment with GT in terms of Cyan and Dark Purple proportions compared with other method. This method particularly stands out in its adeptness at preserving boundaries and detailing, suggesting the advantages of our method in integrating diffusion feature.

D. LIMITATION

We unify DCI image generation and segmentation based on diffusion models, achieving high-quality synthetic images and accurate segmentation results. However, there are still some limitations.

1) LIMITED DETAILS

Our approach generates images at a resolution of 256 × 256, which may result in a limited display of detail. To address this, we can consider a multi-scale generative strategy that first

TABLE 6. The performance of different medical image segmentation methods in DCI datasets.

Method	Dice	mIoU	Precision	Recall	Accuracy
U-Net [53]	0.5751	0.4762	0.6158	0.7279	0.7078
PSPNet [54]	0.6792	0.5462	0.6846	0.7715	0.7214
nnU-Net [55]	0.7688	0.6322	0.7811	0.7623	0.7557
Label-efficient DDPM [39]	0.7418	0.6060	0.7445	0.7814	0.7508
Medsegdiff-v2 [35]	0.8012	0.6775	0.8247	0.7971	0.8113
DiffusionDCI (Ours)	**0.8302**	**0.7131**	**0.8296**	**0.8354**	**0.8164**

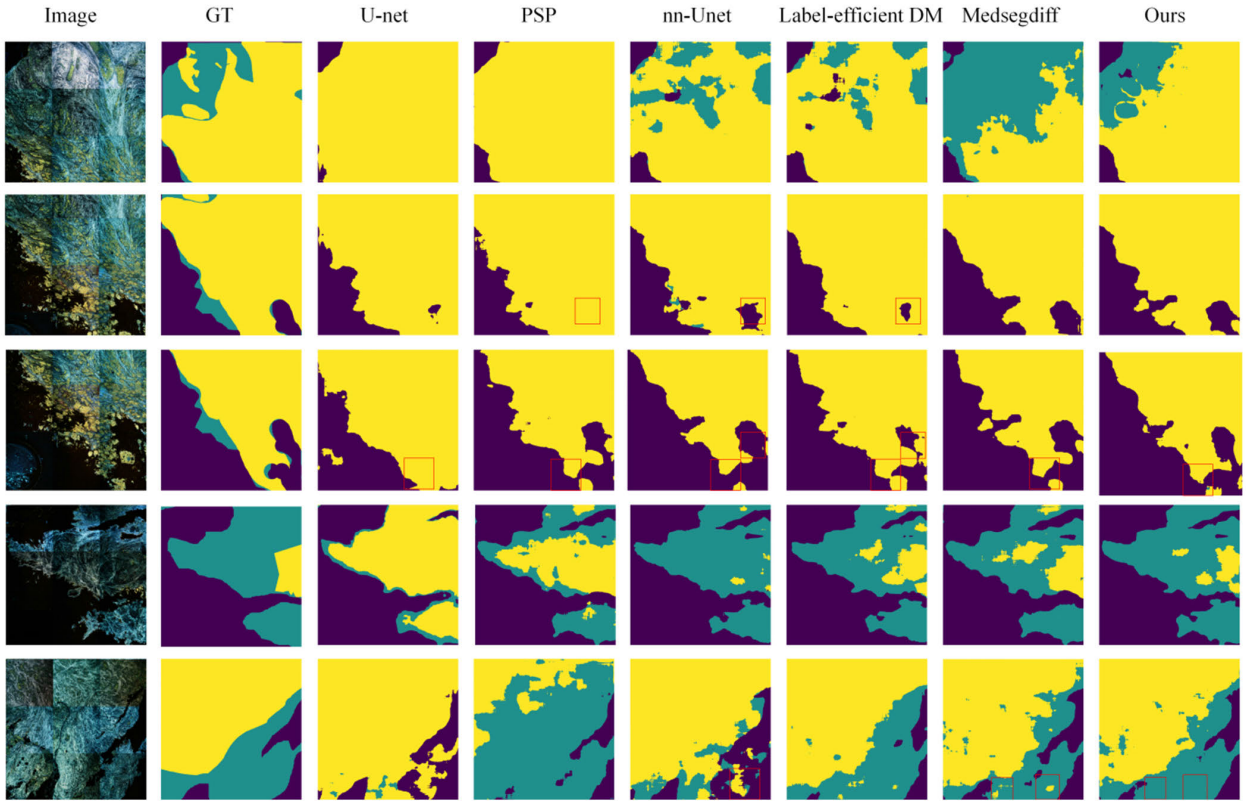


FIGURE 7. Visual comparison of segmentation results with different methods on the DCI dataset. Yellow indicates malignant, Cyan indicates benign and Dark Purple indicates background.

generates a rough image outline and style at low resolution, and then gradually improves the detail level.

2) CONSISTENCY

Our semantic masks are not at a cellular level, leading to sampled results that may not align closely, as visualized in Section V-B. Therefore, our future work will focus on refining semantic masks and optimizing the network structure to enhance semantic consistency.

3) SAMPLING CONTROLLABILITY AND STABILITY

The use of two semantic-guided conditions introduces some redundancy in semantic information and additional noise, adding complexity to the model and compromising the controllability and stability of the results.

Finally, compared with the standard end-to-end image segmentation method, we need extra time to extract and integrate the diffusion features to obtain satisfactory segmentation performance. Therefore, we need to explore more efficient sampling methods for diffusion models to balance sampling time and accuracy.

VI. CONCLUSION

In this paper, we propose a two-stage framework DiffusionDCI for DCI image generation and segmentation. In the first stage, the generated model encompasses two types of semantic guidance: reference image and semantic mask. This dual guidance effectively controls content, structure, and semantic information, leading to the generation of high-quality and semantically relevant DCI images. The second stage focuses on further enhancing image segmentation precision and efficiency by reusing the semantic diffusion features. Experiments on DCI dataset demonstrate that our method notably improves upon previous techniques

in generating high-fidelity, diverse DCI images. The segmentation model accurately identifies and segments cancerous areas, providing crucial support for surgical procedures and improving patient outcomes. In our future work, we aim to enhance the resolution and detail of DCI image generation by employing a multi-scale generative approach, refine semantic masks for improved consistency at the cellular level, and introduce the Latent Diffusion Model for better control and stability in sampling.

REFERENCES

[1] O. Thouvenin, J. Scholler, D. Mandache, M. C. Mathieu, A. B. Lakhdar, M. Darche, T. Monfort, C. Boccara, J.-C. Olivo-Marin, K. Grieve, V. M. Yedid, and E. B. a la Guillaume, "Automatic diagnosis and biopsy classification with dynamic full-field OCT and machine learning," *Res. Square*, vol. 10, no. 3, 2021, Art. no. 034504.

[2] D. R. McCready, "Keeping abreast of marginal controversies," *Ann. Surgical Oncol.*, vol. 11, no. 10, pp. 885–887, Oct. 2004.

[3] C. Apelian, F. Harms, O. Thouvenin, and A. C. Boccara, "Dynamic full field optical coherence tomography: Subcellular metabolic contrast revealed in tissues by interferometric signals temporal analysis," *Biomed. Opt. Express*, vol. 7, no. 4, pp. 1511–1524, 2016.

[4] H. Yang, S. Zhang, P. Liu, L. Cheng, F. Tong, H. Liu, S. Wang, M. Liu, C. Wang, Y. Peng, F. Xie, B. Zhou, Y. Cao, J. Guo, Y. Zhang, Y. Ma, D. Shen, P. Xi, and S. Wang, "Use of high-resolution full-field optical coherence tomography and dynamic cell imaging for rapid intraoperative diagnosis during breast cancer surgery," *Cancer*, vol. 126, no. S16, pp. 3847–3856, Aug. 2020.

[5] Z. Chen, J. Qing, T. Xiang, W. L. Yue, and J. H. Zhou, "Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22710–22720.

[6] Z. Dorjsembe, S. Odonchimed, and F. Xiao, "Three-dimensional medical image synthesis with denoising diffusion probabilistic models," in *Proc. Med. Imag. Deep Learn.*, Switzerland, Sep. 2022, pp. 18–22.

[7] Q. Li, C. Li, C. Yan, X. Li, H. Li, T. Zhang, H. Song, R. Schaffert, W. Yu, Y. Fan, J. Ye, and H. Chen, "Ultra-low dose CT image denoising based on conditional denoising diffusion probabilistic model," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery (CyberC)*, Oct. 2022, pp. 198–205.

[8] D. Stojanovski, U. Hermida, P. Lamata, A. Beqiri, and A. Gomez, "Echo from noise: Synthetic ultrasound image generation using diffusion models for real image segmentation," 2023, *arXiv:2305.05424*.

[9] F. Tang, J. Ding, L. Wang, M. Xian, and C. Ning, "Multi-level global context cross consistency model for semi-supervised ultrasound image segmentation with diffusion model," 2023, *arXiv:2305.09447*.

[10] P. A. Moghadam, S. Van Dalen, K. C. Martin, J. Lennerz, S. Yip, H. Farahani, and A. Bashashati, "A morphology focused diffusion probabilistic model for synthesis of histopathology images," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1999–2008.

[11] T. Park, M. Y. Liu, T. C. Wang, and J. Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2332–2341.

[12] T. Amit, T. Shaharbany, E. Nachmani, and L. Wolf, "SegDiff: Image segmentation with diffusion probabilistic models," 2021, *arXiv:2112.00390*.

[13] A. Alshenoudy, B. Sabrowsky-Hirsch, S. Thumfart, M. Giretzlehner, and E. Kobler, "Semi-supervised brain tumor segmentation using diffusion models," in *Proc. IFIP Adv. Inf. Commun. Technol.*, 2023, pp. 314–325.

[14] X. Guo, Y. Yang, C. Ye, S. Lu, B. Peng, H. Huang, Y. Xiang, and T. Ma, "Accelerating diffusion models via pre-segmentation diffusion sampling for medical image segmentation," in *Proc. Int. Symp. Biomed. Imag.*, 2023, pp. 1–5.

[15] B. Kim, Y. Oh, and J. C. Ye, "Diffusion adversarial representation learning for self-supervised vessel segmentation," 2023, *arXiv:2209.14566*. [Online]. Available: https://arxiv.org/abs/2209.14566

[16] J. Wolleb, R. Sandk, and P. C. Cattin, "Diffusion models for implicit image segmentation ensembles," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, 2022, pp. 1336–1348.

[17] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, and Y. Xu, "MedSegDiff: Medical image segmentation with diffusion probabilistic model," 2022, *arXiv:2211.00611*.

[18] E. B. Asiedu, S. Kornblith, T. Chen, N. Parmar, M. Minderer, and M. Norouzi, "Decoder denoising pretraining for semantic segmentation," 2022, *arXiv:2205.11423*.

[19] E. A. Brempong, S. Kornblith, T. Chen, N. Parmar, M. Minderer, and M. Norouzi, "Denoising pretraining for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 4174–4185.

[20] C. Li, K. Xu, J. Zhu, and B. Zhang, "Triple generative adversarial nets," in *Proc. NIPS*, 2017, pp. 4088–4098.

[21] M. J. M. Chuquicusma, S. Hussein, J. Burt, and U. Bagci, "How to fool radiologists with generative adversarial networks? A visual Turing test for lung cancer diagnosis," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 240–244.

[22] C. Baur, S. Albarqouni, and N. Navab, "MelanoGANs: High resolution skin lesion synthesis with GANs," 2018, *arXiv:1804.04338*.

[23] F. Calimeri, A. Marzullo, C. Stamile, and G. Terracina, "Biomedical data augmentation using generative adversarial neural networks," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2017, pp. 626–634.

[24] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[25] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[27] Y. Wang, L. Qi, Y.-C. Chen, X. Zhang, and J. Jia, "Image synthesis via semantic composition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13729–13738.

[28] T. Zhang, J. Cheng, H. Fu, Z. Gu, Y. Xiao, K. Zhou, S. Gao, R. Zheng, and J. Liu, "Noise adaptation generative adversarial network for medical image analysis," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1149–1159, Apr. 2020.

[29] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 87–8780.

[30] P. Esser, R. Rombach, A. Blattmann, and B. Ommer, "ImageBART: Bidirectional context with multinomial diffusion for autoregressive image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 3518–3532.

[31] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, "Diffusion models for medical anomaly detection," in *Proc. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2022, pp. 35–45.

[32] W. Wang, J. Bao, W. Zhou, D. Chen, D. Chen, L. Yuan, and H. Li, "Semantic image synthesis via diffusion models," 2022, *arXiv:2207.00050*.

[33] A. Shrivastava and P. T. Fletcher, "NASDM: Nuclei-aware semantic histopathology image generation using diffusion models," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2023*, vol. 14225. Cham, Switzerland: Springer, 2023, pp. 786–796.

[34] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models in medical imaging: A comprehensive survey," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102846.

[35] J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, and Y. Xu, "MedSegDiff-v2: Diffusion based medical image segmentation with transformer," 2023, *arXiv:2301.11798*.

[36] K. Pnvr, B. Singh, P. Ghosh, B. Siddiquie, and D. Jacobs, "LD-ZNet: A latent diffusion approach for text-based image segmentation," 2023, *arXiv:2303.12343*.

[37] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. Mach. Learn. Res.*, vol. 139, 2021, pp. 8162–8171.

[38] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 1–12.

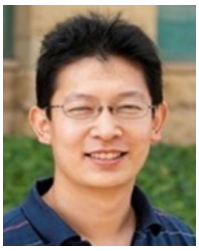
[39] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," 2021, *arXiv:2112.03126*.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[42] P. Ramachandran, B. Zoph, and Q. V. Le, "Swish: A self-gated activation function," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–12.
[43] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1–12.
[44] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.
[45] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proc. Mach. Learn. Res.*, vol. 162, 2022, pp. 16784–16804.
[46] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 36479–36494.
[47] X. Liu, D. H. Park, S. Azadi, G. Zhang, A. Chopikyan, Y. Hu, H. Shi, A. Rohrbach, and T. Darrell, "More control for free! Image synthesis with semantic diffusion guidance," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 289–299.
[48] D. Lee and K. Myung, "ADAM: Method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, Dec. 2014, pp. 1–15.
[49] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–12.
[50] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[51] A. Gilbert, M. Marciniak, C. Rodero, P. Lamata, E. Samset, and K. Mcleod, "Generating synthetic labeled data from existing anatomical models: An example with echocardiography segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2783–2794, Oct. 2021, doi: 10.1109/TMI.2021.3051806.
[52] J. Ye, Y. Xue, P. Liu, R. Zaino, K. C. Cheng, and X. Huang, "A multi-attribute controllable generative model for histopathology image synthesis," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*, vol. 12908. Strasbourg, France, 2021, pp. 613–623.
[53] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany*, vol. 9351. 2015, pp. 234–241.
[54] R. J. Song, F. Zhang, and K. H. Park, "Semantic segmentation based on improved pyramid scene parsing network," *J. Netw. Intell.*, vol. 6, no. 4, pp. 797–806, 2021.
[55] F. Isensee, J. Petersen, A. Klein, and D. Zimmerer, "nnU-Net: Selfadapting framework for U-Net-based medical image segmentation," in *Bildverarbeitung für die Medizin 2019*. Wiesbaden, Germany: Springer, 2019, p. 22.

BIN YANG received the M.S. degree in software engineering from Beijing University of Technology, Beijing, China, in 2020, where he is currently pursuing the Ph.D. degree in software engineering.

His research interests include deep learning, computer vision, and medical image analysis.



JIANQIANG LI (Senior Member, IEEE) received the B.S. degree from Beijing Institute of Technology, in 1996, and the Ph.D. degree from Tsinghua University, in 2004.

He was a Researcher with the National University of Ireland, Galway, from 2004 to 2005. He was with NEC Laboratories China, as a Researcher, from 2005 to 2013. He joined Beijing University of Technology, in 2013, as a Beijing Distinguished Professor. His research interests include Petri nets, data mining, information retrieval, and big data.

JINGYI WANG received the B.S. degree in computer science and technology from Shandong University of Finance and Economics, Jinan, China, in 2021. She is currently pursuing the M.S. degree in software engineering with Beijing University of Technology.

Her research interests include deep learning, weakly supervised learning, and medical image analysis.

RUIQI LI received the B.S. degree in computer science and technology from Hebei University, Baoding, China, in 2021. He is currently pursuing the M.S. degree in software engineering with Beijing University of Technology.

His research interests include deep learning, semi-supervised learning, and medical image analysis.

KE GU (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2009 and 2015, respectively. He is currently a Professor with Beijing University of Technology, Beijing, China. His research interests include industrial vision, environmental perception, image processing, and machine learning.

BO LIU (Senior Member, IEEE) received the B.S. degree from the Department of Automation, Beijing Institute of Technology, Beijing, China, and the M.S. and Ph.D. degrees from the Department of Automation, System Integration Institute, Tsinghua University, Beijing, in 2003 and 2008, respectively. After graduation, she was with NEC Laboratories China, The University of Chicago, Argonne National Laboratory, Beijing University of Technology, and Massey University. Her current research interests include big data, data mining, machine learning, cloud computing, scientific workflow, semantic web, and ontology reasoning.