Electronic Research Archive

ERA, 33(5): 2947–2971.
DOI: 10.3934/era.2025129
Received: 02 March 2025
Revised: 22 April 2025
Accepted: 30 April 2025
Published: 15 May 2025

https://www.aimspress.com/journal/ERA

Research article

DPUSegDiff: A Dual-Path U-Net Segmentation Diffusion model for medical image segmentation

Yazhuo Fan[1], Jianhua Song[1,2,*], Yizhe Lu[1], Xinrong Fu[2], Xinying Huang[2] and Lei Yuan[3]

[1] Key Laboratory of Light Field Manipulation and System Integration Applications in Fujian Province, Minnan Normal University, Zhangzhou 363000, China
[2] College of Physics and Information Engineering, Minnan Normal University, Zhangzhou 363000, China
[3] Jinchuan Group Information & Automation Engineering Co. Ltd, Jinchang 737100, China

* **Correspondence:** Email: songjianhua@mnnu.edu.cn.

Abstract: Denoising diffusion probabilistic models (DDPM) have had remarkable success in image generation. Inspired by this, recent medical image segmentation tasks have started to use diffusion-based methods. These methods leverage iterations and sampling to generate smoother and more representative implicit integration. However, current diffusion-based segmentation models mainly rely on traditional neural networks and seldom focus on effectively interacting semantic and noise information. Moreover, they usually use a single network architecture instead of a hybrid one combining CNN and Transformer. To address limitations, we propose a dual-path U-Net segmentation diffusion (DPUSegDiff) model. It comprises two U-shaped networks based on the edge augmented local encoder (EALE) and the mixed transformer global encoder (MTGE). EALE uses the Sobel operator for local feature extraction, and MTGE has a cross-attention mechanism to facilitate information interaction. To integrate information from both paths selectively and adaptively, we design a bilateral gated transformer module (BGTM) to combine deep semantic information effectively. Experiments on three segmentation tasks—skin lesions, polyps, and brain tumors—show that the proposed DPUSegDiff outperforms other state-of-the-art (SOTA) methods in segmentation performance and generalization ability. The code has been released on GitHub (https://github.com/Fanyyz/DPUSegDiff).

Keywords: medical image segmentation; deep learning; diffusion models; CNN; Transformer; edge detection

1. Introduction

In the domain of computer vision (CV), one of the main challenges is medical image segmentation, including tasks such as skin lesion segmentation in dermoscopic images, polyp segmentation in colonoscopic images, and brain tumor segmentation in MRI images [1]. Segmentation results offer critical insights into anatomical structures, facilitating detailed analysis and significantly aiding physicians in describing lesions, monitoring disease progression, and evaluating the necessity for appropriate treatment [2]. Given the escalating demand for intelligent medical image analysis, the development of accurate and robust segmentation techniques has become increasingly imperative.

Traditional medical image segmentation methods encompass threshold-based approaches [3], region-based techniques [4], edge-based algorithms [5], and clustering-based methodologies [6], among others [7]. While these traditional methods remain effective in simple scenarios, deep learning methods have progressively emerged as the predominant choice due to the inherent challenges of medical images, such as high noise levels, low contrast, and anatomical structures. In the initial stages of deep learning research, convolutional neural networks (CNNs) dominated the semantic segmentation tasks, exemplified by fully convolutional networks (FCNs) [8] and U-Net [9]. CNNs use convolutional kernels that traverse the input image to extract local features, which limits their ability to capture global dependencies due to the restricted receptive field defined by kernel size [10].

To address these limitations, Transformer architectures [11] have been introduced into medical image segmentation. The self-attention mechanism within Transformer establishes connections between tokens, thereby demonstrating superior capability in capturing global contextual information. However, the computational complexity of the self-attention mechanism increases significantly with the length of the input sequence, as it requires calculating correlations between each position and all other positions. This results in a substantial increase in computational time and resource requirements when processing high-resolution feature maps.

Denoising diffusion probabilistic models (DDPM) [12] are a class of deep learning methods based on probabilistic generation. Their core idea lies in two processes—the forward diffusion process and the reverse denoising process—to achieve data generation from noise. Owing to the superior performance of DDPM in image generation, recent studies have extended its application to tasks such as super-resolution [13], deblurring [14], and image restoration [15]. Previously, these areas were typically dominated by generative adversarial networks (GAN) [16]. For example, DAGAN [17] performs image segmentation by training a generator and two discriminators. However, there has been limited exploration of DDPM in semantic segmentation. This is primarily due to the fact that most diffusion methods are applied in domains where absolute ground truth is absent, and the quality and practical value of generated results often depend on subjective evaluation. Medical images frequently exhibit uncertainty, boundary ambiguity, and noise issues [18]. DDPM can generate diverse segmentation predictions through multiple iterations, offering various potential medical interpretations [19]. Moreover, the stepwise denoising process inherent in DDPM facilitates the generation of natural and smooth segmentation outcomes. Building on this foundation, researchers have proposed novel image segmentation methods that leverage diffusion models to address this challenging task [20]. Although these methods have achieved impressive performance, medical images often contain irregular target contours and blurred boundaries. Most existing diffusion models adopt a single feature extraction approach (either CNN or Transformer), leading to a trade-off between modeling local edge features and capturing global spatial relationships [21]. Moreover, current methods typically rely on image

features as guidance but lack an effective feature fusion mechanism during the denoising process [22]. As a result, the guidance information has limited influence on segmentation, especially when it comes to leveraging high-level semantic information.

We have developed a novel hybrid network-based segmentation diffusion model, termed DPUSegDiff, to address these challenges. The proposed model employs two parallel U-Nets during the diffusion process, utilizing CNN and Transformer as the underlying encoder-decoder units, respectively. DPUSegDiff promotes the reuse and reconstruction of contextual information through multi-scale information interaction between the two pathways. To facilitate this rich information exchange, we have designed distinct cross-layer information interaction modules for each pathway. In the CNN-based pathway, we integrate the Sobel operator into the convolutional blocks to create a novel local information fusion and enhancement module. This module effectively fuses and enhances high-frequency information such as boundaries and textures. In the Transformer-based pathway, to incorporate low-level detailed features extracted by the local pathway, we construct a global information fusion and enhancement module guided by cross-attention following the self-attention mechanism. At the bottleneck layer of the network, feature maps typically contain more abstract semantic information. We introduce the bilateral gated transformer module (BGTM) to fuse information from the bottleneck layers of both pathways. Additionally, we design a dual-path fusion module (DPFM) that learns trainable parameters to allocate the final output weights of the two pathways.

In this work, we investigate the efficient and interactive integration of the strengths of convolutional neural networks (CNNs), Transformers, traditional edge detection algorithms, and diffusion models. The contributions of this paper are as follows:

1) We introduce a novel denoising probabilistic diffusion segmentation model that integrates CNNs, Transformers, edge detection techniques, feature fusion mechanisms, and diffusion-based denoising methods. By employing various information interaction mechanisms, this model effectively leverages both local details and global context information.

2) We propose two specialized blocks, EALE and MTGE, to construct dual U-Net networks. The EALE block, built upon convolutional layers and augmented with a Sobel edge detection operator, is designed to extract and enhance local features. The MTGE block, based on Transformer architecture, incorporates an efficient cross-attention mechanism to capture global information. Importantly, these two pathways facilitate continuous information exchange at all stages, rather than operating in isolation.

3) We present the BGTM module, which exploits the global and dynamic properties of Transformers to integrate abstract, coarse-grained, high-level semantic features from deeper network layers.

4) We leverage the stochastic sampling capabilities of the diffusion model to generate multiple segmentation predictions, thereby quantifying the model's uncertainty in specific regions.

Extensive experiments on challenging tasks such as skin lesion, polyp, and brain tumor segmentation consistently demonstrate that our proposed network achieves superior segmentation performance and generalization ability.

2. Related works

2.1. CNN-based method

Image segmentation involves assigning a label to each pixel to determine its category membership. Fully convolutional networks (FCNs) [8] were among the first to introduce a decoder-like architecture

for generating segmentation masks. DeepLabv3+ [23] incorporated multi-scale information through atrous (dilated) convolutions and a spatial pyramid pooling module. U-Net [9] subsequently introduced skip connections and proposed a symmetric encoder-decoder structure, enhancing the model's ability to preserve spatial information. Building on this foundation, U-Net++ [24] introduced dense skip connection paths, while U-Net3+ [25] further advanced the concept with comprehensive skip connections. Other networks have integrated additional modules with U-Net, such as Res-UNet [26] and Attention-UNet [27]. CE-Net [28] proposed a context encoder to capture more high-level semantic information to improve performance. Considering that 3D images contain richer contextual information and exhibit more complex spatial structures compared to 2D images, specialized convolutional methods have been developed for 3D segmentation, including 3D-UNet [29] and V-Net [30]. However, since convolution operations can only focus on local features, the model has certain limitations in capturing long-range dependencies and global contextual information. Although researchers have continuously proposed new methods to expand the receptive field of convolutions, the issue of local perception has not been fundamentally resolved.

2.2. Transformer-based method

Although CNN-based methods have achieved remarkable success in various medical image segmentation tasks, they are known to have limitations in capturing global contextual information, which is crucial for semantic segmentation. Transformer-based methods, on the other hand, excel at addressing this limitation by effectively capturing long-range dependencies. TransUNet [31] incorporates Vision Transformer blocks [32] into the encoder structure and has demonstrated superior segmentation performance. Swin-UNet [33], inspired by the Swin Transformer [34] and U-Net [9], showcases the potential of pure Transformer networks in medical segmentation. MissFormer [35] introduced the self-attention mechanism in skip connections. Although the self-attention mechanism can attend to all features globally to varying degrees, its lack of explicit modeling of local neighborhood structures may lead to insufficient perception of fine-grained edge features and local texture information.

2.3. Boundary-based method

Boundary information also plays a crucial role in medical image segmentation [36]. In medical images, multiple tissues or organs are frequently in close proximity. Preserving distinct boundaries is essential to prevent the blurring of target areas [37], particularly for segmenting small structures such as skin lesions and tumor boundaries [38]. Traditional edge detection techniques [39], especially operator-based methods [40], have been central to this task, with the Sobel operator being notably prominent. The Sobel operator, a first-order derivative edge detection method, enhances boundary information by calculating gradient changes in both horizontal and vertical directions, thereby sharpening object contours. However, most diffusion-based medical image segmentation methods tend to overlook boundary information, treating all regions uniformly, which can compromise the accuracy of target boundary segmentation. This paper advocates for incorporating the Sobel operator into the model to enhance boundary perception.

2.4. Diffusion-based method

DDPM [12] represents a class of generative models based on Markov chains that transform simple distributions, such as Gaussian distributions, into data sampled from complex distributions. Diffusion models exhibit significant potential in medical image segmentation. The operational principle of DDPM in medical image segmentation [18] is shown in Figure 1. In fact, DDPM leverages stochastic sampling processes to generate implicit segmentation sets, thereby enhancing segmentation performance.

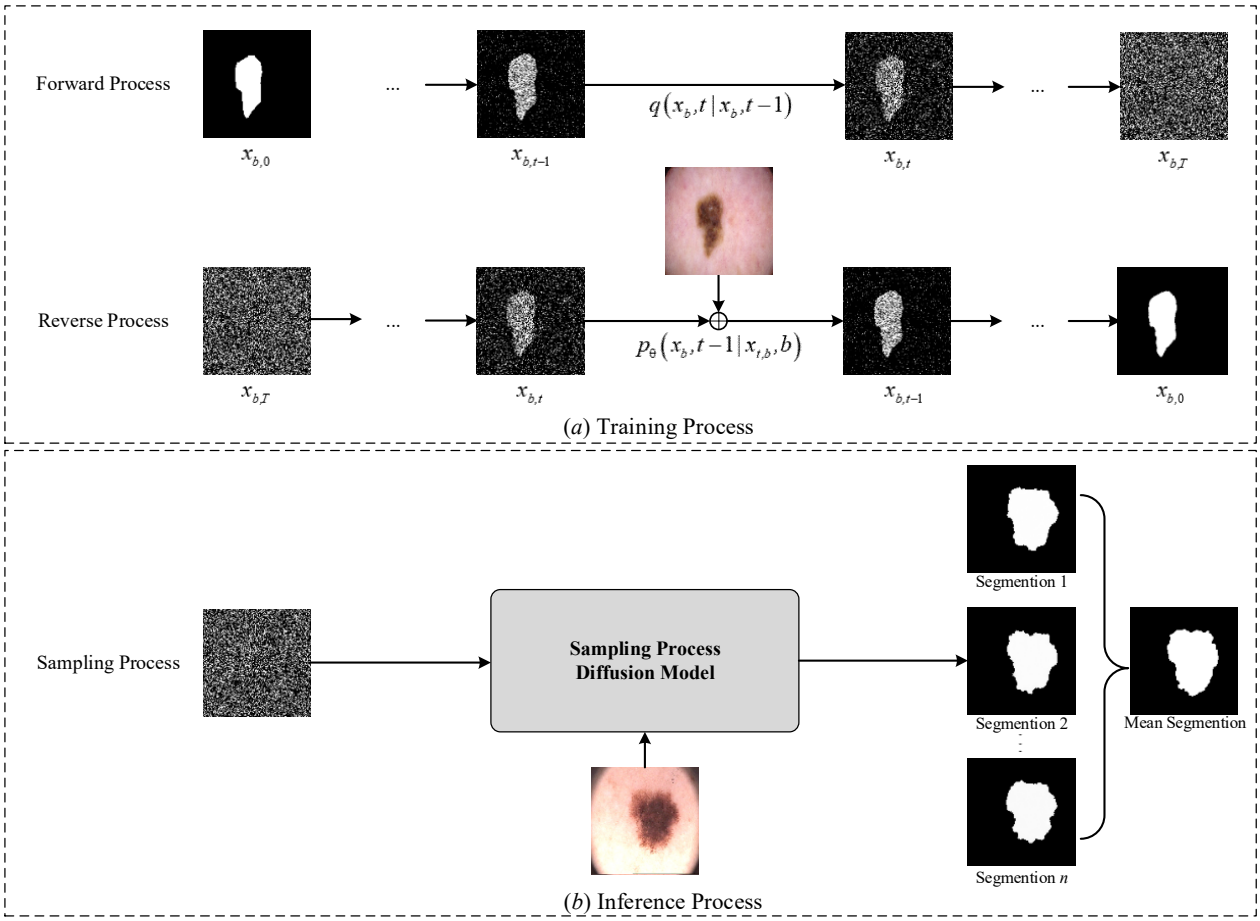

Figure 1. The principle of DDPM. (a) Model training process. (b) Model inference process.

The training process consists of two stages: the forward and the reverse processes. In the forward process, Gaussian noise is progressively added to the original segmentation labels to simulate the degradation from structured labels to random noise. The reverse process learns how to reconstruct clear segmentation labels from any noisy state, guided by the input image. During inference, the process starts from pure Gaussian noise as the initial segmentation map, and under the guidance of the input image, the model progressively denoises and reconstructs the segmentation result until the final prediction is obtained.

SegDiff [20] was pioneering in applying diffusion models to iteratively refine image segmentation. Ensdiff [21] introduced diffusion methods to generate implicit ensembles. MegSegDiff [22] integrated dynamic conditional encoding to adaptively manage varying sampling steps. MedSegDiff-V2 [41]

used Transformer architectures to integrate noise and semantic features. Traditional diffusion models typically use either CNN or Transformer to build the network or directly adopt classic image segmentation architectures (such as U-Net or ViT) as the backbone for the denoising process. However, such single-network structures often suffer from biased capabilities. A single feature extraction approach struggles to simultaneously ensure global semantic consistency and local boundary accuracy, thereby limiting the segmentation performance of the model. In addition, diffusion models need to handle three types of data simultaneously during the denoising process: the noisy image, the time-step encoding, and the original image. This imposes higher modeling demands on the network, requiring it to effectively integrate temporal dynamics, noise distribution, and image content. Existing diffusion model architectures often simply concatenate these inputs or feed them directly into the backbone network, lacking targeted mechanisms for information decoupling and fusion. As a result, the model's understanding and utilization of different information sources are insufficient, which negatively impacts the final reconstruction quality and segmentation performance.

3. Methods

3.1. Diffusion method

DDPM is a class of generative models based on Markov chains, typically used to generate high-quality data samples. The core idea of DDPM involves defining a forward process that gradually adds noise to the data until it reaches a Gaussian distribution and then employing a reverse process to gradually denoise and recover the original data. The process of DDPM in medical image segmentation can be explained using mathematical formulas. In the diffusion process, given a segmentation mask $x_0 \in \mathbb{R}^{H\times W\times 1}$, Gaussian noise is gradually added to it, generating a series of noisy samples until a distribution close to Gaussian noise is obtained, as shown in Eq (1).

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

where x_t is the noisy sample at time step t, β_t is a hyperparameter that changes with t, $\mathcal{N}$ represents a Gaussian distribution with mean μ and covariance matrix $\sum$. x_t can be sampled from x_0 by using the reparameterization trick, as shown in Eq (2).

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I), \tag{2}$$

where $\alpha_t = 1-\beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$, x_t can be represented by Eq (3). After multiple diffusion steps, the segmentation mask x_0 is gradually destroyed, eventually becoming pure noise.

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \;\; \epsilon\sim\mathcal{N}(0,1) \tag{3}$$

In the denoising process, a neural network is trained to progressively denoise through $p_\theta(x_{t-1}|x_t, g)$, ultimately restoring the segmentation mask x_0. The original image $(I \in \mathbb{R}^{H\times W\times C})$ is used as a guide to reconstruct the original structure of the segmentation mask disturbed during the diffusion process, as shown in Eq (4).

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

where the mean $\mu_\theta(x_t, I, t)$ is determined by the noise $\epsilon_\theta(x_t, I, t)$ predicted by the neural network, as shown in Eq (5).

$$\mu_\theta(x_t, I, t) = \left(\frac{1}{\sqrt{\alpha_t}} x_t - \frac{\beta_t}{\sqrt{\bar{\alpha}_t}} \epsilon_\theta(x_t, I, t)\right) \tag{5}$$

Starting from pure noise $x_T \sim \mathcal{N}(0, I)$, each step iterates as shown in Eq (6).

$$x_{t-1} = \mu_\theta(x_t, I, t) + \sigma_t z,\ z \sim \mathcal{N}(0, I), \tag{6}$$

where σ_t is determined by the noise intensity. When $t = 0$, the denoised prediction is obtained.

During actual training, the model learns to predict the noise ϵ added to the mask rather than directly predicting the segmentation mask. The noise prediction loss is shown in Eq (7).

$$\mathcal{L} = \mathbb{E}_{x_0, t, \epsilon}[||\epsilon - \epsilon_\theta(x_t, I, t)||^2] \tag{7}$$

Algorithm 1 Inference Algorithm

Input total diffusion steps T, image I

$x_T \sim \mathcal{N}(0, I_{n \times n})$

for $t = T, T\text{-}1, \ldots, 1$ do

$z \sim N(0, I_{n \times n})$

$\beta = \frac{10^{-4}(T-t) + 2 * 10^{-2}(t-1)}{T-1}$

$\alpha_t = 1 - \beta_t$

$\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$

$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t$

$x_{t-1} = \alpha_t^{-\frac{1}{2}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\right) \int_\theta \left((x_t, I, t)\right) + [t > 1] \tilde{\beta}_t^{\frac{1}{2}} z$

return x_0

Algorithm 2 Training Algorithm

Input total diffusion steps T, images and segmentation

Masks dataset $D = \{(I_k, M_k)\}_k^K = 1$

repeat

Sample $(I_i, M_i) \sim D, \epsilon \sim N(0, I_{n \times n})$

Sample t ~ Uniform({1, ..., T})

$\beta_t = \frac{10^{-4}(T-t) + 2 * 10^{-2}(t-1)}{T-1}$

$\alpha_t = 1 - \beta_t$

$\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$

Take gradient step on $\nabla_\theta \|\epsilon - \epsilon_\theta(x_t, I_i, t)\|, x_t = \sqrt{\bar{\alpha}_t} M_i + \sqrt{1 - \bar{\alpha}_t} \epsilon$

Until convergence

Algorithm 1 represents the inference process, while Algorithm 2 describes the training process. The primary difference between the two is that during training, the noisy images are generated by progressively adding noise to the ground truth. However, during inference, the noisy images are randomly generated through Gaussian sampling.

3.2. Architecture

Our proposed DPUSegDiff model is illustrated in Figure 2. Unlike traditional medical image segmentation methods that directly input raw image data to predict the corresponding segmentation label map, our diffusion model utilizes a noisy segmentation mask, the original image, and a time-guided embedding t as inputs to learn the denoising process.

The structure of DPUSegDiff comprises a dual-path U-Net. One path employs the edge augmented local encoder (EALE) as its fundamental unit to extract image edges and fine details, whereas the other path utilizes the mixed transformer global encoder (MTGE) to capture global and contour-aware features. EALE and MTGE internally integrate a Sobel-guided local information fusion module and a cross-self-attention-guided global information fusion module, respectively. At each stage of upsampling or downsampling, the dual paths share information flow via the embedded local and global fusion modules. In the deeper layers of the network, more abstract and coarse-grained features are fused through the bilateral gated transformer module (BGTM), which leverages the global and dynamic properties of the Transformer to integrate features rich in high-level semantic information. At the end of the model, we designed a dual-path fusion module (DPFM) to selectively integrate the output features of the two U-Net paths, followed by a convolution layer to generate the final prediction. Specifically, DPUSegDiff models each time step of the denoising process through the designed neural network, learning to predict the noise in order to achieve progressive reconstruction. Let the input image be denoted as $g \in \mathbb{R}^{H\times W\times C}$, the segmentation label as $x_t \in \mathbb{R}^{H\times W\times 1}$, and the noise variable as $\epsilon \in \mathcal{N}(0, I)$. The operation mechanism of the diffusion model is described in Section 4.1, and the modeling approach of the diffusion model is shown in Figure 2.

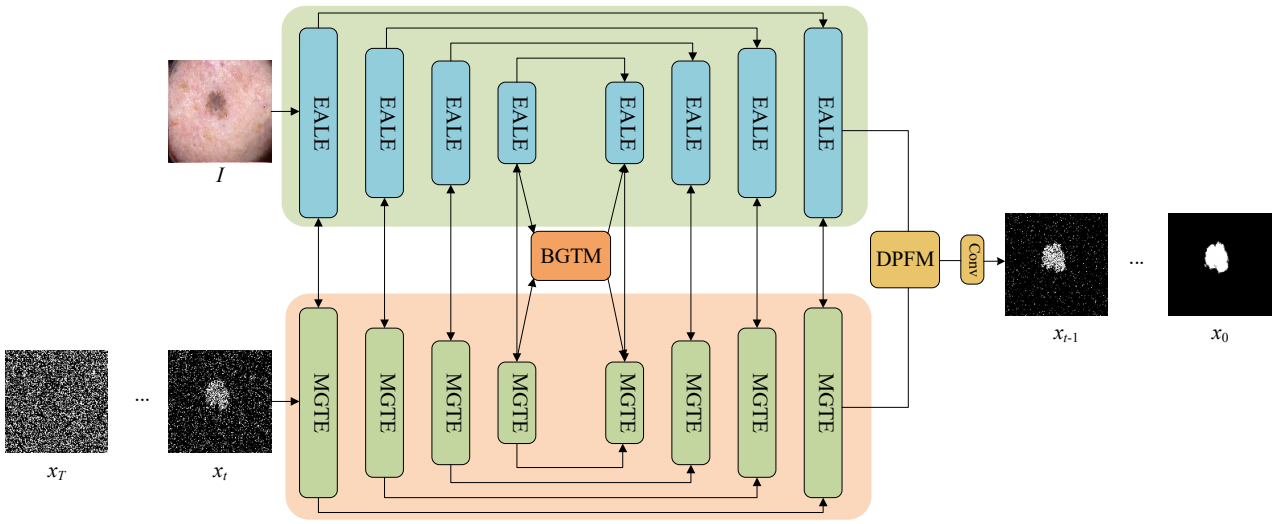

Figure 2. An illustration of DPUSegDiff.

3.2.1. Edge augmented local encoder (EALE)

In the U-Net path focused on local feature extraction, the encoder and decoder share a symmetric structure with a total of 8 stages. Each stage is composed of a ResNet [42] and a local feature fusion module, as shown in Figure 3. This fundamental encoding unit has two advantages: (i) ResNet, with its residual units, enables the convolutional blocks to generate more discriminative feature representations; (ii) the local feature fusion module utilizes the Sobel edge detection operator to extract richer edge information while integrating global semantic information from the other path. We use g^i, $x_t^i \in \mathbb{R}^{H\times W\times C}$ to represent the feature map from the previous stage and the feature map from the other path, respectively. The formulation of this fundamental unit is shown in Eqs (8) and (9).

$$\hat{g}^i=\mathrm{GN}(\mathrm{Convolution}(g^{i-1}))+g^{i-1}, \tag{8}$$

$$\hat{g}_t^i=\mathrm{Convolution}\left(\hat{g}^i\cdot(\mathrm{scale}+1)\cdot\mathrm{shift}\right), \tag{9}$$

where GN represents group norm. The scale and shift matrices are obtained by linearly projecting the timestep *t*, which serves to dynamically suppress or enhance the input features. The adjusted features are then passed through a convolutional block for further feature extraction, as shown in Eq (10).

$$g_t^{edge}=\hat{g}_t^i\otimes\left\{\mathrm{Sobel}\left(\mathrm{Sigmoid}\left(\hat{g}_t^i\right)\right)\right\}, \tag{10}$$

where $\otimes$ represents element-wise multiplication, and g_t^i is first compressed into a single-channel feature using a 1 × 1 convolution before applying a sigmoid activation function, as shown in Eq (11). Subsequently, the Sobel operator is used to extract gradient information (edge features).

$$\hat{x}_t^i=\mathrm{Sigmoid}\left(x_t^i\cdot(\mathrm{scale}+1)\cdot\mathrm{shift}\right), \tag{11}$$

x_t^i undergoes a similar process. Then $\hat{x}_t^i$ and g_t^{edge} are multiplied element-wise and added together to obtain the output, as shown in Eq (12).

$$g_t^i=g_t^{edge}\otimes\hat{x}_t^i+g_t^{edge} \tag{12}$$

The proposed method employs convolution, the Sobel operator, and the sigmoid activation function. The main branch is dedicated to extracting local details of the input image g^{i-1}, such as edges and textures. Meanwhile, the Transformer extracts global features x_t^i to fill in the long-range dependencies that might be missed by the Sobel or convolution operations.

In this encoder, the input ${g_t}^{i-1}$ represents the original image features, which are first passed through two convolutional layers to extract shallow local features. These features are then enhanced using a combination of the sigmoid activation function and the Sobel edge detection operator, aiming to emphasize local edges and texture information in the image. Meanwhile, the other input path, x_t^i, represents the noisy image features. This branch also undergoes normalization via a sigmoid function and is then element-wise multiplied with the enhanced ${g_t}^{i-1}$ features, enabling effective fusion of the original and noisy image information. To prevent information loss and strengthen feature propagation, residual connections are incorporated into all operations within the module.

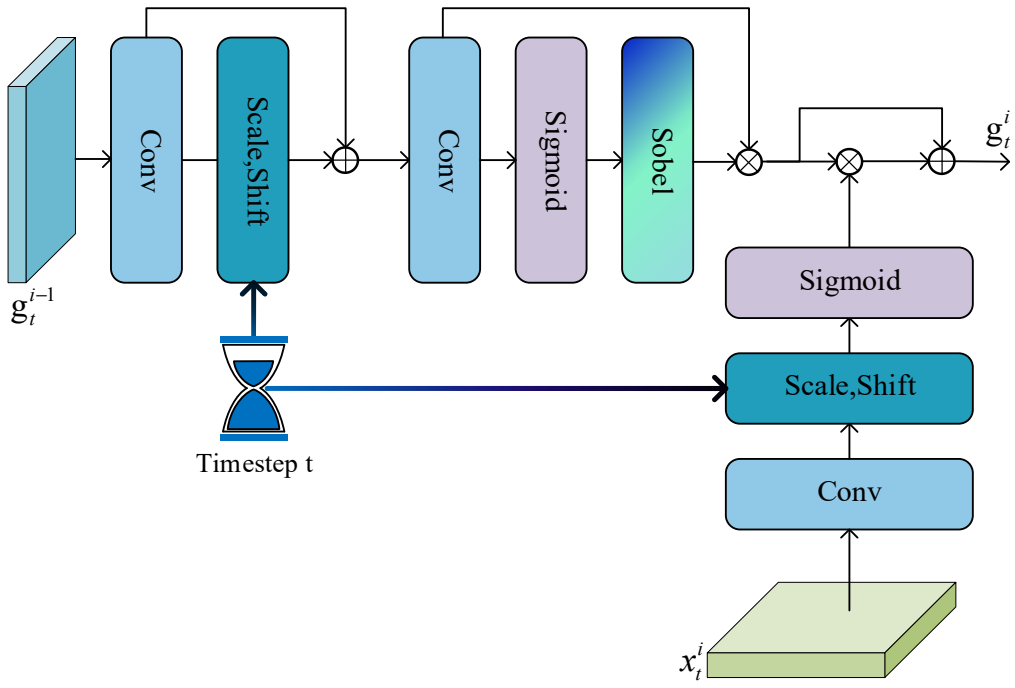

Figure 3. An illustration of EALE.

3.2.2. Mixed transformer global encoder (MTGE)

The global path encoder is constructed based on a Transformer architecture. Figure 4(b) is the main structure of MTGE. In each stage, efficient Transformer blocks are concatenated with self-attention and cross-attention mechanisms. Similar to the local path, the global path also employs Scale and Shift to enhance the feature representation. According to [43], we project a linear matrix, Gate, to adjust the output features of the multilayer perceptron to align the diffusion steps, as shown in Figure 4(a). In Figure 4(a), the time step t is linearly transformed into shift, scale, and gate by a multilayer perceptron. They are categorized as additive $\alpha(\times)$ and multiplicative $\beta(+)$ adjustments based on their respective mechanisms. Figure 4(c) is the efficient self/cross-attention structure. If the input features are used as Q, K, V, it forms self-attention; if the input features are used as Q, and features from the skip connection are used as K, V, it forms cross-attention.

In each stage, self-attention allows each position in the input features to interact with other positions for capturing long-range dependencies. On the other hand, cross-attention derives its Query, Key, and Value from different input features. This mechanism allows the model to align local features with global features, resulting in more precise segmentation boundaries. The computation process of cross-attention is analogous to that of self-attention, as shown in Figure 4(c). Specifically, the input feature $z = [z_1, z_2, \cdots, z_h]$ is divided into multiple heads along the channel dimension. Each head independently performs a simplified self-attention operator SA, and the results are then concatenated back together. The processing procedure is shown in Eqs (13)–(15).

$$E - MHSA(z) = Concat(SA_1(z_1), SA_2(z_2), \cdots, SA_h(z_h))W^p, \tag{13}$$

$$SA(X) = Attention(X \cdot W^Q, P_s(X \cdot W^K), P_s(X \cdot W^V)), \tag{14}$$

$$Attention(Q,K,V) = Soft\,max(\frac{QK^T}{d_k})V, \tag{15}$$

where $Attention$ refers to the self-attention operation. Q, K, and V represent the query, key, and value matrices, respectively, while d_k is the dimension of the key. Generally, shallow feature maps have large spatial resolutions, and attention operations result in computational and memory costs that grow quadratically.

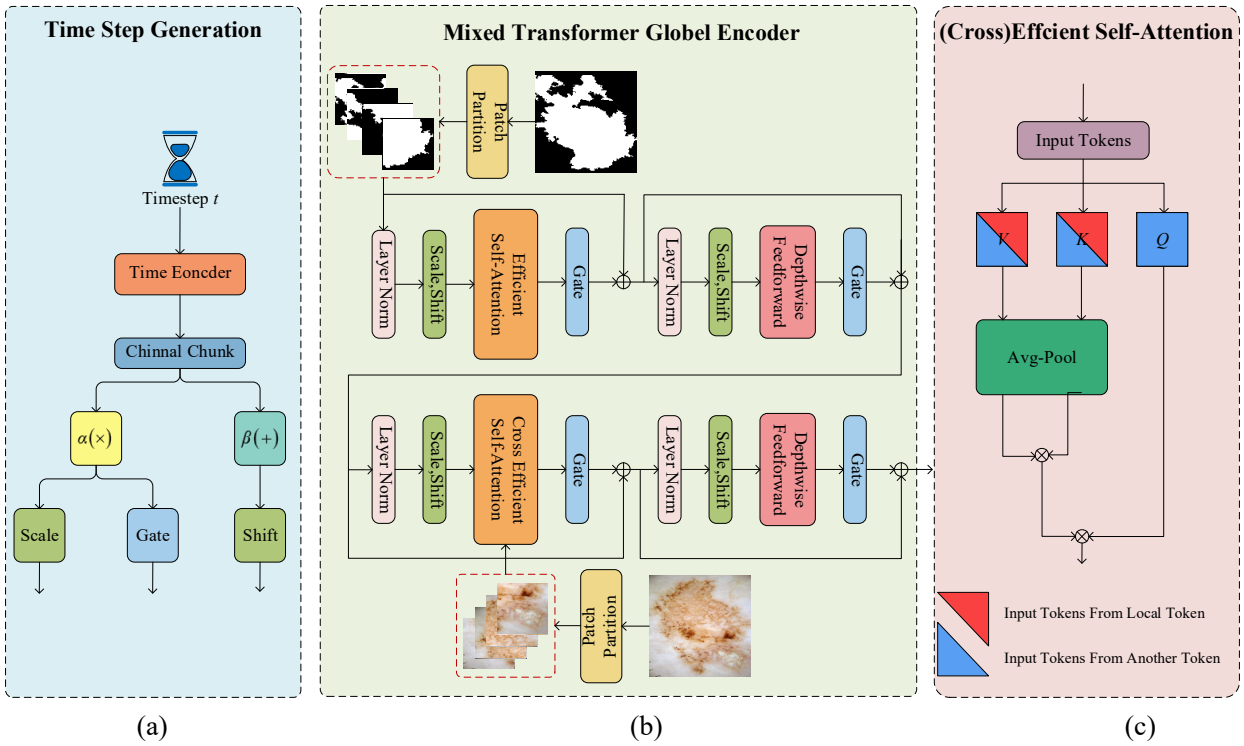

Figure 4. MTGE structure. (a) The time step generation. (b) Mixed Transformer global encoder. (c) Efficient self-attention structure.

To reduce memory usage and computational costs, we introduce a pooling reduction factor P_s at the shallow stages. P_s applies pooling operations only to the K and V matrices, which significantly reduces these overheads while preserving important global context information. Meanwhile, we introduce depthwise separable convolutions [44] (depthwise convolution and pointwise convolution) to replace the traditional multilayer perceptron (MLP). This is because depthwise separable convolution offer the following two advantages: (i) a lower parameter count and computational complexity, which helps improve the model's computational efficiency and training speed; and (ii) compared to MLP, which only models features along the channel dimension, depthwise separable convolution can model local features while preserving spatial structure information. This is particularly crucial for tasks such as medical image analysis, where boundary and structural sensitivity are essential. Depthwise convolution applies $k \times k$ $(k > 1)$ convolution kernels to each channel. The features within the $k \times k$ kernels are aggregated to compute new features. The computation can be expressed as Eqs (16) and (17).

$$Y_{\text{depth}}(i,j,k) = \sum_{p=-P}^{P} \sum_{q=-Q}^{Q} X\,(i+p, j+q, c) \cdot W_{depth}(p,q,c), \tag{16}$$

$$Y(i,j,k) = \sum_{c=1}^{C} Y_{\text{depth}}(i,j,k) \cdot W_{\text{point}}(c,k), \tag{17}$$

where X represents input feature map, W_{depth} and W_{point} represent the depthwise convolution kernel and pointwise convolution kernel, respectively. i,j,p,q are spatial position indices, and c,k are channel position indices.

3.2.3. Bilateral gated transformer module (BGTM)

The deep features in the bottleneck layer encapsulate more abstract and higher-dimensional information. Relying solely on traditional convolution operations makes it challenging to effectively integrate all potential information. Thus, we developed a bilateral gated Transformer structure, as shown in Figure 5, to fully merge the features from both paths. Specifically, Q and K are obtained from the linear projection of feature maps from the same pathway, whereas V originates from the alternate pathway. A gating mechanism is incorporated, where features from the opposite path are processed through a sigmoid activation function and then multiplied with the original path features. This facilitates dynamic modulation of the weight distribution for the original features. The precise calculation is detailed as Eq (18).

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{d_k}\right) \times \left(V \cdot Si\,g\,moid\left(\hat{V}\right) + V\right), \tag{18}$$

where V represents the original features, and $\hat{V}$ represents the features from the contrasting path. The gate captures the importance relationship between the features from the contrasting path and the current features, assigning different attention weights to the original features, thereby enhancing the expression of key features while suppressing redundant or irrelevant ones.

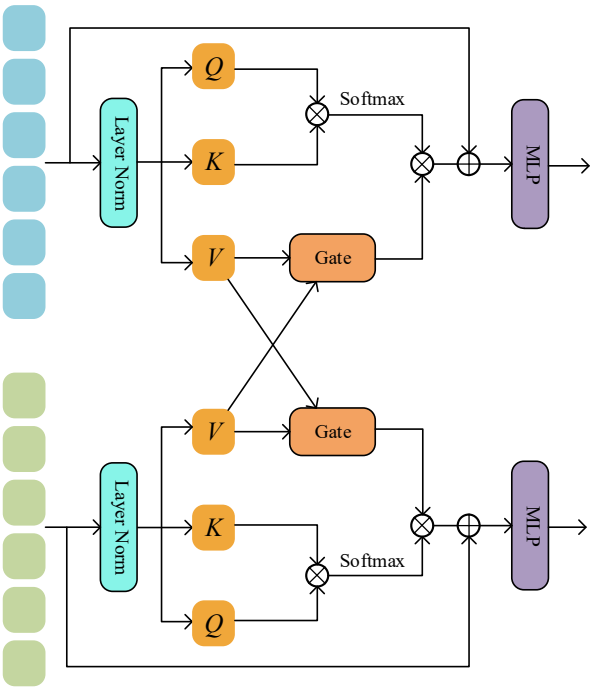

Figure 5. An illustration of BGTM.

3.2.4. Dual-path fusion module (DPFM)

The design of dual-path networks is not commonly seen in existing research, as the information transfer between paths and the fusion of features at the end of the paths present certain challenges. We propose the dual-path fusion module (DPFM), which can merge and reallocate the output features from the two branching paths. It uses learnable parameters $\alpha(\theta)$ to optimize the entire architecture via backpropagation. The output features from one path are multiplied by parameter $\frac{\alpha(\theta)}{2}$, while the output features from the other path are multiplied by $\frac{[1-\alpha(\theta)]}{2}$. The formula can be expressed as Eq (19).

$$I_F = WFM\left(I_x, I_g\right) = \frac{\alpha(\theta)}{2} * I_x + \frac{1-\alpha(\theta)}{2} * I_g, \tag{19}$$

where I_x and I_g are the output feature maps from the EALE and MTGE, respectively. The parameter θ is constrained within the range of 0 to 1 to ensure numerical stability. I_F is the output after fusion, which is then passed through a convolution layer to produce the final prediction results.

4. Experiments

4.1. Implementation details

The proposed DPUSegDiff is fully implemented in PyTorch, and experiments are conducted in a single GPU setup with an NVIDIA 3090 GPU (24 GB). During the training of the diffusion model, the initial learning rate is 0.0001, the batch size is 16, the Adam optimizer is used for backpropagation, and it adopts a hybrid loss function combining Dice loss and cross-entropy loss. To enhance the model's robustness, the total number of diffusion steps, T, is set to 250. During the inference stage, to improve the segmentation accuracy, 10 samples are taken for each input image, and the results are averaged to obtain the final segmentation mask. During the experiments, we resized the input resolution of all datasets to 128 × 128 and applied various combinations of data preprocessing using the Albumentations library, including but not limited to scaling, translation, rotation, brightness adjustment, RGB channel enhancement, Gaussian noise, and Gaussian blur. To ensure a unified and comprehensive evaluation of performance across all datasets, we adopted four metrics as our evaluation criteria: Dice similarity coefficient (DSC), intersection over union (IoU), recall, and specificity. The calculation methods of the four evaluation metrics are as follows:

1) DSC: The segmentation capability of a model is typically measured using the Dice similarity coefficient (also known as F1 score), as shown in Eq (20). This coefficient represents the similarity between two samples, with a range of [0, 1], where a higher value indicates better model performance.

$$Dice = \frac{2TP}{2TP+FN+FP}, \tag{20}$$

where TP represents true positives, FP represents false positives, and FN represents false negatives.

2) IoU: Intersection over union (IoU) is a standard performance evaluation metric for image segmentation problems. It is also used to measure the similarity between two samples. The IoU score is calculated as shown in Eq (21).

$$IoU = \frac{TP}{TP+FP+FN} \tag{21}$$

The range of IoU is also [0, 1]. It is important to note that Dice calculates the ratio of the intersection between the predicted and ground truth regions to the union of those regions, while IoU calculates the weighted proportion of the intersection between the predicted and ground truth regions relative to the total area. The value of DSC is usually higher than that of IoU because DSC gives more weight to the intersection, enhancing the contribution of true positives (TP).

3) Recall: Recall measures the ability of the model to correctly identify all positive samples. Its calculation formula is shown in Eq (22).

$$Recall = \frac{TP}{TP+FN} \tag{22}$$

4) Specificity: Specificity measures the model's ability to correctly identify negative samples (background). The formula is shown in Eq (23).

$$Specificity = \frac{TN}{TN+FP} \tag{23}$$

4.2. Task 1: Skin lesion segmentation

4.2.1. Skin disease dataset

We first validated the proposed DPUSegDiff for skin lesion segmentation. Skin lesion segmentation is an important task in medical image analysis, often used to assist in the diagnosis of diseases such as malignant melanoma. This experiment involves the ISIC 2018 [45] and PH^2 datasets [46]. ISIC 2018 is a widely used skin lesion dataset that includes a rich variety of lesion areas, presenting significant challenges. It helps evaluate the model's performance in complex scenarios, such as blurred boundaries and noise interference. The ISIC 2018 dataset comprises 2594 RGB images, which are divided into training, validation, and test sets at 70%, 10%, and 20%, respectively. The PH^2 dataset, consisting of merely 200 images, is relatively small. Consequently, we refrain from splitting the PH^2 dataset for training purposes and instead employ it to evaluate the generalization capabilities of the ISIC 2018 dataset.

4.2.2. Segmentation performance

Table 1 presents the performance of various methods on the skin lesion segmentation task. The column labeled ISIC 2018 lists the evaluation metrics for each model on the ISIC 2018 dataset. From the experimental results, it can be observed that the proposed DPUSegDiff method significantly outperforms existing mainstream approaches across all metrics.

Specifically, compared to the classic CNN-based U-Net, DPUSegDiff achieves a 4% improvement in DSC, 2% in IoU, 5% in recall (Rec), and 2% in specificity (SP). When compared to the popular Transformer-based architecture TransUNet, DPUSegDiff shows a 5% gain in DSC, 4% in IoU, 5% in Rec, and 2% in SP. Even against the advanced hybrid method UCTransNet, which integrates both CNN and Transformer structures, DPUSegDiff still achieves improvements of 2% in

DSC, 3% in Rec, and 2% in SP. Furthermore, compared to the current state-of-the-art diffusion model MedSegDiff, DPUSegDiff achieves a 1% increase in DSC.

Table 1. Segmentation metrics for skin disease datasets.

Methods	ISIC 2018				ISIC 2018 → PH^2			
	DSC (%)	IOU (%)	Rec (%)	SP (%)	DSC (%)	IOU (%)	Rec (%)	SP (%)
U-Net [9]	85.45	80.69	82.18	96.97	89.36	84.33	87.43	95.88
DAGAN [17]	88.07	83.53	85.38	95.88	90.01	84.25	88.15	96.40
TransUNet [31]	84.99	80.11	82.47	96.53	89.98	83.70	88.04	94.27
Swin-UNet [33]	89.46	84.29	87.26	97.98	91.03	84.78	88.79	95.64
DeepLabv3+ [23]	88.20	83.77	86.12	97.70	90.62	85.03	87.94	98.32
Att-UNet [27]	85.66	81.26	82.88	98.63	90.03	84.76	87.36	96.40
UCTransNet [47]	88.38	84.35	84.72	96.29	90.93	84.08	87.62	95.35
MissFormer [35]	86.31	82.17	85.24	96.58	85.50	80.71	84.33	94.17
EnsDiff [21]	87.75	83.54	85.34	98.12	91.17	85.68	88.28	97.74
Segdiff [20]	88.78	84.18	86.46	98.46	90.18	84.51	88.10	98.06
MedSegDiff [22]	89.17	84.35	86.88	98.36	90.73	84.64	88.42	98.67
DPUSegDiff	90.36	84.72	87.38	98.25	91.67	85.57	88.57	98.13

These quantitative results clearly demonstrate that the proposed DPUSegDiff possesses stronger feature representation capabilities and more precise target region recognition in the skin lesion segmentation task. Whether compared with traditional CNN-based methods, Transformer-based methods, or hybrid models that combine the strengths of both, DPUSegDiff consistently shows a clear advantage. At the same time, DPUSegDiff also outperforms diffusion models that adopt a single network structure, highlighting the higher effectiveness and robustness of our design in terms of information fusion and denoising mechanisms.

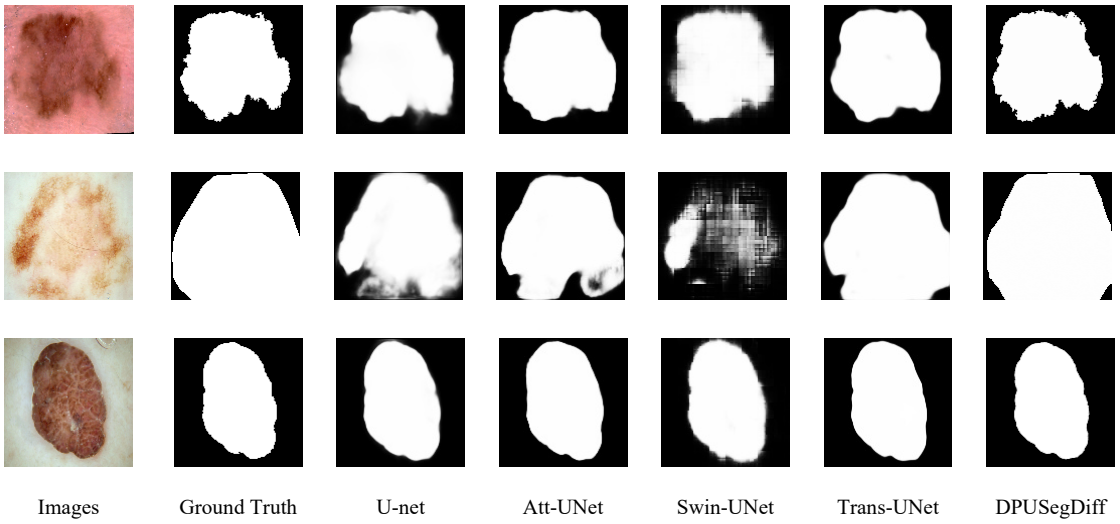

Figure 6. Visualization of segmentation results for skin disease datasets.

Figure 6 illustrates the visual segmentation results of different models on representative samples. As shown in the figure, for lesion regions with clear boundaries and regular shapes, all models achieve satisfactory segmentation performance. However, in more challenging cases—such as lesions with blurred edges or complex structures—DPUSegDiff is able to more accurately reconstruct the target regions, with segmentation boundaries that closely match the ground truth. Other methods commonly suffer from over-segmentation, under-segmentation, or mis-segmentation issues, indicating limitations in their generalization ability under complex conditions.

To further validate the generalization ability of each model, we conducted transfer testing on the PH^2 dataset using model weights trained on the ISIC 2018 dataset. The results show that DPUSegDiff consistently outperforms other methods across all evaluation metrics, further confirming its excellent generalization capability and robustness.

In summary, both quantitative analysis and qualitative evaluation demonstrate that DPUSegDiff achieves superior segmentation performance compared to existing methods, highlighting its broad applicability and practical value in the task of skin lesion segmentation.

4.3. Task 2: Polyp segmentation

4.3.1. Polyp dataset

The second evaluation dataset comprises the colonoscopy images from the CVC-ClinicDB [48] dataset, which are used for early detection and diagnosis of colorectal cancer. CVC-ClinicDB is a colorectal cancer dataset that has been validated by numerous studies and includes standardized segmentation annotations. It has been used by multiple research institutions, which can provide a reliable benchmark for evaluating our model. The experiment employs 612 images in total, with 428 images for training, 61 for validation, and 123 for testing.

4.3.2. Segmentation performance

Table 2. Segmentation metrics for Polyp datasets.

Methods	ClinicDB			
	DSC (%)	IoU (%)	Rec (%)	SP (%)
U-Net [9]	89.28	84.32	85.15	98.13
AttU-Net [27]	89.51	84.24	85.24	99.13
Res-UNet [26]	91.10	85.60	89.67	99.34
U-Net++ [24]	88.94	83.33	84.53	97.71
Deeplabv3+ [23]	90.33	84.75	89.29	98.54
BCDU-Net [49]	89.01	83.62	87.46	97.57
CE-Net [28]	89.35	83.94	88.20	98.66
Trans-UNet [31]	90.30	84.98	86.55	99.02
Swin-UNet [33]	89.47	83.78	87.84	99.58
EnsDiff [21]	90.57	84.22	89.69	99.58
Segdiff [20]	90.26	84.85	88.71	99.47
DPUSegDiff	91.85	86.59	90.55	99.64

Table 2 quantitatively presents the segmentation performance of various models on the polyp dataset. Figure 7 qualitatively illustrates their segmentation results from a visual perspective. Compared to the classic method U-Net, our approach improves DSC and IoU by 2% and Rec by 5%. When compared to the mainstream method, TransUNet, it achieves a 1% gain in DSC and IoU, and a 4% increase in Rec. These quantitative and qualitative results demonstrate the effectiveness and superiority of our proposed method.

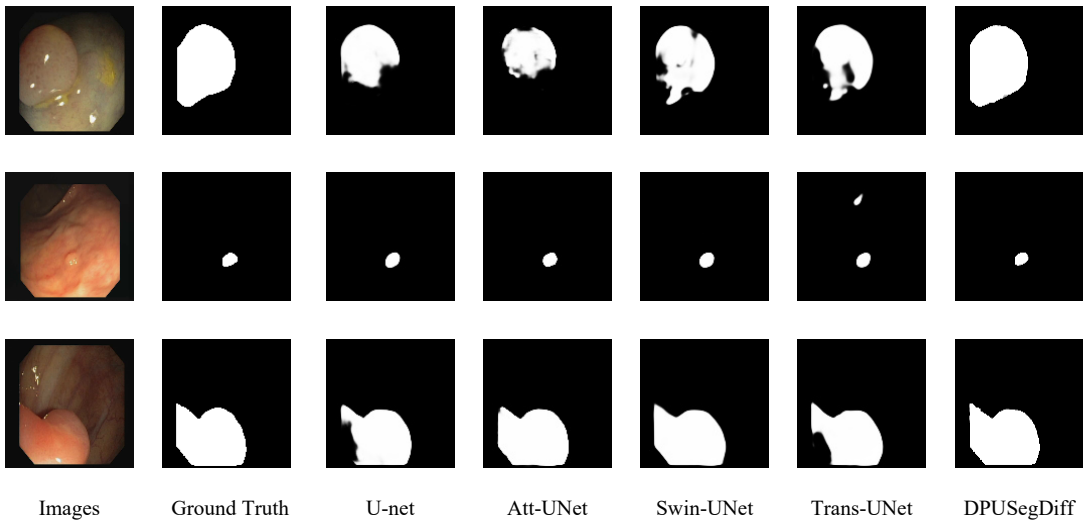

Figure 7. Visualization of segmentation results for Polyp datasets.

4.4. Task 3: Brain tumor segmentation

4.4.1. Brain tumor dataset

The third evaluation centers on the BraTS brain tumor segmentation challenge dataset. This dataset comprises MRI scans from 369 unique patients. Unlike other datasets, the BraTS dataset offers four different MRI modalities for each case: FLAIR, T1w, T1gd, and T2w. These multimodal datasets provide rich texture and structural information, which aids in enhancing segmentation performance. We organized the dataset following the previously used training, validation, and testing splits, and subsequently processed the 2D slices to align with the model's input format.

4.4.2. Segmentation performance

Table 3 provides a quantitative analysis of the evaluation metrics for different methods on the BraTS dataset [50]. The results indicate that DPUSegDiff outperforms all compared models across the board. Specifically, on the brain tumor dataset, our method achieves a significant improvement of 11% in both DSC and IoU, and a 13% increase in Rec compared to the CNN-based U-Net. When compared with the Transformer-based Swin-UNet, DPUSegDiff also shows a 4% improvement in DSC.

Table 3. Segmentation metrics for brain tumor datasets.

Methods	BraTs			
	DSC (%)	IoU (%)	Rec (%)	SP (%)
U-Net [9]	77.63	71.39	72.95	95.14
AttU-Net [27]	79.18	73.62	75.26	97.83
Res-UNet [26]	78.42	72.37	74.58	96.24
U-Net++ [24]	82.33	75.52	78.34	97.61
Trans-UNet [31]	86.54	78.76	82.61	98.65
Swin-UNet [33]	85.72	78.17	81.57	98.39
EnsDiff [21]	87.26	81.55	83.45	99.23
SegDiff [20]	86.86	80.53	82.98	99.47
DPUSegDiff	89.18	82.63	85.36	99.78

Figure 8 presents the qualitative visual segmentation results of different models on test samples. As shown in the figure, although all methods can roughly segment the tumor regions, DPUSegDiff demonstrates superior performance in capturing tumor details and produces sharper and more accurate boundaries. These results strongly demonstrate that DPUSegDiff significantly improves both the accuracy and stability of segmentation in brain tumor segmentation tasks.

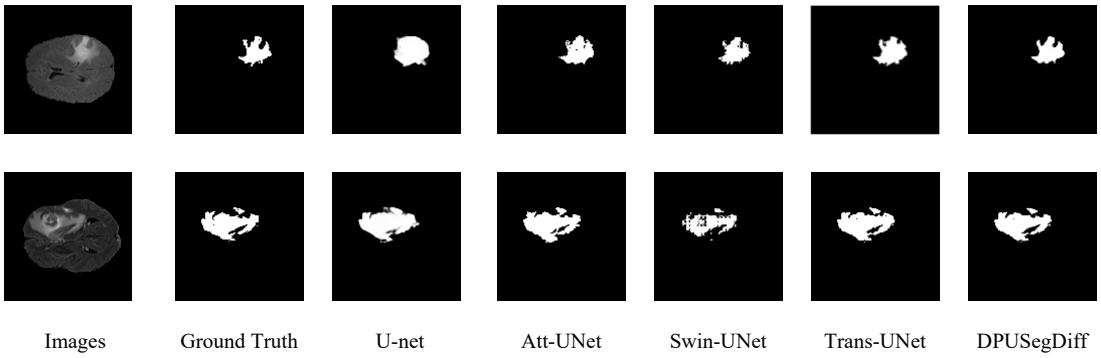

Figure 8. Visualization of segmentation results for brain tumor datasets.

4.5. Ablation study

We conducted a comprehensive ablation study to validate the effectiveness of the proposed modules. The results are presented in Table 4. The symbol √ indicates that the module is deployed in the network, whereas × signifies that the module is not used. Specifically, EALE represents the scenario where the Sobel edge detection operator is not used, and MTGE signifies the scenario without the cross-attention mechanism. As indicated in the table, both EALE and MTGE significantly improved fusion performance. Moreover, the bilateral gated Transformer and DPFM also proved to be effective in enhancing the segmentation performance.

Table 4. Ablation study of DPUSegDiff on each component.

Ablation components				ISIC	Clinicdb
EALE	MTGE	Transformer	DPFM	DSC (%)	DSC (%)
×	√	√	√	88.68	89.63
√	×	√	√	88.79	88.95
√	√	×	√	89.75	90.32
√	√	√	×	89.53	90.94
√	√	√	√	90.36	91.85

5. Discussion

According to [21], implicit integration is a key concept in diffusion model-based methods, primarily related to the sampling strategy and stability during the generation process. It typically refers to the process of implicitly optimizing the target distribution through multiple iterative sampling steps during the reverse process of the diffusion model (generating data from noise). The importance of multiple sampling runs in implicit integration is rooted in the following reasons: (i) Each sampling can be considered an implicit optimization step, and (ii) multiple runs enable the exploration of various generation paths and assist in avoiding entrapment in local optima.

In Figure 9, we present the original image, Ground Truth, two different sampled segmentation masks, as well as the mean and variance maps. From the figure, it is evident that each sampling introduces some regions of uncertainty. This uncertainty typically arises from the model's noise prediction bias or the inherent randomness in the generation process. By conducting multiple samples on the same input and averaging the outcomes, we can significantly diminish the randomness of the predictions, reduce the model's uncertainty, and make the final segmentation results more stable and closer to the ground truth.

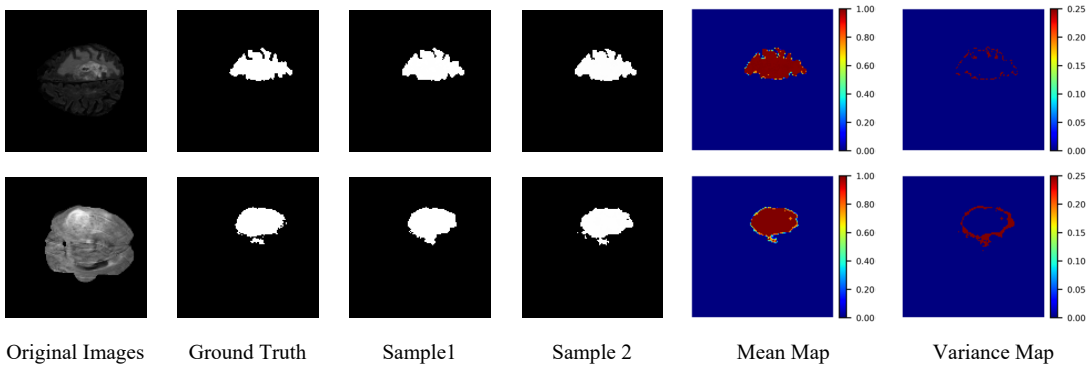

Figure 9. Visualization of multiple sampling results and implicit integration results.

Figure 10 illustrates multiple batches and depicts the relationship between the sampling iterations in the ensemble and the Dice score. We can observe that as the sampling iterations increase, the Dice score gradually rises and then stabilizes.

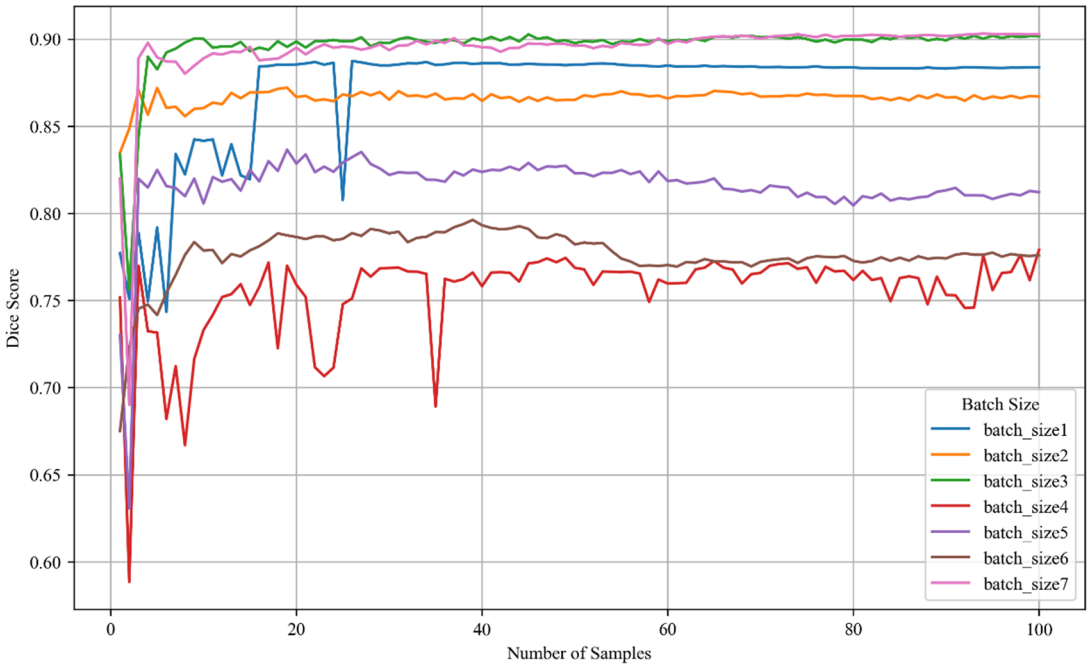

Figure 10. The trend of segmentation performance (average Dice score) with respect to the sampling iterations.

We randomly selected several batches from the test dataset, each with a batch size of 8, and performed 100 independent samplings for each batch. The segmentation results from each sampling were integrated with the previously obtained results, and the average Dice score was calculated accordingly. This process was used to evaluate the model's stability and performance variation under different numbers of sampling iterations.

6. Conclusions

This paper introduces a dual-path U-Net diffusion segmentation model with rich information interaction, called DPUSegDiff. The model leverages a Markov chain denoising process and generates implicit ensemble predictions with multiple outcomes. Specifically, the two paths of DPUSegDiff apply CNN and Transformer modules, focusing on the extraction of local and global features. To enhance the information flow between the two pathways, we have integrated information fusion modules within each pathway, tailored to their respective emphases. These fusion modules incorporate traditional edge detection algorithms and cross-attention mechanisms, thereby sensitively enhancing the information extracted by the dual-path encoder. The proposed BGTM selectively integrates deeper features from the bottleneck layer, effectively bridging the two pathways. Extensive experiments conducted on three challenging tasks—skin lesion, polyp, and brain tumor—consistently demonstrate that the proposed DPUSegDiff achieves superior segmentation performance and generalization capability.

Although DPUSegDiff demonstrates significant advantages in image segmentation tasks, our work has certain limitations. Inference speed remains one of the main limitations of current diffusion models. Specifically, diffusion models typically require hundreds or even thousands of iterative denoising steps to reconstruct high-quality predictions from pure noise, which leads to considerable

computational cost and inference latency in practical applications. Although some recent methods have attempted to accelerate the inference process by reducing the number of sampling steps, these approaches often suffer from accuracy degradation or reduced generalization ability. As such, it remains challenging to fully overcome the speed bottleneck without sacrificing performance. Therefore, we believe that optimizing inference efficiency will continue to be an important research direction for diffusion models in the future.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by the Natural Science Foundation of Fujian Province (Grant Nos. 2024J01821, 2024J01820), Research Project on Undergraduate Higher Education Teaching of Fujian Province (Grant No. FBJY20230083), Model Courses of Ideological and Political Education in Graduate Programs at Minnan Normal University (Grant No. SZ202406), Research Project on Postgraduate Education and Teaching Reform of Minnan Normal University (Grant No. YJG202316) and the Principal Foundation of Minnan Normal University (Grant No. KJ18010)

Author contributions

Yazhuo Fan was responsible for proposing the research idea, designing the core structure of the medical image segmentation model, implementing the algorithm, conducting the experiments, and drafting the initial manuscript. **Jianhua Song** supervised the overall research, provided methodological support, contributed to the model design and optimization, and revised the manuscript extensively. **Yizhe Lu** assisted in the model development and parameter tuning, and contributed to the analysis and visualization of the experimental results. **Xinrong Fu** participated in data preprocessing, experimental pipeline setup, and model evaluation, and contributed to interpreting the results. **Xinying Huang** conducted literature review, supported model validation, and contributed to writing part of the manuscript. **Lei Yuan** provided clinical insights and domain expertise, supported medical image dataset preparation, and offered constructive suggestions for improving the final manuscript.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: A survey, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968

2. R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, A. K. Nandi, Medical image segmentation using deep learning: A survey, *IET Image Process.*, **16** (2022), 1243–1267. https://doi.org/10.1049/ipr2.12419
3. U. Ilhan, A. Ilhan, Brain tumor segmentation based on a new threshold approach, *Procedia Comput. Sci.*, **120** (2017), 580–587. https://doi.org/10.1016/j.procs.2017.11.282
4. A. Pratondo, C. K. Chui, S. H. Ong, Integrating machine learning with region-based active contour models in medical image segmentation, *J. Visual Commun. Image Represent.*, **43** (2017), 1–9. https://doi.org/10.1016/j.jvcir.2016.11.019
5. P. Singh, Y. P. Huang, An ambiguous edge detection method for computed tomography scans of Coronavirus Disease 2019 cases, *IEEE Trans. Syst. Man Cybern.: Syst.*, **54** (2024), 352–364. https://doi.org/10.1109/TSMC.2023.3307393
6. P. Singh, Y. P. Huang, AKDC: Ambiguous kernel distance clustering algorithm for COVID-19 CT scans analysis, *IEEE Trans. Syst. Man Cybern.: Syst.*, **54** (2024), 6218–6229. https://doi.org/10.1109/TSMC.2024.3418411
7. P. Singh, S. S. Bose, A quantum-clustering optimization method for COVID-19 CT scan image segmentation, *Expert Syst. Appl.*, **185** (2021), 115637. https://doi.org/10.1016/j.eswa.2021.115637
8. E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (2017), 640–651. https://doi.org/10.1109/TPAMI.2016.2572683
9. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, Lecture Notes in Computer Science*, Springer, **9351** (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
10. Y. Fan, J. Song, L. Yuan, Y. Jia, HCT-Unet: Multi-target medical image segmentation via a hybrid CNN-transformer Unet incorporating multi-axis gated multi-layer perceptron, *Vis. Comput.*, **41** (2024), 3457–3472. https://doi.org/10.1007/s00371-024-03612-y
11. A. Pratondo, C. K. Chui, S. H. Ong, Robust edge-stop functions for edge-based active contour models in medical image segmentation, *IEEE Signal Process. Lett.*, **23** (2016), 222–226. https://doi.org/10.1109/LSP.2015.2508039
12. J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*, Curran Associates Inc., (2020), 6840–6851.
13. H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, et al., SRDiff: Single image super-resolution with diffusion probabilistic models, *Neurocomputing*, **479** (2022), 47–59. https://doi.org/10.1016/j.neucom.2022.01.029
14. J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, P. Milanfar, Deblurring via stochastic refinement, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2022), 16272–16282. https://doi.org/10.1109/CVPR52688.2022.01581
15. B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, DiffIR: Efficient diffusion model for image restoration, in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, (2023), 13049–13059. https://doi.org/10.1109/ICCV51070.2023.01204

16. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial networks, preprint, arXiv:1406.2661.
17. B. Lei, Z. Xia, F. Jiang, X. Jiang, Z. Ge, Y. Xu, et al., Skin lesion segmentation via generative adversarial networks with dual discriminators, *Med. Image Anal.*, **64** (2020), 101716. https://doi.org/10.1016/j.media.2020.101716
18. A. Rahman, J. M. J. Valanarasu, I. Hacihaliloglu, V. M. Patel, Ambiguous medical image segmentation using diffusion models, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2023), 11536–11546, https://doi.org/10.1109/CVPR52729.2023.01110
19. A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, et al., Diffusion models in medical imaging: A comprehensive survey, *Med. Image Anal.*, **88** (2023), 102846. https://doi.org/10.1016/j.media.2023.102846
20. T. Amit, T. Shaharbany, E. Nachmani, L. Wolf, SegDiff: Image segmentation with diffusion probabilistic models, preprint, arXiv:2112.00390.
21. J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, P. C. Cattin, Diffusion models for implicit image segmentation ensembles, preprint, arXiv:2112.03145.
22. J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, et al., MedSegDiff: Medical image segmentation with diffusion probabilistic model, preprint, arXiv.2211.00611.
23. L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in *Computer Vision – ECCV 2018*, Springer, (2018), 833–851. https://doi.org/10.1007/978-3-030-01234-2_49
24. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. 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 (DLMIA)*, Springer, **11045** (2018), 3–11. https://doi.org/10.1007/978-3-030-00889-5_1
25. H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, et al., UNet 3+: A full-scale connected UNet for medical image segmentation, in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, (2020), 1055–1059, https://doi.org/10.1109/ICASSP40776.2020.9053405
26. X. Xiao, S. Lian, Z. Luo, S. Li, Weighted Res-UNet for high-quality retina vessel segmentation, in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, IEEE, (2018), 327–331, https://doi.org/10.1109/ITME.2018.00080
27. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, et al., Attention U-Net: Learning where to look for the pancreas, preprint, arXiv:1804.03999.
28. Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, CE-Net: Context encoder network for 2D medical image segmentation, *IEEE Trans. Med. Imaging*, **38** (2019), 2281–2292. https://doi.org/10.1109/TMI.2019.2903562
29. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: Learning dense volumetric segmentation from sparse annotation, in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, **9901** (2016), 424–432. https://doi.org/10.1007/978-3-319-46723-8_49
30. F. Milletari, N. Navab, S. A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in *2016 Fourth International Conference on 3D Vision (3DV)*, IEEE, (2016), 565–571. https://doi.org/10.1109/3DV.2016.79

31. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, et al., TransUNet: Transformers make strong encoders for medical image segmentation, preprint, arXiv:2102.04306.
32. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929.
33. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, et al., Swin-Unet: Unet-like pure transformer for medical image segmentation, in *Computer Vision – ECCV 2022 Workshops*, Springer, **13803** (2023), 205–218. https://doi.org/10.1007/978-3-031-25066-8_9
34. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, (2021), 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986
35. X. Huang, Z. Deng, D. Li, X. Yuan, MISSFormer: An effective medical image segmentation transformer, preprint, arXiv:2109.07162.
36. Z. Zhu, X. He, G. Qi, Y. Li, B. Cong, Y. Liu, Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI, *Inf. Fusion*, **91** (2023), 376–387. https://doi.org/10.1016/j.inffus.2022.10.022
37. R. Wang, S. Chen, C. Ji, J. Fan, Y. Li, Boundary-aware context neural network for medical image segmentation, *Med. Image Anal.*, **78** (2022), 102395. https://doi.org/10.1016/j.media.2022.102395
38. Y. Lin, D. Zhang, X. Fang, Y. Chen, K. Cheng, H. Chen, Rethinking boundary detection in deep learning models for medical image segmentation, preprint, arXiv.2305.00678.
39. N. Mathur, S. Mathur, D. Mathur, A novel approach to improve Sobel edge detector, *Procedia Comput. Sci.*, **93** (2016), 431–438. https://doi.org/10.1016/j.procs.2016.07.230
40. A. Elnakib, G. Gimel'farb, J. S. Suri, A. El-Baz, Medical image segmentation: A brief survey, in *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*, Springer, (2011), 1–39. https://doi.org/10.1007/978-1-4419-8204-9_1
41. J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, Y. Xu, MedSegDiff-V2: Diffusion-based medical image segmentation with Transformer, preprint, arXiv:2301.11798.
42. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
43. W. Peebles, S. Xie, Scalable diffusion models with transformers, preprint, arXiv:2212.09748.
44. F. Chollet, Xception: Deep learning with depthwise separable convolutions, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2017), 1800–1807. https://doi.org/10.1109/CVPR.2017.195
45. N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, et al., Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC), preprint, arXiv:1902.03368.
46. T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, J. Rozeira, PH^2 - A dermoscopic image database for research and benchmarking, in *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, (2013), 5437–5440. https://doi.org/10.1109/EMBC.2013.6610779
47. H. Wang, P. Cao, J. Wang, O. R. Zaiane, UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer, preprint, arXiv:2109.04335.

48. J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, *Comput. Med. Imaging Graphics*, **43** (2015), 99–111. https://doi.org/10.1016/j.compmedimag.2015.02.007
49. R. Azad, M. Asadi-Aghbolaghi, M. Fathy, S. Escalera, Bi-directional ConvLSTM U-Net with densely connected convolutions, preprint, arXiv:1909.00166.
50. B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al., The multimodal brain tumor image segmentation benchmark (BRATS), *IEEE Trans. Med. Imaging*, **34** (2014), 1993–2024. https://doi.org/10.1109/TMI.2014.2377694

AIMS Press