frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 24 September 2025
DOI 10.3389/fmed.2025.1661984



OPEN ACCESS

EDITED BY
Linfeng Li,
Capital Medical University, China

REVIEWED BY
Zhixiong Huang,
Dalian Nationalities University, China
Zhaojin Fu,
Beijing Information Science and Technology University, China

*CORRESPONDENCE
Yue Luo
✉ luoyue@cdutcm.edu.cn

RECEIVED 08 July 2025
ACCEPTED 02 September 2025
PUBLISHED 24 September 2025

CITATION
Huang Y and Luo Y (2025) Multi-interactive feature embedding learning for medical image segmentation. *Front. Med.* 12:1661984. doi: 10.3389/fmed.2025.1661984



# Multi-interactive feature embedding learning for medical image segmentation

Yijia Huang[1] and Yue Luo[2]*

[1]School of Public Health, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan, China, [2]School of Intelligent Medicine, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan, China

Medical image segmentation task can provide the lesion object semantic information, but ignores edge texture details from the lesion region. Conversely, the medical image reconstruction task furnishes the object detailed information to facilitate the semantic segmentation through self-supervised learning. The two tasks are supplementary to each other. Therefore, we propose a multi-interactive feature embedding learning for medical image segmentation. In the medical image reconstruction task, we aim to generate the detailed feature representations containing rich textures, edges, and structures, thus bridging the low-level details lost from segmentation features. In particular, we propose an adaptive feature modulation module to efficiently aggregate foreground and background features to obtain a comprehensive feature representation. In the medical segmentation task, we propose a bi-directional fusion module fusing all important complementary information between the two tasks. Besides, we introduce a multi-branch visual mamba to capture structural information at different scales, thus enhancing model adaptation to different lesion regions. Extensive experiments on four datasets demonstrate the effectiveness of our framework.

KEYWORDS

**medical image segmentation, self-supervised learning, adaptive feature modulation module, bi-directional fusion module, multi-branch vision mamba**

## 1 Introduction

Medical image segmentation tasks (1–5) focus on extracting lesion regions from complex medical images, thereby assisting doctors to perform subsequent disease diagnosis, treatment planning and efficacy assessment. In particular, skin lesion segmentation and cell boundary detection tasks enable precise localization of key tissues or lesions, which supports in early diagnosis and clinical assisted decision making by visualizing lesion results (6). Therefore, in public health management, deep learning-based medical image segmentation methods can effectively improve the efficiency of group patient lesion detection. These methods can help public health departments to better monitor and predict the disease spread, thereby promoting disease prevention and treatment.

Existing medical segmentation methods (7, 8) construct complex network structures to improve performance, but ignore texture and boundary detail information about lesion regions in medical images. U-Net (9) introduces encoder-decoder structure, and designs skip connections to combine the different-level semantic information. UNet++ (10) adds dense jump paths and nested decoders to enhance multiscale feature learning. MFSNet (11) combines multi-scale feature extraction and attention mechanisms, which

further improves segmentation performance. However, medical image segmentation task emphasizes on extracting high-level semantic features, resulting in the loss of pixel-level detail information. In contrast, the medical image reconstruction task can provide pixel-level detail information (e.g., texture and boundaries) to the medical image segmentation task through a self-supervised learning strategy, thus obtaining more accurate segmentation results.

Moreover, since convolutional neural network (CNN)—based segmentation methods (12–14) rely on local receptive fields and convolutional structures, it is difficult to effectively capture the non-local relations and structural ambiguity features present in the lesion region. Therefore, Transformer-based segmentation methods (15–18) aim to improve modeling ability for global context, thus enhancing semantic consistency and regional integrity. For example, TransUNet (19) combines the local feature learning of CNN and the global context learning of Transformer advantages. TransFuse (20) designs a two-branch network to capture local and global features, and then fuses them using a fusion module in the decoding stage. This architectural design enhances the model's capability to capture fine-grained boundaries and structural information, thereby improving segmentation accuracy. Although Transformer-based methods can help to recognize organ contours, lesion shapes, and spatial layouts by capturing distant dependencies in medical images through a self-attention mechanism, they require high computational and memory costs. Compared with Transformer-based architectures, Mamba (21, 22) offers lower computational overhead while maintaining strong long-sequence modeling and structural awareness. This is especially valuable in medical image segmentation, where accurate delineation of anatomical structures requires modeling long-range dependencies and preserving fine-grained spatial details. By efficiently extracting spatially hierarchical features, Mamba enables real-time and resource-constrained applications while ensuring precise boundary segmentation.

In this paper, we propose a multi-interactive feature embedding learning (MFEL) for medical image segmentation. Specifically, MFEL consists of a feature interaction-driven image reconstruction (FIIR) and a feature-embedded representation image segmentation (FRIS). On the one hand, FIIR reconstructs the foreground image, background image and medical image through self-supervised learning, thus extracting complete pixel-level features. In particular, an adaptive feature modulation module effectively enhances foreground and background feature representation via the learned modulation parameters, thereby obtaining a more comprehensive and fine-grained pixel-level feature information. On the other hand, FRIS aims to fuse the two different-level features between the reconstruction and segmentation tasks, thereby improving the performance of segmentation task. In particular, a bi-directional fusion module is designed to fuse the feature representations from two tasks, which enhances the information interaction. Moreover, a multi-branch vision mamba utilizes the parallel branching structure and linear state space modeling capability, improving model semantic understanding about different lesion regions.

Our contributions can be summarized as follows:

- We explore an MFEL framework between medical image reconstruction task and medical image segmentation task, and then achieve superior segmentation performance.
- An adaptive feature modulation module is proposed to construct modulation parameters from foreground and background features, thus obtaining a comprehensive pixel-level feature representation.
- A bi-directional fusion module is introduced to establish complementary relationships between structural details and deep semantics, thus enabling feature information interaction between two different-level tasks.
- Multi-branch vision mamba is designed to combine state-space modeling and multi-branch parallel mechanism, efficiently modeling the multi-scale structural information from lesion regions.

## 2 Related work

### 2.1 Medical image segmentation methods

Convolutional neural networks (CNNs) have achieved remarkable success in medical image segmentation by leveraging hierarchical representations and strong inductive biases (23–26). Recent methods enhance segmentation performance by integrating boundary cues and multi-scale features. DCSAU-Net (27) introduces a split attention mechanism with semantic retention, while U-Net v2 (28) incorporates boundary information to refine local detail representations. Transformer-based architectures address CNNs' limitations in modeling long-range dependencies. These models exhibit strong global context awareness and have demonstrated competitive performance in medical image segmentation (19, 29–35). CASF-Net (36) employs dual-branch modeling to combine global semantics and fine-grained features. CSWin-UNet (18) utilizes cross-shaped window attention to improve spatial interactions with low computational cost. Hybrid designs, such as TBConvL-Net (37) and MobileUNETR (38), further balance local detail extraction and global reasoning. Since medical image segmentation as a high-level vision task focuses on extracting semantic structural information, the pixel-level details are ignored. In contrast, we introduce the medical image reconstruction task to learn fine-grained feature representations through self-supervised learning, thus bridging the shortcomings from the semantic segmentation task.

### 2.2 Self-supervised learning methods

Self-supervised learning methods have been widely applied in tasks such as image reconstruction (39, 40), inpainting (41–43), and enhancement (44, 45). For example, Self-path (46) introduces a region-aware contrastive learning framework, which enforces consistency between local and global representations. This strategy effectively enhances feature discrimination and contextual modeling for downstream segmentation tasks. DSFormer (47), designed for multi-contrast MRI reconstruction, proposes a dual-domain self-supervised Transformer architecture. It performs joint reconstruction and context restoration in both k-space

and image space, achieving collaborative modeling of structural information and significantly improving reconstruction quality and generalization. MiM (48) targets 3D medical image analysis by proposing a hierarchical Mask-in-Mask masking mechanism. Through a coarse-to-fine masking strategy combined with residual reconstruction, it guides the model to learn rich semantic structures and fine spatial details, thereby improving its adaptability to downstream tasks such as segmentation and classification. In contrast, the medical image reconstruction task guides the model to focus on pixel-level content (e.g., texture, structure, edges), thus compensating the loss of important details in the cell and skin lesion segmentation tasks.

## 2.3 Vision mamba

Mamba (21) is a novel sequence modeling architecture built upon State Space Models. It enables efficient inference while modeling long-range dependencies. Unlike traditional self-attention mechanisms, Mamba introduces learnable state space kernels and applies linear operations in a sliding-window manner. This design supports global modeling while significantly reducing computational complexity, achieving linear time and space costs. VMamba (22) extends Mamba to vision tasks by introducing a 2D Selective Scan mechanism, which aggregates spatial context from multiple directions with linear complexity, achieving superior accuracy and efficiency over Vision Transformers. Compared with CNNs, Mamba is not limited by local receptive fields and can capture global sequential and contextual information. Compared with Transformers, Mamba avoids the high computational overhead of self-attention in long sequences, achieving better efficiency and performance. These advantages make Mamba particularly suitable for high-resolution or 3D medical image tasks. In medical image segmentation (49–51), accurately capturing the spatial structure and contextual relationships of lesions is critical for performance. Mamba's strength in long-range modeling and computational efficiency provides strong support for this task. Recently, Mamba has been increasingly applied in medical scenarios. U-VM-UNet (52) integrates sparse gating and low-rank decomposition to design an efficient visual selective scan module, achieving strong segmentation results across datasets. Mamba-Sea (53) proposes a global-to-local sequence augmentation mechanism and builds a pure SSM-based framework, improving generalization in cross-domain segmentation tasks. VM-UNetV2 (54) combines Vision Mamba with the UNet v2 (28) architecture and introduces a semantic and detail injection module, showing better performance than conventional models in skin and polyp segmentation. SMM-UNet (55) constructs selective and multi-scale fusion Mamba modules to enhance feature representation at different scales while keeping the network compact. CAMS (56) completely removes convolution and attention mechanisms, adopting a pure Mamba encoder and dual decoder structure to balance global modeling and fine-grained detail recovery in cardiac image segmentation. Therefore, we adopt a multi-branch mamba structure to establish long-distance dependency capturing global relationships and effectively aggregating contextual information, thus enhancing global semantic representation.

# 3 Methods

Medical image segmentation task aims to extract the lesion object semantic information, but ignores the pixel-level detail information. In contrast, medical image reconstruction task focuses on mining low-level content information. Therefore, we combine the medical image reconstruction task and the medical image segmentation task, which is jointly optimized to improve the segmentation performance. Our MFEL framework is shown in Figure 1, which includes a feature interaction-driven image reconstruction (FIIR) and a feature-embedded representation image segmentation (FRIS). The specific details are as follows.

## 3.1 Feature interaction-driven image reconstruction

FIIR employs self-supervised learning to obtain fine-grained feature representations, thereby enhancing the segmentation feature representations. It consists of three components: foreground image reconstruction (FIR), background image reconstruction (BIR), and medical image reconstruction (MIR). Specifically, FIR generates foreground feature, BIR provides background feature, and MIR obtains fine pixel-level feature. Foreground feature contains the key object information (e.g., edges, textures, structures), and background feature includes the irrelevant environment information. In this way, the two features can enhance the pixel-level fine-grained feature representation during medical image reconstruction.

### 3.1.1 Foreground and background feature extraction

The medical image $I_s$ first can be divided into foreground image $I_f$ and background image $I_b$ by the segmentation mask $I_m$. $I_m$ labels the foreground information as 1 and the background information as 0. Therefore, $I_f$ and $I_b$ can be formulated as follows:

$$I_f = I_s \otimes I_m, I_b = I_s \otimes (1 - I_m)\,. \tag{1}$$

Then, $I_f$ and $I_b$ are respectively fed into the foreground encoder $E_f^i$ and the background encoder $E_b^i$ to extract the foreground feature $F_f^i$ and the background feature $F_b^i$, where $i = 1, 2, 3$ denotes the layer index. Finally, $F_f^i$ and $F_b^i$ are input to the foreground decoder $D_f$ and the background decoder $D_b$ to reconstruct the foreground image $I_f'$ and the background image $I_b'$. The self-supervised foreground reconstruction loss $L_f$ and self-supervised background reconstruction loss $L_b$ focus on extracting foreground and background information of the medical segmentation image, which can be formulated as:

$$L_f = \left\| I_f' - I_f \right\|_1, L_b = \left\| I_b' - I_b \right\|_1, \tag{2}$$

where $\|\cdot\|_1$ represents the $l_1$ norm.

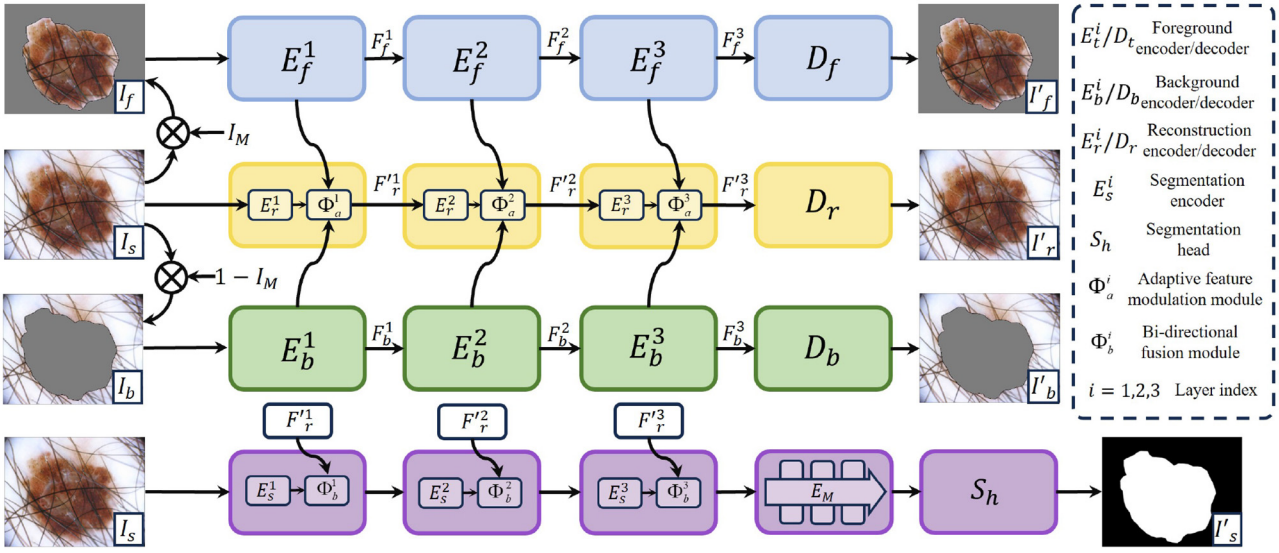


FIGURE 1
The overall framework illustration of the proposed MFEL. MFEL consists of FIIR and FRIS. FIIR aims to extract pixel-level features through self-supervised learning, thus helping the segmentation task to obtain finer-grained information. FRIS fuses semantic segmentation features and fine-grained reconstruction features to generate a more comprehensive feature representation.

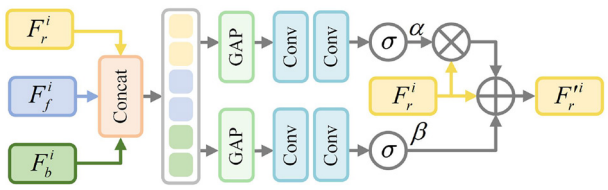


FIGURE 2
Architecture of adaptive feature modulation module.

### 3.1.2 Pixel-level fine-grained feature generation

As shown in Figure 2, $F_f^i$ and $F_b^i$ from each layer are fed into the adaptive feature modulation module $\Phi_a^i$, thereby helping SIR to obtain more significant foreground and background features. Specifically, $F_f^i$, $F_b^i$ and the initial pixel-level feature $F_r^i$ first perform channel feature concatenation to generate the fusion feature $F_u$, and then the global semantic features are extracted by using global average pooling. Further, we utilize dual-stream convolutional blocks to generate calibration parameters $\alpha$ and $\beta$ to guide $F_r^i$. This calibration process can be represented as:

$$\hat{F}_r^i = (1+\alpha) \times F_r^i + \beta. \tag{3}$$

Next, the calibrated pixel-level fine-grained feature $\hat{F}_r^i$ is fed into the reconstruction decoder to reconstruct the medical image. Finally, the medical image reconstruction loss $\mathcal{L}_s$ ensures that the pixel-level fine-grained features can reconstruct a complete segmentation image, which can be expressed as follows:

$$\mathcal{L}_s = \left\| I_r' - I_s \right\|_1, \tag{4}$$

where $I_s$ denotes a medical image, and $I_r'$ represents a reconstructed medical image.

## 3.2 Feature representation reinforcement learning

### 3.2.1 Bi-directional fusion module via hierarchical guidance

In Section 3.1, we obtain pixel-level fine-grained feature $\hat{F}_r^i$ from FIIR. Specifically, as shown in Figure 3, $I_s$ is first fed into the segmentation encoder $E_s^i$ to extract the segmentation semantic feature $F_s^i$. Then, $F_s^i$ and $\hat{F}_r^i$ are input to the bi-directional fusion module $\Phi_b^i$ to obtain a complete feature representation with strong semantics and rich details. Specifically, we compute respectively the cross-attention weights between $F_s^i$ and $\hat{F}_r^i$, thereby jointly modeling the complementary relationship between the reconstruction branch and the semantic branch. In this process, $\hat{F}_r^i$ uses semantic clues to guide $F_s^i$ to enhance structural perception, while $F_s^i$ employs textural details to enhance the spatial resolution of $\hat{F}_r^i$. In particular, First, $F_s^i$ generates the query vector $Q_s$ and $\hat{F}_r^i$ generates the key-value pair $(K_r, V_r)$. Similarly, $Q_r$ is obtained via $\hat{F}_r^i$, and $(K_s, V_s)$ is generated through $F_s^i$. The two-stream cross-modal attention is computed as follows:

$$\text{Attn}_r = \text{Softmax}\left(\frac{Q_s K_r^\top}{\sqrt{d}}\right), \text{Attn}_s = \text{Softmax}\left(\frac{Q_r K_s^\top}{\sqrt{d}}\right), \tag{5}$$

where $\sqrt{d}$ is the channel dimension of each attention head and *Softmax* is an activation function. $\text{Attn}_r$ denotes the segmentation-guided attention map, and $\text{Attn}_s$ represents the reconstruction-guided attention map. Subsequently, we adopt feature aggregation and residual concatenation to generate two enhanced features $\tilde{F}_r^i$ and $\tilde{F}_s^i$, which can be formulated as:

$$\tilde{F}_r^i = C_1^1\left(\text{Attn}_t \times V_r\right) + \hat{F}_r^i, \tilde{F}_s^i = C_1^1\left(\text{Attn}_r \times V_s\right) + \hat{F}_s^i, \tag{6}$$

where $C_1^1$ denotes one convolutional layer with $1 \times 1$ kernel. Finally, we fuse $\tilde{F}_r^i$ and $\tilde{F}_s^i$ to generate the refined segmentation

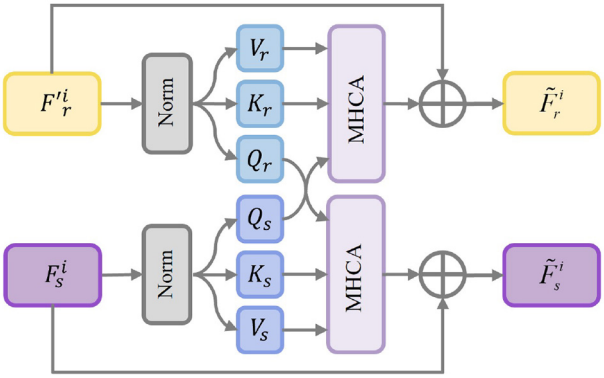


FIGURE 3
Architecture of bi-directional fusion module.

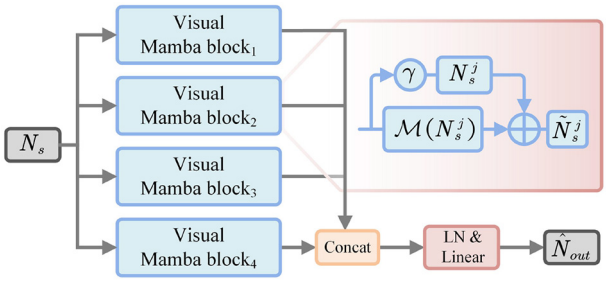


FIGURE 4
Architecture of multi-branch vision mamba.

feature $\bar{F}^i_s$ through concatenation and convolution operations, thus enhancing the feature representation ability.

### 3.2.2 High-level semantic feature mining via multi-branch vision mamba

Multi-branch vision mamba is constructed to force the model to mine high-level semantic information, thus improving the feature representation. Specifically, as shown in Figure 4, $\bar{F}^3_s$ firstly is fed into $E_M$ to perform flattening and normalization, thereby generating segmentation sequence feature $N_s$. Then, we divide $N_s$ into four groups to learn the important representations of different sub-regions, which can be represented as:

$$[N^1_s, N^2_s, N^3_s, N^4_s] = Split(N_s). \tag{7}$$

Next, each subsequence is respectively fed into the weight-sharing Mamba module to perform state modeling, and then refine the representation by residual operations:

$$\widetilde{N}^j_s = \mathcal{M}\left(N^j_s\right) + \gamma \cdot N^j_s, \quad j \in \{1, 2, 3, 4\} \tag{8}$$

where $\gamma$ is a scaling factor. $\mathcal{M}(\cdot)$ denotes the Mamba function. Then, the updated subsequence $N_j$ is performed to feature concatenation from the channel dimension, thus generating the enhanced sequence representation:

$$\widetilde{N}_s = Concat(\widetilde{N}^1_s, \widetilde{N}^2_s, \widetilde{N}^3_s, \widetilde{N}^4_s). \tag{9}$$

where $Concat(\cdot)$ denotes the feature concatenation operation. Subsequently, $\widetilde{N}_s$ is normalized and linearly transformed to project to the original feature dimension, which can be expressed as:

$$\hat{N}_{out} = Proj(LN(\widetilde{N}_s)), \tag{10}$$

where $LN(\cdot)$ represents layer normalization, and $Proj(\cdot)$ indicates linear projection.

Therefore, we utilized the state-space mechanism of Mamba to capture long-distance contextual information. Then, multi-branch decomposition is used to enhance the feature representation between different sub-regions. In this way, multi-branch vision mamba establishes the dependency between global semantics and local details, thus helping the model to improve the segmentation accuracy of key objects.

## 3.3 Model training

### 3.3.1 Image reconstruction head

To constrain the difference at the pixel level between the reconstructed image and the segmentation image, the image reconstruction head $D_f$, $D_r$ and $D_b$ adopt the reconstruction loss $\mathcal{L}_{rec}$, which can be defined as follows:

$$\mathcal{L}_{rec} = \mathcal{L}_s + \mathcal{L}_f + \mathcal{L}_b, \tag{11}$$

where $\mathcal{L}_s$, $\mathcal{L}_f$ and $\mathcal{L}_b$ denote foreground reconstruction loss, background reconstruction loss and medical image reconstruction loss, respectively.

### 3.3.2 Semantic segmentation head

The BCE loss $\mathcal{L}_{bce}$ aims to predict the per-pixel classification accuracy. The Dice loss $\mathcal{L}_{dice}$ can measure the overall overlap region between the prediction mask $I'_s$ and the ground truth $I_{gt}$. Thus, we jointly $\mathcal{L}_{bce}$ and $\mathcal{L}_{dice}$ constrain the segmentation head $S_h$, which can be expressed as:

$$\mathcal{L}_{mask} = \mathcal{L}_{bce}(I'_s, I_{gt}) + \mathcal{L}_{dice}(I'_s, I_{gt}). \tag{12}$$

Finally, the total training loss can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{mask}. \tag{13}$$

# 4 Experiments

In this section, we present a comprehensive overview of our experiments. We begin by introducing the datasets used in the study, followed by detailed descriptions of the experimental settings and implementation details. We then report the results of comparison experiments against state-of-the-art methods. In addition, we perform ablation studies to assess the impact of each key component. These experiments are designed to validate the effectiveness of the proposed method and to provide insights into the contribution of different modules to the overall performance.

## 4.1 Experimental settings

### 4.1.1 Datasets

**GLAS** (57) dataset consists of 165 microscopy images of colorectal adenocarcinoma tissue sections at stage T3 or T4, stained with H&E. Each image has a resolution of 128 × 128 pixels and is collected from a different patient. Due to variations in cancer progression, the lesions exhibit significant differences in shape and distribution. Meanwhile, since all samples originate from the same type of tissue, the surrounding environments are relatively consistent. Additionally, some cells are damaged or ruptured during the sampling process, resulting in large inter-cell variability. These factors make the dataset highly challenging. According to the official split, the training set contains 85 images and the test set contains 80 images. This dataset is mainly used to assess the model's capability in segmenting dense lesion regions and small targets.

**ISIC2016** (58) and **ISIC2017** (59) datasets were released by the International Skin Imaging Collaboration (ISIC) in 2016 and 2017, respectively. They were used as official datasets for the skin lesion analysis challenges held in those years. The goal of these datasets is to raise global awareness of skin disease diagnosis and to improve the detection of melanoma and other benign or malignant lesions. Both datasets contain a large number of samples and include various types of skin lesions. Due to the diversity of lesion types and the wide range of patient backgrounds, the samples show high variability in texture, color, and structure. In addition, some mild lesions look very similar to normal skin, which makes it hard to identify lesion boundaries. This increases the difficulty of the segmentation task. In this study, we evaluate the segmentation performance of our model using the ISIC2016 and ISIC2017 datasets. Both datasets follow the official training and testing splits: ISIC2016 includes 900 training images and 379 testing images, while ISIC2017 consists of 2,000 training images and 600 testing images. All images are resized to 256 × 256 pixels to ensure consistency during the experiments.

**PH2** (60) is a public dataset designed for dermoscopic image segmentation and classification. It aims to support research on computer-aided diagnosis of melanocytic lesions. The images were collected at the dermatology department of Pedro Hispano Hospital in Portugal. All images were captured under the same conditions using the Tuebinger mole analyzer system with 20× magnification. The dataset contains 200 dermoscopic images of melanocytic lesions, including 80 common nevi, 80 atypical nevi, and 40 melanomas. PH2 serves as a reliable benchmark for evaluating lesion detection, segmentation, and classification algorithms. In our experiments, all images were resized to 256 × 256 pixels. It is worth noting that we used PH2 as an external validation dataset. We tested it directly using the model trained on ISIC2016 to assess the effectiveness of our method and its potential for future clinical applications.

### 4.1.2 Metrics

In the quantitative analysis, we adopt widely used evaluation metrics in the field of medical image segmentation. Specifically, we use Precision, Recall, F1, and Intersection over Union (IoU) to assess the performance of the proposed model. Here, TP denotes true positives, FP denotes false positives, TN denotes true negatives, and FN denotes false negatives. These metrics jointly provide a comprehensive evaluation of the model's accuracy and completeness from multiple perspectives.

Precision measures the proportion of true positives among all regions predicted as positive (e.g., lesion areas). It reflects the model's ability to control FP. In medical image segmentation, a high Precision means the model can avoid wrongly identifying normal areas as lesions, which helps reduce the risk of misdiagnosis. When Precision is high, most of the predicted lesion regions are actually correct, and the FP rate is low. This is especially important in cases with small lesions or strong background noise. In such situations, Precision is a key metric to evaluate how well the model limits over-segmentation. The formula is given as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

Recall measures the model's ability to detect all positive targets. It shows how many of the actual positive pixels are correctly identified. In medical image segmentation, a high Recall means the model can successfully detect most lesion areas, which helps reduce missed detections and is important for clinical diagnosis support. The formula is given as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{15}$$

F1 is the harmonic mean of Precision and Recall. It is used to evaluate both the accuracy and completeness of the model. When there is a large gap between Precision and Recall, F1 provides a more balanced result. In segmentation tasks, F1 is especially useful for assessing model performance under class imbalance, such as small lesions against large background regions. A higher F1 indicates that the model achieves a good balance between accuracy and completeness. The formula is given as:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN} \tag{16}$$

IoU is one of the most widely used metrics in image segmentation. It measures the overlap between the predicted region and the GT. It is defined as the ratio of the intersection area to the union area of the prediction and the GT. IoU directly reflects how well the predicted boundary matches the actual boundary. A higher IoU means the predicted region aligns more closely with the GT, indicating better segmentation accuracy. Compared to F1, IoU is more sensitive to small differences and is suitable for evaluating boundary localization. The formula is given as:

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{17}$$

### 4.1.3 Implementations

We use NVIDIA GeForce RTX 4090 GPU to train and inference the model. The network framework is Pytorch. EPOCH is set to 150, and Batch is 4. The optimizer is Adam that uses momentum strategy to steadily update the model parameters. We employ warm-up and cosine annealing schedulers to achieve slow startup in the early stages and fine convergence in the later. The initial learning rate is 1e-3 and gradually decays to 1e-5.

## 4.2 Comparison with SOTA methods

To ensure a more comprehensive and reliable evaluation of model performance, we compare our method against state-of-the-art (SOTA) approaches from the past four years across different network architectures. These comparisons highlight the advantages of our model. Specifically, we select representative CNN-based methods including MsRED (25), MFSNet (11), DCSAU (27), and U-Net V2 (28); Transformer-based methods including BAT (30), FAT-Net (31), SSFormer (32), and CASF-Net (36); and a recent Mamba-based method, U-vm-unet (52). Extensive comparisons are conducted on four public datasets.

TABLE 1 Qualitative comparison results on the GLAS dataset.

| Method | F1(%) | IoU(%) | Precision(%) | Recall(%) |
|---|---|---|---|---|
| BAT (30) | 83.93 | 73.47 | 84.85 | 84.43 |
| FAT-Net (31) | 86.45 | 76.14 | **88.23** | 84.75 |
| MsRED (25) | 85.92 | 75.32 | 87.20 | 84.69 |
| MFSNet (11) | 86.33 | 75.95 | 81.70 | **89.20** |
| SSFormer (32) | 71.60 | 59.13 | 74.17 | 74.00 |
| CASF-Net (36) | 85.83 | 76.08 | 88.05 | 75.20 |
| DCSAU (27) | 88.28 | 79.03 | 87.67 | 88.32 |
| U-vm-unet (52) | 82.07 | 69.60 | 74.39 | 86.60 |
| U-Net V2 (28) | **88.90** | **80.86** | 87.31 | 89.16 |
| Ours | **89.73** | **82.07** | **91.01** | **89.18** |

Red indicates the best performance, and blue indicates the second best.

### 4.2.1 GLAS

#### 4.2.1.1 Qualitative comparisons

As in Table 1, we compare several representative methods from recent years on the GLAS dataset. The results show that Ours achieves the highest scores in F1, IoU, and Precision, and ranks second in Recall, slightly behind MFSNet. Ours reaches 82.07 in IoU, which shows a clear advantage over other methods. This indicates that the predicted lesion regions by Ours have better overlap with the GT and more accurate boundary localization. The Precision score reaches 91.01, suggesting that Ours effectively reduces false positives, which is important in scenarios where over-segmentation should be avoided. Considering that the GLAS dataset contains complex gland structures, a high proportion of small targets, and blurry boundaries, IoU and Precision are key metrics to evaluate real segmentation quality. Some methods achieve higher Recall but perform worse in IoU and Precision, which may be caused by over-segmentation. In contrast, Ours maintains high Recall while achieving high accuracy, showing strong boundary modeling ability and overall robustness.

#### 4.2.1.2 Quantitative comparisons

Figure 5 shows the visual comparison results on the GLAS dataset. In sample (a), the lesion cell has clear boundaries and appears hollow due to structural damage. SSFormer and U-vm-unet make obvious errors in this case, leading to inaccurate boundary prediction and incorrect segmentation of the cell structure. In samples (c) and (d), the lesion boundaries are blurry.

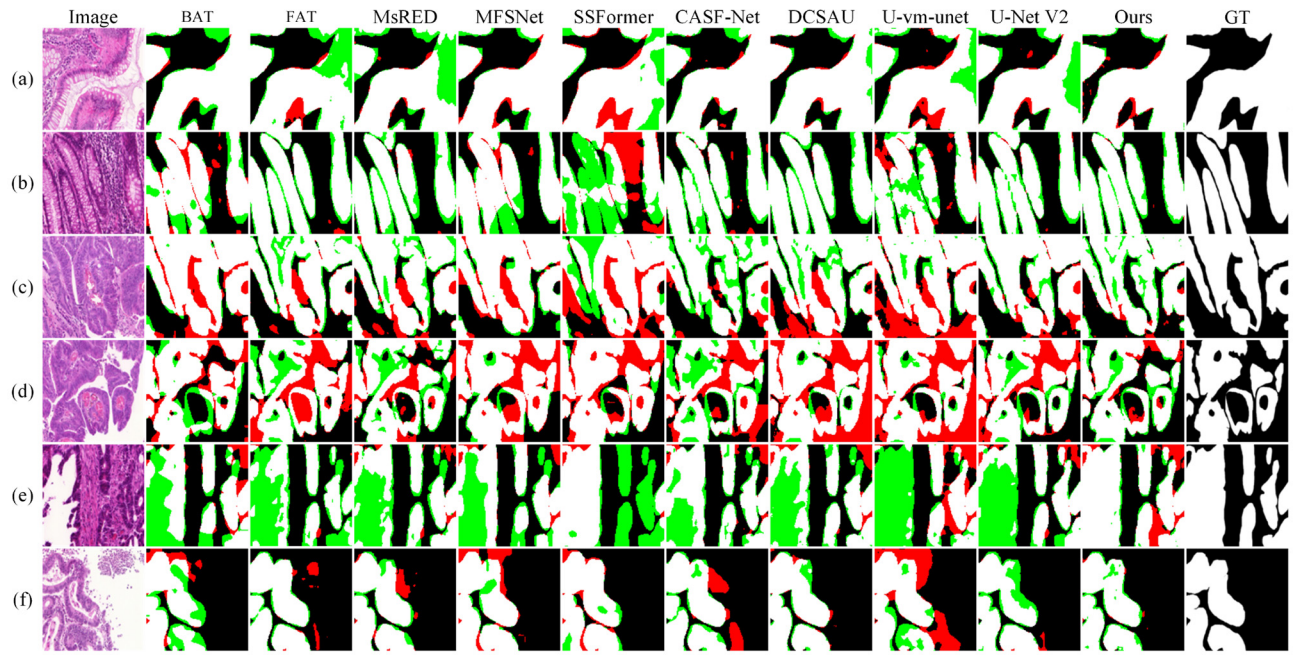


FIGURE 5
Quantitative comparison results on the GLAS dataset. Green regions indicate areas missed with respect to the GT, while red regions represent incorrectly predicted areas compared to the GT.

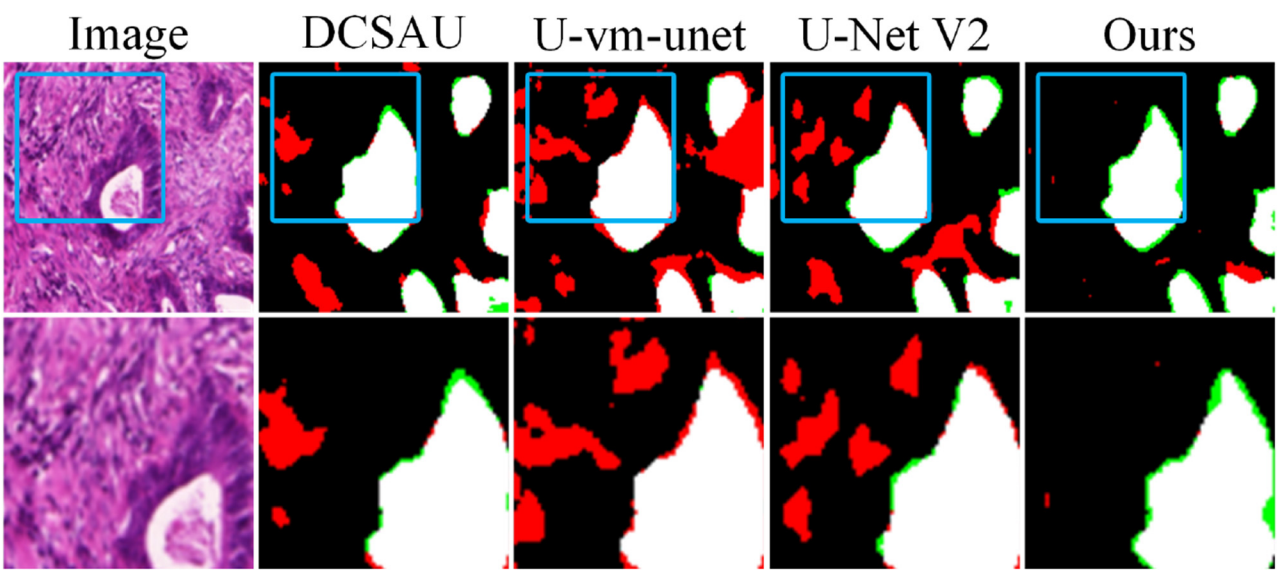


FIGURE 6
Detail comparison on the GLAS dataset.

Most methods fail to extract the target contours correctly and show severe missegmentation. In contrast, although Ours also shows some boundary errors, it preserves the overall shape of the target more completely.

In samples (e) and (f), the white regions represent the internal cell structure and the external background, respectively. These two samples come from different experimental conditions. For sample (e), methods like U-Net V2 miss part of the structure on the left side and mistakenly classify it as background. In sample (f), these methods show incomplete cell boundaries. In comparison, Ours gives results that are closer to the ground truth in both samples, showing better generalization. However, it is worth noting that Ours still makes a mistake in identifying the cell on the right side of sample (e), which suggests that there is still room to improve robustness across different environments.

To evaluate model performance on small targets and in noisy environments, we conducted local zoom-in comparisons on representative samples, as shown in Figure 6. DCSAU, U-vm-unet, and U-Net V2 often misidentify background textures as lesion regions, especially when boundaries are blurred or targets are irregular. This suggests limited robustness to noise and weak discrimination in challenging cases. In contrast, Ours better distinguishes true lesions from noisy backgrounds and successfully detects small, low-contrast targets. Despite minor boundary errors, it shows stronger resistance to noise and improved sensitivity to fine-grained lesion structures.

## 4.2.2 ISIC2016

### 4.2.2.1 Qualitative comparisons

On the ISIC2016 dataset, we compare our method with several representative approaches from recent years and evaluate segmentation performance from multiple aspects. As in Table 2, Ours achieves the best results across all four metrics: F1, IoU, Precision, and Recall, demonstrating strong overall performance. Specifically, the F1 reaches 94.14 and the IoU reaches 89.40, which shows a clear improvement over other methods. This indicates that our model provides a better balance between segmentation accuracy and region coverage, and can more precisely recover lesion shapes. The Precision reaches 95.48, showing stable control over false positives and helping reduce the misclassification of normal skin areas. The Recall reaches 93.45, ensuring high detection rates for lesion regions, which is important in clinical settings where missed detections must be minimized. The ISIC2016 dataset contains many benign and malignant skin lesions with blurry boundaries and similar textures, making segmentation more challenging. Compared to Ours, U-Net V2 achieves a similar Recall but lower Precision, which may cause over-segmentation. DCSAU shows good Precision, but its Recall is lower, which leads to missed

TABLE 2 Qualitative comparison results on the ISIC2016 dataset.

| Method | F1(%) | IoU(%) | Precision(%) | Recall(%) |
|---|---|---|---|---|
| BAT (30) | 91.22 | 84.99 | 93.36 | 91.32 |
| FAT-Net (31) | 91.58 | 85.42 | 92.36 | 92.79 |
| MsRED (25) | 91.61 | 85.51 | 93.37 | 91.90 |
| MFSNet (11) | 92.57 | 86.17 | 93.85 | 91.33 |
| SSFormer (32) | 91.37 | 85.63 | 90.18 | 93.22 |
| CASF-Net (36) | 91.46 | 85.50 | 92.26 | 88.22 |
| DCSAU (27) | 92.72 | 86.42 | 91.42 | **94.05** |
| U-vm-unet (52) | 92.79 | 86.54 | 93.92 | 91.68 |
| U-Net V2 (28) | **93.02** | **86.96** | **96.83** | 93.14 |
| Ours | **94.14** | **89.40** | **95.48** | **93.45** |

Red indicates the best performance, and blue indicates the second best.

lesion areas. In contrast, Ours maintains a better balance across all four metrics, indicating stronger segmentation ability under challenges such as background similarity, boundary ambiguity, and class imbalance commonly found in dermoscopic images.

#### 4.2.2.2 Quantitative comparisons

Figure 7 presents the visual comparison results on the ISIC2016 dataset. In clinical diagnosis, accurate boundary identification of lesions is important for evaluating the development stage and malignancy of the disease. However, in samples (a)–(d), many baseline methods show varying degrees of boundary errors, such as incomplete contours or blurred edges. In contrast, Ours performs more consistently in boundary modeling and produces results that are closer to the ground truth, which is of higher clinical value.

In sample (e), there is a dark skin area in the lower left region with texture similar to the lesion. FAT, U-vm-unet, and U-Net V2 all misclassify this area as a lesion, resulting in obvious false segmentation. MFSNet successfully captures the main region

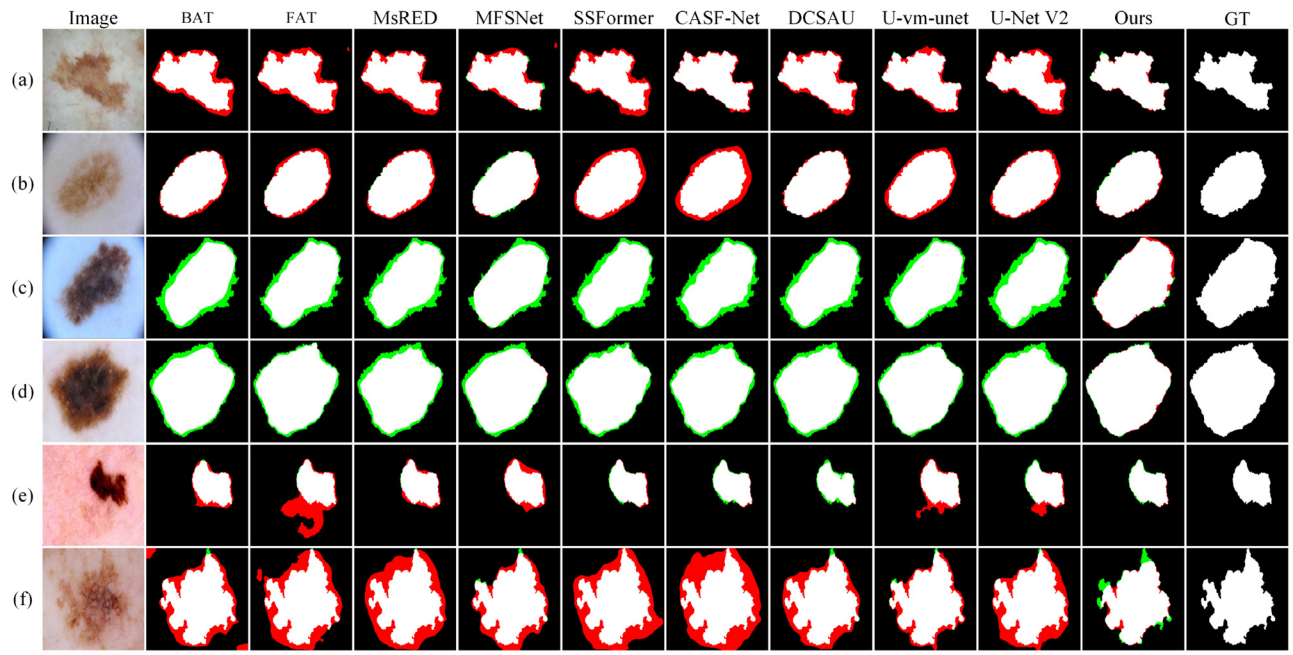


FIGURE 7
Quantitative comparison results on the ISIC2016 dataset. Green regions indicate areas missed with respect to the GT, while red regions represent incorrectly predicted areas compared to the GT.

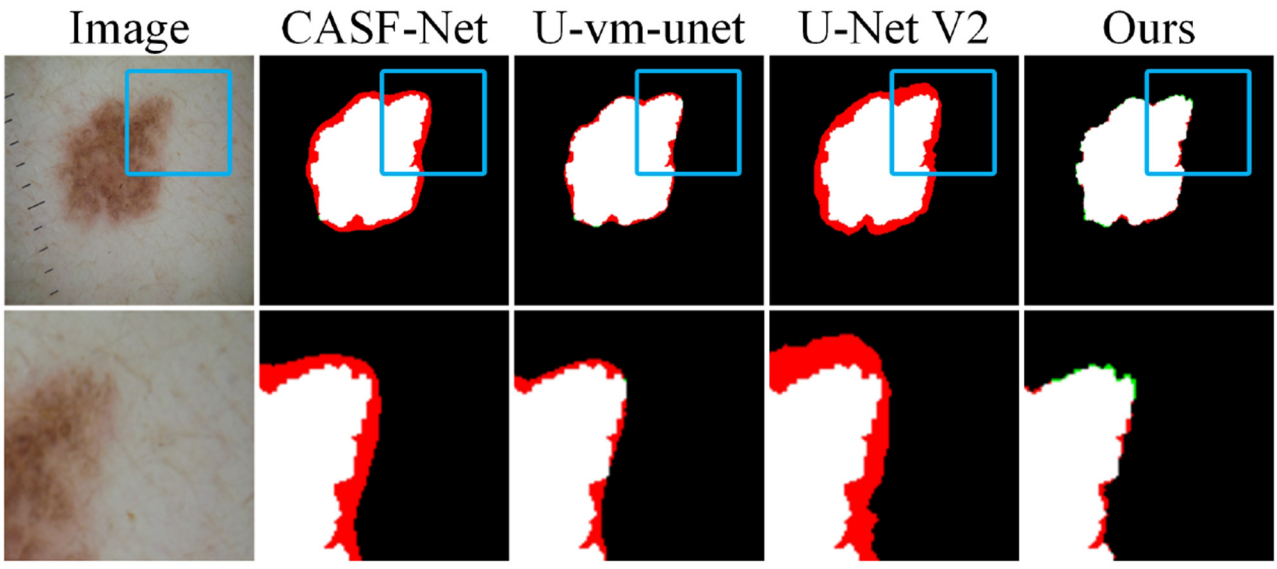


FIGURE 8
Detail comparison on the ISIC2016 dataset.

but misses parts near the boundary, which affects the overall contour quality.

Sample (f) contains a lesion with complex boundaries and fine internal structure. The lesion is located near the image edge, and the background interference is strong. These factors make boundary detection more difficult. Most baseline methods show shifted or broken contours in this case. Although Ours also makes some errors, its prediction is still the closest to the ground truth and better preserves both the overall shape and boundary continuity.

To further evaluate the model's ability to handle blurry boundaries, we selected a group of representative samples and performed local zoom-in comparisons, as shown in Figure 8. The results show that CASF-Net and U-Net V2 produce relatively coarse boundary predictions. Their outputs often show broken or expanded contours, which do not match the ground truth accurately. In contrast, Ours shows better alignment with the ground truth boundaries and performs more stably in preserving fine structural details. These results further demonstrate that our method has stronger fine-grained boundary perception and achieves higher localization accuracy for targets with unclear edges.

TABLE 3 Qualitative comparison results on the PH2 dataset.

| Method | F1(%) | IoU(%) | Precision(%) | Recall(%) |
|---|---|---|---|---|
| BAT (30) | 89.24 | 81.62 | **96.33** | 84.99 |
| FAT-Net (31) | 90.66 | 83.54 | 86.13 | **97.14** |
| MsRED (25) | 88.61 | 80.65 | 84.35 | 95.97 |
| MFSNet (11) | **91.42** | **84.19** | 89.12 | 93.84 |
| SSFormer (32) | 90.77 | 83.98 | 89.04 | 94.65 |
| CASF-Net (36) | 90.85 | 83.86 | 86.92 | **96.60** |
| DCSAU (27) | 87.33 | 77.51 | 90.27 | 84.58 |
| U-vm-unet (52) | 86.87 | 76.79 | 86.43 | 87.32 |
| U-Net V2 (28) | 90.70 | 82.98 | 92.88 | 95.28 |
| Ours | **94.44** | **89.70** | **93.73** | 95.37 |

Red indicates the best performance, and blue indicates the second best.

### 4.2.3 PH2

#### 4.2.3.1 Qualitative comparisons

Table 3 shows the test results on the PH2 dataset, which is used as an external validation set. The model is trained on the ISIC2016 dataset. As shown, Ours achieves the highest scores in the two key metrics, F1 and IoU, with values of 94.44 and 89.70 respectively. These results clearly outperform other methods and demonstrate strong overall segmentation ability and good generalization performance across datasets.

Although the Recall of Ours is not the highest among all methods, it remains at a high level. It is worth noting that the Recall of Ours is slightly lower than that of FAT-Net's 97.14 and CASF-Net's 96.60. This may be due to the fact that lesions in the PH2 dataset are more regular in shape and have relatively clearer boundaries. FAT-Net and CASF-Net tend to enlarge the predicted regions to increase the recall rate. However, this strategy often leads to lower precision and causes a drop in both IoU and F1. In contrast, Ours keeps a good balance. It maintains a reasonable recall

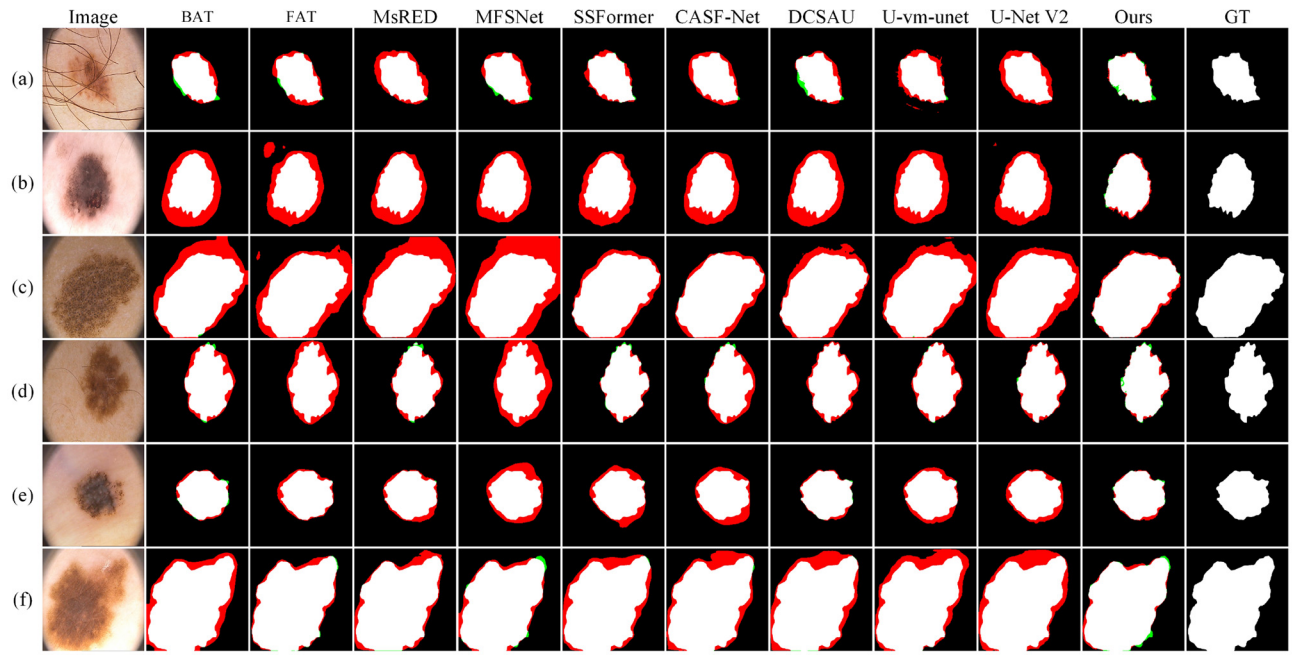


FIGURE 9
Quantitative comparison results on the PH2 dataset. Green regions indicate areas missed with respect to the GT, while red regions represent incorrectly predicted areas compared to the GT.

while avoiding over-segmentation, which helps improve boundary accuracy and overall model stability.

#### 4.2.3.2 Quantitative comparisons

Figure 9 shows the segmentation results of several samples from the PH2 dataset. Overall, most methods can outline the general shape of the lesion, but there are still clear differences in boundary details and the handling of interference regions. In samples (a) and (b), where the lesion boundaries are relatively clear, U-Net V2 and CASF-Net produce coarser edges. In contrast, Ours generates contours that better align with the ground truth, with smoother and more complete boundaries, especially in the transition areas around the lesion. In sample (f), the lesion is large and structurally complex. Methods such as BAT and U-Net V2 show varying degrees of over-segmentation, with a large number of false positive areas (in red). Although Ours also has some prediction errors, its boundaries are more compact and the over-segmentation is significantly reduced.

We also select a group of samples for local zoom-in comparison, as shown in Figure 10. In these samples, the lesion regions are located within a liquid environment, and bubbles above the lesions introduce interference. This causes CASF-Net, U-vm-unet, and U-Net V2 to produce severe misclassifications. Although Ours also shows some boundary inaccuracies due to the blurred edges, its prediction remains the closest to the ground truth.

### 4.2.4 ISIC2017

#### 4.2.4.1 Qualitative comparisons

Table 4 shows the evaluation results on the ISIC2017 dataset. Ours ranks first in three key metrics: F1, IoU, and Recall, with scores of 88.10, 80.06, and 94.84, respectively. These results show that our model achieves strong overall segmentation quality and high lesion detection sensitivity. In particular, the Recall score is significantly higher than other methods, indicating that our model is more sensitive to lesion regions and can reduce missed detections. This is useful for clinical applications that require high recall. Compared to methods such as U-Net V2 and MFSNet, Ours maintains a high Recall while achieving a better balance in IoU and F1, showing better boundary modeling ability and practical value.

However, in terms of Precision, Ours performs relatively lower, with a score of 83.13, which is clearly below methods like U-Net V2's 96.26 and MFSNet's 91.91. The ISIC2017 dataset contains more complex lesions with blurry boundaries and irregular shapes. While trying to capture lesion regions more completely, the model may also include neutral areas near the lesion boundary or non-lesion areas with similar appearance. This increases the false positive rate and leads to a lower Precision score.

TABLE 4 Qualitative comparison results on the ISIC2017 dataset.

| Method | F1(%) | IoU(%) | Precision(%) | Recall(%) |
|---|---|---|---|---|
| BAT (30) | 84.85 | **76.23** | 86.64 | **88.75** |
| FAT-Net (31) | 84.79 | 76.06 | 89.08 | 85.93 |
| MsRED (25) | 84.43 | 75.79 | 91.21 | 83.61 |
| MFSNet (11) | 85.42 | 74.55 | **91.91** | 79.79 |
| SSFormer (32) | 83.43 | 71.30 | 81.51 | 85.54 |
| CASF-Net (36) | 84.20 | 72.71 | 85.14 | 84.51 |
| DCSAU (27) | **85.92** | 75.32 | 83.93 | 88.01 |
| U-vm-unet (52) | 85.26 | 74.93 | 89.51 | 81.39 |
| U-Net V2 (28) | 85.00 | 73.90 | **96.26** | 82.86 |
| Ours | **88.10** | **80.06** | 83.13 | **94.84** |

Red indicates the best performance, and blue indicates the second best.

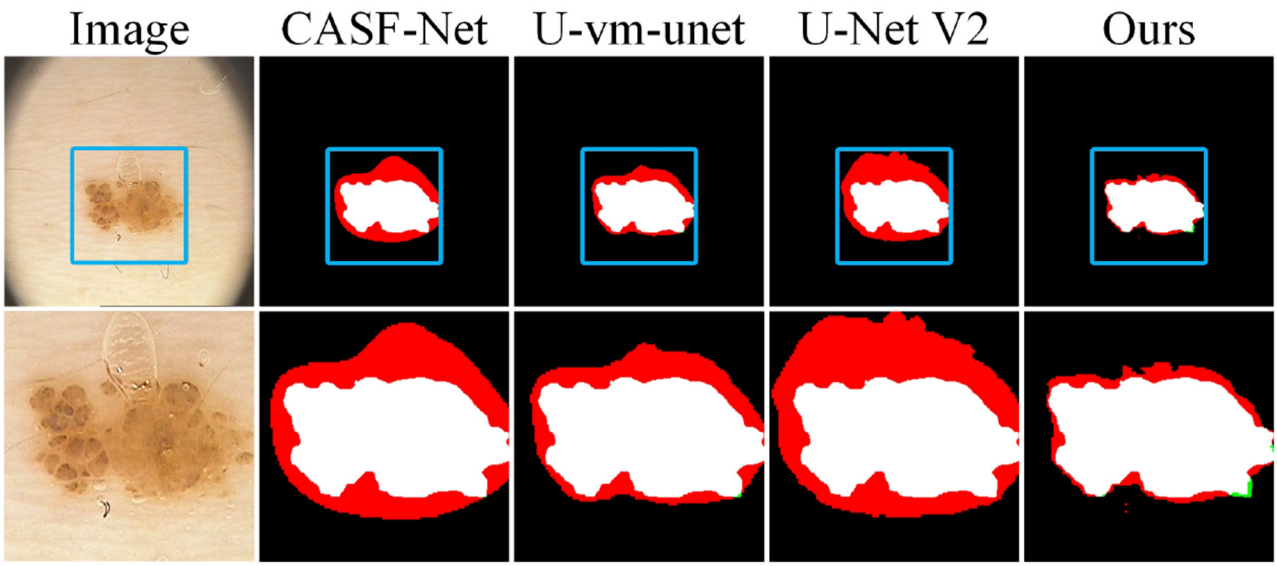


FIGURE 10
Detail comparison on the PH2 dataset.

#### 4.2.4.2 Quantitative comparisons

As shown in Figure 11, the samples in the ISIC2017 dataset often have more blurred boundary information. This leads to boundary prediction errors across all compared methods. In sample (b), the lesion boundaries are highly similar to the surrounding skin texture, causing all models to misidentify the boundary. In sample (d), although the lesion boundary is relatively clear, the surrounding skin is more complex. As a result, U-vm-unet mistakenly includes the ruler at the bottom as part of the lesion. In sample (e), the lesion gradually darkens from left to right. Most methods can accurately detect the boundary on the right where the contrast is high, but fail to identify the blurry boundary on the left. In contrast, Ours achieves a result that is closest to the ground truth.

In Figure 12, we present a local zoom-in comparison. The blurred and small-sized lesion increases the difficulty of segmentation. Compared with DCSAU and two other methods,

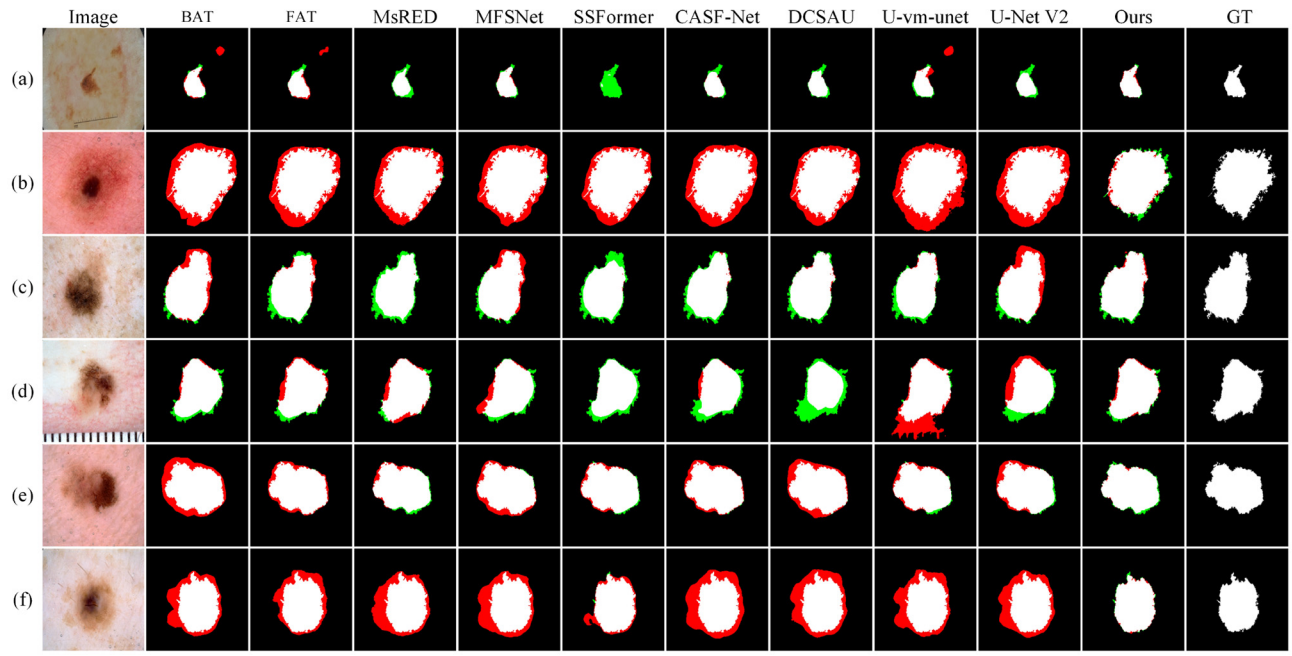


FIGURE 11
Quantitative comparison results on the ISIC2017 dataset. Green regions indicate areas missed with respect to the GT, while red regions represent incorrectly predicted areas compared to the GT.

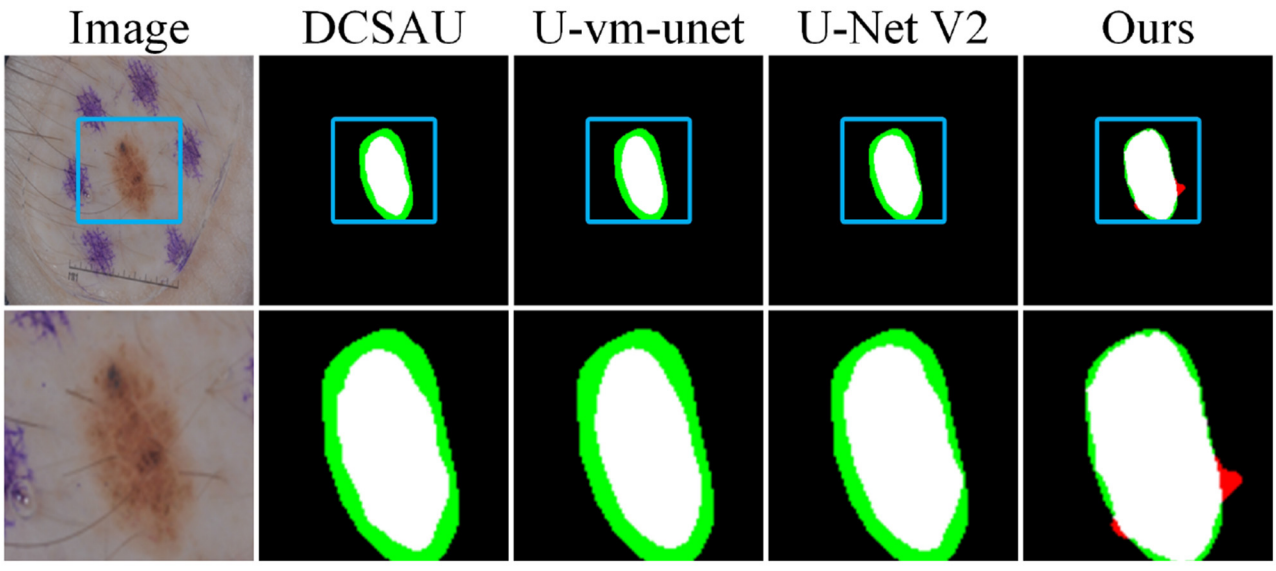


FIGURE 12
Detail comparison on the ISIC2017 dataset.

Ours shows better performance in boundary prediction. However, Ours is still affected by the surrounding environment and mistakenly identifies hair in the lower-left area as part of the boundary. This indicates that there is still room for improvement in handling fine-grained features.

## 4.3 Ablation studies

### 4.3.1 GLAS

To evaluate the contribution of each module in the model, we conducted a systematic ablation study on the GLAS dataset. The quantitative results obtained after removing different components are presented in Table 5, and the corresponding visual segmentation outputs are shown in Figure 13, offering a clear view of how each module affects the final performance.

As can be seen from the visual results, w/o FIIR (1) leads to evident deficiencies along the object boundaries and causes incomplete structural predictions. This demonstrates that FIIR plays an important role in enhancing pixel-level detail and supporting the extraction of informative segmentation features. When the adaptive feature modulation module $\Phi_a^i$ is removed, and then replaced with feature summary (2) or concat (3), the model tends to produce over-segmentation. This is reflected in the increased number of false positives in the predicted maps. Although the recall remains relatively high, reaching 96.10 and 96.31 respectively, the precision drops significantly to 78.18 and 66.37, suggesting that the model becomes less capable of regulating foreground and background responses effectively.

Moreover, we adopt summary (4) and concat (5) operations to replace the bi-directional fusion module $\Phi_b^i$. The predicted structures remain mostly intact, but the boundaries are less precise, indicating that this module still contributes to enhancing local detail and structural consistency. Further, the removal of the multi-branch vision mamba module $E_M$ (6) results in a decrease in both IoU and F1, and the predicted boundaries become less distinct. This shows that $E_M$ plays a critical role in aggregating hierarchical features and is particularly helpful in capturing complex object shapes.

Among all the configurations, the complete model (7) achieves the best overall performance. It obtains an F1 of 89.73, an IoU of 82.07, a precision of 91.01, and a recall of 89.18. Its visual results are also the most aligned with the ground truth annotations. These observations confirm that the synergy between the proposed modules leads to significant improvements in both segmentation accuracy and visual quality.

TABLE 5 Ablation studies results on the GLAS dataset.

| Method | F1(%) | IoU(%) | Precision(%) | Recall(%) |
|---|---|---|---|---|
| (1) w/o FIIR | 79.68 | 67.90 | 81.94 | 80.65 |
| (2) w/o $\Phi_a^i$ summary | 84.58 | 75.42 | 78.18 | **96.10** |
| (3) w/o $\Phi_a^i$ concat | 76.27 | 64.35 | 66.37 | **96.31** |
| (4) w/o $\Phi_b^i$ summary | 88.64 | 80.31 | 89.77 | 87.79 |
| (5) w/o $\Phi_b^i$ concat | **89.60** | **81.91** | **89.87** | 88.72 |
| (6) w/o $E_M$ | 85.01 | 75.05 | 84.18 | 87.77 |
| (7) Ours | **89.73** | **82.07** | **91.01** | 89.18 |

Red indicates the best performance, and blue indicates the second best.

TABLE 6 Ablation studies results on the ISIC2016 dataset.

| Method | F1(%) | IoU(%) | Precision(%) | Recall(%) |
|---|---|---|---|---|
| (1) w/o FIIR | 90.81 | 84.33 | 91.76 | 91.87 |
| (2) w/o $\Phi_a^i$ summary | 91.64 | 85.45 | 86.82 | **97.09** |
| (3) w/o $\Phi_a^i$ concat | **93.75** | **88.86** | 89.19 | **98.48** |
| (4) w/o $\Phi_b^i$ summary | 91.51 | 85.14 | **95.48** | 88.16 |
| (5) w/o $\Phi_b^i$ concat | 88.77 | 81.19 | 89.25 | 91.17 |
| (6) w/o $E_M$ | 90.79 | 84.04 | 95.23 | 87.37 |
| (7) Ours | **94.14** | **89.40** | **95.48** | 93.45 |

Red indicates the best performance, and blue indicates the second best.

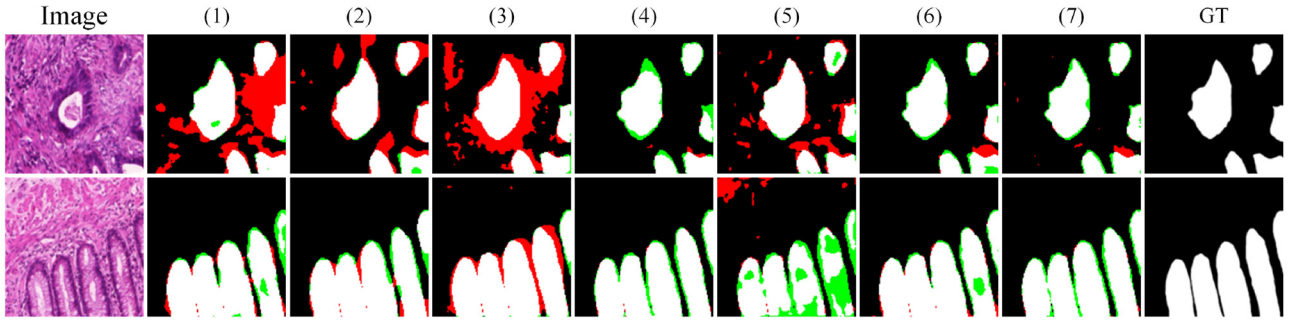


FIGURE 13
Quantitative comparative results of ablation experiments on the GLAS dataset.

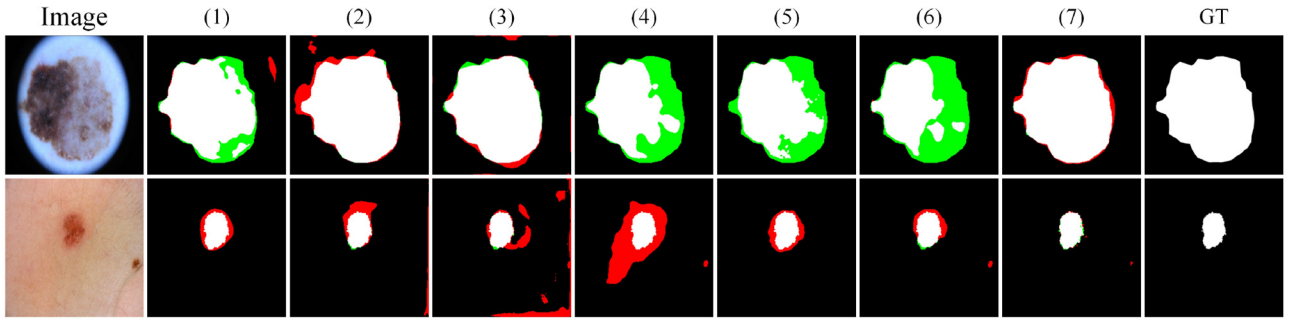


FIGURE 14
Quantitative comparative results of ablation experiments on the ISIC2016 dataset.

### 4.3.2 ISIC2016

We further validate the effect of each model component by conducting ablation experiments on the ISIC2016 dataset. Table 6 reports the numerical performance under different ablation settings, while Figure 14 illustrates the corresponding segmentation outputs for visual comparison. w/o FIIR (1) leads to a noticeable decline in IoU and F1, which drops to 84.33 and 90.81, respectively. Despite the recall and precision being reasonably balanced, the visual outputs exhibit weaker boundary fidelity, particularly in areas with low contrast, where the predicted masks tend to deviate from the lesion margins. Interestingly, we adopt w/o $\Phi_a^i$ concat (3) to produce the highest recall at 98.48, suggesting that the model becomes more permissive in capturing lesion pixels. However, this also comes at the cost of increased false positives, as reflected in the relatively lower precision and the presence of redundant red areas in the predicted masks. w/o $\Phi_a^i$ summary (2) causes the prediction accuracy to decrease, reinforcing that the absence of the modulation structure compromises the foreground-background balancing mechanism.

To verify the effectiveness of the bi-directional fusion module $\Phi_b^i$, we use $\Phi_b^i$ summary (4) and $\Phi_b^i$ concat (4) instead of $\Phi_b^i$. Specifically, w/o $\Phi_b^i$ concat (5) has a clearer negative effect, with IoU decreasing to 81.19, accompanied by more pronounced boundary irregularities in the visualization. w/o $E_M$ on IoU and F1 metrics scores lower than the full model. This suggests that although the primary structure still functions, the lack of high-low feature interaction leads to reduced segmentation confidence near ambiguous regions. With all components intact, the full model (7) achieves the strongest performance across all metrics F1 reaches 94.14, IoU improves to 89.40, and both precision and recall are maximized. The output masks are tightly aligned with the lesion contours, even under challenging conditions such as blurry or low-contrast boundaries, confirming the complementary nature of all proposed modules.

## 5 Conclusion

In this paper, we propose a multi-interactive feature embedding learning method for medical image segmentation. The core idea is to establish information interaction between the reconstruction task and the segmentation task, thus achieving superior segmentation performance. Specifically, an adaptive feature modulation module can efficiently fuse foreground and background features, thereby extracting pixel-level fine-grained features. Then, a bi-directional fusion module integrates important feature information between two different tasks, enhancing semantic understanding and detail retention. Finally, a multi-branch visual mamba effectively captures structural details by extracting multi-scale features in parallel, thus improving the model capability in terms of local texture and global semantics. Extensive experiments demonstrate that the proposed method can accurately segment the lesion region compared to other state-of-the-art segmentation methods.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Ethics statement

Ethical approval was not required for the studies on humans in accordance with the local legislation and institutional requirements because only commercially available established cell lines were used. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

## Author contributions

YH: Conceptualization, Data curation, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. YL: Funding acquisition, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was

funded by National Natural Science Foundation of China (Youth Fund, 81904324).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

1. Al-Masni MA, Al-Shamiri AK, Hussain D, Gu YH. A unified multi-task learning model with joint reverse optimization for simultaneous skin lesion segmentation and diagnosis. *Bioengineering*. (2024) 11:1173. doi: 10.3390/bioengineering11111173

2. Al-Absi AA, Fu R, Ebrahim N, Al-Absi MA, Kang DK. Brain tumour segmentation and grading using local and global context-aggregated attention network architecture. *Bioengineering*. (2025) 12:552. doi: 10.3390/bioengineering12050552

3. Dong Z, Yuan G, Hua Z, Li J. Diffusion model-based text-guided enhancement network for medical image segmentation. *Expert Syst Appl*. (2024) 249:123549. doi: 10.1016/j.eswa.2024.123549

4. Ding W, Li Z. Curriculum consistency learning and multi-scale contrastive constraint in semi-supervised medical image segmentation. *Bioengineering*. (2023) 11:10. doi: 10.3390/bioengineering11010010

5. Fu Z, Li J, Hua Z, Fan L. Deep supervision feature refinement attention network for medical image segmentation. *Eng Appl Artif Intell*. (2023) 125:106666. doi: 10.1016/j.engappai.2023.106666

6. Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider A, Landman BA, et al. The medical segmentation decathlon. *Nat Commun*. (2022) 13:4128. doi: 10.1038/s41467-022-30695-9

7. Han Z, Jian M, Wang GG. ConvUNeXt: an efficient convolution neural network for medical image segmentation. *Knowl Based Syst*. (2022) 253:109512. doi: 10.1016/j.knosys.2022.109512

8. Müller D, Kramer F. MIScnn: a framework for medical image segmentation with convolutional neural networks and deep learning. *BMC Med Imaging*. (2021) 21:1–11. doi: 10.1186/s12880-020-00543-7

9. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. Springer: New York (2015). p. 234–241. doi: 10.1007/978-3-319-24574-4_28

10. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: a nested u-net architecture for medical image segmentation. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4*. Springer: New York (2018). p. 3–11. doi: 10.1007/978-3-030-00889-5_1

11. Basak H, Kundu R, Sarkar R. MFSNet: a multi focus segmentation network for skin lesion segmentation. *Pattern Recognit*. (2022) 128:108673. doi: 10.1016/j.patcog.2022.108673

12. Ma J, Yuan G, Guo C, Gang X, Zheng M. SW-UNet: a U-net fusing sliding window transformer block with CNN for segmentation of lung nodules. *Front Med*. (2023) 10:1273441. doi: 10.3389/fmed.2023.1273441

13. Amin J, Azhar M, Arshad H, Zafar A, Kim SH. Skin-lesion segmentation using boundary-aware segmentation network and classification based on a mixture of convolutional and transformer neural networks. *Front Med*. (2025) 12:1524146. doi: 10.3389/fmed.2025.1524146

14. Liu X, Tan H, Wang W, Chen Z. Deep learning based retinal vessel segmentation and hypertensive retinopathy quantification using heterogeneous features cross-attention neural network. *Front Med*. (2024) 11:1377479. doi: 10.3389/fmed.2024.1377479

15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst*. (2017) 30.

16. Meng W, Liu S, Wang H. AFC-Unet: attention-fused full-scale CNN-transformer unet for medical image segmentation. *Biomed Signal Process Control*. (2025) 99:106839. doi: 10.1016/j.bspc.2024.106839

17. Ren S, Li X. HResFormer: hybrid residual transformer for volumetric medical image segmentation. *IEEE Trans Neural Netw Learn Syst*. (2025) 36:10558–66. doi: 10.1109/TNNLS.2024.3519634

18. Liu X, Gao P, Yu T, Wang F, Yuan RY. CSWin-UNet: transformer UNet with cross-shaped windows for medical image segmentation. *Inf Fusion*. (2025) 113:102634. doi: 10.1016/j.inffus.2024.102634

19. Chen J, Mei J, Li X, Lu Y, Yu Q, Wei Q, et al. TransUNet: rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Med Image Anal*. (2024) 97:103280. doi: 10.1016/j.media.2024.103280

20. Zhang Y, Liu H, Hu Q. Transfuse: fusing transformers and CNNS for medical image segmentation. In: *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part I 24*. Springer: New York (2021). p. 14–24. doi: 10.1007/978-3-030-87193-2_2

21. Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:231200752*. (2023). doi: 10.48550/arXiv.2312.00752

22. Liu Y, Tian Y, Zhao Y, Yu H, Xie L, Wang Y, et al. VMamba: visual state space model. *Adv Neural Inf Process Syst*. (2024) 37:103031–63.

23. Oktay O, Schlemper J, Folgoc LL, McDonagh S, Kainz B, Glocker B, et al. Attention u-net: learning where to look for the pancreas. *arXiv preprint arXiv:180403999*. (2018). doi: 10.48550/arXiv.1804.03999

24. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. (2021) 18:203–11. doi: 10.1038/s41592-020-01008-z

25. Dai D, Dong C, Xu S, Yan Q, Li Z, Zhang C, et al. Ms RED: a novel multi-scale residual encoding and decoding network for skin lesion segmentation. *Med Image Anal*. (2022) 75:102293. doi: 10.1016/j.media.2021.102293

26. Wei S, Hu Z, Tan L. Res-ECA-UNet++: an automatic segmentation model for ovarian tumor ultrasound images based on residual networks and channel attention mechanism. *Front Med*. (2025) 12:1589356. doi: 10.3389/fmed.2025.1589356

27. Xu Q, Ma Z, He N, Duan W. DCSAU-Net: a deeper and more compact split-attention U-Net for medical image segmentation. *Comput Biol Med*. (2023) 154:106626. doi: 10.1016/j.compbiomed.2023.106626

28. Peng Y, Chen DZ, Sonka M. U-net v2: rethinking the skip connections of u-net for medical image segmentation. In: *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*. Houston, TX: IEEE (2025). p. 1–5. doi: 10.1109/ISBI60581.2025.10980742

29. Feng S, Wang H, Han C, Liu Z, Zhang H, Lan R, et al. Weakly supervised gland segmentation with class semantic consistency and purified labels filtration. *Proc AAAI Conf Artif Intell*. (2025) 39:2987–95. doi: 10.1609/aaai.v39i3.32306

30. Wang J, Wei L, Wang L, Zhou Q, Zhu L, Qin J. Boundary-aware transformers for skin lesion segmentation. In: *Medical Image Computing and Computer*

*Assisted Intervention-mICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part I 24*. Springer: New York (2021). p. 206–16. doi: 10.1007/978-3-030-87193-2_20

31. Wu H, Chen S, Chen G, Wang W, Lei B, Wen Z. FAT-Net: feature adaptive transformers for automated skin lesion segmentation. *Med Image Anal.* (2022) 76:102327. doi: 10.1016/j.media.2021.102327

32. Shi W, Xu J, Gao P. SSformer: a lightweight transformer for semantic segmentation. In: *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*. Shanghai: IEEE (2022). p. 1–5. doi: 10.1109/MMSP55362.2022.9949177

33. Sun J. MedFusion-TransNet: multi-modal fusion via transformer for enhanced medical image segmentation. *Front Med.* (2025) 12:1557449. doi: 10.3389/fmed.2025.1557449

34. Bakkouri I, Bakkouri S. UGS-M3F: unified gated swin transformer with multi-feature fully fusion for retinal blood vessel segmentation. *BMC Med Imaging.* (2025) 25:77. doi: 10.1186/s12880-025-01616-1

35. Zeng L, Zhu M, Wu K, Li Z. Medical image segmentation via sparse coding decoder. In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE: Hyderabad, India (2025). p. 1–5. doi: 10.1109/ICASSP49660.2025.10889260

36. Zheng J, Liu H, Feng Y, Xu J, Zhao L. CASF-Net: cross-attention and cross-scale fusion network for medical image segmentation. *Comput Methods Programs Biomed.* (2023) 229:107307. doi: 10.1016/j.cmpb.2022.107307

37. Iqbal S, Khan TM, Naqvi SS, Naveed A, Meijering E. TBConvL-Net: a hybrid deep learning architecture for robust medical image segmentation. *Pattern Recognit.* (2025) 158:111028. doi: 10.1016/j.patcog.2024.111028

38. Perera S, Erzurumlu Y, Gulati D, Yilmaz A. MobileUNETR: a lightweight end-to-end hybrid vision transformer for efficient medical image segmentation. In: *European Conference on Computer Vision*. Springer: New York (2025). p. 281–99. doi: 10.1007/978-3-031-91721-9_18

39. Beliy R, Gaziv G, Hoogi A, Strappini F, Golan T, Irani M. From voxels to pixels and back: self-supervision in natural-image reconstruction from FMRI. *Adv Neural Inf Process Syst*. (2019) 32.

40. Zhang Y, Hao J, Zhou B. Dual-domain multi-path self-supervised diffusion model for accelerated MRI reconstruction. *arXiv preprint arXiv:250318836*. (2025). doi: 10.48550/arXiv.2503.18836

41. Chen L, Bentley P, Mori K, Misawa K, Fujiwara M, Rueckert D. Self-supervised learning for medical image analysis using image context restoration. *Med Image Anal.* (2019) 58:101539. doi: 10.1016/j.media.2019.101539

42. Zhang Z, Xu R, Liu M, Yan Z, Zuo W. Self-supervised image restoration with blurry and noisy Pairs. *Adv Neural Inf Process Syst*. (2022) 35:29179–91.

43. Thakkar JD, Bhatt JS, Patra SK. Self-supervised learning for medical image restoration: investigation and finding. In: *International Conference on Machine Intelligence and Signal Processing*. Springer: New York (2022). p. 541–552. doi: 10.1007/978-981-99-0047-3_46

44. Li S, Meng W, Liu C, Long C, He S. S4 FD: self-supervision-enhanced semisupervised fault diagnosis for complex industrial processes. *IEEE Trans Ind Inform*. (2025) 21:3585–94. doi: 10.1109/TII.2024.3523590

45. Mammadov A, Folgoc LL, Adam J, Buronfosse A, Hayem G, Hocquet G, et al. Self-supervision enhances instance-based multiple instance learning methods in digital pathology: a benchmark study. *arXiv preprint arXiv:250501109*. (2025). doi: 10.1117/1.JMI.12.6.061404

46. Koohbanani NA, Unnikrishnan B, Khurram SA, Krishnaswamy P, Rajpoot N. Self-path: self-supervision for classification of pathology images with limited annotations. *IEEE Trans Med Imaging*. (2021) 40:2845–56. doi: 10.1109/TMI.2021.3056023

47. Zhou B, Dey N, Schlemper J, Salehi SSM, Liu C, Duncan JS, et al. DSFormer: a dual-domain self-supervised transformer for accelerated multi-contrast MRI reconstruction. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA: IEEE (2023). p. 4966–75. doi: 10.1109/WACV56688.2023.00494

48. Zhuang J, Wu L, Wang Q, Fei P, Vardhanabhuti V, Luo L, et al. MiM: mask in mask self-supervised pre-training for 3D medical image analysis. *IEEE Trans Med Imaging*. (2025). doi: 10.1109/TMI.2025.3564382

49. Liu Z, Zhang Y, Wang B, Yang Y, Cai L. SFma-Unet: a mamba-based spatial-frequency fusion network for medical image segmentation. In: *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Hyderabad: IEEE (2025). p. 1–5. doi: 10.1109/ICASSP49660.2025.10889117

50. Liu S, Lin Y, Liu D, Wang P, Zhou B, Si F. Frequency-enhanced lightweight vision mamba network for medical image segmentation. *IEEE Trans Instrum Meas.* (2025) 74:1–12. doi: 10.1109/TIM.2025.3527526

51. Sun J, Chen K, Wu X, Xu Z, Wang S, Zhang Y. MSM-UNet: a medical image segmentation method based on wavelet transform and multi-scale Mamba-UNet. *Expert Syst Appl.* (2025) 288:128241. doi: 10.1016/j.eswa.2025.128241

52. Wu R, Liu Y, Liang P, Chang Q. Ultralight vm-unet: parallel vision mamba significantly reduces parameters for skin lesion segmentation. *arXiv preprint arXiv:240320035*. (2024). doi: 10.1016/j.patter.2025.101298

53. Cheng Z, Guo J, Zhang J, Qi L, Zhou L, Shi Y, et al. Mamba-sea: a mamba-based framework with global-to-local sequence augmentation for generalizable medical image segmentation. *IEEE Trans Med Imaging*. (2025). doi: 10.1109/TMI.2025.3564765

54. Zhang M, Yu Y, Jin S, Gu L, Ling T, Tao X. VM-UNET-V2: rethinking vision mamba UNet for medical image segmentation. In: *International Symposium on Bioinformatics Research and Applications*. Springer: New York (2024). p. 335–46. doi: 10.1007/978-981-97-5128-0_27

55. Li G, Huang Q, Wang W, Liu L. Selective and multi-scale fusion mamba for medical image segmentation. *Expert Syst Appl.* (2025) 261:125518. doi: 10.1016/j.eswa.2024.125518

56. Khan A, Asad M, Benning M, Roney C, Slabaugh G. CAMS: convolution and attention-free mamba-based cardiac image segmentation. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Tucson, AZ: IEEE (2025). p. 1893–903. doi: 10.1109/WACV61041.2025.00191

57. Sirinukunwattana K, Pluim JPW, Chen H, Qi X, Heng PA, Guo YB, et al. Gland segmentation in colon histology images: the glas challenge contest. *Med Image Anal.* (2017) 35:489–502. doi: 10.1016/j.media.2016.08.008

58. Gutman D, Codella N, Celebi ME, Helba B, Marchetti M, Mishra N, et al. ISIC challenge 2016: skin lesion analysis towards melanoma detection. *arXiv preprint arXiv:160501397*. (2016). doi: 10.48550/arXiv.1605.01397

59. Berseth M. ISIC 2017-skin lesion analysis towards melanoma detection. *arXiv preprint arXiv:170300523*. (2017). doi: 10.48550/arXiv.1703.00523

60. Conoci S, Rundo F, Petralta S, Battiato S. Advanced skin lesion discrimination pipeline for early melanoma cancer diagnosis towards PoC devices. In: *2017 European Conference on Circuit Theory and Design (ECCTD)*. Catania: IEEE (2017). p. 1–4. doi: 10.1109/ECCTD.2017.8093310