University of Reading

Edge-guided cross-modal fusion network for multi-resolution breast cancer segmentation in smart digital pathology

Article

Accepted Version

Li, T., Song, S., Wang, Q., Fong, S., Song, W., Gao, J., Pan, Y., Dey, N. and Sherratt, R. S. ORCID: https://orcid.org/0000-0001-7899-4445 (2026) Edge-guided cross-modal fusion network for multi-resolution breast cancer segmentation in smart digital pathology. IEEE Transactions on Consumer Electronics, 72 (1). pp. 342-352. ISSN 1558-4127 doi: 10.1109/TCE.2025.3646038 Available at https://centaur.reading.ac.uk/127665/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1109/TCE.2025.3646038

Publisher: IEEE

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

University of Reading

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Edge-Guided Cross-Modal Fusion Network for Multi-Resolution Breast Cancer Segmentation in Smart Digital Pathology

Tengyue Li, Shuangli Song, Qiong Wang, Simon Fong, Wei Song, Juntao Gao, Yi Pan, *Member, IEEE*, Nilanjan Dey, *Senior Member, IEEE*, and Robert Simon Sherratt, *Fellow, IEEE*

***Abstract*—Accurate segmentation of carcinoma in situ and invasive carcinoma in Whole Slide Images (WSIs) is crucial for improving breast cancer diagnostics in smart healthcare systems. Existing methods that rely solely on Hematoxylin and Eosin (H&E) staining lack molecular boundary-specific markers and struggle with resolution limitations. To address these challenges, we propose a breast cancer segmentation framework that fuses multi-resolution semantic features from H&E images with edge information from Cytokeratin 5/6 (CK5/6) immunohistochemical staining. The model integrates three modules: a multi-resolution semantic segmentation branch, an edge detection module aligned with H&E images, and a multi-scale fusion module. By combining multi-modal information and selectively zooming in on key regions, the method enhances the diagnostic process of medical practitioners, making the system more accurate and suitable for deployment in an Internet of Medical Things (IoMT) platform. Evaluations on the Breast Cancer Semantic Segmentation (BCSS) and the Chinese People's Liberation Army (PLA) General Hospital datasets show segmentation similarity coefficients of 81.28% and 93.16%, respectively. This approach offers an effective solution for user-facing digital pathology systems and supports clinical decision-making in consumer-centric smart healthcare.**

This research was supported by grants from the Beijing Natural Science Foundation (No. L253018), the North China University of Technology research start-up fund (11005136024XN147-14, 11005136025XN076-038), the Youth Research Special Project of NCUT (Project No. 2025NCUTYRSP015), the Special project in support of the National Natural Science Foundation of China in 2024 (110051360024XN151-97), Guangzhou Development Zone Science and Technology Project (2023GH02), the Macau FDCT (0032/2022/A, 0019/2025/RIB1), and MYRG2022-00271-FST.*(Corresponding author: Tengyue Li.)*

Tengyue Li, Shuangli Song, Wei Song are with the Department of Artificial Intelligence and Computer Science, North China University of Technology, Beijing, 100144, China (e-mail: litengyue@ncut.edu.cn; ssl@mail.ncut.edu.cn; sw@ncut.edu.cn).

Qiong Wang is with the Department of Pathology, the First Medical Center, Chinese People's Liberation Army General Hospital, Beijing, 100039, China (e-mail: qwang301@163.com).

Simon Fong is with the Department of Computer and Information Science, University of Macau, Macau, 999078, China (e-mail: ccfong@umac.mo).

Juntao Gao is with the Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, 100084, China (e-mail: jtgao@tsinghua.edu.cn).

Yi Pan is with the Faculty of Computer Science and Control Engineering, Shenzhen University of Advanced Technology, Shenzhen, 518055, China (e-mail: panyi@suat-sz.edu.cn).

Nilanjan Dey is with the Department of Computer Science and Engineering, Techno International New Town, Chakpachuria, West Bengal 700156, India (e-mail: nilanjan.dey@tint.edu.in)

R. Simon Sherratt is with the Department of Biomedical Engineering, University of Reading, RG6 6AY Reading, U.K. (e-mail: r.s.sherratt@reading.ac.uk).

***Index Terms*—Breast pathology, histopathological image segmentation, multi-modal, Internet of Medical Things**

I. INTRODUCTION

THE distinction between carcinoma in situ and invasive carcinoma is an important task in pathological diagnosis within the breast cancer diagnosis and treatment system. With the rapid development of digital case technologies, Whole Slide Images (WSIs) have provided new technical support for breast cancer diagnosis [1]. To accurately differentiate between carcinoma in situ and invasive carcinoma, medical practitioners need to comprehensively analyze WSIs stained with various techniques, such as the commonly used H&E stain, CK5/6 stain, and Smooth Muscle Myosin Heavy Chain (SMMHC) stain. Fig. 1 shows the characteristics of slides stained with different methods. Each staining technique provides different pathological information, which complements one another and plays a crucial role in tumor segmentation [2]. However, analyzing WSIs is a tedious and time-consuming task that relies heavily on the clinical experience and expertise of pathologists. Therefore, there is an urgent need for an automated system to distinguish between invasive carcinoma and carcinoma in situ regions in WSIs with a single click.

Due to the increasing incidence of breast cancer worldwide, the demand for professional medical services is growing, particularly in regions with limited healthcare resources. This has driven the development of telemedicine technologies. The Internet of Medical Things (IoMT) combines traditional healthcare services with emerging artificial intelligence technologies, creating a new model for breast cancer diagnosis [3], [4]. Through this model, medical professionals can obtain accurate delineation results of invasive and carcinoma in situ in WSIs, eliminating the cumbersome diagnostic process of the past, saving both time and effort. At the same time, specialists around the world can collaborate to analyze data and provide remote treatment for patients [5]. However, many deep learning models, especially black-box models, suffer from a lack of transparency, making the interpretability and reliability of these models a challenge [6]. Therefore, there is an urgent need for a medical image segmentation model with high accuracy and interpretability in consumer IoMT-based breast cancer diagnostic systems.

Currently, most tumor segmentation tasks in WSI primarily rely on H&E-stained slides. However, the determination of whether cancer cells have infiltrated is mainly based on the absence of myoepithelial cells. Due to the limited color richness of H&E staining, myoepithelial cells are easily confused with other cells. Therefore, in actual clinical diagnosis, doctors typically need to observe both H&E-stained slides and various immunohistochemical (IHC) stained slides. In situ carcinoma typically preserves the integrity of myoepithelial cells, and IHC slides can specifically highlight the boundaries of in situ carcinoma, which are absent in invasive carcinoma. This boundary information not only assists in the precise localization of tumors but also significantly improves segmentation accuracy on H&E-stained images. Additionally, traditional single-resolution methods often struggle to capture the fine features of breast cancer cells and the global characteristics of tumors.

This paper proposes a multimodal WSI fusion segmentation method, which simulates the pathologist's analysis workflow of multiple stained slides and integrates H&E semantic information with CK5/6 edge information to improve segmentation accuracy. Specifically, the method consists of three core modules: the semantic branch module, the edge detection module, and the multi-scale fusion module. The semantic segmentation module uses a multi-resolution structure to achieve collaborative learning of both global and local features of the H&E image. Attention heatmaps are used to magnify high-resolution details of key regions, simulating the pathologist's focused observation of suspicious areas, thereby enhancing the model's interpretability. The edge detection module extracts edge features from CK5/6-stained images aligned with the H&E images, combining CK5/6 molecular expression information as a basal cell biomarker to enhance the recognition of tissue boundaries. Finally, the multi-scale fusion module guides the edge information from CK5/6 images to optimize the segmentation results of the H&E images. This method not only improves the accuracy of tumor boundary identification but also visualizes key regions through heatmaps, enhancing diagnostic transparency and reliability.

Fig. 2 illustrates the framework of our consumer IoMT-based breast cancer diagnostic system. Tissue samples are obtained from patients/consumers via biopsy, then stained and digitally scanned to generate pathological slide data, which are uploaded to the cloud through smart devices. The tumor segmentation model processes the WSIs and outputs the lesion segmentation results. Medical practitioners can remotely access the visual interface for diagnosis via devices such as tablets, PCs, or AR glasses. The attention heatmap generated by the model highlights the boundary regions of tissue, visually demonstrating the model's focus on lesion boundary segmentation, which validates the model's learning accuracy and enhances trust in the model from both doctors and

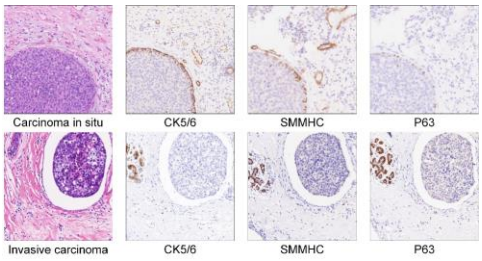

Fig. 1. CK5/6 marks the basal cell layer, defining carcinoma in situ boundaries; SMMHC highlights smooth muscle components at the invasive edge; P63 emphasizes myoepithelial cell distribution, aiding in tumor-normal tissue boundary differentiation.

Fig. 2. The framework for medical image segmentation in IoMT-based diagnostic systems.

consumers. Consumers can interact with medical practitioners via a mobile app or web interface, improving the remote healthcare experience. The system provides remote connectivity and support services for consumer electronic devices, enabling these devices to access the healthcare system via the internet, facilitating remote monitoring, diagnosis, and interaction with doctors, thereby having a profound impact on the consumer electronics industry.

The main contributions of this work are: (1) A novel multimodal segmentation framework based on the collaborative analysis of H&E and CK5/6 staining is developed, extracting biomarker-driven edge features from the registered CK5/6 images. (2) An interpretable multi-resolution H&E semantic segmentation module is proposed employing a multi-resolution structure design to enable collaborative learning of global and local features from H&E images.

II. Related Work

In early studies, WSI tissue segmentation was mainly achieved using more traditional classifiers [7]. However, these methods showed significant limitations when dealing with Complex tissue images make it difficult to accurately segment the target tissues. In recent years, with the rapid development of Convolutional Neural Networks (CNNs), such as UNet [8], SegNet [9], and DeepLabv3+ [10], researchers have gradually applied them to the field of WSI tissue image segmentation.

The U-Net architecture proposed by Ronneberger et al. [8], with its distinctive encoder-decoder design and skip connection mechanism, has achieved remarkable results in medical image segmentation tasks. Consequently, many researchers have adopted the U-Net as a baseline model for WSI tissue segmentation. For instance, Saltz et al. [11] utilized U-Net to map tumor-infiltrating lymphocytes in H&E images across 13 TCGA cancer types, demonstrating U-Net's strong capability in capturing multi-scale features and preserving spatial details in the images. However, despite the advantages of the U-Net structure, it still faces challenges in handling complex textures and fine details.

In recent years, various modifications to the U-Net architecture have been proposed. Zhao et al. [12] introduced SCAU-Net, which integrates both spatial and channel attention modules, enhancing the model's ability to capture gland boundaries. Building on this, Wen et al. [13] further fused traditional image processing techniques by combining Gabor filters with a cascaded squeeze-attention module, enabling the network to explicitly learn texture features at different scales and orientations, addressing the limitations of the original U-Net in texture information extraction. To further optimize target region localization accuracy, Lu et al. [14] proposed a two-stage framework, BreasTDLUSeg, based on a multi-scale attention mechanism, which achieves precise localization and segmentation of breast terminal duct lobular units.

While these modifications have enhanced the model's ability to capture texture features, most of the existing methods rely on single-resolution inputs, making it difficult to effectively balance global context with local details. As a result, researchers have started exploring multi-level information from WSI images. For example, Abdel-Nabi et al. [15] and Schmitz et al. [16] proposed Ms3LcU-Net and msY-Net, respectively, which use multi-branch path designs to fuse multi-scale features. To address the spatial alignment issue in multi-resolution fusion, Van Rijthoven et al. [17] introduced HookNet, which integrates multi-resolution features through a hook mechanism. Furthermore, to mitigate the problem of interference from irrelevant information in multi-resolution fusion, Dong et al. [18] employed a recursive zoom-in strategy. This method filters suspicious regions at an initial

resolution and then zooms in on these regions to acquire finer local details. Considering that different tissue types require optimal magnification at varying levels, Deng et al. [19] proposed Omni-Seg, which utilizes scale-aware and class-aware controllers to adaptively adjust feature extraction and segmentation strategies based on tissue type and magnification.

III. Methods

In this section, we introduce our proposed method and the definition of the loss function. The overall framework of our proposed model is illustrated in Fig. 3. It consists of three main branches: a semantic segmentation branch, an edge detection branch, and a multi-scale fusion branch. Through the IoMT platform, the automatic segmentation of invasive and carcinoma in situ in WSIs is achieved.

The H&E semantic branch enables collaborative learning of global and local features through multi-resolution. This branch consists of two encoder-decoder sub-branches: the target branch and the detail branch. (1) The target branch processes H&E-stained slides at 10x magnification. (2) The detail branch processes H&E-stained slides at 40x magnification. By zooming in on key regions in the 10x magnification slides it provides high-resolution, fine-grained information.

The CK5/6 edge detection branch is designed to learn edge features. This branch consists of two encoder-decoder sub-branches: the semantic segmentation branch and the edge detection branch. (1) The semantic segmentation branch segments CK5/6-stained slides, providing cross-modal semantic information for the aligned H&E images. (2) The edge detection branch extracts edge structure features from the CK5/6 slides, converted to grayscale, offering complementary edge guidance for H&E segmentation.

The multi-scale fusion branch integrates features from different modalities to guide H&E image segmentation with CK5/6 edge information. During the training phase, precise segmentation is achieved by adjusting the loss weight

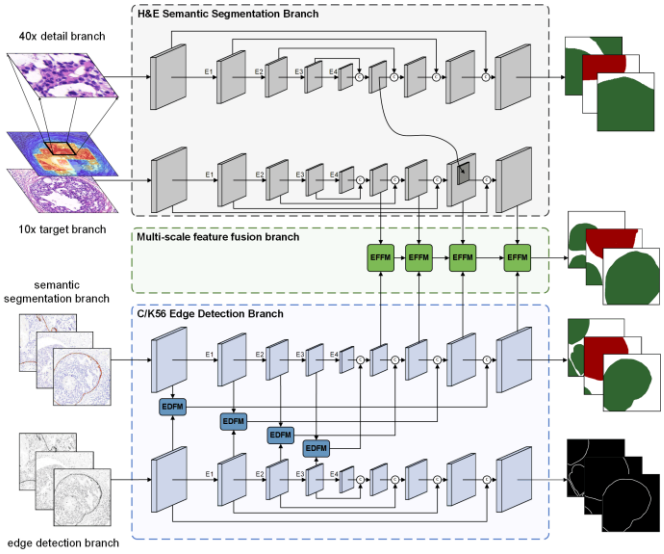

Fig. 3 The tumor semantic segmentation model consists of the H&E semantic segmentation module, the CK5/6 edge detection module, and the multi-modal fusion module. EDFM is used to fuse the CK5/6 dual-branch information in the CK5/6 edge detection module, while EFFM is used to fuse H&E semantic information and CK5/6 edge information, enabling effective integration of multimodal information.

parameters of the CK5/6 branch.

A. H&E semantic segmentation branch

In the H&E semantic segmentation branch, we employed a multi-resolution input strategy to enhance model performance. Specifically, 10x magnification H&E-stained slides are first processed using the VGG-UNet network to generate an initial semantic segmentation result. Then, Grad-CAM is used to generate a saliency heatmap $S \in [0,1]^{H\times W}$ where each pixel value $S(i,j)$ represents the importance of the corresponding region to the model's decision. we use a sliding window (128×128 pixels) to search for significant peak regions $R = \{(x_t, y_t), (x_b, y_b)\}$.

Fig. 4 compares different zoom strategies. (a) concentric zoom, where the central region is zoomed in concentrically; (b) global zoom, where the entire field is zoomed in and segmented; (c) our strategy, which selectively zooms into key regions, avoiding irrelevant areas and focusing on regions most important for detailed feature extraction. Fig. 5 illustrates the multi-resolution fusion module. Finally, the tissue regions corresponding to R are extracted from the original WSI and zoomed in 40x for subsequent detailed feature extraction.

The H&E semantic segmentation module consists of two main branches: the target branch and the detail branch. The input to the target branch is $X_{aim} \in \mathrm{R}^{C\times H\times W}$, and the input to the detail branch is $X_{high} \in \mathrm{R}^{C\times H\times W}$, where $H \times W$ represents the size of the feature map, and C represents the number of channels, with all having three channels. The resolution of the target branch is r_{aim}, and the resolution of the detail branch is r_{high}. The encoder parts of both branches use the VGG-UNet.

To ensure alignment between the target and detail branches, an appropriate fusion layer is selected. During the encoder downsampling process, the resolution changes as $SRF = 2^d r$, where d represents the downsampling depth of the encoder, and r is the resolution of the input patch (in μm/px). During the decoder process, the resolution changes as $SRF = 2^{d_e - d_f} \cdot r$, where d_e represents the encoder depth and d_f represents the decoder depth. To ensure proper fusion at the same resolution, the SRF ratio between the two branches must satisfy:

$$\frac{SRF_{aim}}{SRF_{high}} = 2^{d_{aim}-d_{high}} \cdot \frac{r_{high}}{r_{aim}} \quad (1)$$

When both branches have the same resolution, the ratio must satisfy:

$$\frac{SRF_{aim}}{SRF_{high}} = 1 \quad (2)$$

To fuse features from different resolutions, a multi-resolution fusion module is employed. Since the target and detail branches operate at different depths in the decoder, we employed a Squeeze-and-Excitation (SE) module to adjust the channel numbers.

Considering that the key regions are determined based on the resolution of the original input features, it is necessary to rescale the coordinates of these regions during fusion, as X_{aim} and X_{high} are at different scales. The rescaled coordinates of the key region are given as (top_left_x, top_left_y, bottom_right_x, bottom_right_y). Then, the corresponding region is cropped from X_{aim} and fused with the channel-reduced X_{high} to obtain the fused features. These fused features are then smoothed, and the updated result is used to refine X_{aim} , completing the feature fusion process.

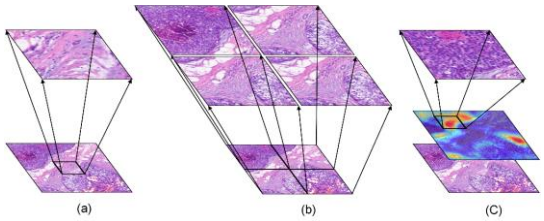

Fig. 4. illustrates three different zoom-in strategies. (a) high-magnification zoom of the central region to match the original view; (b) global zoom followed by region division for analysis; (c) our method, selectively zooming in on key regions to maintain the original field of view.

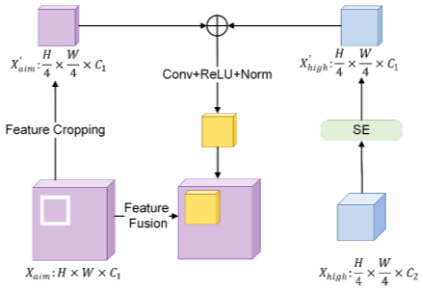

Fig. 5. Fusion module of the target and detail branches. The detail branch is first processed with a Squeeze-and-Excitation (SE) block and then fused with the target branch within the key region.

B. Edge Detection Branch for CK5/6

The C/K56 edge detection branch consists of two main inputs: $X_{sem} \in \mathrm{R}^{3\times H\times W}$,representing the three-channel H&E-stained slides, and $X_{edge} \in \mathrm{R}^{2\times H\times W}$, the grayscale-converted C/K56 image. Both inputs are processed using an encoder-decoder architecture.

To fully exploit the semantic information and edge characteristics of C/K56-stained slides, we propose an edge feature fusion module. As shown in Fig. 6. C/K56 staining exhibits significant coloring properties at the edges of carcinoma in situ regions, while the shallow features of the encoder contain rich low-level information (e.g., textures, edges). Therefore, we perform cross-modal fusion of features from the semantic segmentation branch and the edge detection branch during the encoding phase. This module comprises a channel attention mechanism, a gated attention mechanism, and a depthwise separable convolution.

The four encoder features from the semantic segmentation

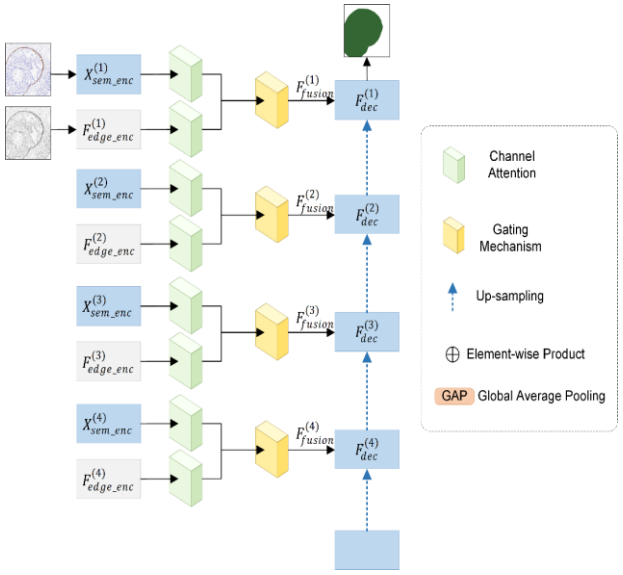

Fig. 6. Edge detection fusion module. The four encoding layers from the CK5/6 semantic branch and the edge branch are first refined using a Channel Attention mechanism. Then, they are fused through a Gating Mechanism. Finally, the four fused features are concatenated with the upsampled features from the decoder of the semantic branch.

branch $\{X_{sem_enc}^{(l)}\}_{l=1}^{4}$ and the edge detection branch $\{F_{edge_enc}^{(l)}\}_{l=1}^{4}$ are first passed through a channel attention mechanism to filter important feature channels, generating enhanced features x'_{sem} and x'_{edge}.

Then, the enhanced features x'_{sem} and x'_{edge} are concatenated along the channel dimension to form a new feature representation F_{concat}. This combined feature is then fed into an attention gating mechanism, which adaptively adjusts spatial weights, highlighting critical regions such as the edges of carcinoma in situ, while suppressing background noise.

Finally, a depthwise separable convolution is applied to reduce computational complexity while maintaining feature representation capability. The fusion module generates fused features $\{F_{fusion}^{(l)}\}_{l=1}^{4}$ at four encoder levels. These features are recursively fused with the decoder of the C/K56 semantic segmentation branch through skip connections:

$$F_{dec}^{(l)} = Up\left(F_{dec}^{(l+1)}\right) + F_{fusion}^{(l)} \tag{3}$$

Where Up denotes bilinear upsampling. Through multi-level fusion, the model leverages both shallow detail information and deep semantic information, enhancing edge detection capabilities. The semantic segmentation branch generates the C/K56 semantic segmentation result $Y_{sem} \in R^{H\times W}$, while the edge detection branch produces the grayscale C/K56 semantic segmentation result $Y_{edge} \in R^{H\times W}$.

C. Multi-Scale Feature Fusion Branch

The branch aims to integrate the semantic segmentation information of H&E-stained slices with the edge detection information of C/K56-stained slices. This branch designs a cross-modal feature fusion module with a hierarchical progressive structure, as shown in Fig.7. The module utilizes feature maps from four different scales of the H&E semantic segmentation decoder, denoted as $\{X_{sem_dec}^{(l)}\}_{l=1}^{4}$, and corresponding scale feature maps from the C/K56 edge detection decoder, denoted as $\{X_{edg_dec}^{(l)}\}_{l=1}^{4}$, Through iterative fusion with the upper-level fusion result F_{fusion}^{l-1} multi-scale features are refined layer by layer.

To effectively fuse the aforementioned two sets of features, we propose a multi-scale feature fusion module that includes a spatial attention mechanism, an axial cross-attention mechanism, and a feature concatenation channel attention mechanism. First, the spatial attention mechanism dynamically adjusts the input feature maps $X_{sem_dec} \in$

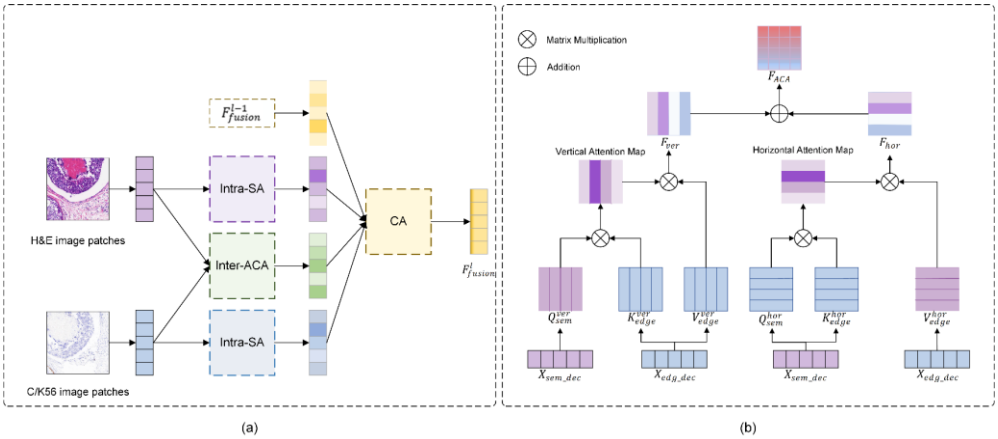

Fig. 7. The left panel (a) illustrates the multi-modal fusion module, where H&E semantic features and CK5/6 edge features are used to learn key tumor regions through a spatial attention mechanism. Meanwhile, the axial attention mechanism learns the interaction between the two modalities. Finally, the three features are weighted and fused with the results from the previous layer. The right panel (b) demonstrates the axial attention mechanism, where H&E semantic features are used as Q, and CK5/6 edge features as K and V, with computations performed in the axial and horizontal directions, respectively.

$R^{C\times H\times W}$ and $X_{edge_dec} \in R^{C\times H\times W}$,to obtain the attention-weighted feature maps $X^l_{sem_dec} \in R^{C\times H\times W}$ and $X^l_{edge_dec} \in R^{C\times H\times W}$.

The multi-head attention mechanism in Transformer models incurs high computational costs. Additionally, global self-attention lacks directional bias, making it prone to interference from irrelevant areas, which hinders fine-grained alignment. To address these issues and improve tumor segmentation accuracy, we adopted the axial attention mechanism to fuse multimodal data. This process involves attention calculations in both horizontal and vertical directions. The horizontal attention weight is calculated as follows:

$$A_{hor} = \sigma\left(\left(X^l_{sem_dec} \cdot W^{hor}_Q\right)\left(X^l_{edge_dec} \cdot W^{hor}_K\right)^T\right)/\sqrt{d_K} \quad (4)$$

$$F_{hor} = A_{hor} \cdot \left(X^l_{edge_dec} \cdot W^{hor}_V\right) \quad (5)$$

$$F^l_{ACA} = F_{hor} + F_{ver} \quad (6)$$

The horizontally weighted feature map $F_{hor} \in R^{C\times H\times W}$ and vertically weighted feature map $F_{ver} \in R^{C\times H\times W}$ are obtained through horizontal and vertical attention. The W^{hor}_Q, W^{hor}_K, W^{hor}_V, are the linear projection matrices for the Query, Key, and Value, respectively These two feature maps are then summed to produce the final output of the Axial Cross Attention module, denoted as F^l_{ACA}

Finally, the spatially attention-weighted feature maps X'_{sem_dec} and X'_{edge_dec}, the axial cross-attention output F_{ACA}, and the upper-level fusion result F^{l-1}_{fusion} are concatenated. The F^l_{fusion} is obtained through a channel attention mechanism.

D. Loss Function

To train the model proposed in this paper, we used three loss functions: L_{seg}, L_{edge} and L_{fusion}, to ensure that semantic segmentation and edge detection can learn collaboratively.

First, the loss function L_{seg} for the H&E semantic segmentation branch is primarily used to optimize the semantic segmentation performance of H&E images:

$$L_{seg} = \left(0.5 \cdot L_{aim_dice} + 0.5 \cdot L_{aim_CE}\right) + 0.5 \cdot L_{high_CE} \quad (7)$$

Here, L_{aim_dice} is the Dice loss function for the target branch, Both L_{aim_dice} and L_{high_CE} are cross-entropy losses, used to measure the accuracy of class predictions.

Next, the loss function L_{edge} for the CK5/6 edge detection branch is designed to enhance the accurate localization of tumor boundaries, and it is defined as:

$$L_{edge} = 0.5 \cdot L_{sem_CE} + 0.5 \cdot L_{edge_dice} \quad (8)$$

where L_{sem_CE} and L_{edge_dice} represent the cross-entropy loss for the CK5/6 semantic segmentation branch and the Dice loss for the CK5/6 edge semantic segmentation, respectively.

Finally, we combine the semantic segmentation and edge detection losses to construct the final semantic segmentation network loss function L_{fusion}:

$$L_{fusion} = L_{seg} + \gamma \cdot L_{edge} \quad (9)$$

Here, γ is a hyperparameter used to control the impact of the edge detection loss on the overall training process, ensuring that both semantic segmentation and edge information are optimized collaboratively.

IV. Experimental Results

In this chapter, we will present our experiments from six aspects: Datasets, Parameter analysis, Evaluation metrics, Comparison with other methods, Ablation study, and Model Complexity Analysis.

A. Datasets

We used two WSI breast cancer tumor datasets to evaluate our model: the Breast Cancer Semantic Segmentation (BCSS) Dataset and a breast cancer dataset collected from the Chinese PLA General Hospital.

1) Our breast cancer dataset: We collected whole-slide imaging (WSI) data from 73 breast cancer patients at the Chinese PLA General Hospital. This dataset includes H&E staining, CK5/6 staining, and SMMHC staining WSIs for each patient. All slides were prepared following the standardized procedures of the PLA Pathology Department and digitized using a Jiangfeng scanner (model KFPBL00500108015) at a spatial resolution of 0.25 μm/px under consistent scanning parameters.

The dataset spans patients aged 25 to 72, with an average age of 46.6 years, primarily concentrated between 30 and 50 years old. It covers various tumor types, including invasive carcinoma, DCIS (ductal carcinoma in situ), mixed types, lobular carcinoma, mucinous carcinoma, and papillary carcinoma, with a ratio of invasive carcinoma to carcinoma in situ of approximately 1.2:1. Among invasive carcinomas, grade I accounts for 2.4%, grade II for 76.2%, and grade III for 4.8%. Annotation work was performed by a team of expert doctors at the PLA General Hospital using QuPath software. From the H&E-stained slides of the 73 breast cancer patients, we extracted 1,201 regions of interest (ROIs), with a similar number (1,201) of ROIs extracted from the CK5/6-stained WSIs. Two categories were annotated: carcinoma in situ and invasive carcinoma.

We evaluated the performance of our model by splitting the dataset into training, validation, and test sets in a 6:2:2 ratio.

2) BCSS dataset: The BCSS dataset is a large-scale dataset annotated based on breast cancer WSIs from The Cancer Genome Atlas (TCGA). It includes annotations from pathologists, pathology residents, and medical students, covering over 20,000 annotated regions of breast cancer tissue. All slides were digitized at a spatial resolution of 0.25 μm/px. The dataset includes five annotated classes: Tumor, Stroma, Inflammatory, Necrosis, and Other (e.g., ducts, lobules, and other specific tissue types). We applied the same preprocessing steps as those used for the H&E branch of the PLAGH dataset, generating 1,484 H&E patches for validating the performance of the H&E semantic segmentation branch.

B. Parameter Analysis

In this subsection, we discuss the adjustable parameter γ. γ is used in Equ. (9) to balance the loss between the semantic segmentation branch and the edge detection branch, ensuring that both branches contribute maximally when working together. To evaluate the impact of γ on model performance, in our experiments, we set γ to 0.75, 0.5, 0.25, and 0.1. The experiments were conducted using our dataset as the baseline, and the corresponding evaluation results are shown in Table I.

From the results in Table I, it can be observed that as γ increases from 0.1 to 0.5, both the Dice and mIoU values improve, indicating that appropriately increasing the weight of the edge detection branch provides valuable edge information to the semantic segmentation branch, thereby enabling more accurate tumor boundary localization. However, when γ continues to increase, the performance metrics begin to decline. This suggests that excessively high values of γ cause the model to overly rely on the edge detection branch during training, weakening the learning of global semantic features and resulting in worse semantic segmentation performance.

Based on this analysis, we set the default value of γ to 0.5 in our method to ensure that the edge detection branch effectively assists the semantic segmentation branch, significantly improving the semantic segmentation results.

C. Implementation Details

We evaluated the performance of our model on the breast cancer tumor semantic segmentation task using both the BCSS dataset and our dataset. To comprehensively assess the model's performance, we first calculated the true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) for each class. Based on these statistics, we further computed three key evaluation metrics: accuracy (ACC), mean Intersection over Union (mIoU), and Dice coefficient.

In the BCSS dataset, we considered five categories: Tumor, Stroma, Lymphocytes, Necrosis, and Other. In our dataset, two categories were considered: carcinoma in situ and invasive carcinoma.

TABLE I
TEST THE SEMANTIC SEGMENTATION PERFORMANCE OF THE MODEL UNDER DIFFERENT VALUES OF THE PARAMETER γ

	Acc(%)	Dice(%)	mIoU(%)
γ=0.75	94.30	92.89	88.81
γ=0.5	95.01	93.16	89.89
γ=0.25	95.20	93.52	89.58
γ=0.1	94.90	92.85	89.81

TABLE II
EVALUATION RESULTS OF DIFFERENT MULTI-RESOLUTION METHODS ON THE BCSS DATASET AND OUR DATASET

Dataset	Method	Acc(%)	Dice(%)	mIoU(%)	p-value
BCSS	HookNet	88.89±0.24	69.79±0.55	63.41±0.36	P<0.001
	msY-Net	86.18±0.31	65.41±0.51	59.20±0.64	P<0.001
	H&E branch	**90.48±0.23**	**81.28±0.26**	**70.64±0.34**	P<0.001
Our dataset	HookNet	92.85±0.27	84.38±0.44	78.07±0.42	P<0.001
	msY-Net	92.52±0.31	83.41±0.39	75.01±0.51	P<0.001
	H&E branch	**94.54±0.22**	**87.01±0.21**	**88.34±0.35**	P<0.001

We implemented all the models proposed in this paper using the PyTorch framework and trained them on an NVIDIA RTX 3090 24GB GPU. The Adam optimizer was used to train both the proposed and comparative methods, with an initial learning rate set to 0.0001. The learning rate was dynamically adjusted using the CosineAnnealingLR strategy, smoothly decaying from the initial value to the minimum value. The batch size was set to 2.

D. Comparison with Other Methods

To validate the effectiveness of our method, we compared it with several state-of-the-art semantic segmentation approaches on both the BCSS dataset and our collected dataset. Each model was independently run five times under identical experimental conditions. These methods include CNN-based semantic segmentation models such as UNet [8], UNet++ [20], and Attention UNet [21]; hybrid models combining Transformer and UNet, such as Swin-UNet [22], UCTransNet [23], and nnWNet [24]; and models that integrate state space models with UNet, such as VM-UNet [25]. We also compared with methods specifically designed for multi-resolution H&E semantic segmentation, such as HooKNet [17] and msY-Net [11].

In Table II, we compare different multi-resolution methods with our proposed H&E semantic segmentation branch (H&E branch). Both HookNet [17] and msY-Net [16] perform semantic segmentation at 40× resolution and enhance their results by incorporating global information from 10×. In contrast, our method conducts semantic segmentation at 10× and selectively zooms into key regions at 40× to supplement fine-grained details. The experimental results show that our method achieves Dice scores of 81.28% and 87.01% on the BCSS dataset and our collected dataset, respectively. On the BCSS dataset, our method outperforms other methods by approximately 11.49% and 15.87%, and on the collected

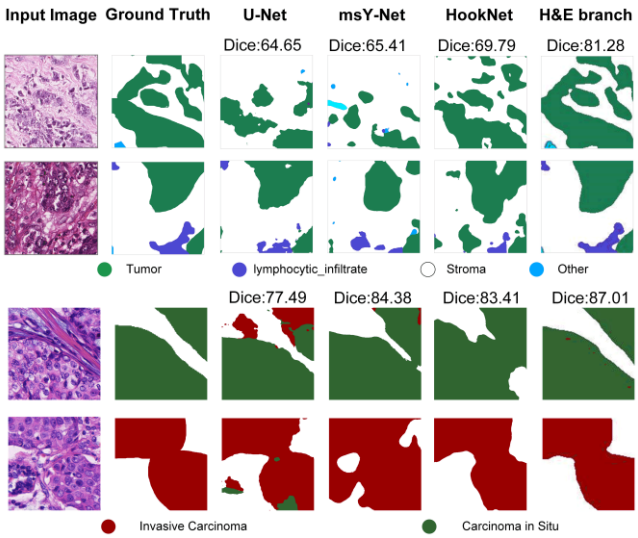

Fig. 8. Visualization of results for different Multi-resolution Methods on the BCSS dataset and Our dataset. The first two rows correspond to the BCSS dataset, while the last two rows correspond to the collected dataset.

dataset by 2.63% and 3.60%, respectively. The visual results in Fig. 8 also confirm that our semantic segmentation results are more accurate. When semantic segmentation is performed at high resolution (40×), global context can be supplemented by 10× inputs, but the abundance of fine details may introduce noise and artifacts, making semantic segmentation more challenging. Our method, by segmenting at 10× and refining key regions with 40× details, not only significantly reduces computational cost but also achieves a better balance between global context and local detail.

In Table III, we present the evaluation results of single-resolution methods and our proposed multi-modal semantic segmentation model on both datasets. Table IV reports the per-class evaluation results of different semantic segmentation methods across both datasets. First, we compare the semantic segmentation performance of the UNet model at 40× and 10× resolutions. The results show that at 10×, the Dice scores improve by 9.93% and 4.97% on the BCSS dataset and our collected dataset, respectively, confirming the challenges of semantic segmentation at high resolution. Therefore, we choose to perform semantic segmentation at 10×. Specifically, under 10× resolution, our method achieves Dice scores of 81.28% and 93.16% on the BCSS and collected datasets, respectively. Compared with CNN-based methods (UNet, UNet++, Attention UNet), our method improves Dice scores by 6.70%, 5.95%, and 1.86% on the BCSS dataset, and by 10.70%, 7.89%, and 7.35% on the collected dataset, respectively. Compared with transformer-based hybrid methods (Swin-Unet, UCTransNet), our method shows improvements of 5.91% and 11.87%. 0n BCSS, and 7.01% and 7.80% on the collected dataset. Compared to VM-UNet, which combines state-space models and UNet, our method outperforms it by 7.15% on BCSS and 7.78% on the collected dataset. As shown in Fig. 9, with the help of CK5/6 edge information, our tumor semantic segmentation boundaries become more precise, further confirming the effectiveness of edge features in improving semantic segmentation accuracy.

To further validate the effectiveness of the proposed method, we conducted statistical significance analysis on the tumor segmentation task for breast cancer by comparing it with state-of-the-art models, using Dice and mIoU metrics. Paired t-tests were used to assess the significance of performance differences. The experimental results show that the p-values for the Dice and mIoU score differences between the proposed method and the comparative models are both less than 0.001, indicating highly significant improvements with strong clinical application potential.

E. Ablation study

To validate the effectiveness of the proposed core components, we conducted an ablation study, evaluating the contributions of the Multi-Resolution Fusion Module (MRFM), the Edge Detection Module (EDM), and the Multi-Scale Feature Fusion Module (MSFFM). Using UNet as the baseline, we incrementally incorporated these components and compared five models:

1) UNet: The baseline model performing semantic segmentation on 10× images.

2) UNet + MRCFM: A variant with the center region at 40× magnification concatenated with features from the 10× image.

3) UNet + MRFM: The baseline model with the proposed Multi-Resolution Fusion Module for fusing 40× and 10× features.

4) UNet + MRFM + EDM: This model incorporates the Edge Detection Module with the UNet + MRFM architecture but without deeper integration of edge information.

5) Proposed model: The complete model integrating UNet, MRFM, EDM, and the Multi-Scale Feature Fusion Module (MSFFM).

TABLE III
EVALUATION RESULTS OF DIFFERENT SINGLE RESOLUTION METHODS ON THE BCSS DATASET AND OUR DATASET

Method	BCSS			Our dataset			
	Acc(%)	Dice(%)	mIoU(%)	Acc(%)	Dice(%)	mIoU(%)	p-value
U-Net_40x	86.06±0.31	64.65±0.37	58.97±0.54	86.07±0.26	77.49±0.28	74.24±0.48	P<0.001
U-Net_10x	88.17±0.23	74.58±0.29	61.72±0.28	92.01±0.24	82.46±0.33	82.72±0.32	P<0.001
UNet++	83.09±0.17	75.33±0.35	63.23±0.37	92.21±0.21	85.18±0.46	85.76±0.28	P<0.001
AttenUNet	84.87±0.19	79.42±0.24	66.64±0.28	92.43±0.19	85.81±0.45	83.81±0.26	P<0.001
Swin-Unet	89.68±0.24	75.37±0.36	65.62±0.32	93.14±0.18	86.15±0.26	83.11±0.34	P<0.001
UCTransNet	80.12±0.26	69.41±0.31	58.18±0.42	90.09±0.22	85.36±0.23	79.87±0.26	P<0.001
VM-UNet	81.48±0.29	74.13±0.30	60.07±0.46	92.45±0.21	85.38±0.36	82.68±0.41	P<0.001
nnWNet	85.16±0.12	78.69±0.18	65.37±0.35	92.83±0.16	85.49±0.38	85.48±0.24	P<0.001
Proposed	**90.48±0.23**	**81.28±0.26**	**70.64±0.34**	**95.01±0.13**	**93.16±0.27**	**89.89±0.26**	P<0.001

TABLE IV
EVALUATION RESULTS OF DIFFERENT SEMANTIC SEGMENTATION METHODS FOR EACH CATEGORY ON THE TWO DATASETS

Method	BCSS					Our dataset	
	Tumor	Stroma	Inflammatory	Necrosis	Other	Carcinoma in situ	Invasive carcinoma
U-Net_40x	89.56±0.39	81.67±0.60	55.54±0.65	50.34±0.59	46.12±0.72	67.51±0.26	87.47±0.24
HookNet	84.70±0.32	73.86±0.51	68.84±0.58	81.00±0.51	40.54±0.68	82.16±0.55	86.60 ±0.38
msY-Net	84.08±0.29	69.40±0.47	64.41±0.55	76.14±0.47	33.03±0.63	83.44±0.45	83.38±0.39
U-Net_10x	89.23±0.25	81.15±0.38	72.90±0.42	66.72±0.36	62.96±0.48	73.80±0.36	91.12±0.39
UNet++	88.89±0.34	81.72±0.35	75.61±0.51	66.64±0.39	63.80±0.47	78.93±0.57	91.43±0.40
AttenUNet	90.59±0.17	84.12±0.54	78.72±0.72	73.34±0.50	70.34±0.67	81.94±0.34	89.68±0.20
Swin-Unet	88.16±0.26	80.98±0.34	75.37±0.52	63.54±0.39	68.82±0.50	81.44±0.39	91.18±0.22
UCTransNet	87.21±0.31	77.47±0.29	64.83±0.52	65.16±0.46	52.39±0.49	81.49±0.27	89.24±0.30
VM-UNet	88.29±0.30	80.54±0.29	72.91±0.47	67.13±0.45	61.76±0.43	80.29 ±0.42	90.47±0.39
nnWNet	90.27±0.21	83.16±0.22	**81.33±0.37**	66.91±0.60	**71.79±0.37**	79.13±0.69	91.83±0.11
H&E branch	**92.02±0.25**	**84.23±0.24**	77.74±0.42	**81.02±0.22**	71.43±0.37	85.75±0.36	88.26±0.14
Proposed						**92.12±0.34**	**94.21±0.14**

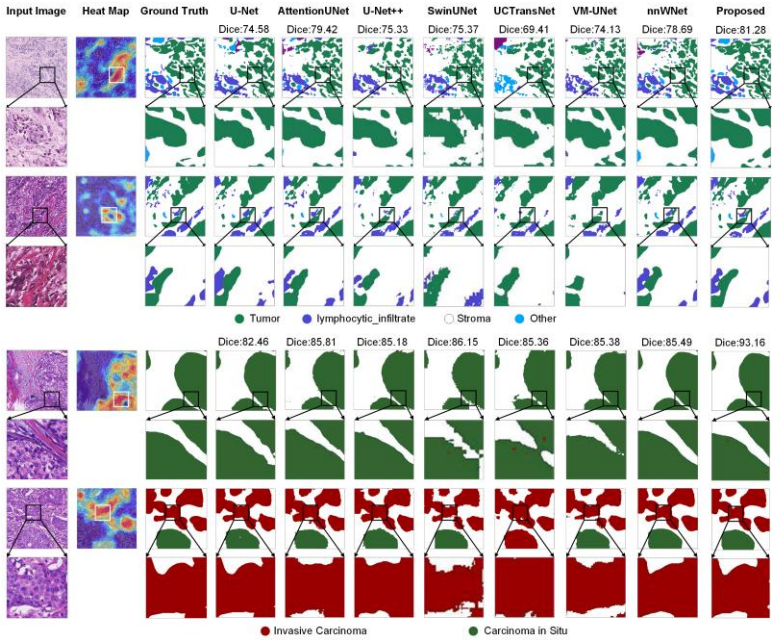

Fig. 9. Visualization of results for different single-resolution Methods on the BCSS dataset and our dataset. The first two rows correspond to the BCSS dataset, while the last two rows correspond to the collected dataset.

TABLE V
EVALUATION RESULTS OF ABLATION EXPERIMENTS ON OUR DATASET.

Method	Acc(%)	Dice(%)	mIoU(%)
UNet	92.03	82.53	82.74
UNet+MRCFM	94.18	85.65	85.25
UNet+MRFM	94.54	87.01	88.34
UNet+MRFM+EDM	94.74	91.11	89.01
Proposed	95.01	93.16	89.89

Table V shows the ablation study results, emphasizing each module's contribution. Compared to direct segmentation on 10× H&E images, UNet + MRFM improves local feature representation with added 40× details. UNet + MRCFM uses region-guided zoom-in, focusing on challenging areas and outperforming concentric cropping. Edge information enhances boundary precision, and the proposed model surpasses UNet + MRFM + EDM, confirming the effectiveness of CK5/6 edge guidance and our multimodal fusion strategy.

The confusion matrix in Fig. 10 further demonstrates that, in the baseline model (U-Net), the limited resolution at low magnification makes it difficult to capture the overall structure and spatial distribution of cancer cells, leading to confusion between in situ carcinoma and invasive carcinoma. While

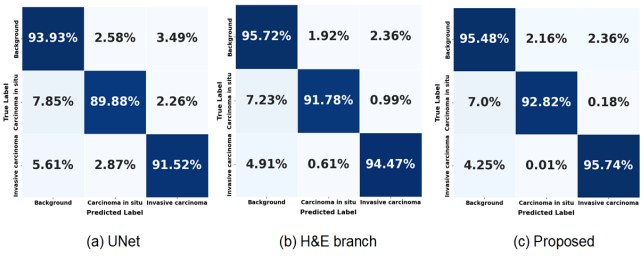

Fig. 10. Confusion matrices for the models UNet, H&E branch, and Proposed model.

TABLE VI
THE COMPLEXITY ANALYSIS OF THE PROPOSED METHOD COMPARED TO OTHER METHODS

Method	Params(M)	FLOPs(G)	InferenceTime(s)
U-Net	7.85	112.81	0.0410
UNet++	8.59	203.49	0.0409
AttenUNet	8.73	116.04	0.0570
Swin-Unet	27.14	131.51	0.0294
UCTransNet	33.79	558.24	0.1455
VM-UNet	7.56	**94.98**	0.0296
nnWNet	**7.04**	171.27	**0.0165**
H&E branch	35.31	452.58	0.0844
Proposal	78.50	1064.96	0.1484

UNet+MRFM incorporates high-resolution details and low-resolution global context, enhancing its ability to identify tumor regions, the absence of edge information guidance results in blurred tumor boundaries, often confusing them with surrounding normal tissue or background, thus affecting segmentation accuracy.

F. Model Complexity Analysis

To analyze the model complexity, we compared the proposed model with other state-of-the-art models in terms of parameter count (Params), floating-point operations (FLOPs), and inference time on an NVIDIA RTX 3090 GPU. Table VI shows the results of the complexity analysis. Among CNN-based models, U-Net had the smallest Params, FLOPs, and inference time. In contrast, Transformer-based models generally had longer inference times. Compared to existing methods, our proposed model has the longest inference time, mainly due to its more complex structural design: in addition to processing multi-resolution information through a dual-branch structure, it also integrates two branches for edge-guided information, achieving optimal segmentation performance. Despite the longer inference time, it remains within an acceptable and reasonable range, demonstrating its feasibility for clinical application.

V. Conclusion

This paper presents a computationally efficient multimodal tumor segmentation method tailored for IoMT platforms. By fusing multi-resolution semantic features from H&E-stained images with edge information from spatially aligned CK5/6-stained slides, the method improves boundary delineation and segmentation precision. The architecture integrates three lightweight modules: a dual-branch H&E semantic segmentation module that focuses on diagnostically challenging regions, an edge detection module leveraging CK5/6 images, and a multi-scale fusion module for refined prediction. Experimental results on the Breast Cancer Semantic Segmentation dataset and images from the Chinese People's Liberation Army General Hospital dataset demonstrate superior performance compared to state-of-the-art methods.

Looking ahead, we aim to deploy this method on IoMT platforms to enhance the efficiency of hospital resource utilization and improve the work efficiency of healthcare professionals, further contributing to the promotion of patient health. Despite the significant advancements in segmentation performance achieved by our model, two challenges remain: (1) Future research will explore the integration of additional modalities, such as radiological data or pathology reports, to broaden its application in consumer-centric smart healthcare systems; (2) The complexity of the model needs to be further simplified to better meet the needs of portable medical services in hospitals. To enhance the scalability and privacy protection of the system, we plan further optimizations based on a cloud-edge collaborative architecture. By employing federated learning, we aim to store and train data on edge devices, with the locally trained model weights being uploaded to the cloud for aggregation. The resulting global model will then be distributed back to the nodes, enabling the automatic segmentation of medical images.

References

[1] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *J. Pathol. Inform.*, vol. 7, no. 1, p. 29, Jan. 2016, doi: 10.4103/2153-3539.186902.

[2] D. C. Zaha, "Significance of immunohistochemistry in breast cancer," *World J. Clin. Oncol.*, vol. 5, no. 3, p. 382, 2014, doi: 10.5306/wjco.v5.i3.382.

[3] Y. Zhao, S. Wang, Y. Zhang, Y. Ren, Y. Zhang, and S. Pang, "Dual Encoder Cross-Shape Transformer Network for Medical Image Segmentation in Internet of Medical Things for Consumer Health," *IEEE Trans. Consum. Electron.*, pp. 1–1, 2025, doi: 10.1109/TCE.2025.3526801.

[4] S. Sai, M. Prasad, A. Upadhyay, V. Chamola, and N. Herencsar, "Confluence of Digital Twins and Metaverse for Consumer Electronics: Real World Case Studies," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 3194–3203, Feb. 2024, doi: 10.1109/TCE.2024.3351441.

[5] W. Gao and Z. Zhao, "Self-Supervised Multi-Source Heterogeneous Data Fusion Using Encode and Decode Attention for Intelligent Medical Device Communication Analysis," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 1318–1325, Feb. 2024, doi: 10.1109/TCE.2023.3321331.

[6] K. Doulani, A. Rajput, A. Hazra, M. Adhikari, and A. K. Singh, "Explainable AI for Communicable Disease Prediction and Sustainable Living: Implications for Consumer Electronics," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 2460–2467, Feb. 2024, doi: 10.1109/TCE.2023.3325155.

[7] D. Komura and S. Ishikawa, "Machine Learning Methods for Histopathological Image Analysis," *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 34–42, 2018, doi: 10.1016/j.csbj.2018.01.001.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 18, 2015, *arXiv*: arXiv:1505.04597. doi: 10.48550/arXiv.1505.04597.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," Oct. 10, 2016, *arXiv*: arXiv:1511.00561. doi: 10.48550/arXiv.1511.00561.

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," Aug. 22, 2018, *arXiv*: arXiv:1802.02611. doi: 10.48550/arXiv.1802.02611.

[11] J. Saltz *et al.*, "Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images," *Cell Rep.*, vol. 23, no. 1, pp. 181-193.e7, Apr. 2018, doi: 10.1016/j.celrep.2018.03.086.

[12] P. Zhao, J. Zhang, W. Fang, and S. Deng, "SCAU-Net: Spatial-Channel Attention U-Net for Gland Segmentation," *Front. Bioeng. Biotechnol.*, vol. 8, p. 670, July 2020, doi: 10.3389/fbioe.2020.00670.

[13] Z. Wen, R. Feng, J. Liu, Y. Li, and S. Ying, "GCSBA-Net: Gabor-Based and Cascade Squeeze Bi-Attention Network for Gland Segmentation," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 4, pp. 1185–1196, Apr. 2021, doi: 10.1109/JBHI.2020.3015844.

[14] Z. Lu *et al.*, "BreasTDLUSeg: A coarse-to-fine framework for segmentation of breast terminal duct lobular units on histopathological whole-slide images," *Comput. Med. Imaging Graph.*, vol. 118, p. 102432, Dec. 2024, doi: 10.1016/j.compmedimag.2024.102432.

[15] H. Abdel-Nabi, M. Z. Ali, and A. Awajan, "A multi-scale 3-stacked-layer coned U-net framework for tumor segmentation in whole slide images," *Biomed. Signal Process. Control*, vol. 86, p. 105273, Sept. 2023, doi: 10.1016/j.bspc.2023.105273.

[16] R. Schmitz *et al.*, "Multi-scale fully convolutional neural networks for histopathology image segmentation: From nuclear aberrations to the global tissue architecture," *Med. Image Anal.*, vol. 70, p. 101996, May 2021, doi: 10.1016/j.media.2021.101996.

[17] M. Van Rijthoven, M. Balkenhol, K. Siliņa, J. Van Der Laak, and F. Ciompi, "HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images," *Med. Image Anal.*, vol. 68, p. 101890, Feb. 2021, doi: 10.1016/j.media.2020.101890.

[18] N. Dong, M. Kampffmeyer, X. Liang, Z. Wang, W. Dai, and E. P. Xing, "Reinforced Auto-Zoom Net: Towards Accurate and Fast Breast Cancer Segmentation in Whole-slide Images," July 29, 2018, *arXiv*: arXiv:1807.11113. doi: 10.48550/arXiv.1807.11113.

[19] R. Deng *et al.*, "Omni-Seg: A Scale-Aware Dynamic Network for Renal Pathological Image Segmentation," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 9, pp. 2636–2644, Sept. 2023, doi: 10.1109/TBME.2023.3260739.

[20] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," July 18, 2018, *arXiv*: arXiv:1807.10165. doi: 10.48550/arXiv.1807.10165.

[21] O. Oktay *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas," May 20, 2018, *arXiv*: arXiv:1804.03999. doi: 10.48550/arXiv.1804.03999.

[22] H. Cao *et al.*, "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," May 12, 2021, *arXiv*: arXiv:2105.05537. doi: 10.48550/arXiv.2105.05537.

[23] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-wise Perspective with Transformer," Jan. 25, 2022, *arXiv*: arXiv:2109.04335. doi: 10.48550/arXiv.2109.04335.

[24] Y. Zhou, L. Li, L. Lu, and M. Xu, "nnWNet: Rethinking the Use of Transformers in Biomedical Image Segmentation and Calling for a Unified Evaluation Benchmark," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, June 2025, pp. 20852–20862. doi: 10.1109/CVPR52734.2025.01942.

[25] J. Ruan, J. Li, and S. Xiang, "VM-UNet: Vision Mamba UNet for Medical Image Segmentation," Nov. 08, 2024, *arXiv*: arXiv:2402.02491. doi: 10.48550/arXiv.2402.02491.

Tengyue Li received the PhD degree in computer science from the University of Macau. She is currently a lecturer with the North China University of Technology. She has published more than 20 SCI papers, including IoTJ, Information Fusion, Knowledge base System. She contributed six province level scientific research projects. Her research focuses on the development the methods for applying AI on medical images.

Shuangli Song is currently pursuing the M.S. degree with the Department of Artificial Intelligence and Computer Science, North China University of Technology. Her research focuses on the development of methods for applying AI to medical images.

Qiong Wang received the Ph.D. degree in pathology and pathophysiology. She is currently an Associate Chief Technician with the Department of Pathology, the First Medical Center, Chinese PLA General Hospital, Beijing, China. She has over ten years of experience in clinical tumor pathology testing and research. She serves as a committee member of the Molecular Diagnostics Branch of the Chinese Research Hospital Association and the Tumor Molecular Pathology Group of the Chinese Anti-Cancer Association.

Simon James Fong graduated from La Trobe University in Australia with a First Class Honours BEng degree in Computer Systems and a PhD in Computer Science in 1993 and 1998, respectively. He is currently an Associate Professor in the Department of Computer and Information Science at the University of Macau and a Senior Visiting Scholar at Tsinghua University, Beijing. Dr. Fong has published over 500 peer-reviewed international conference and journal papers, primarily in the areas of medical applications and data mining. He actively serves as a SIG Chair for IEEE ComSoc e-Health and as Editor-in-Chief of the Medical Data Mining journal.

Wei Song received the B.E. degree from Northeastern University, Shenyang, China, in 2005, and the M.Eng. and Ph.D. degrees from Dongguk University, Seoul, South Korea, in 2008 and 2013, respectively. He is currently a Professor of computer science and technology with the North China University of Technology, Beijing, China. His research interests include environmental perception, object recognition, semantic segmentation, 3-D reconstruction, and LiDAR applications.

Juntao Gao is a research associate professor working with Beijing National Research Center for Information Science and Technology, Tsinghua University and Center for Synthetic & Systems Biology, Tsinghua University. His work has been reported by Nature Methods as 'Research Highlights'. Honor: Silver Medals, Exhibition of Inventions Geneva, 2023; The 35th, 33rd Large Instrument and Equipment Utilization Efficiency Awards, Tsinghua University, P. R. China.

Yi Pan (Member, IEEE) received the B.Eng. and M.Eng. degrees in computer engineering from Tsinghua University, Beijing, China, in 1982 and 1984, respectively, and the Ph.D. degree in computer science from the University of Pittsburgh, Pittsburgh, PA, USA, in 1991. He is currently the Dean and Chair Professor with the Faculty of Computer Science and Control Engineering, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences Shenzhen, Guangdong, China. He has authored or coauthored more than 300 academic articles in SCI indexed journals, including more than 100 papers in top IEEE/ACM transactions/journals. His publications have been cited more than 29000 times, and his current H-index is 101. His research interests include parallel and distributed processing systems, Internet technology, and bioinformatics.

Nilanjan Dey (Senior Member, IEEE) received the B.Tech., M.Tech. in information technology from West Bengal Board of Technical University and Ph.D. degrees in electronics and telecommunication engineering from Jadavpur University, Kolkata, India, in 2005, 2011, and 2015, respectively.
Currently, he is Associate Professor with the Techno International New Town, Kolkata and a visiting fellow of the University of Reading, UK.
He is the Editor-in-Chief of International Journal of Ambient Computing and Intelligence, Associate Editor of IEEE Transactions on Technology and Society and series Co-Editor of Springer Tracts in Nature-Inspired Computing and Data-Intensive Research from Springer Nature and Advances in Ubiquitous Sensing Applications for Healthcare from Elsevier etc.
He is a Fellow of IETE and member of IE, ISOC etc.

Robert Simon Sherratt (Fellow, IEEE) received a B.Eng. from Sheffield City Polytechnic in 1992, M.Sc. from The University of Salford in 1993, and PhD from The University of Salford in 1996. In 1996, he was appointed as a Lecturer in Electronic Engineering with the University of Reading, where he is currently a Professor of Biosensors. His research area is wearable devices, mainly for healthcare and emotion detection.
Eur Ing Professor Sherratt was awarded the 1st place IEEE Chester Sall Award in 2004, 2nd place in 2014 and 2021, 3rd place in 2015 2016 for best papers in the IEEE Transactions on Consumer Electronics.