J. Meas. Sci. Instrum., 2025, 16(3): 359-370.
DOI: 10.62756/jmsi.1674-8042.2025035
http://jmsi.publish.founderss.cn/
jmsi@nuc.edu.cn

A medical image segmentation model based on SAM with an integrated local multi-scale feature encoder

DI Jing*, ZHU Yunlong, LIANG Chan

School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

***Corresponding author:** DI Jing (18139873188@163.com)

Received: September 29, 2024 **Revised:** November 7, 2024 **Accepted:** November 13, 2024

Abstract: Despite its remarkable performance on natural images, the segment anything model (SAM) lacks domain-specific information in medical imaging. and faces the challenge of losing local multi-scale information in the encoding phase. This paper presents a medical image segmentation model based on SAM with a local multi-scale feature encoder (LMSFE-SAM) to address the issues above. Firstly, based on the SAM, a local multi-scale feature encoder is introduced to improve the representation of features within local receptive field, thereby supplying the Vision Transformer (ViT) branch in SAM with enriched local multi-scale contextual information. At the same time, a multiaxial Hadamard product module (MHPM) is incorporated into the local multi-scale feature encoder in a lightweight manner to reduce the quadratic complexity and noise interference. Subsequently, a cross-branch balancing adapter is designed to balance the local and global information between the local multi-scale feature encoder and the ViT encoder in SAM. Finally, to obtain smaller input image size and to mitigate overlapping in patch embeddings, the size of the input image is reduced from 1 024×1 024 pixels to 256×256 pixels, and a multidimensional information adaptation component is developed, which includes feature adapters, position adapters, and channel-spatial adapters. This component effectively integrates the information from small-sized medical images into SAM, enhancing its suitability for clinical deployment. The proposed model demonstrates an average enhancement ranging from 0.038 7 to 0.319 1 across six objective evaluation metrics on BUSI, DDTI, and TN3K datasets compared to eight other representative image segmentation models. This significantly enhances the performance of the SAM on medical images, providing clinicians with a powerful tool in clinical diagnosis.

Key words: segment anything model (SAM); medical image segmentation; encoder; decoder; multiaxial Hadamard product module (MHPM); cross-branch balancing adapter

0 Introduction

Medical images play a crucial role in the diagnosis and treatment of diseases. In many clinical applications such as computer-aided diagnosis, treatment planning, and disease progression monitoring, medical image segmentation is necessary as a key step, and its task is to accurately extract and identify specific anatomical structures, including organs, lesions, and tissues, from multimodal medical images[1-3]. This process typically requires segmentation algorithms with high robustness and accuracy to handle the complexity and heterogeneity of medical images, ensuring precise and effective clinical decision-making.

Recently, deep learning has demonstrated promising potential in medical image segmentation. It can learn complex image features and deliver accurate segmentation results across various tasks, from segmenting specific anatomical structures to identifying pathological regions[4,5]. Inspired by fully convolutional networks (FCN) [6] originally designed for image classification, Ronneberger et al. [7] proposed U-Net, dedicated to medical image segmentation. U-Net is extensively utilized in medical image segmentation tasks due to its unique U-shaped architecture, which effectively integrates contextual information. Additionally, it offers high training efficiency and operates efficiently with relatively small datasets. Chen et al. [8] incorporated the Kolmogorov-Arnold network (KAN) into the U-Net and proposed the U-KAN model, enhancing the U-Net's nonlinear modeling ability and interpretability in the field of medical image segmentation. However, the local receptive field of convolutional neural network (CNN) in U-Net limits it to capturing only local information, resulting in the neglect of global context and spatial details. This also restricts its ability to learn important features such as texture, shape, and size. Wu et

al. [9] proposed FAT-Net, incorporating additional transformer branches to capture long-term dependencies and global contextual information efficiently. Wang et al.[10] introduced UCTransNet, which combines global attention transformers with locality-focused CNNs, effectively improving the accuracy and efficiency of medical image segmentation by balancing global features with local details. Based on Swin Transformers and residual CNNs, Yuan et al. [11] developed CTC-Net, further enhancing the representation of long-term dependencies for medical image segmentation. To preserve precise shape information while capturing global long-term dependencies, Li et al. [12] introduced TCNet for the automatic segmentation of malignant thyroid nodules. However, these methods often fall short of fully exploring local-global cues at each scale and modeling interactions among consensus regions across multiple scales, limiting their ability to adapt to variations in the size, shape, and position of target objects. To address these limitations, Pang et al.[13] proposed a novel intra and inter attention with mutual consistency learning network (IIAM). Despite these advances, both traditional U-Net models and transformer-based approaches still require task-specific customization for different medical imaging modalities, which restricts their generalizability and increases training costs.

Unlike U-Net-based segmentation models, segment anything model (SAM)[14], developed by MetaAI Labs, comprises a Vision Transformer (ViT) [15] encoder, a prompt encoder, and a mask decoder. The encoder is trained on the large-scale SA-1B dataset[16] utilizing the ViT architecture, which endows SAM with robust generalization capabilities. It incorporates three distinct prompting methods—point prompting, box prompting, and text prompting—allowing for improved performance in natural image segmentation. Inspired by SAM, Xiong et al.[17] proposed EfficientSAM, a model that reduces the complexity of SAM while maintaining high segmentation accuracy. SAM can integrate multiple medical image segmentation tasks into a unified framework, which greatly facilitates clinical applications[18]. However, medical images differ significantly from natural images, making it challenging to directly apply SAM for medical image segmentation and fully leverage the advantages of its large-scale natural image pre-training[19]. To adapt to the task of medical image segmentation, Ma et al. [20] proposed MedSAM, which enhances the model's segmentation performance in medical images through targeted adaptations. Subsequently, Cheng et al.[21] proposed SAM-Med2D for 2D medical images, which achieves improved results in medical image segmentation tasks by fine-tuning the encoder and decoder of the original SAM. Nevertheless, SAM-Med2D shows some limitations in segmenting medical images with complex morphology, small size, or low contrast. Hu et al. [22] proposed BreastSAM to enhance the segmentation of breast tumor ultrasound images by fine-tuning the encoder of SAM. However, the fine-tuning operations are all based on the ViT encoder of SAM. Although ViT is highly effective in processing global information, it lacks a clear mechanism to restrict the attention scope at each location, making it difficult to capture subtle lesion cues and local features in medical images through simple fine-tuning alone and often leading to the neglect of crucial local multi-scale information.

In response to the above issues, we propose a medical image segmentation model based on SAM with an integrated local multi-scale feature encoder (LMSFE-SAM), which can effectively incorporate medical image information into SAM. To achieve the above objective, a local multi-scale feature encoder is designed to operate in parallel with the ViT encoder. This encoder integrates a multiaxial Hadamard product module (MHPM) with linear complexity, which enables it to accurately capture subtle localized visual features in medical images. Additionally, to reduce computational complexity, we employ a strategy of downsizing the input image sequence, which effectively shortens the sequence length processed by the ViT encoder. To further balance the local and global information between the local multi-scale feature encoder and the ViT encoder, we also propose a cross-branch balancing adapter. Finally, to optimize the network model's ability to understand medical images, we construct a multidimensional information adaptation component to fine-tune the ViT encoder. This allows the distilled medical image information to be more effectively integrated into SAM, thereby improving the performance and accuracy of medical image segmentation.

1 Methods

In this work, we propose the LMSFE-SAM, a medical image segmentation model based on SAM with an integrated local multi-scale feature encoder. This model is developed based on the principles of SAM. The model primarily comprises a local multiscale feature encoder, a cross-branch balancing adapter, a prompt encoder, and a mask decoder, with a multidimensional information adaptation component integrated into the ViT encoder. To begin with, in the encoding stage, the local multi-scale encoder integrated with MHPM and

the ViT encoder simultaneously extracts image features. Then, to balance the local and global information obtained by both encoders, we introduce a cross-branch balancing adapter, which provides the ViT encoder with richer image feature information. After that, to reduce the input size and avoid overlapping patch embedding, the size of input image is reduced from 1 024×1 024 pixels to 256×256 pixels and fed to the encoder to extract image features. Subsequently, the ViT encoder is fine-tuned to fit smaller-sized medical images utilizing a multidimensional information adaptation component consisting of a feature adapter, a position adapter, and a channel-spatial adapter. In the final stage, the mask decoder predicts the segmentation mask by fusing the image embeddings and prompt embeddings generated by the two encoders. The mask decoder consists of two Transformer layers and a dynamic mask prediction head. Fig.1 illustrates the overall architecture.

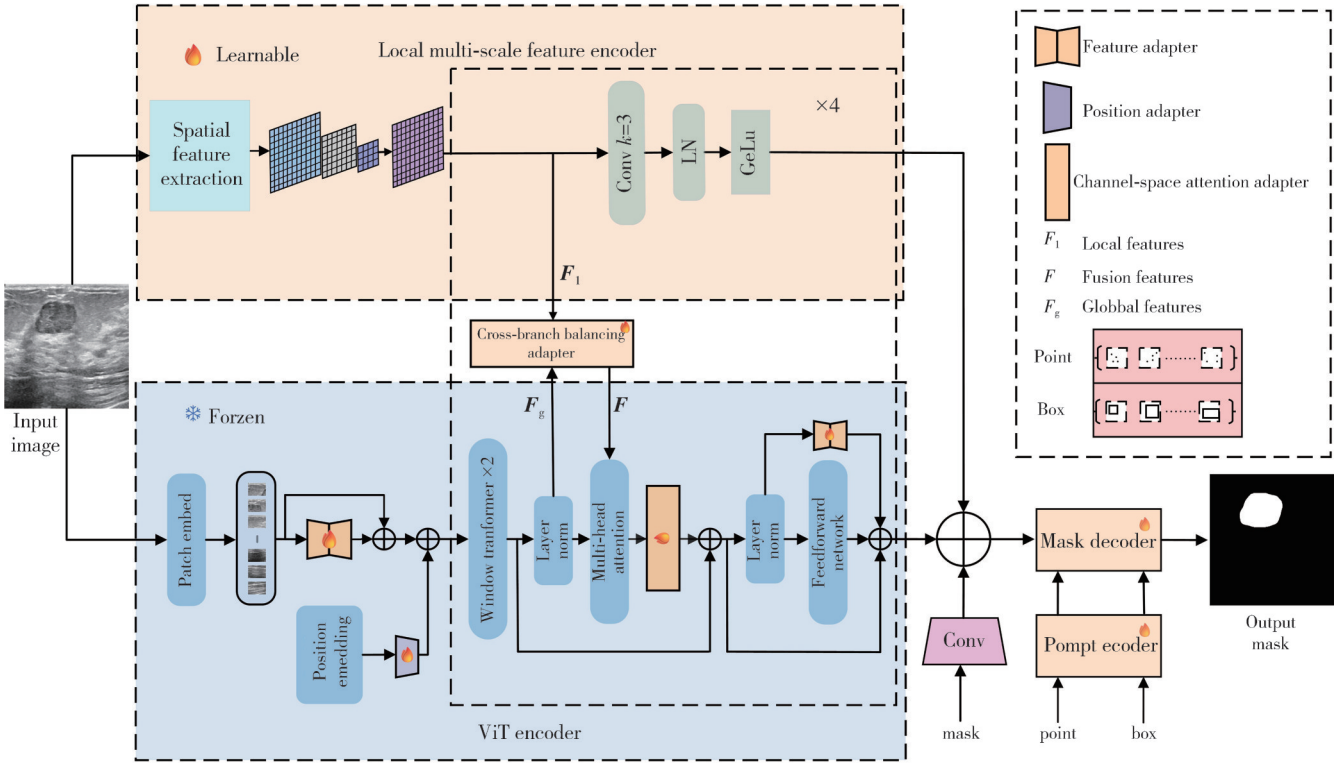

Fig. 1 Overview of model framework

1.1 Local multi-scale feature encoder

Since medical image lesion tissues have complex and diverse local features such as edges and textures, the ViT encoder cannot extract these local features effectively. We design a local multi-scale feature encoder, which is based on the MHPM and works in parallel with the ViT encoder, to complement rich local multi-scale feature information and improve segmentation performance. The input image X first undergoes initial spatial feature extraction via a CNN backbone network. This is followed by down-sampling of the features using three serially connected convolutional kernels and a batch norm layer, enabling the extraction of richer spatial information. The multi-scale max-pooling method then processes the down-sampled features, which includes three parallel MHPM modules along with max-pooling operations with kernel sizes of 5×5, 9×9, and 13×13, respectively. This setup extracts multi-scale features while preserving the diversity of local spatial features. After multi-scale pooling, the features are concatenated and then down-sampled to further refine the image feature representation. Before the final convolution operation, the initially extracted features merge with the current features to enhance the overall feature representation. Eqs. (1) — (3) outline the detailed process as

$$F(\boldsymbol{x})=f_1(f_3(f_1(Cat(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3)))), \tag{1}$$

$$f_{spp}(\boldsymbol{x})=Cat(GP(MP_{5\times5}(\boldsymbol{x}), GP(MP_{9\times9}(\boldsymbol{x}), GP(MP_{13\times13}(\boldsymbol{x})))), \tag{2}$$

$$\boldsymbol{F}_1=f_1(Cat(f_3(f_1(Cat(F(\boldsymbol{x}),f_{spp}(\boldsymbol{x})))),f_3(\boldsymbol{x}))), \tag{3}$$

where $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$ are the feature vectors extracted by the CNN backbone network; $Cat(\cdot)$ denotes the concatenation operation along the channel dimensions; $GP(\cdot)$ denotes the MHPM operation; $MP_{k\times k}(\cdot)$ denotes the max pooling with $k\times k$ kernel size; $f_{spp}(\cdot)$ denotes the

spatial pyramid pooling operation; $f_1(\cdot)$ and $f_3(\cdot)$ denote the convolutional layers with kernel sizes of $k=1$ and $k=3$, respectively; and $F_l \in \mathbf{R}^{H\times W\times C}$ denotes the output features of the local branches.

Fig.2 illustrates the proposed local multi-scale feature encoder.

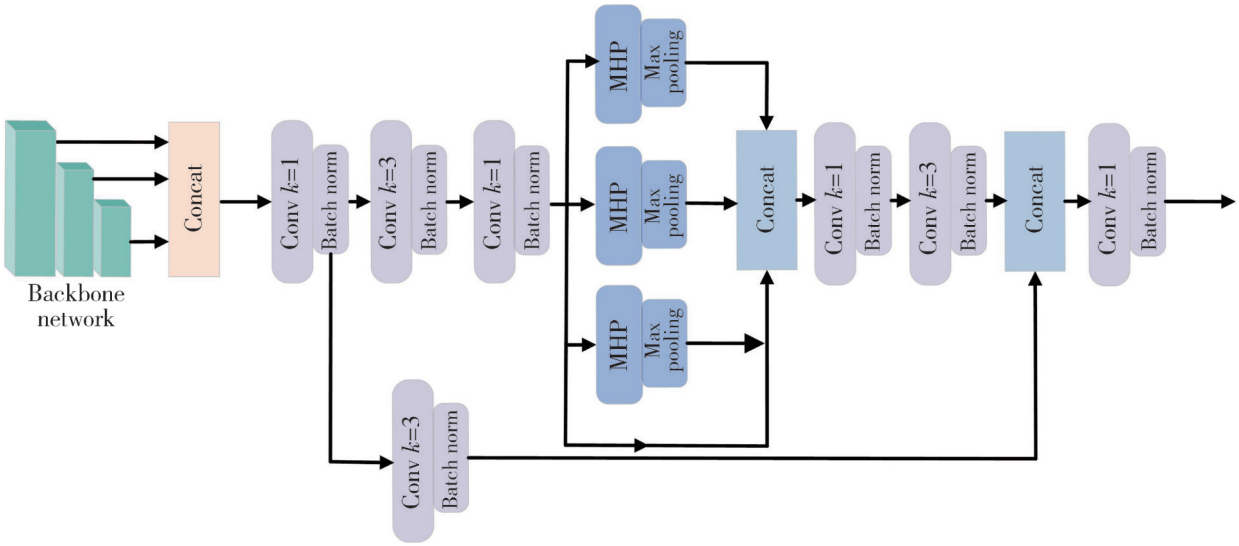

Fig. 2 Structure schematic of local multi-scale feature encoder

1.1.1 MHPM

To address the quadratic complexity problem associated with multi-head self-attention, we put forward the MHPM, which is devised for the local multi-scale feature encoder. The deep fusion of features is achieved through a fine-grouping strategy, facilitating the extraction of more abundant and representative feature embeddings. Algorithm 1 presents the structural particulars.

Algorithm 1: Grouped MHPM

Input: $\boldsymbol{x}$, the feature map with shape $[B,H,W,C]$, dim_{in}, dim_{out}
Output: $\boldsymbol{y}$, the feature map with shape $[B,H,H,W]$
1: Initialization:
hyperparameters a and b to 8.
Tensors $\boldsymbol{P}_{xy}$ with shape $[1,C//4,a,b]$
Tensors $\boldsymbol{P}_{zx}$ with shape $[1,1,C//4,a]$
Tensors $\boldsymbol{P}_{zy}$ with shape $[1,1,C//4,b]$
2: $\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4 = torch.chunk(LN(\boldsymbol{x}),4,dim=1)$
3: $\boldsymbol{x}_1 = \boldsymbol{x}_1 * DW(BI(\boldsymbol{P}_{xy}))$
4: $\boldsymbol{x}_4 = DW(\boldsymbol{x}_4)$
5: $\boldsymbol{x}_2 = (\boldsymbol{x}_2 \cdot permute(0,3,1,2) * DW(BI(\boldsymbol{P}_{zx}))) \cdot permute(0,2,3,1)$
6: $\boldsymbol{x}_3 = (\boldsymbol{x}_3 \cdot permute(0,2,1,3) * DW(BI(\boldsymbol{P}_{zy}))) \cdot permute(0,2,3,1)$
7: $y = DW(LN(torch.cat([\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4],dim=1)))$
Note: *torch.chunk*, splits operations;
DW, Depth-wise separable convolution;
LN, layer-norm;
$*$, convolution; //, exact division;
torch.cat, connection operations;
BI, bilinear interpolation.

The MHPM consists of layer-norm (LN), depth-wise separable convolution (DW), and bilinear interpolation (BI). Specifically, the input features are divided into four groups along the channel dimension. The first three groups undergo the Hadamard product (HP) operation along the height, width, and channel-height/channel-width axes, respectively. The final group processes the feature map exclusively using a DW operation. Finally, the outputs from all groups are concatenated along the channel dimension. The kernel size of all convolutional kernels used in the DW operation is set to 3. A detailed description of the HP procedure is given by

$$\boldsymbol{y} = \boldsymbol{x} \otimes ((BI(\boldsymbol{P}, \boldsymbol{x}_{\text{shape}}) * K_{\text{DW}}) K_{\text{PW}}), \tag{4}$$

where $*$ denotes the convolution operation; K_{DW} and K_{PW} are the convolution kernels for depth-wise separable convolution and point-by-point convolution, respectively; $BI(\cdot,\cdot)$ is the bilinear interpolation operation; $\otimes$ denotes the Hadamard product; and $\boldsymbol{x}_{\text{shape}}$ denotes the shape of the input features.

1.2 Cross-branch balancing adapter

In the encoding stage, the local multi-scale feature encoder and ViT encoder provide local and global information, respectively. However, simple fusion fails to adequately capture the overall features of the image. Therefore, we introduce a cross-branch balancing adapter to harmonize local and global information, enhancing the representation of image features. Let $F_l \in \mathbf{R}^{H\times W\times C}$ and $F_g \in \mathbf{R}^{H\times W\times C}$ be the feature representations of the local multi-scale feature encoder and ViT encoder, respectively. As illustrated in Fig.3, the cross-branch balancing adapter balances the local feature information extracted by the local multiscale feature encoder with the global feature information provided by ViT encoder. First, we perform summation operations on F_l and F_g, followed by executing max pooling and average pooling along the channel

dimensions to obtain the enhanced local features $\boldsymbol{H}_{\max}$ and global features $\boldsymbol{H}_{\text{mean}}$, which can be respectively expressed as

$$\boldsymbol{H}_{\max}=MP(\boldsymbol{F}_{\text{g}}+\boldsymbol{F}_{\text{l}}), \tag{5}$$

$$\boldsymbol{H}_{\text{mean}}=AP(\boldsymbol{F}_{\text{g}}+\boldsymbol{F}_{\text{l}}). \tag{6}$$

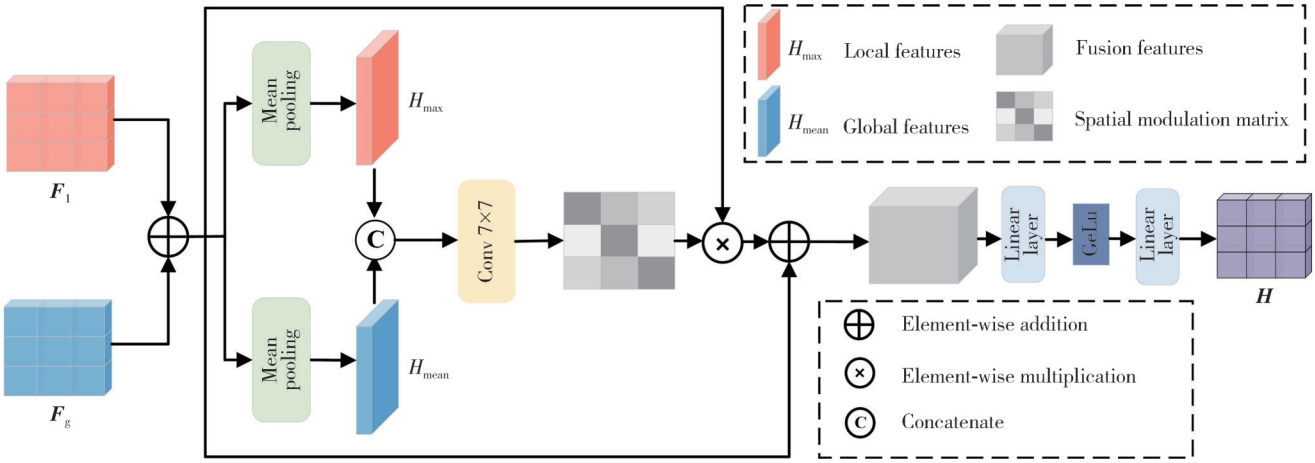

Fig. 3 Cross-branch balancing adapter structure schematic

Next, the two enhanced features are concatenated and processed through a convolutional layer with a kernel size of 7, generating a spatial modulation matrix $\boldsymbol{W}$. Using the modulation weights $\boldsymbol{W}$, the fused local and global features are modulated. Finally, two linear layers followed by a GeLu activation function generate the balanced overall image features $\boldsymbol{H}$, providing richer feature information to the ViT encoder. A detailed description of the complete process is

$$\boldsymbol{W}=\sigma(f(Cat(\boldsymbol{H}_{\max},\boldsymbol{H}_{\text{mean}})), \tag{7}$$

$$\boldsymbol{H}=\partial h_2(G(h_1(\boldsymbol{W}\odot(\boldsymbol{F}_{\text{g}}+\boldsymbol{F}_{\text{l}})))), \tag{8}$$

where $f(\cdot)$ denotes a convolutional layer with $k=7$; σ denotes a sigmoid activation function; $\odot$ denotes element-by-element multiplication; $h_1(\cdot)$ denotes a linear layer with decreasing number of features; $G(\cdot)$ denotes a GeLu activation function; $h_2(\cdot)$ denotes a linear layer with increasing number of features; and ∂ denotes a scaling factor, which is used to adjust for the effects of the cross-branch balancing adapter.

1.3 Multidimensional information adaptation component

To improve the adaptability of the image encoder in SAM for small medical images, while efficiently incorporating medical domain information into the ViT encoder, we introduce a multidimensional information adaptation component. This component comprises a position adapter, a feature adapter, and a channel-spatial adapter. These adapters can fine-tune the image encoder along the channel and spatial dimensions, allowing for efficient adjustment of the ViT encoder with minimal parameters. This enables efficient feature representation and domain adaptation, ultimately enhancing the model's performance in medical image segmentation.

1.3.1 Position adapter

The position adapter's role is to modify the positional embedding in the ViT encoder to accommodate varying input image sizes, thereby improving the model's performance. Specifically, the max pooling with a stride of $s=2$ and a kernel size of $k=2$ obtains the same resolution as the embedded sequence. A convolutional layer with a kernel size of $k=3$ then adjusts the positional embedding to help the ViT encoder handle smaller inputs. This is succeeded by two-dimensional layer normalization and the GeLu activation function to further enhance the model's adaptability. The procedure for normalizing the 2D layer is expressed as

$$y=\gamma\frac{x-\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}x_{:,:,i,j}}{\sqrt{\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(\boldsymbol{x}_{:,:,i,j}-\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{x}_{:,:,i,j}\right)^2+\varepsilon}}+\beta, \tag{9}$$

where H and W denote the tensor height and width, respectively; $\boldsymbol{x}$ and $\boldsymbol{x}_{:,:,i,j}$ represent the input feature tensor with shape (N,C,H,W), respectively; ε is a very small value used to prevent divide-by-zero errors; γ and β are learnable parameters used for scaling and panning, respectively; $\sum_{i=1}^{H}\sum_{j=1}^{H}$ represents a double summation, among which the outer summation $\sum_{i=1}^{H}$ sums over the height dimension H, and the inner summation $\sum_{j=1}^{H}$ sums

over the width dimension W.

1.3.2 Feature adapter

The feature adapter performs four operations: downward linear projection, activation function, upward linear projection, and a skip connection. Eq. (10) outlines the process of the feature adapter as

$$\boldsymbol{y}=\begin{cases}\boldsymbol{x}+\boldsymbol{W}_2\cdot act(\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{b}_1)+\boldsymbol{b}_2, & \text{skip\_connect is true},\\ \boldsymbol{W}_2\cdot act(\boldsymbol{W}_1\cdot+\boldsymbol{b}_1)+\boldsymbol{b}_2, & \text{skip\_connect is false},\end{cases} \tag{10}$$

where $\boldsymbol{x}$ is the input feature; $\boldsymbol{W}_1$ and $\boldsymbol{b}_1$ are the weight matrix and bias of the first layer, respectively; $\boldsymbol{W}_2$ and $\boldsymbol{b}_2$ are the weight matrix and bias of the second layer; and $act(\cdot)$ is the activation function (GeLu by default).

1.3.3 Channel-spatial adapter

The channel-spatial adapter primarily consists of adaptive average pooling, a channel adapter, a spatial adapter, and a skip connection. Adaptive average pooling processes the input features into a shape of $(B,C,1,1)$, preparing them for subsequent further processing. Subsequently, in the channel dimension, two linear layers, combined with ReLU and Sigmoid activations, generate new channel features. Fig.4 illustrates the structure of the channel-spatial adapter.

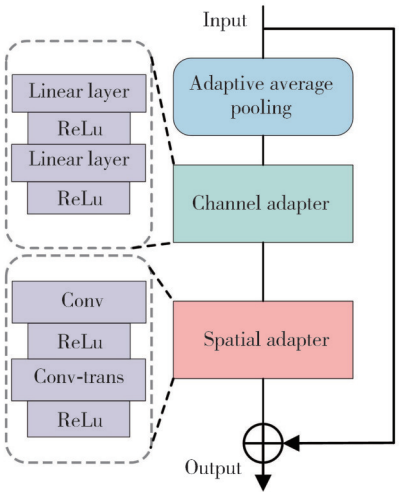

Fig. 4 Structure diagram of channel-spatial adapter

In the spatial dimension, convolutional layers with a kernel size of $k=3$ process the channel-adapted features, followed by transposed convolutional layers with a kernel size of $k=4$. Each convolutional and transposed convolutional layer applies a ReLU activation function. Skip connections follow each adapter layer to maintain the integrity of the original features and enhance the model's expressive capacity.

2 Experimental results and analysis

2.1 Datasets

In this study, the BUSI[23], DDTI[24], TN3K[25], and TG3K[25] datasets, comprising breast and thyroid ultrasound images, were utilized for the experimental evaluation. Table 1 presents detailed dataset information.

Table 1 Overview of datasets

Dataset	Number of slices	Number of masks	Training set	Testing set	Category
BUSI[23]	647	647	518	129	Breast
DDTI[24]	637	637	—	637	Thyroid
TN3K[25]	3 494	3 494	2 795	699	Thyroid
TG3K[25]	3 585	3 585	3 585	—	Thyroid
Total	8 363	8 363	6 898	1 465	—

The BUSI dataset consists of 780 images with pixel-level breast cancer annotations, divided into three groups: benign, malignant, and normal, with the average size of the images of 500×500 pixels. This study selected a total of 647 cases, encompassing both benign and malignant conditions. The thyroid nodule dataset consists of DDTI, TN3K, and TG3K, with 637, 3 494, and 3 585 ultrasound images with pixel thyroid nodule annotations, respectively. Since DDTI, TG3K, and TN3K share the same segmentation target, TG3K and TN3K were used as the training set in this study, with a portion of TN3K reserved as the test set, while DDTI served as a test set to evaluate the model's generalization ability.

2.2 Experimental setup

The experiments were carried out on a Windows operating system utilizing Python 3.11 and the PyTorch 2.0.1 framework. The hardware setup included an Nvidia GeForce RTX 4060 GPU. The model was initialized using pre-trained weights from SAM-Med2D for the LMSFE-SAM network. During the training, the parameters of the ViT encoder were frozen, while only the parameters of the local multi-scale feature encoder, the multidimensional information adaptation component, the cross-branch balancing adapter, and the prompt encoder and mask decoder were updated. To adapt SAM from the natural domain to the medical domain and to reduce the memory cost of the GPU, the size of the input image was resized to 256×256 pixels. The network optimizer used was adaptive moment estimation (Adam), the initial learning rate was 1×10^{-4}, the training period was set to 15 epochs, and the batch size was set to 2.

2.3 Evaluation metrics

In our work, six widely used evaluation metrics were

employed: the Dice similarity coefficient, accuracy, specificity, sensitivity, intersection over union, and Hausdorff distance.

The Dice similarity coefficient, also referred to as the overlap index, quantifies the similarity between two sets of samples, and is mathematically defined as

$$Dice = \frac{2TP}{FP + FN + 2TP}. \tag{11}$$

Intersection over union, a metric that measures the similarity between two sets of samples, is mathematically defined as

$$IoU = \frac{TP}{FP + TP + FN}. \tag{12}$$

Accuracy, which measures the proportion of correct predictions made by the model, is mathematically defined as

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}. \tag{13}$$

Sensitivity, known as the true positive rate, measures the model's ability to correctly identify positive samples. It is mathematically defined as

$$Se = \frac{TP}{TP + FN}. \tag{14}$$

Specificity, which measures the model's ability to correctly identify negative samples, is mathematically defined as

$$Sp = \frac{TN}{TN + FP}. \tag{15}$$

Hausdorff distance, which measures the greatest distance between the farthest points in two sets, is mathematically defined as

$$HD(A,B) = \max\left\{\max_{a\in A}\left\{\min_{b\in B}\|a-b\|\right\}, \max_{b\in B}\left\{\min_{a\in A}\|b-a\|\right\}\right\}, \tag{16}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively; A and B represent two distinct sets of points; and $\|\cdot\|$ is the distance paradigm between point sets A and B.

2.4 Analysis of quantitative results

In this study, three medical image datasets, BUSI, DDTI, and TN3K, were selected and compared with eight state-of-the-art segmentation models, including the foundation models such as SAM[14], MedSAM[20], SAM-Med2D[21], EdgeSAM[26], and the U-architecture network models such as U-Net[7], UCTransNet[10], TransResUnet[27], and U-KAN[8].

Table 2 presents the experimental results for each model on the BUSI dataset, where the bold font indicats the optimal result. The data show that LMSFE-SAM outperforms the other models, with 0.892 7 for *Dice*, 0.811 9 for *Acc*, 0.895 9 for *Sp*, 0.995 7 for *Se*, 0.981 0 for *IoU*, and 11.758 0 mm for *HD*. Compared to the second-best model, SAM-Med2D, LMSFE-SAM shows improvements in the metrics by 0.008 9 for the *Dice*, 0.000 3 for *Acc*, 0.012 5 for *Sp*, 0.020 8 for *Se*, and 0.015 8 for *IoU*. Additionally, the HD is optimized by 2.629 9 mm owing to LMSFE-SAM's effective adaptation of local multi-scale feature encoders through the cross-branch balancing adapter. This design compensates for the ViT encoder's limitations in capturing local dependencies, enhancing overall segmentation performance.

Table 2 Quantitative comparison results on seeable BUSI dataset

Model	*Dice*	*IoU*	*Se*	*Sp*	*Acc*	*HD*/mm
U-Net[7]	0.613 6	0.578 2	0.737 6	0.987 1	0.975 9	24.463 2
UCTransNet[10]	0.720 8	0.608 7	0.821 7	0.933 9	0.928 4	44.992 6
TransResUnet[27]	0.801 2	0.710 9	0.853 4	0.975 4	0.966 2	19.840 8
U-KAN[8]	0.715 7	0.609 4	0.843 6	0.970 4	0.956 8	25.629 5
SAM[14]	0.424 6	0.328 1	0.674 5	0.705 1	0.756 9	78.736 4
EdgeSAM[26]	0.618 6	0.516 2	0.780 6	0.795 7	0.843 2	39.455 2
MedSAM[20]	0.838 7	0.739 9	0.869 1	0.989 3	0.958 5	20.455 1
SAM-Med2D[21]	0.883 8	0.796 1	0.875 1	0.983 2	0.980 7	14.387 9
LMSFE-SAM	**0.892 7**	**0.811 9**	**0.895 9**	**0.995 7**	**0.981 0**	**11.758 0**

Table 3 shows the evaluation results for each model on the TN3K dataset, where the bold font indicates the optimal result.

Table 3 Quantitative comparison results on seeable TN3K dataset

Model	*Dice*	*IoU*	*Se*	*Sp*	*Acc*	*HD*/mm
U-Net[7]	0.650 1	0.533 4	0.753 4	0.975	0.966 4	25.253 2
UCTransNet[10]	0.737 5	0.619 6	**0.813 2**	0.950 3	0.928 2	30.232 1
TransResUnet[27]	0.805 8	0.706 5	0.796 7	0.981	0.962 7	29.184 7
U-KAN[8]	0.762 8	0.659 2	0.773 3	0.973 0	0.950 3	19.422 2
SAM[14]	0.529 1	0.458 9	0.471 8	0.584 8	0.652 5	131.228 8
EdgeSAM[26]	0.696 7	0.562 1	0.668 6	0.979 6	0.950 4	48.764 9
MedSAM[20]	0.797 1	0.688 8	0.795 8	0.963	0.912 2	34.898
SAM-Med2D[21]	0.767 2	0.638 5	0.656 2	0.983 9	0.927 6	28.512 3
LMSFE-SAM	**0.831 5**	**0.722 7**	0.797 2	**0.989 7**	**0.986 7**	**17.346 8**

LMSFE-SAM achieves the best results across all five metrics (*Dice*, *IoU*, *Sp*, *Acc*, and *HD*). Specifically, LMSFE-SAM improves by 0.025 7 in *Dice*, 0.016 2 in *IoU*, 0.008 7 in *Sp*, and 0.002 4 in *Acc*, and optimizes the *HD* by 2.075 4 mm compared to the second-best model, TransResUnet. These results highlight LMSFE-SAM's superior performance in predicting lesion regions on the diverse TN3K dataset, demonstrating a closer alignment with the corresponding labels.

To assess the generalization ability of LMSFE-SAM

on new medical image datasets, the DDTI dataset was used for testing. The evaluation results for each model are presented in Table 4, where the bold font indicates the optimal result.

Table 4 Quantitative comparison results on seeable DDTI dataset

Model	*Dice*	*IoU*	*Se*	*Sp*	*Acc*	*HD*/mm
U-Net[7]	0.678 0	0.512 9	0.682 8	0.983 4	0.953 4	25.829 6
UCTransNet[10]	0.762 1	0.641 7	0.803 4	0.976 5	0.959 9	28.120 8
TransResUnet[27]	0.806 7	0.710 8	0.820 7	0.985 3	**0.971 3**	23.366 2
U-KAN[8]	0.746 4	0.623 3	0.794 2	0.974 3	0.953 3	30.809 8
SAM[12]	0.558 8	0.401 2	0.626 7	0.710 1	0.649 9	69.682 1
EdgeSAM[26]	0.472 5	0.379 2	0.845 1	0.725 5	0.726 8	79.767 2
MedSAM[20]	0.741 8	0.592 4	0.780 9	0.975 3	0.919 3	27.577 2
SAM-Med2D[21]	0.819 2	0.704 2	0.757 4	**0.992 3**	0.960 4	**17.725 9**
LMSFE-SAM	**0.840 8**	**0.761 0**	**0.861 9**	0.984 0	0.967 5	18.914 4

The experimental results show that LMSFE-SAM achieves the highest performance in *Dice*, *IoU*, and *Se* metrics. Compared to the second-best model, SAM-Med2D, LMSFE-SAM improves these metrics by 0.021 6 for *Dice*, 0.056 8 for *IoU*, and 0.104 5 for *Se*. However, due to the absence of clearly defined nodule boundaries in the DDTI dataset and the insufficient exposure of such complex images during the training, the model's performance is only slightly inferior to the optimal model in terms of *Sp*, *Acc*, and *HD* metrics. Overall, these results highlight the superior performance of the LMSFE-SAM model in medical image segmentation.

2.5 Analysis of qualitative results

Fig.5 illustrates the qualitative segmentation results of various models, including U-Net, UCTransNet, TransResUnet, U-KAN, and LMSFE-SAM. Specifically, the images in the first and fourth raws show the results from the breast ultrasound dataset, while the images in other rows display the results from the thyroid nodule dataset.

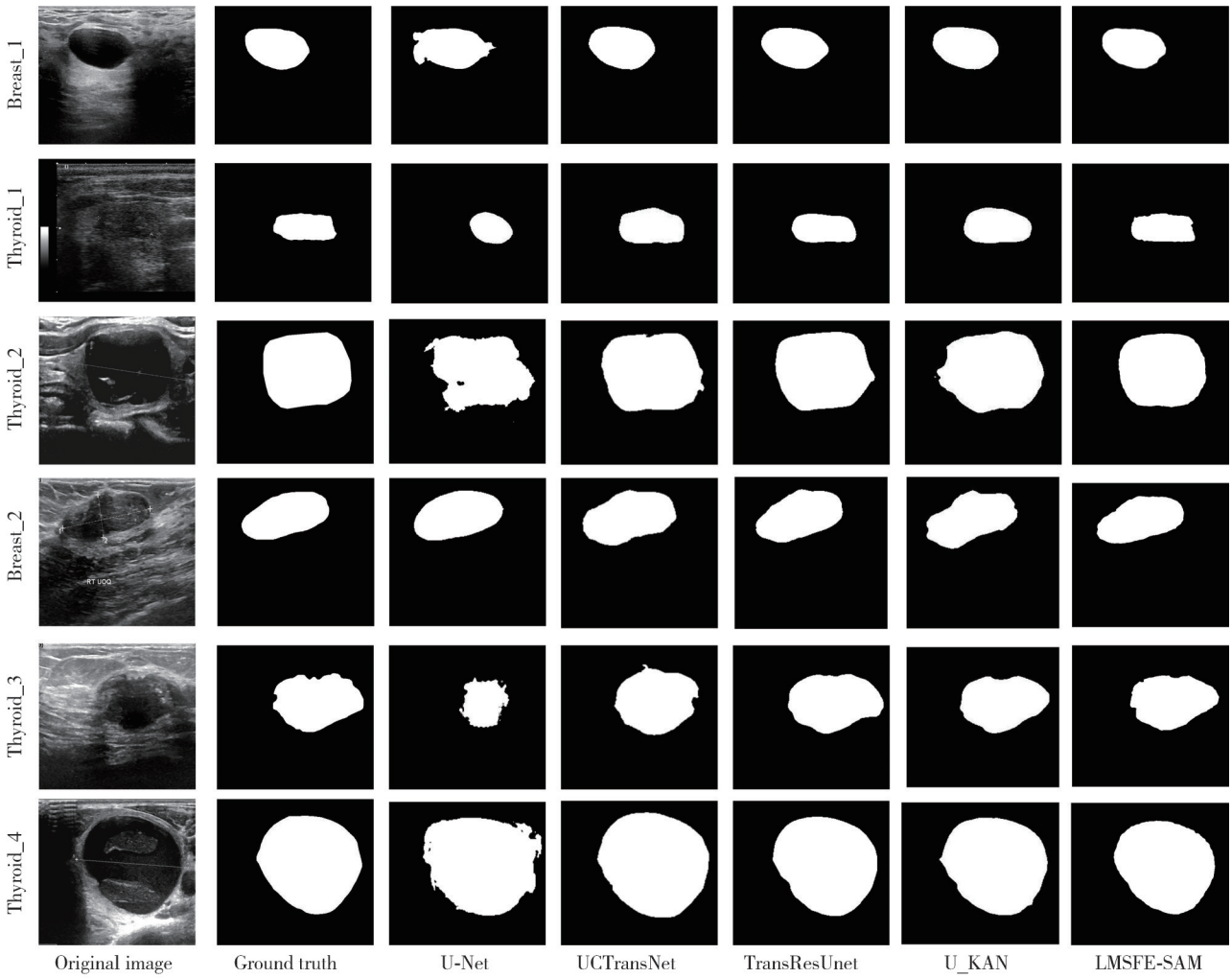

Fig. 5 Qualitative comparison between LMSFE-SAM and U-shaped architecture model

As shown in Fig. 5, the segmentation results of LMSFE-SAM outperforms that of the other networks. In the second rows, the contrast between the target region and the background in the original image segmentation was low. As a result, U-Net fails to fully segment the target region, and UCTransNet,

TransResUnet, and U-KAN struggle to accurately capture the boundary information. This limitation reduces their ability to effectively represent the nodal tissue. In the fourth row, U-Net, UCTransNet, and U-KAN face challenges in capturing boundary information of breast tumor images with complex boundaries. In the fifth row, U-Net and UCTransNet have difficulty in accurately segmenting images with complex boundary structures. Additionally, U-shaped network models often require retraining for specific datasets, which increases training costs. In contrast, LMSFE-SAM demonstrates significant advantages in maintaining the integrity of the target area and accurately identifying the lesion area. It also greatly reduces training costs, making it more practical for clinical applications.

Fig. 6 presents the qualitative segmentation results of LMSFE-SAM compared with the foundation models, including SAM, EdgeSAM, MedSAM, SAM-Med2D, and LMSFE-SAM. The images in the first and fourth rows are from the breast ultrasound dataset, while the images in other rows are from the thyroid nodule dataset.

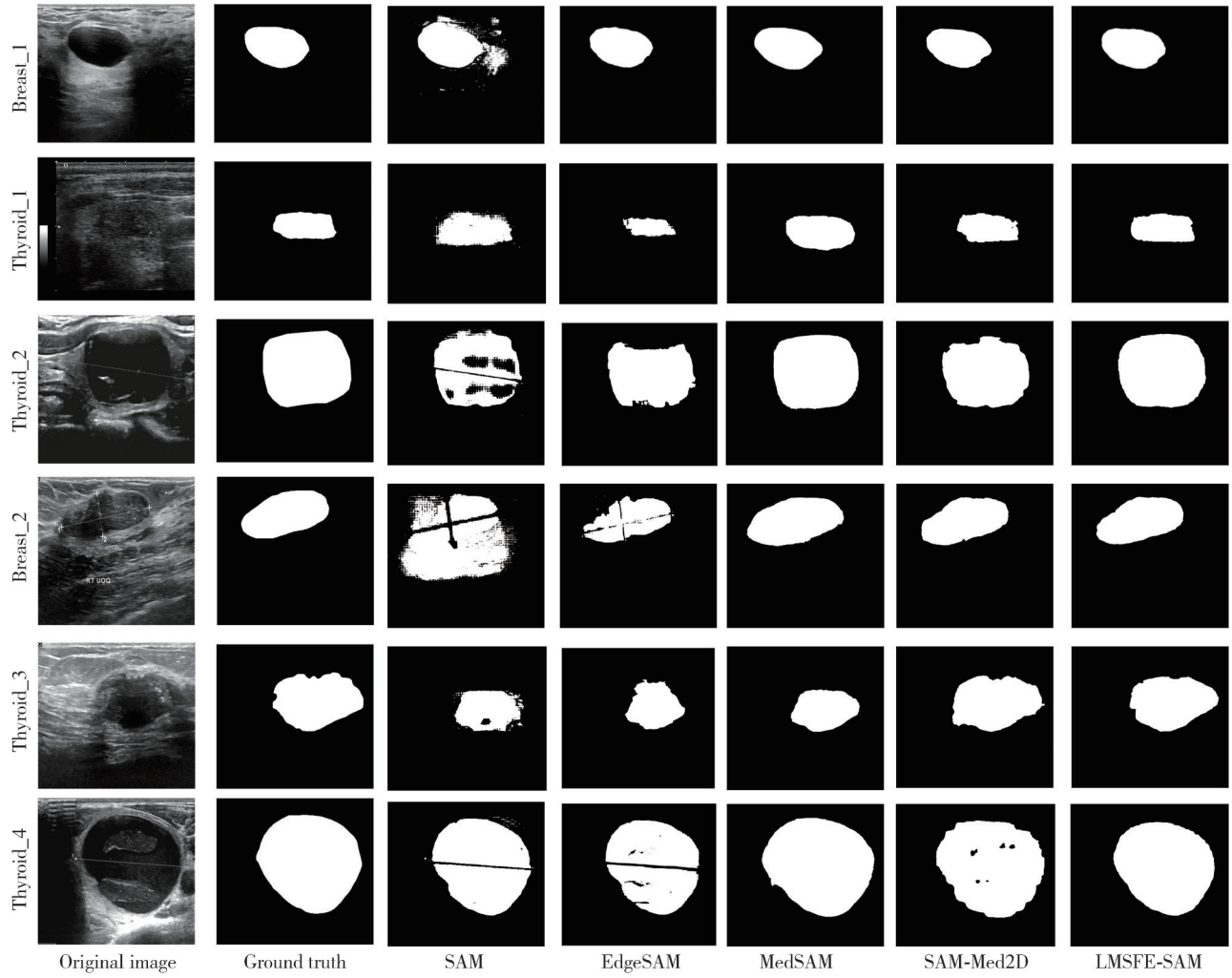

Fig. 6 Qualitative comparison between LMSFE-SAM and foundation mode

It can be seen that in the second, third and fourth rows, without the customized adaptation of SAM to medical image orientation, both SAM and EdgeSAM are unable to segment the complete lesion area. Fine-tuning SAM on medical image datasets through the introduction of adapters enhances its segmentation capability. As illustrated in the second row, both MedSAM and SAM-Med2D show significant improvements, demonstrating a better ability to capture the lesion areas in medical images. However, as shown in the third row, SAM-Med2D fails to accurately capture the boundary information of the thyroid nodule, while MedSAM struggls to extract complete lesion information. Additionally, in the sixth row, the target region segmented by SAM-Med2D exhibits noticeable noise and incomplete boundary regions. These shortcomings stem from the limited effectiveness of simple fine-tuning, which is insufficient for SAM to thoroughly learn and represent the intricate and detailed information present in medical images. In contrast, the model proposed in this study accurately locates and segments lesion regions,

even in noisy images with complex boundary structures, demonstrating excellent performance in medical image segmentation tasks.

3 Ablation experiments

Using two evaluation metrics, *Dice* and *IoU*, the effectiveness of different modules in LMSFE-SAM were illustrated through ablation experiments on three datasets, BUSI, DDTI, and TN3K. The specific results of the ablation experiments are presented in Table 5, where LM encoder refers to the local multi-scale feature encoder. ViTA denotes the multidimensional information adaptation component in the ViT encoder, and CBA refers to the cross-branch balancing adapter. Fig.7 further demonstrates the role of the CBA and MHPM modules through qualitative results.

As demonstrated by the experimental results in Table 5, incorporating ViTA into ablation experiment 1 shows significant improvements in *Dice* and *IoU* scores: 0.226 1 and 0.233 6 on BUSI, 0.064 0 and 0.021 1 on DDTI, and 0.169 6 and 0.146 0 on TN3K, respectively, compared to the baseline model. These findings indicate that ViTA effectively adapts the information in medical images to the SAM model, significantly enhancing its segmentation performance on medical images.

Table 5 Ablation study results of LMSFE-SAM

Method	Component				BUSI		DDTI		TN3K	
	LM encoder	MHPM	ViTA	CBA	*Dice*	*IoU*	*Dice*	*Iou*	*Dice*	*IoU*
Baseline	×	×	×	×	0.536 0	0.429 5	0.517 5	0.405 1	0.427 6	0.322 9
Ablation 1	×	×	√	×	0.762 1	0.663 1	0.581 5	0.426 2	0.597 2	0.468 9
Ablation 2	√	×	×	×	0.766 8	0.657 5	0.742 8	0.618 4	0.767 1	0.673 8
Ablation 3	√	×	×	√	0.782 8	0.675 9	0.749 2	0.624 7	0.788 7	0.676 7
Ablation 4	√	√	×	√	0.862 6	0.765 6	0.757 0	0.632 5	0.796 3	0.686 2
LMSFE-SAM	√	√	√	√	**0.893 0**	**0.812 3**	**0.840 8**	**0.761 0**	**0.831 5**	**0.722 7**

In ablation experiment 2, integrating the LM encoder into the baseline model results in *Dice* and *IoU* score improvements of 0.230 8 and 0.228 0 on the BUSI dataset, 0.225 3 and 0.213 3 on the DDTI dataset, and 0.339 5 and 0.350 9 on the TN3K datasets, respectively. These results demonstrate that the LM encoder significantly enhances the extraction of comprehensive and enriched features from medical images.

In ablation experiment 3, the LM encoder and CBA were introduced into the baseline model. Compared to the previous experiments, both *Dice* and *IoU* scores show further improvements, indicating that the CBA effectively integrates the outputs of the ViT encoder and the LM encoder, and achieving an optimal balance between local and global feature information. This highlightes the CBA's capability to enhance the fusion of these diverse feature types, leading to more accurate segmentation results.

In ablation experiment 4, the MHPM module was added to the LM encoder. The experimental results demonstrate that the MHPM module further enhances the representation of local information, thereby improving the segmentation performance of the model.

In the experimental results of the complete model LMSFE-SAM presented in this paper, the overall improvements in *Dice* and *IoU* are 0.357 0 and 0.382 8, 0.323 3 and 0.355 9, and 0.403 9 and 0.399 8, respectively, based on the baseline network, which verifies the effectiveness of the four modules designed in this study and greatly improves the segmentation accuracy of the network. The segmentation performance of the network reaches its highest level.

As shown in Fig. 7, without the CBA module or MHPM module, the local feature information and global feature information output from the LM encoder and ViT encoder will lose part of the boundary information during the fusion process.

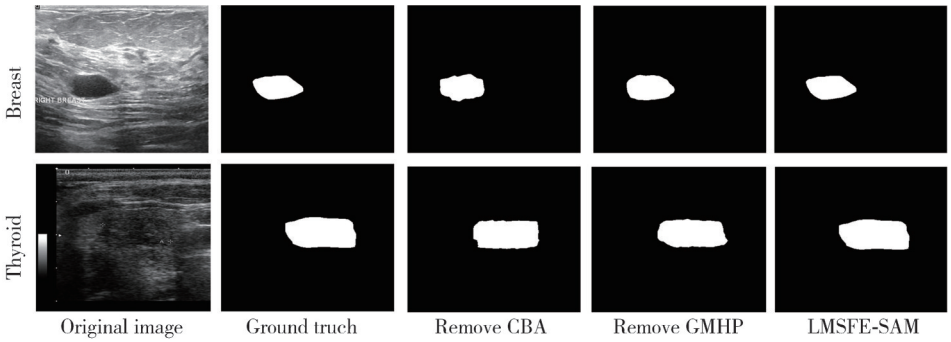

Fig. 7 Qualitative analysis of ablation experiments of CBA and GMHP modules

4 Conclusions

This paper presents a medical image segmentation model based on SAM with an integrated local multi-scale feature encoder, LMSFE-SAM. To this end, a local multi-scale feature encoder is designed to capture more local detail information, and based on it, the MHPM module is introduced to effectively reduce the quadratic complexity and noise interference. Additionally, a cross-branch balancing adapter is designed to balance the local and global dependencies between the local multi-scale feature encoder and the ViT encoder at different feature scales, thereby maximizing the preservation of both local and global information. To further enhance the adaptability of the model to medical images, we add a multidimensional information adaptation component to the ViT encoder, which successfully incorporates the medical image domain information into the SAM, and solves the problem of poor performance of the SAM in medical image segmentation tasks and the limitations of the ViT encoder in the extraction of local feature information. The experimental outcomes reveal that LMSFE-SAM outperforms the eight high-performing segmentation models, attaining enhancements of 0.379 6, 0.291 0, 0.323 1, 0.268 7, and 0.376 5 respectively in the average metrics of *Dice*, *Acc*, *Sp*, *Se*, and *IoU* across the BUSI, DDTI, and TN3K datasets. This highlights the model's superiority in segmentation refinement, particularly in handling fuzzy boundary detection and noise suppression. These significant performance enhancements demonstrate that LMSFE-SAM excels in capturing fine details and extracting local information. Nevertheless, the model demonstrates constraints in dealing with medical images featuring complex tissue structures and high background heterogeneity. Future endeavors will center on extending the model's application to a wider Spectrum of medical images and scenarios, with the aim of enhancing its versatility and robustness.

Acknowledgement

This work was supported by Natural Science Foundation Programme of Gansu Province (No. 24JRRA231), National Natural Science Foundation of China (No. 62061023), and Gansu Provincial Science and Technology Plan Key Research and Development Program Project (No. 24YFFA024).

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] LITJENS G, KOOI T, BEJNORDI B E, et al. A survey on deep learning in medical image analysis. Medical Image Analysis, 2017, 42: 60-88.

[2] ZHOU S K, GREENSPAN H, DAVATZIKOS C, et al. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE, 2021, 109(5): 820-838.

[3] WU S Y, ZHUO G P, ZHU J D, et al. Segmentation and classification algorithm of retinal arteriovenous vessel based on improved U-Net. Journal of North University of China (Natural Science Edition), 2023, 44(1): 79-85.

[4] CHAI R, LUO Y B, QIN P L, et al. Colon cancer gland segmentation network based on edge fusion and multi scale feature enhancement. Journal of North University of China (Natural Science Edition), 2025, 46(4): 411-421.

[5] FABIAN I, PAUL F J, KOHL S A A, et al. NNU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 2021, 18(2): 203-211.

[6] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation//2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 3431-3440.

[7] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation// Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. October 5-9, 2015, Munich, Germany. Cham: Springer International Publishing, 2015: 234-241.

[8] LI C X, LIU X Y, LI W Y, et al. U-KAN makes strong backbone for medical image segmentation and generation. 2024: 2406. 02918. https://arxiv. org/abs/2406. 02918v3.

[9] WU H S, CHEN S H, CHEN G L, et al. FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. Medical Image Analysis, 2022, 76: 102327.

[10] WANG H, CAO P, WANG J, et al. UCTransNet: rethinking the skip connections in U-net from a channel-wise perSpective with transformer//AAAI Conference on Artificial Intelligence, February 22 — March 1, 2022, Vancouver, Canada. New York: AAAI Press, 2022, 36(3): 2441-2449.

[11] YUAN F, ZHANG Z X, FANG Z J. An effective CNN and transformer complementary network for medical image segmentation. Pattern Recognition, 2023, 136: 109228.

[12] LI G, CHEN R Y, ZHANG J, et al. Fusing enhanced transformer and large kernel CNN for malignant thyroid

nodule segmentation. Biomedical Signal Processing and Control, 2023, 83: 104636.
[13] PANG C, LU X Q, LIU X, et al. IIAM: intra and inter attention with mutual consistency learning network for medical image segmentation. IEEE Journal of Biomedical and Health Informatics, 2024, 28(10): 5971-5983.
[14] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything//2023 IEEE/CVF International Conference on Computer Vision, October 1-6, 2023, Paris, France. New York: IEEE, 2023: 3992-4003.
[15] DOSOVITSKIY A. An image is worth 16x16 words: transformers for image recognition at scale. (2010-10-22) [2024-08-26].https://doi.org/10.48550/arXiv.2010.11929.
[16] HE S, BAO R, LI J, et al. Computer-vision benchmark segment-anything model (SAM) in medical images: accuracy in 12 datasets. (2023-04-18) [2024-09-01]. https://doi.org/10.48550/arXiv.2304.09324.
[17] XIONG Y Y, VARADARAJAN B, WU L M, et al. EfficientSAM: leveraged masked image pretraining for efficient segment anything//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 16-22, 2024, Seattle, WA, USA. New York: IEEE, 2024: 16111-16121.
[18] ZHANG Y C, SHEN Z R, JIAO R S. Segment anything model for medical image segmentation: current applications and future directions. Computers in Biology and Medicine, 2024, 171: 108238.
[19] HUANG Y H, YANG X, LIU L, et al. Segment anything model for medical images? Medical Image Analysis, 2024, 92: 103061.
[20] MA J, HE Y T, LI F F, et al. Segment anything in medical images. Nature Communications, 2024, 15(1): 654.
[21] CHENG J, YE J, DENG Z, et al. SAM-MED2D. (2023-08-30) [2024-04-10]. https://doi.org/10.48550/arXiv.2308.16184.
[22] HU M Z, LI Y H, YANG X F. BreastSAM: adapting the segmentation anything model for breast tumor segmentation in ultrasound imaging//Medical Imaging 2024: Ultrasonic Imaging and Tomography, February 18-22, 2024, San Diego, California, USA. Bellingham: SPIE, 2024, 12932: 182-196.
[23] AL-DHABYANI W, GOMAA M, KHALED H, et al. Dataset of breast ultrasound images. Data in Brief, 2020, 28: 104863.
[24] PEDRAZA L, VARGAS C, NARVÁEZ F, et al. An open access thyroid ultrasound image database//10th International Symposium on Medical Information Processing and Analysis, October 14-16, 2015, Cartagena de Indias, Colombia, USA. Bellingham: SPIE, 2015, 9287: 188-193.
[25] GONG H F, CHEN J X, CHEN G Q, et al. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. Computers in Biology and Medicine, 2023, 155: 106389.
[26] ZHOU C, LI X T, LOY C C, et al. EdgeSAM: prompt-In-the-loop distillation for on-device deployment of SAM. 2023: 2312.06660. https://arxiv.org/abs/2312.06660v2.
[27] REZA S, AMIN O B, HASHEM M M A. TransResUNet: improving U-Net architecture for robust lungs segmentation in chest X-rays//2020 IEEE Region 10 Symposium, June 5-7, 2020, Dhaka, Bangladesh. New York: IEEE, 2020: 1592-1595.

融合局部多尺度特征编码器的SAM医学图像分割模型

邸 敬*, 朱云龙, 梁 婵

兰州交通大学 电子与信息工程学院, 甘肃 兰州 730070

摘 要: 尽管分割一切模型(Segment anything model, SAM)在自然图像分割任务上表现出色,但它缺乏针对医学影像领域的特定专业知识,并且在编码阶段存在局部多尺度信息丢失的问题。为此,提出了一种融合局部多尺度特征编码器的SAM医学图像分割模型(ASM with a local multi-scale feature encoder, LMSFE-SAM)以处理上述问题。首先,在SAM的基础上,引入了一种局部多尺度特征编码器,提升对局部感受野特征的表达能力,为SAM中的Vision Transformer(ViT)分支提供图像的局部多尺度上下文信息。同时,以轻量级的方式,在局部多尺度特征编码器中加入了多轴Hadamard积模块(Multiaxial Hadamard product module, MHPM),减少了平方复杂度和噪声干扰。其次,设计了一种跨分支平衡适配器,用于平衡局部多尺度特征编码器与SAM中ViT图像编码器的局部和全局信息。最后,为减少输入图像尺寸大小并避免重叠补丁嵌入,将输入图像的尺寸从1 024×1 024像素缩小到256×256像素,并构建了包含特征适配器、位置适配器和通道-空间适配器的多维度信息适配组件,将小尺寸医学图像信息融入SAM中,更利于临床应用。与其他8种具有代表性的图像分割模型相比,所提出的模型在BUSI、DDTI和TN3K三个数据集上的6项客观评价指标平均提升了0.038 7~0.319 1,显著改善了SAM在医学图像中的表现,为临床医生提供了有力的辅助诊断工具。

关键词: 分割一切模型;医学图像分割;编码器;解码器;多轴Hadamard积模块;跨分支平衡适配器

引用格式: DI Jing, ZHU Yunlong, LIANG Chan. A medical image segmentation model based on SAM with an integrated local multi-scale feature encoder. Journal of Measurement Science and Instrumentation, 2025, 16(3): 359-370. DOI: 10.62756/jmsi.1674-8042.2025035