Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 16, No. 1, 2025

Yolo-MLSAM: SAM Based Breast Cancer Microcalcification Cluster-Segmentation Method

Hongguang Chen [1], Banteng Liu [2] and Ke Wang [2, *]

[1] School of Information Engineering, Huzhou university, Huzhou, China
[2] College of Information Science and Technology, Zhejiang Shuren University, Hangzhou, China
[*] **Corresponding author**: Ke Wang (Email: wangke1992@zju.edu.cn)

Abstract: Although the HQ-SAM model has achieved some results in improving the accuracy of fuzzy boundary segmentation, it is still difficult to achieve accurate segmentation in medical image processing, especially in the face of small targets such as breast cancer microcalcification clusters, in addition, high labor costs make Prompt operation cumbersome, in order to solve these problems. A novel segmentation method of breast cancer microcalcification cluster based on SAM was proposed. The method first uses Yolov8 neural network model to accurately obtain the lesion region, then uses MLSAM model to perform more detailed semantic segmentation of the lesion region, and finally realizes semi-automatic annotation function, greatly reducing the cost and complexity of manual participation. The experimental results show that compared with the HQ-SAM model, the new method has significantly improved the segmentation performance, and the dice similarity coefficient reaches 81.78%.

Keywords: Medical image; Breast cancer; Object detection; Semantic segmentation; Semiautomatic annotation.

1. Introduction

Breast cancer remains the most prevalent malignancy among women globally, with early detection being crucial for reducing its high mortality rate [1]. Despite advances in screening technologies and therapeutic approaches, the disease's high incidence and pathological complexity continue to pose significant clinical challenges [2]. Full-field digital mammography (FFDM), as a mainstream non-invasive screening modality, employs low-dose radiation to generate high-resolution breast tissue images, proving particularly effective for detecting early-stage lesions. The growing demand for breast cancer screening has propelled automated image analysis into a research focus for breast mass detection [3].

Traditional breast cancer segmentation methodologies exhibit notable limitations in clinical applications. These approaches demonstrate high sensitivity to image contrast variations, often failing to accurately identify lesions with subtle tissue differentiation [4]. Moreover, their performance significantly degrades when handling irregularly shaped lesions with indistinct boundaries, particularly for small early-stage tumors that lack distinctive morphological features [5]. To address these challenges, deep learning techniques have emerged as promising solutions in medical image segmentation. While convolutional neural networks (CNNs) have achieved remarkable progress in this domain, two critical limitations persist: 1) Progressive information loss during convolution and pooling operations compromises edge detection accuracy for blurred lesion boundaries; 2) Limited generalization capability across diverse clinical scenarios [6]. Recent advancements introduce the Segment Anything Model (SAM), which demonstrates exceptional zero-shot generalization capabilities and shows substantial potential for medical image segmentation applications [7].

2. Related work

2.1. Medical image segmentation based on deep learning

FFDM (Full-Field Digital Mammography) is a critical tool for breast cancer screening, and deep learning models have been widely applied to the automated segmentation of masses and microcalcifications in breast images. Models such as U-Net and Mask R-CNN have demonstrated excellent performance in breast image segmentation tasks, especially in images with ambiguous lesion boundaries and irregular shapes, significantly improving segmentation accuracy. Punn et al. [8] proposed the RCA-IUnet model, which integrates residual connections, cross-spatial attention mechanisms, and Inception modules to enhance the segmentation ability of tumors in breast ultrasound images. This architecture is particularly adept at handling complex and noisy medical images, significantly improving segmentation precision while maintaining model efficiency, with notable improvements in detail capture and feature extraction. Ning et al. [9] introduced an improved model named SMU-net, which incorporates saliency-guided and shape-awareness mechanisms. By leveraging the saliency information from significant regions in the image, the model enhances segmentation accuracy and improves the recognition of complex shapes and blurred boundaries, thereby better capturing lesion morphological features and exhibiting outstanding segmentation performance in complex and low-contrast medical images. Dar et al. [10] proposed the EfficientU-Net model, which optimizes the network architecture to improve segmentation efficiency and accuracy, particularly excelling in handling complex lesions in ultrasound images. This model combines lightweight design, reducing computational resources while significantly enhancing both segmentation and classification performance, thereby contributing to the automation of breast cancer diagnosis. Despite the significant accuracy improvements of U-Net in breast tumor and lesion segmentation, its

adaptability remains limited when dealing with complex shapes or blurred boundaries. In particular, segmentation performance suffers when handling lesions with irregular shapes. To address these limitations, some researchers have focused on using Mask R-CNN and its variants, incorporating multi-scale feature extraction and attention mechanisms to further improve segmentation accuracy and robustness, thus better handling complex segmentation tasks. Shen et al. [11] proposed an instance segmentation method that combines attention mechanisms with Mask R-CNN. By integrating attention mechanisms, this model efficiently captures important features, thereby enhancing object recognition and segmentation performance in complex scenarios. Li et al. [12] proposed the EMDFNet network, which enhances Mask R-CNN's detection performance through multi-scale feature extraction. This network provides more accurate detection and segmentation of multiple targets in complex scenarios, especially improving model precision and robustness when dealing with images containing multiple objects and intricate backgrounds.

2.2. Medical image segmentation based on deep learning

U-Net, Mask R-CNN, and other mainstream neural network models are typically optimized for specific datasets, but such adjustments often limit their generalization capability. When these models are applied to contexts outside the training data, their performance may deteriorate, leading to restricted applicability. Meta introduced SAM, a universal large model capable of one-click segmentation of arbitrary objects in photos or videos, which enhances the model's adaptability and practicality across various scenarios. Some researchers have tested SAM's foundational model on medical images to explore its potential applications in the medical field, aiming to validate its performance in handling complex medical images and assess its suitability and accuracy for medical image segmentation. Mazurowski et al. [13] investigated SAM's application in medical image analysis, experimentally evaluating its performance with medical image data, analyzing the model's applicability and limitations, and confirming SAM's potential in medical image segmentation. They also proposed optimization directions to improve its effectiveness in the healthcare domain. Zhang et al. [14] systematically assessed the performance of the SAM model on 12 public medical image datasets, covering multiple organs and different imaging modalities. The study found that SAM underperformed in handling medical images with weak boundaries and low contrast, but segmentation results could be significantly improved by adding manual prompts (e.g., bounding boxes). Zhang et al. [15] proposed SAMed, which was specifically optimized for medical image segmentation. It incorporated a low-rank approximation (LoRA) fine-tuning strategy, enabling SAMed to efficiently process medical images. Training strategies such as warmup and the AdamW optimizer significantly accelerated model convergence and improved segmentation accuracy, though boundary handling remained somewhat imprecise. Ke et al. [16] introduced HQ-SAM, which, while maintaining the original SAM model's zero-shot segmentation ability, significantly improved segmentation precision through minimal adjustments. By incorporating high-quality output tokens (HQ-Output Token) and a global-local feature fusion mechanism, HQ-SAM produced more refined segmentation boundaries and performed well with complex object structures. However, this model still relies on manual prompts, and its application in medical image segmentation has limitations.

To overcome the limitations of the HQ-SAM model, this paper proposes a new model called Yolo-MLSAM, which combines the improved Yolov8 neural network with the Multi-Level SAM (MLSAM) segmentation model. The main contributions are as follows:

1) Introduction of attention mechanisms: The CBAM attention mechanism is added to the Yolov8 neural network to enhance the recognition of breast cancer microcalcification cluster lesion areas.

2) Elimination of manual prompt dependence: By using the region boxes obtained from the Yolov8 neural network, the brightest points are extracted and used as point prompts for the MLSAM model input.

3)Multi-level feature fusion: Features from the shallow, mid, and deep layers of Vision Transformer (ViT) are combined to fuse different levels of feature information, capturing both low-level details and high-level semantic features simultaneously.

3. Yolo-MLSAM-based Segmentation Method for Breast Cancer Microcalcification Clusters

The novel segmentation method for breast cancer microcalcification clusters proposed in this study, based on Yolo-MLSAM, primarily consists of two components: the region proposal network based on Yolov8 and the image segmentation based on the MLSAM model. The overall framework of the method is illustrated in Figure 1.

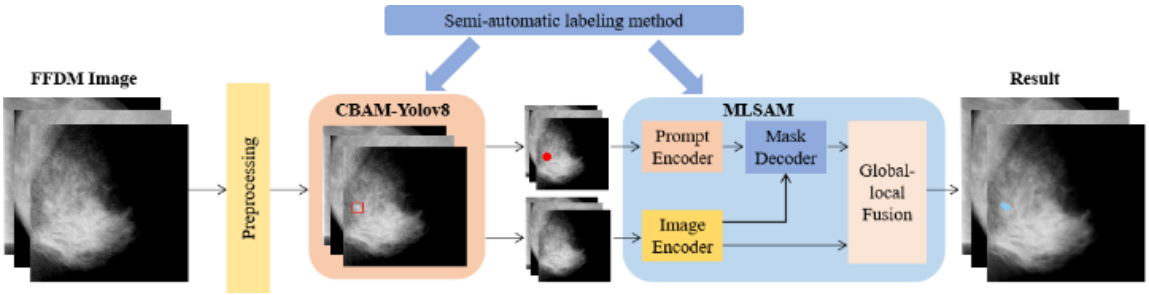

Fig. 1 Flowchart of Yolo-MLSAM breast cancer microcalcification cluster segmentation method

3.1. Regional candidate network based on CBAM-Yolov8

To reduce the manual annotation cost of constructing Prompt engineering, Yolov8 is utilized for medical diagnostic image processing to delineate specific lesion areas and extract the brightest point from each lesion region as input for the MLSAM model. This approach aims to address two main

issues: first, to accurately identify and locate lesion regions in medical images, thereby improving diagnostic accuracy; second, to extract the highest brightness point from each lesion region and use these points as input for the MLSAM model, thus eliminating the reliance on manual prompts and further enhancing overall segmentation efficiency.

To improve Yolov8's sensitivity to small target objects, the Mosaic data augmentation method [17,18] is introduced. The FFDM suspicious region localization algorithm based on Yolov8 utilizes FFDM data to construct the dataset and performs lesion detection through the Yolov8 network. The Yolov8 network employs convolutional layers, pooling layers, etc., to extract data features, and improves the network's localization ability through the construction of CIOU loss. The specific implementation process is as follows:

Construct the FFDM dataset. Let the total number of patients be N, with each patient containing four FFDM images taken at different positions. The sample image dataset and label dataset are constructed with individual patients as the smallest unit.

$$I = (m_1, m_2, m_3, \cdots, m_N) \tag{1}$$

$$L = (l_1, l_2, l_3, \cdots, l_N) \tag{2}$$

Here, m represents the image data, and l denotes the corresponding label data.

$$L_{SIoU} = 1 - IoU + \alpha \cdot d^2 + \beta \cdot v \tag{3}$$

Here, IoU represents the intersection over union (IoU) between the predicted box and the ground truth box, α controls the impact of the distance loss on the overall loss, d denotes the normalized distance between the boundary box centers, β is the angle weight coefficient, and v represents the angular loss.

Inserting CBAM after two feature extraction layers can enhance the feature representation ability in the early stages of the network. By adding CBAM at these positions, the network can more effectively focus on and strengthen key features, while filtering out irrelevant information. This helps improve the model's precision in target extraction, especially when dealing with complex backgrounds or multi-scale targets. The specific process is illustrated in Figure 2.

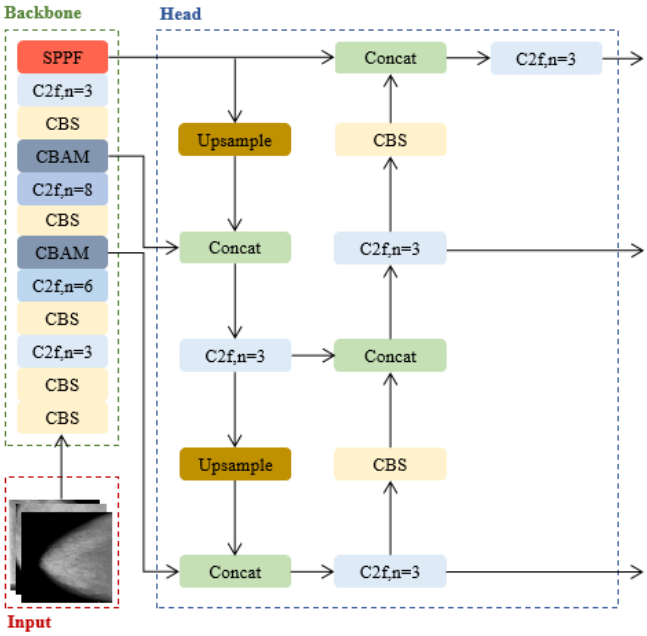

Fig. 2 CBAM-Yolov8 network structure

CBAM consists of two modules: the channel attention module and the spatial attention module [19]. The channel attention module assigns different weights to the feature maps of each channel by focusing on the inter-channel attention regions, enhancing the relevant features while suppressing the irrelevant ones. The spatial attention module, on the other hand, computes spatial attention weights by focusing on the important regions within a single-channel image, emphasizing critical spatial locations. By integrating these two modules, CBAM significantly improves the model's ability to perceive microcalcification clusters in terms of both channel and spatial positions. The detailed description of CBAM is as follows:

1) Channel Attention: The input feature map is F ∈ RC×H×W, where C represents the number of channels, and H and W are the height and width, respectively. First, global average pooling and global maximum pooling are performed. The results of both pooling operations are then fed into a shared Multi-Layer Perceptron (MLP). After applying the ReLU activation function followed by the Sigmoid activation function, the attention weights for each channel are obtained. The specific formula is as follows:

$$F1_{avg} = Avgpool(F) \tag{4}$$

$$F1_{\max} = Maxpool(F) \tag{5}$$

$$W_C(F) = \sigma(MLP(F1_{avg}) + MLP(F1_{\max})) \tag{6}$$

Here, F1avg and F1max represent the results of the max pooling and average pooling operations, respectively, while σ denotes the activation function.

2) Spatial Attention: First, global max pooling and average pooling are applied along the channel dimension of the input feature map. The results of these two pooling operations are then concatenated along the channel axis. A convolution operation is subsequently performed, followed by an activation function to generate the spatial attention weights. Finally, the spatial attention weights are multiplied by the input feature map, enhancing or suppressing the spatial locations within the feature map.

$$F2_{avg} = Avgpool(F) \tag{7}$$

$$F2_{\max} = Maxpool(F) \tag{8}$$

$$M_S(F) = \sigma\left(f^{7\times7}\left(F2_{avg}; F2_{\max}\right)\right) \tag{9}$$

Here, F2avg and F2max represent the results of the average pooling and max pooling operations, respectively, while f7×7 denotes a 7×7 convolution.

3) CBAM Attention Feature Map: The channel attention feature map is multiplied by the spatial attention map, refining the feature representation. This allows the model to focus on the most informative features in both dimensions, thereby enhancing the ability of Yolov8 to detect microcalcification clusters in breast cancer. The final output of CBAM is:

$$F_C = W_C(F)\cdot F \tag{10}$$

$$F_{out} = M_S(F_C)\cdot F_C \tag{11}$$

To extract the brightest point in each lesion region, we first initialize the maximum brightness value Imax=−∞ and its corresponding coordinates (xmax,ymax). Then, we iterate through all the pixels in the region, and whenever a pixel value I(x,y)>Imax is found, we update Imax and the coordinates (xmax,ymax). The point with the highest brightness, (xmax,ymax), will be extracted and used as the key point for input into MLSAM.

3.2. Image segmentation network based on MLSAM model

Although the HQ-SAM (High-Quality Segment Anything Model) is capable of handling complex samples in medical image segmentation, it still faces challenges such as boundary errors and missed detections in certain cases. To address this issue, this paper proposes an improved segmentation algorithm, MLSAM. This algorithm enhances the model's ability to capture local details and integrate global information by introducing multi-scale fusion of shallow, middle, and deep features, thereby improving segmentation accuracy. As shown in Figure 3.

The key improvement in MLSAM lies in the utilization of the multi-layer outputs from the ViT model, combined with the results of the mask decoder for global feature fusion [20]. Specifically, the model extracts multi-scale features from different network layers: shallow features from the 6th output block of the ViT, middle features from the 12th output block, and deep features from the 24th output block. These features play distinct roles at different levels: shallow features primarily preserve fine-grained local information, such as edges and textures; middle features capture local details while beginning to integrate global contextual information; and deep features mainly represent global semantic information and macro-structural patterns [21].

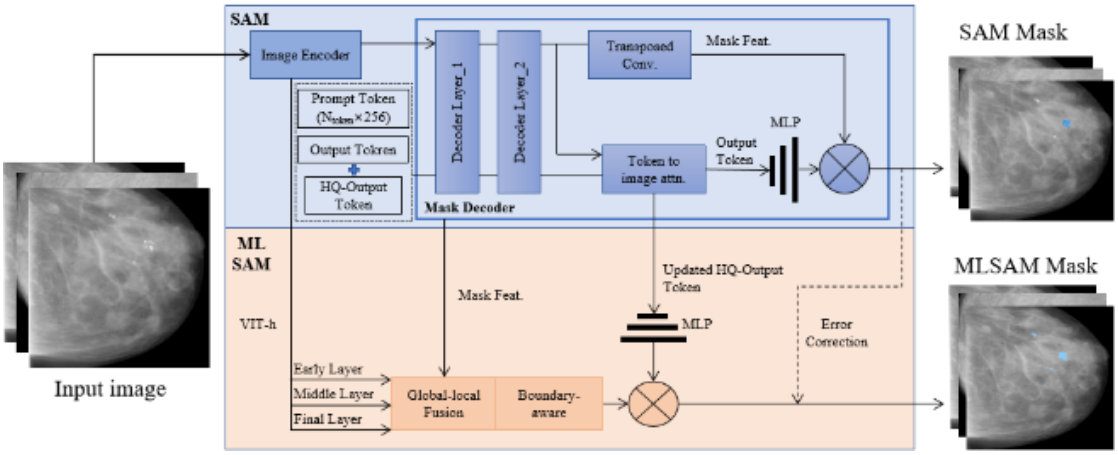

Fig. 3 Segmentation flow chart based on MLSAM model

Moreover, MLSAM further strengthens the capture and processing of middle-level features through the multi-layer self-attention mechanism in the ViT framework. Compared to low-level features, middle-level features are better suited for segmenting complex shapes, irregular boundaries, or blurred targets, improving segmentation accuracy. The multi-scale feature extraction and fusion strategy not only enhances the precision of the segmentation task but also strengthens the model's ability to understand objects at different scales.

To effectively fuse these multi-scale features, shallow, middle, and deep features are first upsampled to a spatial size of 256×256 through transposed convolution to ensure alignment in spatial dimensions. Then, these multi-scale features are globally fused with the output mask from the mask decoder. The mask decoder generates high-resolution masks through the image encoder, providing spatial information for the target regions, further enhancing segmentation precision.

Ultimately, through the global fusion involving shallow, middle, and deep features, along with the mask generated by the mask decoder, MLSAM generates high-quality features. These features combine rich local details and global semantic information, significantly improving the model's segmentation accuracy and robustness.

3.3. Semi-automatic annotation based on Yolo-MLSAM model

To improve the efficiency of data annotation and model training in medical image segmentation tasks, and to reduce the high human and time costs associated with traditional data annotation workflows [22, 23], we introduce a semi-automatic annotation method based on Yolo-MLSAM. This method combines annotated and unannotated data to enhance model performance, improve overall research efficiency, and reduce annotation costs. The entire semi-automatic annotation process can be divided into three stages:

1)Assisted Manual Stage: The Yolo-MLSAM model is used to perform initial analysis and processing of FFDM images, detecting the breast cancer microcalcification clusters and generating preliminary annotation masks. The segmentation results are typically presented as masks, where a pixel value of "1" indicates pixels that need to be processed, and "0" represents non-lesion background areas. By analyzing

these masks, boundary pixels with irregular shapes are identified, and their X-axis and Y-axis coordinates are extracted. The bounding box of the shape is determined using a max-min value updating mechanism to optimize the shape boundaries and improve mask accuracy. The method for expressing the coordinates of all non-zero pixels in the mask is as follows:

$$S = \{(x, y) | mask[y][x] = 1\} \tag{12}$$

Here, S represents a set containing a group of two-dimensional coordinates (x,y), where (x,y) denotes a two-dimensional coordinate point, and mask[y][x] represents the pixel value at the y row and x column of the mask.

2) Semi-Automatic Stage: To enhance the model's performance in FFDM image annotation, a pre-trained model is deployed to automate the initial annotation process. The results generated in this stage undergo review by medical experts, which not only helps correct potential misjudgments made by the model but also focuses on specific lesion areas that the model may have failed to accurately segment. This step significantly improves the accuracy and completeness of the annotations and strengthens the identification of lesion features. The data, after being carefully corrected by experts, is fed back into the model for subsequent training iterations, with the aim of enhancing the model's prediction capabilities and reducing reliance on expert intervention. Through this iterative optimization process, the precision and efficiency of semi-automatic annotation are significantly improved, ultimately enhancing the model's ability to autonomously learn breast cancer image features.

3) Fully Automatic Stage: In the final stage, Yolo-MLSAM is capable of independently generating masks for the images in the dataset without any manual input. Additionally, it can handle complex segmentation tasks, significantly reducing the cost of manual annotation.

4. Experimental results

4.1. Experimental data

The breast mammography X-ray segmentation dataset used in this study is from the First Affiliated Hospital of Zhejiang University. The dataset contains 4,000 two-dimensional FFDM images from 1,000 patients, each annotated with corresponding segmentation masks. Each image was reviewed by two different annotators in two rounds to ensure proper de-identification. Any discrepancies found were arbitrated by a third radiologist to resolve the issues [24].

4.2. Evaluation indicators

The purpose of this paper is to accurately segment the segmentation of the relevant medical image data, so the true positive rate (TPR), false positive rate (FPR), precision (Pr), F1 score, Dice index (Dice), Accuracy (ACC),Mean Average Precision @0.5(mAp@0.5),and mAP@0.5:0.95 are used as the evaluation indicators. The expression of each evaluation index is as follows:

$$\mathrm{TPR} = \frac{TP}{TP + FN} \tag{13}$$

$$\mathrm{FPR} = \frac{FP}{TN + FP} \tag{14}$$

$$\mathrm{Pr} = \frac{TP}{TP + FP} \tag{15}$$

$$F1 = 2 \times \frac{\mathrm{Pr} \times \mathrm{Re}\,call}{\mathrm{Pr} + \mathrm{Re}\,call} \tag{16}$$

$$\mathrm{Dice} = \frac{2TP}{2TP + FP + FN} \tag{17}$$

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

$$mAP@0.5 = \frac{1}{N} \sum_{i=1}^{N} AP_i(IoU \geq 0.5) \tag{19}$$

$$mAP@0.5:0.95 = \frac{1}{10} \sum_{IoU=0.5}^{0.95} mAP_{IoU} \tag{20}$$

Here, true-positive (TP) refers to the count of samples with correctly identified microcalcification clusters, while false-negative (FN) represents the count of samples where microcalcification clusters were incorrectly missed. Similarly, false-positive (FP) denotes the count of samples wrongly identified as having microcalcification clusters. The number of true-negatives (TN) signifies the samples that were correctly determined as lacking microcalcification clusters, C is the number of classes and APi is the average precision of the i class, APi(IoU) is the AP of category i in the specified IoU threshold.

4.3. Model performance evaluation

The model performance evaluation is divided into three parts: evaluation of the region proposal network based on CBAM-Yolov8, evaluation of the performance of the MLSAM model, and evaluation of the performance of the Yolo-MLSAM model.

4.3.1. Regional candidate network performance evaluation based on CBAM-Yolov8

To validate the effectiveness of the CBAM attention mechanism, a comparison was made between the Yolov8 model and its variants with the addition of CBAM, SE, and ECA attention mechanisms. The experimental results are shown in Table 1, with accuracy, true positive rate (TPR), F1 score, mAP@0.5, and mAP@0.5:0.95 as the measurement metrics.

Table 1. Experimental results of attention mechanism comparison

Attention mechanism			Pr%	TPR%	F1%	mAP@0.5%	mAP@0.5:0.95%
CBAM	SE	ECA					
×	×	×	86.76	84.64	86.35	88.51	56.43
√	×	×	**88.14**	**85.42**	**87.46**	**89.32**	**57.62**
×	√	×	87.43	84.86	86.68	88.46	56.35
×	×	√	86.53	84.25	86.24	87.75	55.69

According to the results in Table 1, compared to the original Yolov8 algorithm, the addition of the CBAM

attention mechanism improved the algorithm's accuracy by 1.38%, true positive rate by 0.78%, F1 score by 1.11%, mAP@0.5 by 0.81%, and mAP@0.5:0.95 by 1.19%. After adding the SE attention mechanism, all performance metrics showed some improvement, but the gains were not as significant as those achieved by CBAM. In contrast, after adding the ECA attention mechanism, the accuracy decreased by 0.23%, true positive rate by 0.39%, F1 score by 0.11%, mAP@0.5 by 0.76%, and mAP@0.5:0.95 by 0.74%.

Overall, the results in Table 1 suggest that the ECA attention mechanism not only leads to a decrease in performance but also increases the computational burden of the model. On the other hand, both the CBAM and SE attention mechanisms resulted in performance improvements, with CBAM showing the most significant enhancement, primarily due to its stronger modular design and feature enhancement capability.

For FFDM images, the CBAM-Yolov8 network was employed for object detection, with true positive rate (TPR), false positive rate (FPR), accuracy, Dice coefficient, and precision as evaluation metrics. In the experiments, performance was compared across five confidence thresholds ranging from 0.1 to 0.5. During training, the batch size was set to 8, the number of epochs to 150, and the initial learning rate was set to 0.01, with a cosine annealing method used to adjust the learning rate. The specific experimental results are presented in Table 2.

Table 2. Results of the confidence threshold comparison experiment

Thresholdvalue	TPR%	FPR%	ACC%	Pr%	Dice%
0.1	**97.86**	49.13	67.53	50.54	65.42
0.2	90.63	14.30	**87.35**	**85.56**	**83.52**
0.3	65.23	11.26	80.83	74.37	78.22
0.4	49.97	8.13	78.66	76.56	63.55
0.5	30.21	**7.41**	74.51	68.23	44.32

Based on the results presented in Table 2, it can be observed that when the confidence threshold is set to 0.1, the network achieves a high true positive rate of 97.86%. However, this high true positive rate is accompanied by a significant increase in the false positive rate, while both accuracy and Dice coefficient are relatively low. In this case, it is necessary to balance the trade-off between different metrics to determine the optimal model parameter settings. Ultimately, a confidence threshold of 0.2 was selected. Although the true positive rate decreases slightly under this setting, the false positive rate is substantially reduced, indicating that the model's predictions are more reliable. Specifically, by adjusting the confidence threshold, a balance between true positive and false positive rates can be achieved, leading to more accurate and dependable results. This is not only crucial for improving the performance of the current model but also provides valuable guidance for achieving high-precision segmentation in subsequent SAM models. By optimizing model parameters, computer-aided diagnostic techniques can be applied more effectively, enhancing early detection and diagnostic efficiency for breast diseases, providing better medical services to patients, and offering more effective support and assistance for patients' health and healthcare needs.

4.3.2. Performance evaluation based on MLSAM model

The MLSAM model enhances the image's local details by adding an additional intermediate feature layer. To evaluate the effectiveness of this modification, an ablation study was conducted by comparing different model configurations, as shown in Table 3.

Table 3. Ablation studies of MLSAM feature sources

Model	Global fusion	Decoder mask features	VITencoder			mIoU%	mBIoU%
			Early Layer	Middle Layer	Final Layer		
HQ-SAM	√	√	√	×	√	79.63	71.56
MLSAM	×	√	×	×	×	77.12	69.34
	√	√	×	×	×	77.84	70.86
	√	√	×	×	√	78.82	71.23
	×	√	√	√	√	79.64	71.13
	√	√	√	√	√	**81.12**	**71.89**

As shown in Table 3, the MLSAM model outperforms the HQ-SAM model in both mIoU and mBIoU metrics. Specifically, MLSAM achieved an mIoU of 81.12%, which represents a slight improvement over HQ-SAM's 79.63%. Additionally, in terms of mBIoU, MLSAM's highest value reached 71.89%, surpassing HQ-SAM's 71.56%. This demonstrates that the MLSAM model effectively utilizes global fusion and information from different feature layers, leading to higher segmentation accuracy and improved boundary handling. Notably, it excels in integrating global context information and generating precise masks.

The image encoder used in MLSAM has three different scale versions: Vit-b, Vit-l, and Vit-h, with Vit-b being the smallest model and Vit-h being the largest. To determine the most suitable image encoder, the following experiment was conducted, and the results are shown in Table 4.

Table 4. Image encoder type performance comparison experiment

Image encoder type	Model configuration	TPR%	FPR%	ACC%	Dice%
Vit-b	HQ-SAM	68.35	23.86	73.56	58.62
Vit-l		70.52	20.67	74.34	68.51
Vit-h		73.23	18.62	75.12	74.35
Vit-b	MLSAM	71.34	24.31	73.42	60.21
Vit-l		73.62	23.53	74.35	70.54
Vit-h		**76.34**	**21.35**	**77.58**	**76.34**

As shown in the table above, larger model sizes achieve better fine-grained results in downstream tasks. This finding indicates a positive correlation between the model's scale and its processing capability, where larger models tend to capture and handle details more effectively, leading to improved overall performance. Therefore, in applications requiring high precision, selecting a larger-scale model may be an effective strategy to enhance the quality of the results.

4.3.3. Performance evaluation based on Yolo-MLSAM model

In the performance evaluation of the Yolo-MLSAM model, the dataset was randomly split into 5 subsets, and 5-fold cross-validation was used to assess the model's performance. In each validation iteration, 4 subsets were used for model training, and the remaining 1 subset was used for testing, ensuring the reliability of the evaluation and the model's generalization capability. The evaluation metrics used were true positive rate, false positive rate, accuracy, and Dice coefficient to assess the model's performance. The specific experimental results are shown in Table 5.

Table 5. Segmentation results of HQ-SAM model and Yolo-MLSAM model

Model	Fold number	TPR%	FPR%	ACC%	Dice%
HQ-SAM	1	80.56	21.41	81.13	80.26
	2	81.34	16.34	82.35	81.34
	3	78.35	21.47	79.64	78.25
	4	80.93	19.54	81.67	80.85
	5	79.62	18.85	80.62	79.23
	Average	80.16	19.52	81.08	79.99
Yolo-MLSAM	1	89.31	11.35	90.31	89.21
	2	87.56	13.52	88.25	87.76
	3	85.41	15.42	86.24	83.59
	4	88.57	12.67	89.35	87.58
	5	86.43	14.23	87.52	85.74
	Average	**87.46**	**13.44**	**88.33**	**86.78**

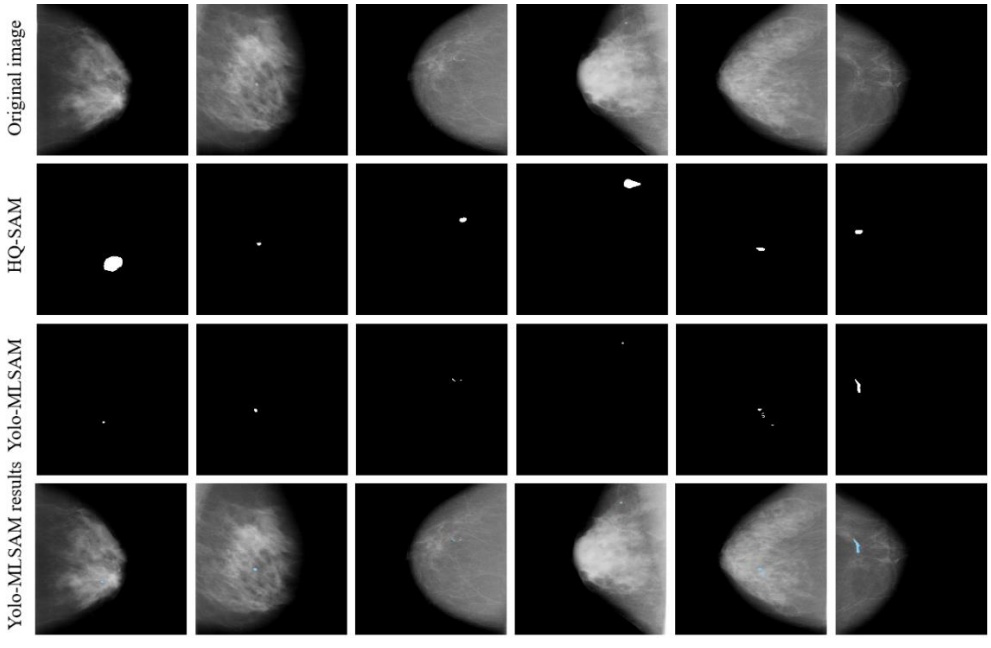

Fig. 4 Comparison between HQ-SAM model and Yolo-MLSAM model

Comparing the segmentation results of the HQ-SAM model and the Yolo-MLSAM model, it is evident that the Yolo-MLSAM model shows significant improvements across all metrics. The true positive rate increased by 7.3%, while the false positive rate decreased by 6.08%. Additionally, the segmentation accuracy and Dice coefficient of the Yolo-MLSAM model improved by 7.25% and 6.79%, respectively. These results demonstrate that the proposed method has a clear advantage over the HQ-SAM model in terms of feature discrimination for glandular tissue and microcalcification clusters. Therefore, the Yolo-MLSAM model allows for more precise segmentation of glandular tissue and microcalcification clusters.

As shown in Figure 4 for 6 sample cases, the HQ-SAM model incorrectly identified the glandular region as the lesion region in the first case, failing to effectively segment the lesion area. In contrast, the Yolo-MLSAM model successfully differentiated the regions. In the remaining five cases, the Yolo-MLSAM model not only identified all lesion areas but also performed more refined segmentation compared to the HQ-SAM model. These segmentation results further validate the superiority and accuracy of the Yolo-MLSAM model over the HQ-SAM model in medical image segmentation tasks.

5. Summary

This study proposes a Yolo-MLSAM-based framework for the segmentation of breast cancer microcalcification clusters, addressing the challenges faced by existing models in accurately segmenting small targets such as microcalcification clusters in medical images and the high cost of manual annotations. The proposed framework achieves more accurate and efficient medical image segmentation. Despite improvements, certain issues and limitations persist, such as suboptimal results when handling more challenging categories. Additionally, the use of a relatively simple pixel intensity strategy to handle regional uncertainties, without considering information from neighboring pixels, affects the accuracy and stability of the segmentation results. Future work will integrate learning from pixel contrasts across different categories to further enhance the model's performance.

References

[1] Brahimetaj R, Willekens I, Massart A, et al. Improved automated early detection of breast cancer based on high resolution 3D micro-CT microcalcification images[J]. BMC cancer, 2022, 22(1): 162.

[2] Naeim R M, Marouf R A, Nasr M A, et al. Comparing the diagnostic efficacy of digital breast tomosynthesis with full-field digital mammography using BI-RADS scoring[J]. Egyptian Journal of Radiology and Nuclear Medicine, 2021, 52: 1-13.

[3] Malliori A, Pallikarakis N. Breast cancer detection using machine learning in digital mammography and breast tomosynthesis: A systematic review[J]. Health and Technology, 2022, 12(5): 893-910.

[4] Abhisheka B, Biswas S K, Purkayastha B. A comprehensive review on breast cancer detection, classification and segmentation using deep learning[J]. Archives of Computational Methods in Engineering, 2023, 30(8): 5023-5052.

[5] Nguyen N V, Huynh H T, Le P L. Deep Learning Techniques for Segmenting Breast Lesion Regions and Classifying Mammography Images[C]//International Conference on Future Data and Security Engineering. Singapore: Springer Nature Singapore, 2023: 471-483.

[6] Amer A, Lambrou T, Ye X. MDA-unet: a multi-scale dilated attention U-net for medical image segmentation[J]. Applied Sciences, 2022, 12(7): 3676.

[7] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4015-4026.

[8] Punn N S, Agarwal S. RCA-IUnet: a residual cross-spatial attention-guided inception U-Net model for tumor segmentation in breast ultrasound imaging[J]. Machine Vision and Applications, 2022, 33(2): 27.

[9] Ning Z, Zhong S, Feng Q, et al. SMU-Net: Saliency-guided morphology-aware U-Net for breast lesion segmentation in ultrasound image[J]. IEEE transactions on medical imaging, 2021, 41(2): 476-490.

[10] Dar M F, Ganivada A. Efficientu-net: a novel deep learning method for breast tumor segmentation and classification in ultrasound images[J]. Neural Processing Letters, 2023, 55(8): 10439-10462.

[11] Shen L, Su J, Huang R, et al. Fusing attention mechanism with Mask R-CNN for instance segmentation of grape cluster in the field[J]. Frontiers in plant science, 2022, 13: 934450.

[12] Li P, Liu C, Li T, et al. EMDFNet: Efficient Multi-scale and Diverse Feature Network for Traffic Sign Detection[J]. arXiv preprint arXiv:2408.14189, 2024.

[13] Mazurowski M A, Dong H, Gu H, et al. Segment anything model for medical image analysis: an experimental study[J]. Medical Image Analysis, 2023, 89: 102918.

[14] Zhang Y, Jiao R. Towards segment anything model (SAM) for medical image segmentation: a survey[J]. arXiv preprint arXiv:2305.03678, 2023.

[15] Zhang K, Liu D. Customized segment anything model for medical image segmentation[J]. arXiv preprint arXiv:2304.13785, 2023.

[16] Ke L, Ye M, Danelljan M, et al. Segment anything in high quality[J]. Advances in Neural Information Processing Systems, 2024, 36.

[17] Yi H, Liu B, Zhao B, et al. Small object detection algorithm based on improved YOLOv8 for remote sensing[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023.

[18] Vasanthi P, Mohan L. Efficient YOLOv8 algorithm for extreme small-scale object detection[J]. Digital Signal Processing, 2024, 154: 104682.

[19] Lu E, Hu X. Image super-resolution via channel attention and spatial attention[J]. Applied Intelligence, 2022, 52(2): 2260-2268.

[20] Xia C, Wang X, Lv F, et al. Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 5493-5502.

[21] Xiao H, Li L, Liu Q, et al. Transformers in medical image segmentation: A review[J]. Biomedical Signal Processing and Control, 2023, 84: 104791.

[22] Bui P N, Le D T, Bum J, et al. Semi-supervised learning with fact-forcing for medical image segmentation[J]. IEEE Access, 2023.

[23] Abhisheka B, Biswas S K, Purkayastha B. A comprehensive review on breast cancer detection, classification and segmentation using deep learning[J]. Archives of Computational Methods in Engineering, 2023, 30(8): 5023-5052.

[24] Nguyen H T, Nguyen H Q, Pham H H, et al. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography[J]. Scientific Data, 2023, 10(1): 277.