YOLO-MAFS: Multi-scale Attention Fusion Network for Precise Breast Cancer Segmentation in Ultrasound Images

Luming Chen, Xinhe Zhang

Abstract—**In recent years, the incidence of breast cancer has continued to rise. Early screening and precise segmentation of lesion regions are essential for improving diagnostic accuracy and patient prognosis. However, limited data availability and the blurred boundaries of lesions often lead to false detections and missed segmentations in existing models. To address these challenges, this paper proposes a novel model for breast cancer ultrasound image segmentation, named YOLO-MAFS. The model integrates group convolution into the C3k2 module, effectively reducing computational complexity while enhancing segmentation accuracy. Additionally, a lightweight feature fusion module, MAFM, is introduced to handle the diverse appearances of lesion regions. To further address the issue of blurred lesion boundaries, the HAFB module is incorporated between the backbone and the detection head. Moreover, the LASH segmentation head is employed to improve the model's ability to capture subtle features. Experimental results on the BUS-BRA dataset demonstrate that YOLO-MAFS achieves improvements of 4.2% and 2.5% in mAP@50 and mAP@50–95, respectively, compared to YOLOv11. The model maintains a lightweight architecture while delivering high accuracy, making it well-suited for lesion segmentation in breast cancer ultrasound imaging.**

Index Terms—**Deep Learning, Breast cancer, Ultrasound images, YOLOv11, Segmentation**

I. Introduction

According to data from the World Health Organization (WHO), the incidence of breast cancer continues to rise. Annually, breast cancer accounts for 12.5% of all newly diagnosed cancers worldwide and remains one of the leading causes of increased mortality among women [1]. Early detection and timely diagnosis play a crucial role in the prevention and treatment of breast cancer. These measures not only effectively mitigate the incidence and mortality rates but also significantly lower treatment costs and improve patient survival rates.

In recent years, due to its advantages of being radiation-free, non-invasive, cost-effective, and convenient, breast ultrasound examination has been increasingly employed in clinical practice. It has become a crucial tool for the screening and diagnosis of breast abnormalities. However, the interpretation of conventional breast ultrasound images largely depends on the subjective expertise of radiologists, making it highly susceptible to variations in individual experience. This not only increases the workload but also lead to inconsistent diagnostic outcomes. To address this issue, Computer-Aided Diagnosis (CAD) systems have been developed to improve diagnostic efficiency, thereby assisting physicians in making more reliable judgements [2]. Accurate segmentation of breast ultrasound (BUS) images is crucial in clinical practice, as it precisely delineates lesion areas, aids radiologists in localizing breast cancer, and provides essential information about tumor morphology [3].

Despite significant advancements in artificial intelligence for breast ultrasound image analysis, current technologies still face challenges in clinical implementation. Key limitations include the scarcity of high-quality annotated datasets, poor model interpretability, and insufficient robustness when handling complex ultrasound image characteristics. Consequently, developing segmentation methods with super generalization capacity, consistent diagnostic reliability, and enhanced clinical applicability remains a critical research priority.

In recent years, deep learning techniques have driven significantly advancements in medical image segmentation, with architectures such as FCN [4] and U-Net [5] demonstrating remarkable performance through their powerful nonlinear representation capabilities. Nevertheless, traditional convolutional neural networks (CNNs) remain limited by their local receptive fields during feature extraction, making it challenging to effectively capture cross-scale global context information and consequently restricting feature diversity. To overcome this limitation, Ronneberger et al. [6] put forward the Attention U-Net model, which integrates attention gate into the skip connections of the conventional U-Net, enabling the network to focus on clinically relevant regions while suppressing irrelevant background features, thereby enhancing segmentation accuracy. Further innovations include Qin et al.'s U^2-Net [7], which introduces a nested U-Net structure with cascaded connections to strengthen multi-scale feature extraction and detail preservation. The TransUNet model, introduced by Chen et al. [8], combines the Vision Transformer [9] with the U-Net architecture. The primary objective of this integration is to address the limitations of U-Net in capturing long-range dependencies, thereby enhancing the model's capacity to model global semantic information.

Manuscript received June 20, 2025; revised September 1, 2025.

Luming Chen is a Postgraduate Student of School of Electronic Information, University of Science and Technology Liaoning, Anshan, 114051 China (e-mail: 18941035980@163.com).

Xinhe Zhang is an associate professor of School of Electronic and Information Engineering, Liaoning University of Science and Technology, Anshan, 114051 China (Corresponding author to provide phone: +86-151-2410-1849; e-mail: xhzhang@ustl.edu.cn).

Moreover, some research efforts have explored the integration of object detection and image segmentation to balance efficiency and accuracy. Su et al. [10] put forward a hybrid method combining YOLO and LOGO. In this method, YOLOv5-L6 is employed to detect and crop breast masses, followed by a dual-branch segmentation process involving local and global branches. Ultimately, the results are fused to improve the accuracy. Li et al. [11] developed an automated breast mass detection and segmentation framework based on YOLOv5 and GOLO-CMSS architectures. Through the coordinated operation of detection and the global-local branches, along with the integration of the multi-scale selection (MSS) and multi-layer feature fusion (MLI) mechanisms, this approach has achieved more accurate localization and segmentation of breast masses.

Although the integration of YOLO with segmentation networks has improved breast mass detection and segmentation performance, several critical challenges remain unresolved. First, the existing feature extraction and fusion module are not adequately optimized for the distinctive characteristics of breast ultrasound images, such as blurred lesion boundaries and significant scale variations. This limitation results in suboptimal feature extraction and ineffective fusion, ultimately compromising segmentation accuracy. Second, in complex clinical scenarios, ultrasound images are affected by factors like equipment differences, imaging angle variations, and tissue background complexity. Existing methods exhibit poor robustness under these conditions, struggling to maintain stable performance.

To further enhance the performance of mass detection and segmentation in breast ultrasound images, this study introduces several architectural enhancements to the YOLOv11 framework [12], proposing a novel YOLO-MAFS model optimized for efficient segmentation tasks. The key contributions include:

1) An efficient multi-scale convolution (EMSConv) was designed to optimize the structure of the C3k2 module in YOLOv11. By incorporating a grouped convolution-based multi-scale feature processing mechanism, the proposed model significantly improves the model's capacity to represent breast masses across varying sizes.

2) A lightweight multi-scale feature fusion module (MSFM) was proposed. In this module, Ghost-Shuffle Convolution (GSConv) was employed to substitute for traditional convolution, thereby enabling low-cost and high-efficiency multi-level semantic information fusion. When integrated into the Neck structure of YOLOv11, this module not only significantly reduces the computational complexity of the model but also enhances its adaptability to complex breast lesions and segmentation performance.

3) To enhance the model's ability to precisely model the boundaries of lesions, a hierarchical attention fusion block (HAFB) was introduced. This module combines local and global attention mechanisms along with reparametrized convolution (RepConv) to effectively aggregate multi-level semantic information, thereby strengthening the modeling of context relationships and improving the accuracy of boundary identification, especially in lesion areas with blurred boundaries or complex shapes.

4) A lightweight asymmetric segmentation head (LASH) was developed to resolve the task-specific conflicts between detection, classification, and mask prediction tasks. By optimizing the mask generation pathway while maintaining inference efficiency, LASH enhances segmentation performance in breast ultrasound images, particularly for detail handling and edge transitions.

II. Method Introduction

A. Constructing YOLOv11

YOLOv11 is a highly efficient object detection algorithm recently introduced by the Ultralytics team, inheriting the core advantages of speed and accuracy from the YOLO series. In addition, YOLOv11 has excellent scalability and adaptability, which makes it show certain potential when dealing with medical images that are diverse and highly complex [13].

B. Constructing YOLO-MAFS

To enhance the performance of YOLOv11 breast cancer ultrasound image segmentation, this paper proposes an innovative improved model based on YOLOv11, whose overall structure is shown in Fig.1.

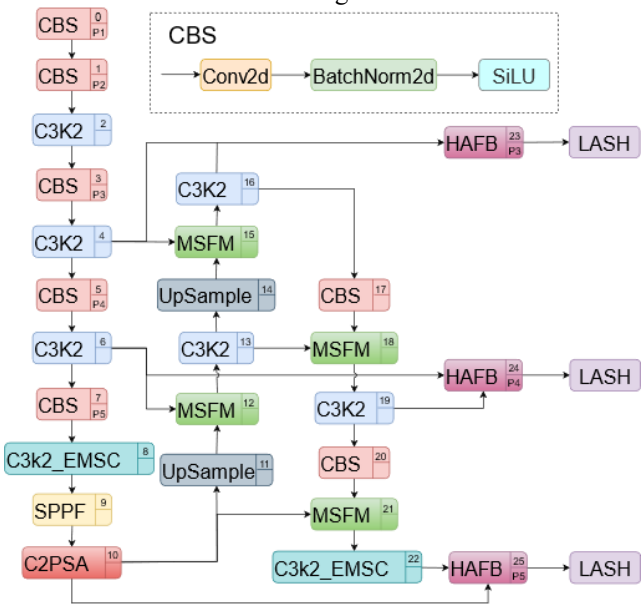

Fig. 1. Structural Diagram of YOLO-MAFS

Firstly, the model extracts feature from the input image via the backbone network. In the structure of the backbone network, the C2k3-EMSC module is introduced. By employing a multi-scale processing mechanism with grouped convolution, the model achieves enhanced perception of breast masses across varying scales. These extracted high-level features are subsequently processed by the MSFM module for refined feature integration. This module is incorporated into the neck architecture of YOLOv11. On the premise of maintaining computational efficiency, it can effectively integrate spatial and semantic information from different hierarchical levels. Subsequently, the hierarchical attention fusion block (HAFB) is introduced to conduct a further fusion of the multi-scale features extracted from the backbone network and the neck structure. HAFB effectively establishes hierarchical feature dependencies by synergistically integrating local and global attention mechanisms across multiple feature levels. It enhances lesion boundary modeling, improves the model's understanding of context, and increases segmentation accuracy and continuity. Finally, the fused features are input into the LASH module for detection and segmentation, generating the final

segmentation results. In the subsequent chapters of this paper, we will present a detailed account of the design principles and performance of modules such as C2k3-EMSC, MSFM, HAFB, and LASH will be presented. This is to validate their effectiveness in the task of breast cancer ultrasound image segmentation.

C. *Efficient Multi-Scale Convolution (EMSConv)*

In YOLOv11, if the C3k parameter is FALSE, C3k2 becomes the original C2F module; if TRUE, the C2F Bottleneck is replaced by the C3k module. Drawing on the GhostNet proposed by Han et al. [14], this research conducts a structural reconstruction of the Bottleneck structure within the C3k2 module. An efficient multi-scale convolution mechanism grounded in grouped convolution (Efficient Multi-Scale Convolution, EMSConv) was introduced. Ultimately, the module C3k2-EMSC was constructed, and its structure is shown in Fig.2.

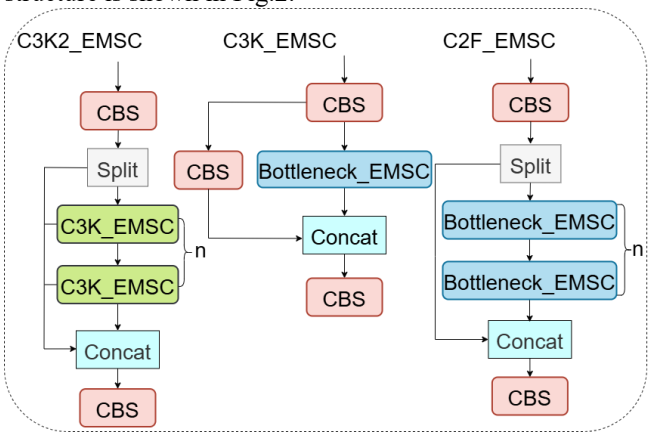

Fig. 2. Structure diagram of C3k2_EMSC

The Bottleneck-EMSC module remarkably reduces the computational complexity in convolution operations by incorporating Group Convolution, and effectively strengthens the feature extraction ability. Its structure is shown in Fig.3. Within this module, EMSConv evenly partitions the input channels into four subgroups. Each subgroup then utilizes convolution kernels of varying scales for feature extraction. The features extracted from each group are concatenated, and the features across different channels are subsequently fused through Pointwise Convolution (PW), thereby enhancing the representational capacity of each channel. The uniform distribution of input channels among multiple convolution groups not only strikes a balance in computational efficiency but also guarantees that each group can capture features independently and comprehensively.

Within the EMSConv module, the choice of four convolution kernels with distinct scales, specifically 1×1, 3×3, 5×5, and 7×7, is grounded in their respective merits in capturing spatial features across diverse scales. The 1×1 convolutional kernels are specifically designed to extract localized fine-grained features through nonlinear transformations, which effectively enhances the model's feature representation capacity. This compact receptive field enables precise capture of pixel-level patterns while maintaining computational efficiency. The 3×3 convolutional kernels can capture more intricate local patterns and extract spatial information at a medium scale by extending the receptive field. In contrast, the 5×5 and 7×7 convolutional kernels, with their larger receptive fields, are capable of capturing broader contextual information and global features. Particularly in scenarios involving larger objects or complex scenes, these kernels facilitate understanding the relationships between image objects and the overall image structure.

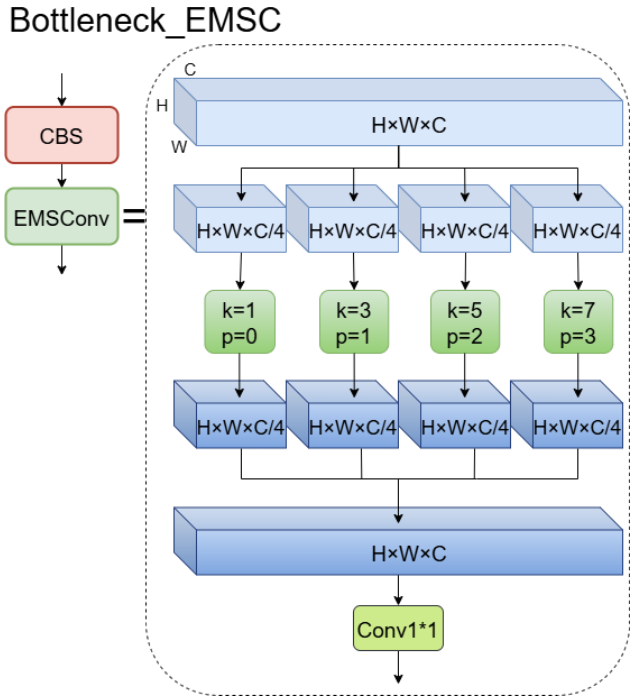

Fig. 3. Structure diagram of Bottleneck_EMSC

As verified experimentally, as shown in TABLE I, when the number of groups is insufficient, the model fails to fully exploit multi-scale features, thereby restricting its feature extraction ability. When the number of groups becomes excessively large, information tends to be overly partitioned. This result in inadequate feature fusion across groups and compromises the effectiveness of the final feature integration. Consequently, dividing the channels into four groups ensures that each convolutional kernel efficiently processes the features of its corresponding channel group. This strategy allows the model to more effectively capture feature details at multiple levels, thereby improving the overall robustness of the model.

TABLE I
TEST RESULTS

Groups	mAP@50(M)
n=3	0.831
n=4	0.852
n=5	0.836

In this study, the C3k2 module in the final layer of the Backbone and Neck in YOLOv11 is replaced with the enhanced C3k2-EMSC module. The replacement is carried out at this position because the deep-layer feature maps exhibit a higher number of channels and richer semantic information, which makes them more suitable for the effective operation of multi-scale architectures.

D. *Multi-Scale Feature Fusion Module (MSFM)*

To improve the efficiency of multi-scale feature fusion in target segmentation tasks, this study introduces a lightweight multi-scale feature fusion module (Multi-Scale Feature Fusion Module, MSFM). This module is integrated into the Neck component of YOLOv11, taking the place of certain

conventional feature fusion units. The design inspiration of this module is derived from the SDI module in UNetV2 [15]. Comprehensive optimizations have been carried out in both the structural design and fusion strategy to better meet the computational efficiency and representation requirements for breast cancer ultrasound image segmentation tasks.

In contrast to the UNetV2 approach, which relies on standard convolution and weighted fusion, the MSFM module replaces traditional convolutions by integrating the lightweight and highly efficient Ghost-Shuffle Convolution (GSConv) [16]. The structure diagram of GSConv is shown in Fig.4.

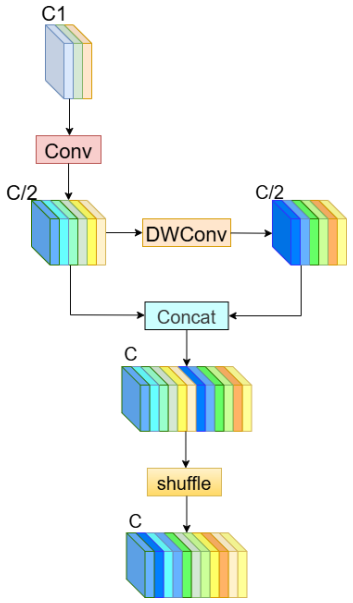

Fig .4. Structure diagram of GSConv

This strategy not only substantially reduce the computational cost of the model but also enhances the representational consistency of cross-scale features by means of the Ghost feature generation mechanism and channel shuffling.

Assume that the input to the MSFM module (shown in Fig.5) is a collection of feature maps $\{x_i\}_{i=1}^{N}$ derived from various scales. First, channel attention is introduced for the higher-scale features, while spatial attention is incorporated for the lowest-scale features. This step aims to boost the intensity of representation for critical information channels or spatial locations. Then, through an average pooling operation (adaptive_avg_pool2d), the features are uniformly adjusted to the same spatial resolution as that of the lowest-scale feature x_i Subsequently, each feature is compressed through an independent GSConv branch to a unified channel dimension C . Finally, element-wise multiplication (Hadamard Product) is performed to achieve fusion:

$$F_{MSFM} = \bigodot_{i=1}^{N} GSConv_i(\tilde{x}_i) \quad (1)$$

Among them, $\tilde{x}_i$ represents the i -th feature map after scale alignment. GSConv stands for a convolutional operation with the ability of lightweight feature transformation.

In contrast to traditional feature concatenation or summation strategies, this multiplicative fusion mechanism exhibits enhanced efficiency and robustness in information integration. By performing element-by-element multiplication operations, the module can more effectively uncover the co-occurrence relationships among features across different scales. This strengthens the synergy and selectivity of feature representation, thereby enhancing the model's ability to perceive complex image content. Furthermore, the incorporation of the lightweight GSConv (Ghost - Shuffle Convolution) as a feature compression unit not only effectively preserves the crucial semantic information within the input features but also significantly reduces the number of parameters and computational overhead of the module. Consequently, the inference efficiency of the overall model is improved.

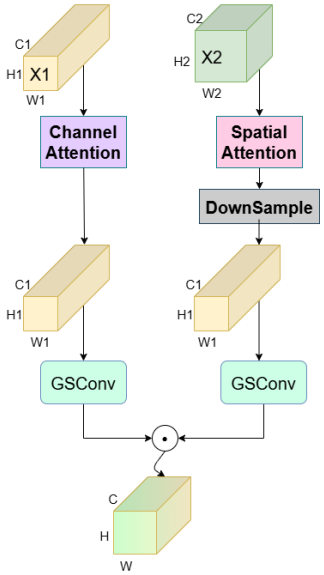

Fig .5. Structure diagram of MSFM

In summary, the MSFM module augments the model's comprehensive modeling capabilities for spatial details and contextual semantics, demonstrating excellent performance in multi-scale feature fusion. After being integrated into the YOLOv11 architecture, the MSFM has exerted a positive influence on improving the quality of feature representation and enhancing the target perception ability. Overall, this underscores its application potential and promotional value as a general multi-scale fusion module in tasks such as object segmentation.

E. Hierarchical Attention Fusion Block (HAFB)

In the realm of deep learning, particularly within computer vision tasks, multi-scale feature fusion and attention mechanisms have been extensively demonstrated to substantially boost the model's capacity to perceive complex scenarios. However, existing methods often encounter the problem of poor coordination between local and global information when fusing features, which in turn limits the model's comprehensive understanding of fine-grained targets and contextual semantics. To address this challenge, this study proposes a novel module - the Hierarchical Attention Fusion Block (HAFB). The structure diagram of HAFB is shown in Fig.6. Specifically, this module is an improvement on the hierarchical attention branch (LocalGlobalAttention) proposed in HCF-Net [17], integrating both local and global attention branches and refining the fusion of input features through a hierarchical feature processing path. This design

not only preserves the spatial details of the input features but also strengthens the semantic representation capacity of the global context via an adaptive weighting mechanism.

HAFB employs a parallel multi-branch architecture to extract features at different scales and semantic hierarchies. This module encompasses three crucial branches: the local branch, the global branch, and the serial convolution branch. Each of these branches is designed to model local details, global context, and structural information, respectively. This design significantly elevates the model's performance in high-precision localization tasks, such as lesion detection.

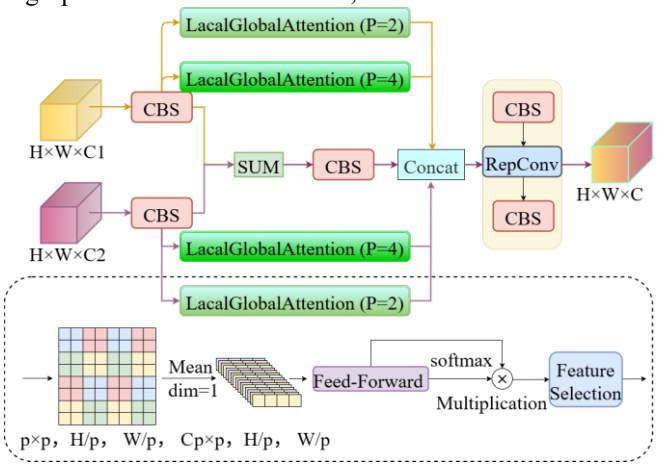

Fig .6. Structure diagram of HAFB

Suppose the input feature tensors are $F_1 \in \mathbb{R}^{H\times W\times C_1}$ and $F_2 \in \mathbb{R}^{H\times W\times C_2}$. Initially, channel compression is carried out via point-wise convolution to acquire the intermediate feature $F_1^*, F_2^* \in \mathbb{R}^{H\times W\times C}$ with a unified dimension. Subsequently, these are fed into five parallel branches for processing, yielding the local attention branches $F_{lacal_1} \in \mathbb{R}^{H\times W\times C}$ and $F_{lacal_2} \in \mathbb{R}^{H\times W\times C}$, the global attention branches $F_{global_1} \in \mathbb{R}^{H\times W\times C}$ and $F_{global_2} \in \mathbb{R}^{H\times W\times C}$ and the convolutional branch $F_{conv} \in \mathbb{R}^{H\times W\times C}$. After summing up the results of all branches along the channel dimension, they are fed into the re-parameterizable convolutional layer (Re-parameterizable Convolution, RepConv [18]). Eventually, the fused feature output $F \in \mathbb{R}^{H\times W\times C}$ is acquired. Among them, RepConv is a module that adopts a multi-branch structure during the training stage and fuses into a single convolution operation during the inference stage. It can significantly improve the inference speed while ensuring performance.

To achieve effective discrimination and interaction between the local and global branches, this study partitions the feature map into non-overlapping regions in the spatial dimension by manipulating the Patch size parameter. Subsequently, the aggregation and displacement operations of local regions are accomplished. Employing the Unfold and Reshape operations, the features F_1^* and F_2^* are partitioned into contiguous non-overlapping blocks, presenting a feature representation of dimension $p\times p, \frac{H}{p}, \frac{W}{p}, C$. Then, average processing is performed on each channel to obtain a feature representation of dimension $p\times p, \frac{H}{p}, \frac{W}{p}$. Introduce this feature into the feed-forward neural network (Feed-Forward Neural Network, FFN) [19] for linear transformation, and map it to the probability distribution of spatial response through the activation function. This distribution is utilized to assign attention weights to each Patch. By doing so, explicit modulation in the spatial dimension is achieved, thereby enhancing the significance of local or global features.

Upon acquiring the weighted Patch representations, this study further incorporates a task-relevant feature selection mechanism to bolster the discriminative power of the model. Let $d = \frac{H\times W}{p\times p}$, and represent the weighted result as $(t_i)_{i=1}^{C^*}$. Here, $t_i \in \mathbb{R}^d$ represents the Patch representation on the i-th channel. To accomplish effective label selection, the following weighting strategy is proposed:

$$\hat{t}_i = P\cdot sim(t_i, \xi)\cdot t_i \tag{2}$$

Where, $\xi \in \mathbb{R}^{C^*}$ represents the task embedding vector, $P \in \mathbb{R}^{C^*\times C^*}$ is the learnable linear projection matrix, and $sim(,)$ denotes the normalized cosine similarity function (with a range of [0, 1]).

In this context, ξ acts as the task embedding, designating which tokens are relevant to the task. Each token t_i is re-weighted according to its relevance to the task embedding (measured by the cosine similarity), thereby effectively emulating token selection. Subsequently, a linear transformation of P is applied to each token for channel selection. Then, Reshape and smooth interpolation operations are carried out. Finally, the features $F_{lacal} \in \mathbb{R}^{H\times W\times C}$ and $F_{global} \in \mathbb{R}^{H\times W\times C}$ are generated. In addition, after directly adding F_1^* and F_2^*, the result is fed into a 3×3 convolutional layer. The output $F_{conv} \in \mathbb{R}^{H\times W\times C}$ of the serial convolutional branch is thus obtained, which enhances the ability to model local structures. Ultimately, the features output by the three branches are fused in RepConv to generate a unified context-aware feature representation.

The HAFB module can be embedded into various backbone networks. In this paper, it is integrated between the head and neck components of the YOLOv11 model. This integration aims to fully harness its advantages in multi-scale and fine-grained information extraction.

F. Lightweight Asymmetric Segmentation Head (LASH)

In this paper, we proposed a lightweight asymmetric segmentation head (LASH) based on the Lightweight Asymmetric Detection Head (LADH) proposed by Zhang et al. [20]. To address the different requirements of regression tasks, mask prediction tasks, and classification tasks, this new head is more suitable for segmenting high-pixel images. It aims to enhance the computational efficiency and task accuracy of the model. Fig. 7 provides an overview of the LASH structure.

During the instance segmentation process, the regression task primarily focuses on the location and bounding box information of the object, necessitating the precise capture of the object's spatial location and dimensions. To extract these boundary features more effectively, compared with the

original detection head of YOLOv11, the LASH model innovatively substitutes two conventional convolutions with three 3×3 depth-wise separable convolutions (DWConv) in the regression task. By doing so, this design allows the model to extract more intricate features layer-by- layer and incorporate information at different scales. As a result, it significantly improves the stability and accuracy of the regression outcomes.

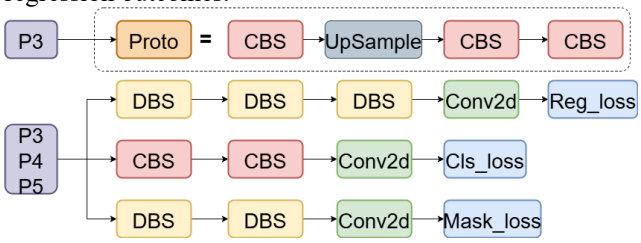

Fig .4. Structure diagram of LASH

Regarding the mask prediction task, this task necessitates precise pixel-level processing. Especially in complex scenarios, the delicate features of the target are of great significance. Within LASH, two conventional convolutions in the mask prediction head are substituted by two 3×3 depth-wise separable convolutions. Furthermore, the employment of two depth-wise separable convolutions can effectively extract the details within an image, prevent overfitting, and improve the capacity to capture details in the segmentation task. This optimization can effectively improve the accuracy and robustness of the mask prediction task. Especially when handling medical images with high noise levels and complex backgrounds, it enables better preservation and segmentation of crucial lesion areas.

In the classification task, LASH continues to adopt traditional convolutional operations instead of depth-wise separable convolutions. Traditional convolutions are capable of effectively extracting high-level features of images globally, meanwhile ensuring a relatively high classification accuracy. In the context of the classification task for breast cancer ultrasound images, this approach proves to be more suitable. The reason is that the conventional convolutions are better able to integrate global information, thereby enhancing the classification accuracy. This is particularly evident when the target regions are relatively ambiguous or have ill-defined boundaries.

In summary, the LASH model optimizes the original detection head of YOLOv11. By reasonably substituting convolutional operations, it can better satisfy the characteristics and demands of diverse tasks. As a result, in the high-pixel image segmentation task, the model can more effectively cope with complex scenarios and capture details, thus enhancing the overall performance of detection and segmentation.

III. Experiment And Results

A. Data and Experimental Setup

In this study, the research objects are derived from the publicly available BUS-BRA [21] dataset. This dataset includes 1,875 anonymized images from 1,064 female patients. These images were acquired by four ultrasonic scanners during a systematic investigation at the National Cancer Institute (Rio de Janeiro, Brazil). The dataset encompasses tumors confirmed by biopsy, which are categorized into 722 benign cases and 342 malignant cases. To tackle the problem of a limited number of datasets and mitigate the influence of overfitting, in this research, some of the 1,875 images in the dataset were augmented through translation and flipping operations. Eventually, 2,552 images were obtained. The images were partitioned according to a ratio of 7:2:1. Specifically, 1786 images were designated for training, 510 for testing, and 256 for validation. Additionally, measures were taken to ensure that each type of tumor was uniformly distributed across the training set, test set, and validation set. Such a distribution facilitates the model in precisely learning the characteristics of each type of tumor and guarantees the accuracy of the model evaluation.

This experiment was conducted on a system running the Windows 10 operating system, equipped with an Intel(R) Core (TM) i9-14900HX processor clocked at 2.20 GHz and an NVIDIA GeForce RTX 4060 graphics card. The deep-learning framework consisted of Python version 3.9.20, PyTorch version 2.5.1, and CUDA version 12.1. During the training phase, the initial learning rate was set to 0.01, the momentum was set at 0.937. The size of the input images was 640×640, the batch size was 16, and the number of epochs was 300.

B. Contrast Experiment

To comprehensively validate the effectiveness of the YOLO-MAFS model proposed in this study in instance segmentation tasks, this section conducts comparative experiments between YOLO-MAFS and mainstream lightweight as well as medium-to-high complexity YOLO series models. These comparative models include YOLOv11n-seg, YOLOv11s-seg, YOLOv8n-seg, YOLOv8s-seg, YOLOv5n-seg, and YOLOv5s-seg. All models were trained and evaluated on the same breast ultrasound image dataset. The evaluation metrics encompassed mAP@50, mAP@50-95, GFLOPs, and the number of model parameters. The experimental results are shown in TABLE II.

As can be observed from TABLE II, YOLO-MAFS exhibits significant advantages in multiple key metrics. In terms of detection accuracy, YOLO-MAFS achieved the highest mAP@50 value of 0.852, representing a 4.2%

TABLE II
Comparison of Model Accuracy and Computational Performance in Task Segmentation

Methodology	Map@50 (B)	Map@50-95 (B)	Map@50 (M)	Map@50-95 (M)	GFLOPs	Parameters
YOLOv11n-seg [12]	0.801	0.567	0.810	0.553	10.2	2834958
YOLOv11s-seg [12]	0.831	0.568	0.830	0.566	35.3	10067590
YOLOv8n-seg [22]	0.807	0.551	0.808	0.553	10.7	2937174
YOLOv8s-seg [22]	0.832	0.559	0.828	0.560	57.3	10482454
YOLOv5n-seg [23]	0.779	0.525	0.784	0.533	**6.9**	**1886015**
YOLOv5s-seg [23]	0.790	0.542	0.797	0.54	26.4	7643158
YOLO-MAFS	**0.843**	**0.575**	**0.852**	**0.578**	10.9	3858286

over the baseline model YOLOv11n-seg and surpassing the moderately complex models YOLOv11s-seg and YOLOv8s-seg by 2.2% and 1.8%, respectively. Meanwhile, under the more stringent mAP@50-95 metric, YOLO-MAFS achieved a relatively high score of 0.578, significantly outperforming YOLOv11n-seg (0.553), YOLOv11s-seg (0.566), and YOLOv8s-seg (0.560). This result fully demonstrates YOLO-MAFS's stronger stability and generalization in multi-scale object detection and detail area recognition. YOLO-MAFS not only exhibits outstanding segmentation accuracy, but also strikes an efficient balance between model scale and inference speed, thereby presenting significant potential for practical deployment. The GFLOPs of YOLO-MAFS is 10.9, which is only marginally higher than those of YOLOv11n-seg (10.2) and YOLOv8n-seg (10.7), yet substantially lower than those of larger-scale models like YOLOv11s-seg (35.3) and YOLOv8s-seg (57.3). Simultaneously, the number of parameters of YOLO-MAFS is 3.86M, which lies within the scope of lightweight models. This suggests that the model exhibits favorable deployment adaptability and is appropriate for applications in resource-constrained edge devices or real-time diagnostic scenarios.

Based on the above analysis, it can be concluded that YOLO-MAFS efficiently integrates multi-scale perception and attention-guided mechanisms within a lightweight architecture. This design enables the model to strike an effective balance between high segmentation accuracy and low computational overhead. The multi-scale strategy allows the model to capture rich contextual information across various spatial resolutions, enhancing its ability to recognize complex structures. Meanwhile, attention guidance focuses computational resources on the most relevant regions, improving precision in challenging scenarios. Together, these strategies make YOLO-MAFS particularly well-suited for accurate and efficient medical image segmentation.

To compare the model performance more clearly, Fig.8 and Fig.9 respectively illustrate the P-R curves of the baseline model YOLOv11n-seg and the improved model YOLO-MAFS on the validation set. Meanwhile, the detection results are presented in the form of a histogram (Fig.10).

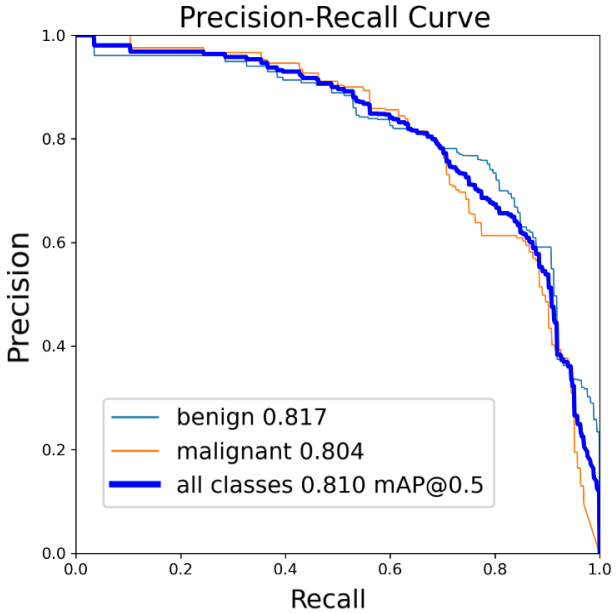

Fig.8. The validation findings of the YOLOv11n-seg model

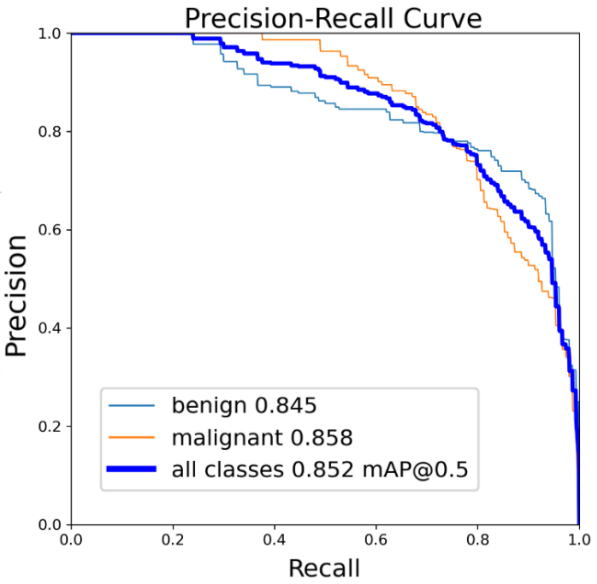

Fig.9. The validation findings of the YOLO-MAFS model

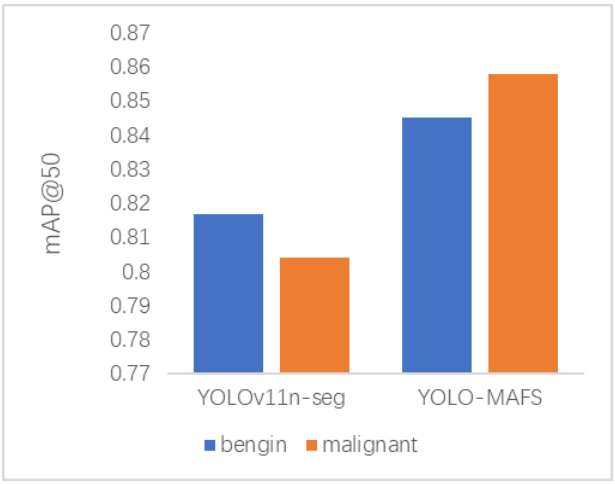

Fig.10. The mAP@50 performance comparison between YOLOv11n-seg and YOLO-MAFS

It can be clearly observed from Fig.10 that in the task of breast cancer segmentation, YOLO-MAFS shows more outstanding performance than YOLOv11n-seg. Specifically, in the detection of benign and malignant masses, the mAP@50 of YOLO-MAFS reached 0.845 and 0.858, respectively. Compared with YOLOv11n-seg, these values were increased by 2.8% and 5.4%, respectively, indicating that YOLO-MAFS achieved relatively significant performance improvements across different categories. This can be partly attributed to the numerous optimized designs in YOLO-MAFS model. In particular, the introduction of HAFB has played a significant role. By strengthening the model's capacity to perceive multi-scale information and its ability to focus on local key features, YOLO-MAFS attains a higher level of accuracy in both the detailed characterization of benign masses and the boundary identification of malignant masses. In the malignant category, the remarkable enhancement of detection performance demonstrates that when confronted with highly challenging lesions, such as those with complex morphologies and blurred margins, YOLO-MAFS exhibits a more robust capacity for feature extraction and discrimination. This advancement not only elevates the overall detection precision of the model but also bolsters its adaptability and dependability when dealing with diverse types of lesions in real-world clinical settings.

TABLE III
ABLATION STUDY ON THE YOLO-MAFS MODEL

Methodology	Map@50 (B)	Map@50-95 (B)	Map@50 (M)	Map@50-95 (M)	GFLOPs	Parameters
+EMSC+MSFM+HAFB+LASH	**0.843**	**0.575**	**0.852**	**0.578**	10.9	3858286
+MSFM+HAFB+LASH	0.831	0.564	0.836	0.567	10.9	3903086
+EMSC+MSFM+HAFB	0.833	0.57	0.831	0.571	12.8	4384486
+EMSC+HAFB+LASH	0.838	0.573	0.839	0.568	11.2	3071582
+EMSC+MSFM+LASH	0.835	0.569	0.828	0.566	8.9	2404558
+EMSC+MSFM	0.83	0.568	0.832	0.571	10.6	2804918
+EMSC+HAFB	0.829	0.574	0.834	0.573	12.3	4284430
+EMSC+LASH	0.828	0.572	0.828	0.567	10.7	3852974
+MSFM+HAFB	0.837	0.575	0.838	0.57	12.6	4334542
+MSFM+LASH	0.829	0.566	0.83	0.569	9.0	2419222
+HAFB+LASH	0.832	0.565	0.828	0.574	10.7	3897774
+EMSC	0.827	0.565	0.829	0.575	10.2	2790158
+MSFM	0.825	0.562	0.825	0.567	10.6	2849462
+HAFB	0.828	0.570	0.828	0.570	10.7	3852974
+LASH	0.818	0.554	0.827	0.565	**8.6**	**2403502**
YOLOv11	0.801	0.567	0.810	0.553	10.2	2834958

C. Ablation Experiment

To assess the functions and indispensability of the EMSConv, MSFM, HAFB, and LASH modules, a series of ablation experiments were carried out in this study. Using YOLOv11 as the baseline model, each module was incrementally added or removed. Comparative tests were then performed within the task of breast ultrasound image segmentation. The evaluation indicators encompass mAP@50, mAP@50-95, GFLOPs, and the quantity of parameters. The detailed results are presented in TABLE III. The experimental findings demonstrate that each module effectively enhances the model's performance in diverse aspects. Upon the introduction of EMSConv, the model's capacity for multi-scale feature representation was notably enhanced. Specifically, the mAP@50 increased from 0.810 to 0.827, thereby validating its superiority in scale perception. With an efficient semantic fusion mechanism, MSFM improves the quality of feature fusion while maintaining low computational cost. This enables the model to approximate the performance of the full-fledged structure even under a lightweight configuration. The HAFB module bolsters the model's comprehension of contextual information and boundary details. Evidently, the mAP@50-95 has been elevated from 0.553 to 0.574. This module is especially conducive to the segmentation of lesions characterized by ambiguous morphologies and intricate boundaries. LASH optimizes the mask generation path. It demonstrates excellent performance in terms of detail preservation and edge transition, and moreover, does not introduce an increase in inference latency.

Further analysis reveals that the absence or isolated utilization of individual modules imposes inherent limitations. Specifically, when the LASH module is removed from YOLO-MAFS, the model demonstrates moderately enhanced performance on certain local metrics; however, the overall parameter count increases substantially, leading to a marked decline in efficiency. When the EMSConv module is excluded, although the mAP@50-95 improves slightly from 0.553 to 0.567, the overall detection accuracy remains inferior to that of the complete model. Integrating EMSConv and MSFM into the baseline model confers distinct advantages in scale adaptation and feature fusion, yet the lack of a boundary refinement mechanism frequently results in edge misalignment or contour distortion. When only the HAFB module is added, improvements in edge extraction and local feature integration are observed; nevertheless, the model's parameter count and computational complexity rise significantly. Furthermore, feature redundancy tends to induce over-segmentation or artifacts, rendering the overall performance suboptimal.

In summary, individual modules typically optimize only specific aspects, making it difficult to simultaneously balance detection accuracy, edge details, and computational efficiency. In contrast, YOLO-MAFS achieves multi-dimensional collaborative optimization through the integration of multiple modules. While maintaining a reasonable number of parameters and GFLOPs, it delivers outstanding performance with a mAP@50 of 0.852 and a mAP@50-95 of 0.578, which strongly validates the necessity and effectiveness of the multi-module collaborative design.

D. Generalization Experiments

To comprehensively evaluate the generalization capability of the proposed YOLO-MAFS model across different data distributions, we conducted experiments not only on the BUS-BRA dataset for training and validation, but also on the BUSI dataset as a cross-domain evaluation. These experiments provide strong evidence of the model's ability to handle diverse data sources in real-world applications.

The experimental results on the BUSI dataset [24] are presented in Fig.11 and Fig.12. Compared with the baseline model, YOLO-MAFS achieved an improvement in mAP@50, increasing from 0.797 to 0.815, which represents a 1.8 percentage point gain.

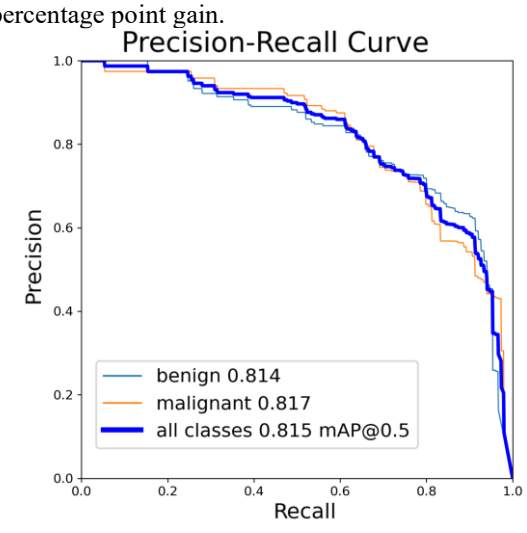

Fig.11. The validation findings of the YOLOv11n-seg model

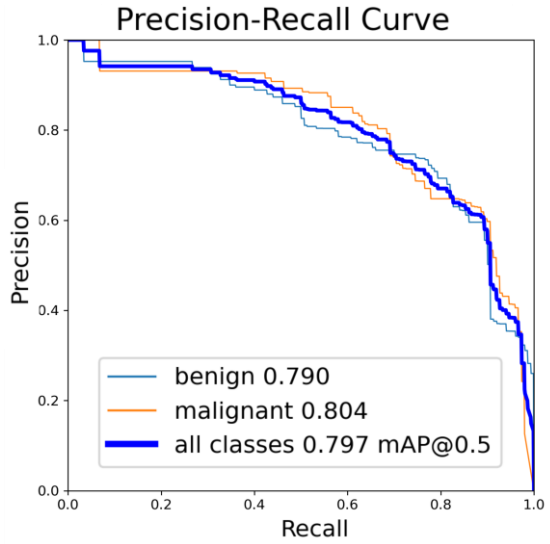

Fig.12. The validation findings of the YOLO-MAFS model

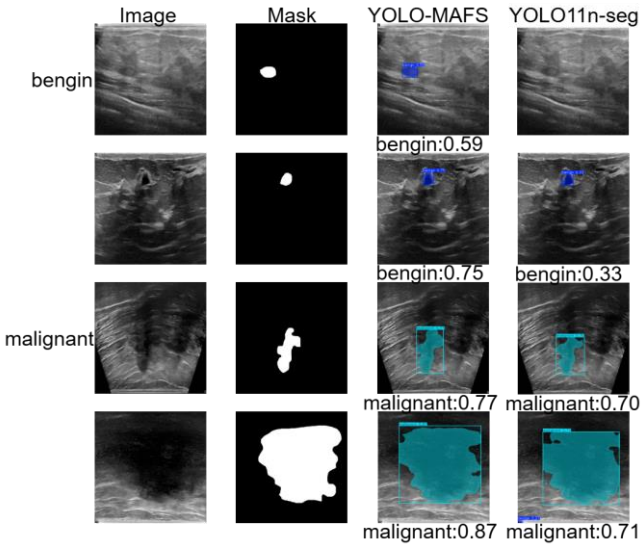

Fig.13. Visual comparison of segmentation results on the BUSI dataset

Furthermore, to further highlight the advantages of YOLO-MAFS, we compared its segmentation results with those of the baseline model (Fig. 13). Visualizations reveal that YOLO-MAFS delivers higher confidence, more precise boundary localization, and fewer missed or false detections in breast tumor ultrasound image segmentation. These outcomes validate YOLO-MAFS's robust performance and cross-dataset stability, underscoring its potential for real-world application.

E. Model Evaluation

To comprehensively evaluate the performance of the YOLO-MAFS model, this study conducts comparative experiments against same-scale models such as YOLOv11n-seg, YOLOv8n-seg, and YOLOv5n-seg, assessing its overall effectiveness in segmentation tasks.

As shown in Fig.14, the segmentation outcomes of diverse models for different types of defects are depicted. The experimental findings demonstrate that, in comparison to the existing benchmark models, the YOLO-MAFS model exhibits superiority in terms of the segmentation accuracy of various defect regions, the capacity to identify boundary contours, and the detection performance of minute targets. Meanwhile, the model still maintains high stability and robustness when dealing with complex backgrounds and diverse defect morphologies, demonstrating excellent generalization ability. These advantageous features comprehensively attest to the application potential and practical significance of the YOLO-MAFS model within the realm of medical image segmentation tasks.

IV. Discussion

Although the proposed model has achieved significant performance improvements in the task of breast cancer ultrasound image segmentation, particularly showcasing robust resilience and adaptability when handling lesions with ambiguous boundaries and intricate morphologies, certain inaccuracies still persist in the actual segmentation outcomes.

The root cause of the problem lies in the fact that, despite the introduction of the HAFB module to enhance boundary and context modeling, it still lacks sufficient discriminative power for fine-grained structures and ambiguous boundaries. This is particularly evident when the boundary features of benign and malignant lesions are highly similar, leading to segmentation deviations.

Therefore, future research will focus on further refining feature extraction mechanisms and enhancing boundary perception capabilities, while actively exploring uncertainty-aware learning methods to more accurately model ambiguous regions that are prone to diagnostic confusion. These targeted enhancements aim to mitigate the issues of missed detections and false positives arising from boundary inaccuracies in medical image segmentation tasks. By addressing such critical challenges, the overarching goal is to substantially improve the model's reliability and elevate its practical effectiveness in real-world clinical applications.

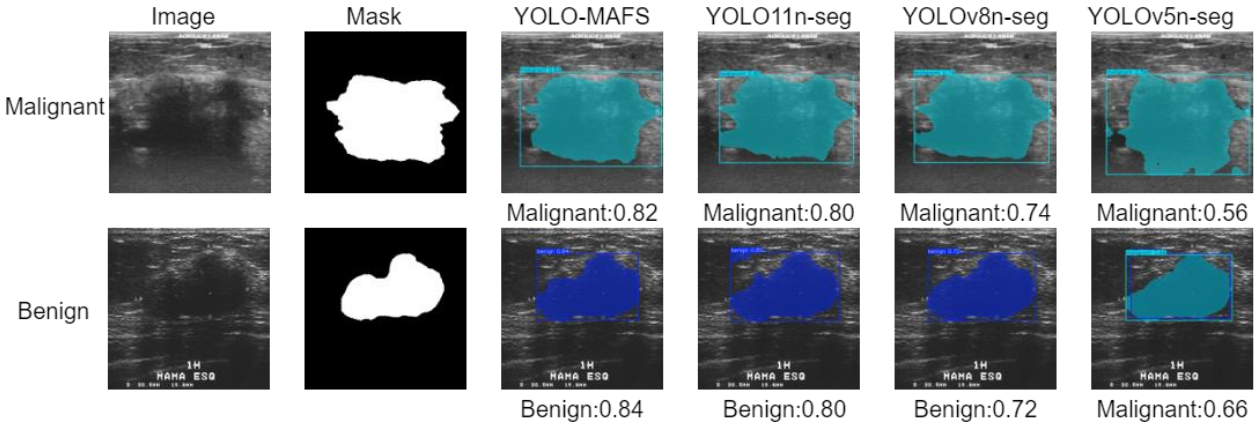

Fig.14. Visual comparison of segmentation results

V. Conclusion

In this study, an innovative model named YOLO-MAFS is proposed for the segmentation task of breast cancer ultrasound images. The model is designed to accurately detect and segment lesions of varying scales with high quality, thereby providing solid technical support for medical-assisted diagnosis. Based on the architecture of YOLOv11, the model has undergone systematic optimization and improvement. Firstly, the Efficient Multi-scale Convolution Module (EMSConv) is incorporated to replace the original C3k2 structure, which enhances the model's multi-scale feature modeling capability. Meanwhile, it effectively reduces the number of parameters and computational burden, thereby improving the model's adaptability to lesions of different sizes. Secondly, a lightweight multi-scale feature fusion module (MSFM) is developed, which adopts Ghost-Shuffle Convolution (GSConv) to achieve efficient semantic fusion. This method can significantly reduce computational complexity while ensuring segmentation accuracy. To strengthen the boundary modeling ability, a Hierarchical Attention Fusion Block (HAFB) is introduced. By integrating local and global attention mechanisms with RepConv, the model's ability to perceive the edges and context of complex lesions is effectively enhanced. Finally, a lightweight asymmetric segmentation head (LASH) is developed. Considering the differences between detection, classification, and mask prediction tasks, an asymmetric branch structure is designed. This structure optimizes the mask generation pathway and improves the representation of segmentation details.

Extensive experimental validation confirms the superior performance of YOLO-MAFS, which achieves significant improvements of 4.2% in mAP@50 and 2.5% in mAP@50-95 compared to YOLOv11. The ablation studies demonstrate that each carefully designed module contributes independently and synergistically to enhance three critical capabilities: multi-scale feature perception, precise boundary modeling, and efficient semantic fusion. Comprehensive generalization experiments validate the robustness of the YOLO-MAFS model across diverse clinical scenarios. Notably, YOLO-MAFS maintains low computational complexity while delivering more accurate breast lesion segmentation than other models of similar scale, effectively reducing both false positives and missed detections. These results underscore its strong generalization ability and substantial clinical potential for breast cancer diagnosis.

Despite these advancements, the model still faces challenges when dealing with lesions that have ambiguous boundaries or complex structures. Future research will focus on developing more refined boundary supervision techniques, enhanced multi-scale attention mechanisms, and optimized mask prediction strategies to further improve the segmentation accuracy and stability for clinically difficult cases.

References

[1] Breast Cancer Facts & Stats 2024 - Incidence, Age, Survival, & More. National Breast Cancer Foundation. (2024). Available: https://www.nationalbreastcancer.org/breast-cancer-facts

[2] R. Wu, X. Lu, Z. Yao, and Y. Ma, "MFMSNet: A Multi-frequency and Multi-scale Interactive CNN-Transformer Hybrid Network for Breast Ultrasound Image Segmentation," *Computers in Biology and Medicine*, vol. 177, Jul. 2024.

[3] G. Lin, M. Chen, M. Tan, L. Chen, and J. Chen, "A dual-stage transformer and MLP-based network for breast ultrasound image segmentation," *Biocybernetics and Biomedical Engineering*, vol. 43, no. 4, pp. 656–671, 2023.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

[6] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," *Medical Image Analysis*, vol. 52, Dec. 2018, pp. 199–214.

[7] X. Qin *et al.*, "U2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognition*, vol. 106, 2020.

[8] J. Chen *et al.*, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[9] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[10] Y. Su, Q. Liu, W. Xie, and P. Hu, "YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms," *Computer Methods and Programs in Biomedicine*, vol. 221, 2022.

[11] W. Li, X. Ye, X. Chen, X. Jiang, and Y. Yang, "A deep learning-based method for the detection and segmentation of breast masses in ultrasound images," *Physics in Medicine & Biology*, vol. 69, no. 15, Jul. 2024.

[12] G. Jocher and J. Qiu, *Ultralytics YOLO11* (Version 11.0.0) [Computer software]. 2024. [Online]. Available: https://github.com/ultralytics/ultralytics.

[13] Siwen Fang, Xinhe Zhang, Bochao Su, and Wenxuan Zhu, "Vehicle-pedestrian Instance Segmentation Algorithm Based on Improved YOLOv8n-seg," *Engineering Letters*, vol.33, no.6, pp1879-1889, 2025.

[14] K. Han, Y. Wang, Q. Tian, J. Guo, and C. Xu, "GhostNet: More features from cheap operations," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1577–1586, 2020.

[15] Y. Peng, D. Z. Chen and M. Sonka, "U-Net V2: Rethinking the Skip Connections of U-Net for Medical Image Segmentation," *2025 IEEE 22nd International Symposium on Biomedical Imaging*, Houston, TX, USA, pp. 1–5, 2025.

[16] H. Li, J. Li, H. Wei *et al.*, "Slim-neck by GSConv: A lightweight-design for real-time detector architectures," *Journal of Real-Time Image Processing*, vol. 21, no. 62, 2024.

[17] S. Xu *et al.*, "HCF-Net: Hierarchical context fusion network for infrared small object detection," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2024.

[18] M. Soudy, Y. Afify, and N. Badr, "RepConv: A Novel Architecture for Image Scene Classification on Intel Scenes Dataset," *International Journal of Intelligent Computing and Information Sciences*, vol. 22, no. 2, pp. 63–73, May. 2022.

[19] G. Bebis and M. Georgiopoulos, "Feed-forward neural networks," *IEEE Potentials*, vol. 13, no. 4, Oct.-Nov. 1994, pp. 27–31.

[20] J. Zhang *et al.*, "Faster and lightweight: An improved YOLOv5 object detector for remote sensing images," *Remote Sensing*, vol. 15, no. 20, 2023.

[21] W. Gómez-Flores, M. J. Gregorio-Calas, and W. C. de Albuquerque Pereira, "BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems," *Medical Physics*, vol. 51, no. 4, pp. 3110–3123, 2024.

[22] M. Sohan *et al.*, "A review on YOLOv8 and its advancements," in *International Conference on Data Intelligence and Cognitive Informatics*, pp. 529–545, 2024.

[23] X. Cong *et al.*, "A review of YOLO object detection algorithms based on deep learning," *Frontiers in Computing and Intelligent Systems*, vol. 4, no. 2, pp. 17–20, 2023.

[24] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, "Dataset of breast ultrasound images," Data in Brief, vol. 28, 2020.