A Cross-Attention and Multilevel Feature Fusion Network for Breast Lesion Segmentation in Ultrasound Images

Guoqi Liu, Yanan Zhou, Jiajia Wang, Zongyu Chen, Dong Liu, *Member, IEEE*, and Baofang Chang

***Abstract*— Lesion segmentation in breast ultrasound images (BUSIs) plays a crucial role in the early diagnosis of diseases. Convolutional neural networks (CNNs) have demonstrated remarkable performance in BUSI segmentation. However, they struggle with capturing long-term dependencies in the images, which can reduce the accuracy of lesion segmentation. We propose a cross-attention and multilevel feature fusion network (CMFF-Net) for breast lesion segmentation, a novel hybrid CNN-transformer architecture. First, we propose a cross-attention feature fusion (CAFF) module that combines the global context information from the transformer with the local spatial information extracted by the CNN, compensating for the lost semantic information for the decoding phase and enhancing the feature representations. Second, we design a feature extraction module (FEM) to refine the local features of CNN. The original features can effectively establish the dependencies between features at the current size latitude. Finally, to improve the network performance, we construct a multilevel feature aggregation (MFA) module that adaptively computes the weights of different feature layers to ensure smoother interaction and integration. Extensive experiments on three public breast ultrasound datasets demonstrate that the proposed CMFF-Net outperforms other state-of-the-art (SOTA) methods.**

***Index Terms*— Breast lesion segmentation, convolutional neural networks (CNNs), parallel biencoder architecture, transformer.**

I. Introduction

BREAST cancer is a disease marked by a high incidence rate and represents a significant health risk for women globally [1]. Early detection and treatment can greatly reduce the mortality rates of breast cancer [2]. At present, convenient and efficient ultrasound examination have become the mainstream method of cancer screening [3]. To improve the efficiency of cancer screening and reduce the workload of doctors, there is a higher demand for automatic segmentation of breast ultrasound images (BUSIs) [4]. However, the similarity in intensity distribution between diseased and nondiseased tissues, the blurring of the edges between the object and the background, and the irregular morphology of the tumor itself pose challenges to the accurate segmentation of the diseased region.

Convolutional neural networks (CNNs) have achieved great success in medical image segmentation due to their powerful nonlinear expression capabilities, particularly with fully convolutional networks (FCNs) [5] and U-Net [6]. In 2018, Almajalid et al. [7] developed a BUSI segmentation framework based on U-Net, which employed preprocessing techniques, including contrast enhancement and speckle reduction, to improve image quality. Subsequently, several improved methods based on the U-Net structure, such as SegNet [8] and U-Net++ [9], have been introduced. However, due to its restricted receptive field, the traditional U-Net has inherent limitations in capturing broader and deeper semantic information. To capture the nonlocal dependencies over the long term and further refine the segmentation accuracy of BUSI, two strategies are proposed: on the one hand, dilated convolution is the most common strategy to expand the receptive field [10], [11] or merging intermediate and high-level features with more task-relevant semantic features. However, Xue et al. [12] pointed out that deeper convolutional layers tend to specialize in local feature extraction, and it is difficult to obtain a truly global view by using dilated convolution on deeper convolutional layers. On the other hand, the strategy is the widespread use of attention mechanisms [13], [14]. Chen et al. [15] proposed an adaptive attention U-Net (AAU-Net), which utilized hybrid adaptive attention to replace the traditional convolution operations. The above two strategies further mitigate the impact of various factors on the segmentation accuracy of breast lesions.

Transformers have demonstrated superior performance in the realm of natural language processing (NLP), researchers have started to use transformers in computer vision tasks, such as image classification [16], [17], [18], [19], image segmentation [20], and object detection [21] and have achieved remarkable results. Although transformers can capture global contextual information well, they pay little attention to local detail features. Moreover, exclusively transformer-based methods can only give satisfactory results on large-scale datasets

Manuscript received 24 February 2024; revised 5 May 2024; accepted 29 May 2024. Date of publication 24 June 2024; date of current version 4 July 2024. This work was supported by the National Natural Science Foundation of China under Grant U1904123 and Grant 61901160. The Associate Editor coordinating the review process was Dr. Yu Yang. *(Corresponding author: Yanan Zhou.)*

Guoqi Liu, Yanan Zhou, Jiajia Wang, Zongyu Chen, and Dong Liu are with the College of Computer and Information Engineering, Henan Normal University, Xinxiang, Henan 453000, China, and also with the Key Laboratory of Artificial Intelligence and Personalized Learning in Education of Henan Province, Xinxiang, Henan 453000, China (e-mail: gqliu@htu.edu.cn; 18437084569@163.com; devinsweet@163.com; chenzongyu1010@163.com; liudong@htu.edu.cn).

Baofang Chang is with the College of Computer and Information Engineering, Henan Normal University, Xinxiang, Henan 453000, China (e-mail: changbaofang@htu.edu.cn).

Digital Object Identifier 10.1109/TIM.2024.3418104

[16]. Recently, some researchers have tried to combine CNN with transformer structures for medical image segmentation. To achieve good results on general datasets, they establish a joint mechanism between CNN focusing on local features and transformer focusing on global context features. These methods fall into three broad categories.

1) Using transformer as encoder and CNN as a decoder to model local and global information. For example, pyramid vision transformers (PVTs) [22] focuses on capturing local information based on global context information, which may easily ignore detailed information such as lesion edge.
2) The integration of CNNs and transformers in a serial architecture, which typically relies on single-scale self-attention mechanisms to capture global context dependencies while ignoring interscale consistency and dependencies. Moreover, the serial network will encounter a decrease in feature reusability as the deepening of the network layer during training. Deep networks with a large number of training parameters are often unstable when dealing with small medical image datasets.
3) The parallel network structure of CNN and transformer, exemplified by multiscale nonlocal feature extraction network (MNEF-Net) [23], usually performs simple fusion (e.g., addition or concatenation) of local and global contexts in a sequential manner, which makes the coarse fusion methods unable to play the advantages of combination.

TransFuse [24] improves the fusion strategy to selectively fuse the features of the two branches for joint prediction. Nonetheless, there are still limitations in extracting richer detailed features, global context information, and accurately identifying lesion regions that closely resemble surrounding tissues.

To overcome these challenges, we design a cross-attention and multilevel feature fusion network (CMFF-Net). It is a parallel dual-branch architecture, which can effectively mitigate the problem of gradient disappearance and feature reduction in serial dual-branch. Moreover, it effectively utilizes the long-term dependencies on the transformer and the local detail representation features of the CNN, which accurately segments the ultrasound breast masses. We adopt the traditional encoder-decoder structure, utilizing pyramid pooling transformer (P2T) [25] and ResNet34 [26] backbones as encoders to extract global and hierarchical features. The decoder comprises the multilevel feature aggregation (MFA) module and upsampling operations. We embed a feature extraction module (FEM) to replace the skip connection, focusing on refining the shallow and deep features extracted by CNN and achieving long-term dependencies. In particular, to further explore the effective coupling between CNN and transformer, we propose a cross-attention feature fusion (CAFF) module, which integrates local feature information from CNN and global context feature information from the transformer. In summary, our contributions are as follows.

1) We propose a parallel encoder framework for BUSI segmentation named CMFF-Net. The framework can capture information at different scales in both local and global dimensions to reduce interference from speckle noise and blurred cancer edge.
2) We engineer a CAFF module, which effectively removes the background noise of the transformer, integrates the transformer's global semantic clues into the pixel of the CNN and supplements the rich global context information during the decoding process.
3) We construct an FEM that employs convolutional blocks with different kernel sizes to capture features at different scales and can effectively integrate local space and high-level semantic context information.
4) To better integrate different levels of features, we design an MFA module to align the output features from different dimensions. This design enhances the sensitivity of the model to alterations across diverse dimensions, thereby markedly enhancing the comprehensive performance of the network.
5) In addition, we conduct extensive experiments on three public ultrasound breast datasets. The experimental results demonstrate the superiority of the model in breast tumor segmentation compared with the existing methods.

II. Related Works

A. CNN-Based Breast Mass Segmentation Methods

In recent years, with the relentless development of deep learning, a succession of CNN models has emerged, including FCN, U-Net, and U-Net++. U-Net++ particularly stood out for its innovative approach to introducing a method of multi-level feature fusion and gradual upsampling, which effectively aggregated the features of different semantic scales in the subnetwork. However, it only contained low-level feature information and ignored the global information of high-level features. In addition, SegNet demonstrated good performance in some medical image segmentation tasks, thanks to its continuous refinement of feature fusion between adjacent regions. Yap et al. [27] used the CNN method to detect breast lesions. They integrated patch-based LeNet, U-Net, and FCN for tumor segmentation.

To further improve the accuracy, Hu et al. [28] proposed to employ dilated convolution within a deeper network to expand the receptive fields in breast tumor segmentation. Cao et al. [29] integrated a set of hybrid dilated convolutions into the network to address the challenges posed by different lesion sizes and shapes. However, merely relying on dilated convolution operations to achieve a larger receptive field does not adequately address the issues caused by adjacent tissues and blurred edges. Chen et al. [15] proposed the AAU-Net framework, which used a hybrid adaptive attention module composed of channel self-attention and spatial self-attention module to replace the convolution operation in traditional U-Net. In addition, Xue et al. [30] developed a global guidance network for breast lesion segmentation by introducing a boundary detection module. Huang et al. [31] used MFA to obtain more semantic information for tumor segmentation and introduced a boundary selection module, thereby refining the contour of breast lesions.

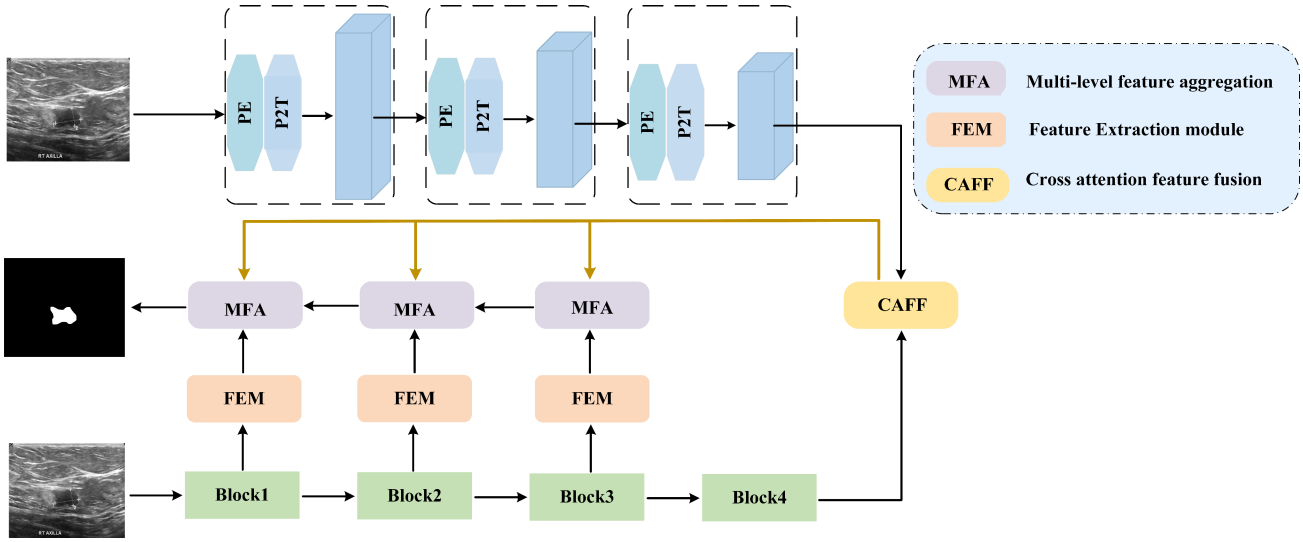


Fig. 1. CMFF-Net. We use P2T and pretrained ResNet34 as the encoder. PE is patch embedding. P2T is a pyramid pooling transformer. FEM is used to refine local features and enhance the correlation between shallower and deeper features. The CAFF aggregates features from CNN and transformer, and the yellow arrow is the CAFF data flow direction. MFA further aggregates features from different levels.

B. Transformer in Medical Image Segmentation

In recent years, with the vigorous development of transformers in the realm of computer vision, more researchers have applied this technology to medical image processing. Dosovitskiy et al. [16] first used the most primitive transformer model for image classification and proposed a structure solely reliant on the self-attention mechanism [vision transformer (ViT)]. Data-efficient image transformer (DeiT) [32] employed the teacher–student strategy grounded in distillation tokens. DeiT was the first to demonstrate that the transformer can be effectively trained and function with medium-sized datasets. Tokens-to-token ViT (T2T-ViT) [33] enabled segment overlapping images, capable of superior spatial context aggregation. To better capture the fine spatial details of different scales, a multilevel ViT, including convolutional vision transformer (CVT) [34] and PVT [35], was proposed. PVT was the inaugural transformer model to feature a pyramid structure.

Compared with PVT, TransFuse introduced a dual-encoder architecture with a parallel structure by integrating CNNs. In particular, TransFuse used a novel fusion method to comprehensively integrate the high-resolution information extracted by CNN and the global information captured by the transformer branch to achieve higher segmentation accuracy. In addition, pyramid pooling has proven to be an effective approach for modeling the long-term feature relationship of deep feature maps. P2T applied the idea of a pyramid pool to the calculation of multihead self-attention (MHSA) in the visual converter, which enhanced its capability to capture rich context information. Pyramid scene parsing network (PSPNet) [36] used pyramid pooling to extract context information from multiple scales in the feature map. Experimental results showed its effectiveness in semantic segmentation. Considering the advantages of pyramid pooling and P2T and taking full advantage of the self-attention mechanism and hierarchical structure, it is worth exploring introducing pyramid pooling into BUSI segmentation.

III. METHODS

In this article, we propose a dual-branch network, termed CMFF-Net, to effectively solve the problem of BUSI segmentation. Fig. 1 provides a schematic of the method. The structure consists of three parts (encoder, decoder, and skip connection). Considering the effectiveness of the P2T in modeling long-term dependencies and obtaining robust context features, as well as the superiority of pyramid pooling, helps the network in efficiently capturing fine and rough details in the input, we use it as one of the two branches to capture the long-distance dependencies of the feature map. We use pretrained ResNet34 as one of the encoders to extract local features and details. To better integrate the feature information, we design CAFF to aggregate the features of these two scales and suppress redundant information and background noise while extracting valuable semantic information. In addition, we employ FEM to refine local details and capture global information. Subsequently, features at various scales are sent to the corresponding decoder. In the decoding stage, we construct the MFA to smoothly fuse features across different levels, thereby enhancing the composite performance of the model.

A. CAFF Module

After extracting both local and global feature information of the dual backbone network, how to effectively couple these two types of features are crucial for the segmentation task. We observe that most of the fusion of CNNs and transformers is through concatenation or addition operations. However, series fusion may lead to the output features are not refined enough, and some information will be lost in the subsequent decoding process. Therefore, to further tap the potential of the CNNs and transformers joint mechanism and compensate for the global spatial information, we propose the CAFF module to guide the global features in the decoding stage. The CAFF module uses cross-attention to process global spatial information in more detail at two scales, thereby significantly improving the segmentation accuracy of breast malignant lesion areas.

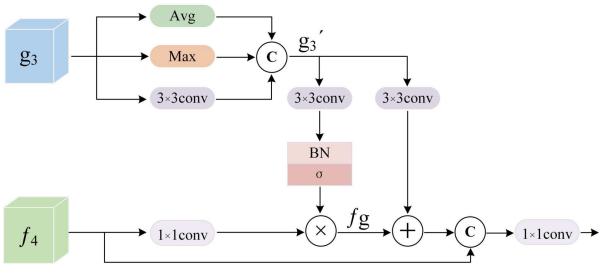


Fig. 2. CAFF module. g_3 represents the feature of P2T, and f_4 represents the feature from ResNet34.

As shown in Fig. 1, we first use P2T as the backbone to extract the feature maps of the four stages, the feature correspondence of each stage g_j, $j \in \{1, 2, 3, 4\}$, when $j = 1$, it is the low-level feature of the encoder output. When $j = \{2, 3\}$, they represent the middle two layers of features through the encoded output. When $j = 4$, it indicates that the output is a high-level feature. The purpose of deploying the CAFF module is to extract rich global context features, and f_4 is the deep feature of CNN, which contains more global features. The deep feature g_4 output from the transformer branch will lose most of the information due to the low resolution, so we only use the first three layers of P2T features. For the CNN side, we use ResNet34 as the backbone to extract a set of feature maps, which can be expressed as f_i, $i \in \{1, 2, 3, 4\}$. f_i represents the shallow to deep features and corresponds to the corresponding encoder output feature layer.

The details of Fig. 2 are as follows. First, we considered that since g_3 we use is a mid-level feature, it might incorporate noise, so we use adaptive average pooling and maximum pooling operations to suppress the noise of g_3. 3×3 CBR [3×3 convolutional layer, batch normalization (BN), and rectified linear unit (ReLU)] operation is applied to the g_3 feature, and three branches are concatenated to the channel to obtain g_3'. The 1×1 CBR operation is applied to the feature graph f_4 and then multiplied by the g_3 after 3×3 convolution, BN, and sigmoid operations to obtain the feature graph f_g. In addition, 3×3 CBR operation is performed on g_3', and then added with feature f_g to suppress background interference that may cause abnormalities. Finally, we use the original feature f_4 to add the feature obtained in the above step to preserve the original information. The features are extracted and denoised using a convolution operation with a 1×1 CBR. The formula is written as follows:

$$g_3' = \mathrm{Cat}\left[\mathrm{AVP}(g_3), \mathrm{Max}(g_3), \mathrm{conv}_3(g_3)\right] \quad (1)$$

$$f_g = \sigma\left(\mathrm{BN}\left(\mathrm{conv}\left(g_3'\right)\right)\right) \times \mathrm{conv}_1(f_4) \quad (2)$$

$$\mathrm{output} = \mathrm{conv}_1\left[\mathrm{Cat}\left[\left(\mathrm{conv}_3\left(g_3'\right) + f_g\right), f_4\right]\right] \quad (3)$$

where $\mathrm{Cat}(\cdot)$ represents the concatenation operation, and $\mathrm{conv}_1(\cdot)$ is the 1×1 CBR. $\mathrm{conv}_3(\cdot)$ is the 3×3 CBR. $\mathrm{AVP}(\cdot)$ is a global adaptive pooling, $\mathrm{Max}(\cdot)$ is the maximum pooling, and σ represents the sigmoid.

As depicted in Fig. 3, the pooling operations effectively eliminate the noise caused by the low-level feature g_3, and the high-level feature f_4 has more global context features. Using the CAFF module, high-level features can be used as prior knowledge to guide the attention of low-level features to the object area. By fusing the features of the two branches, we can capture rich contextual features and then make up for the lost semantic information for the subsequent decoding stage.

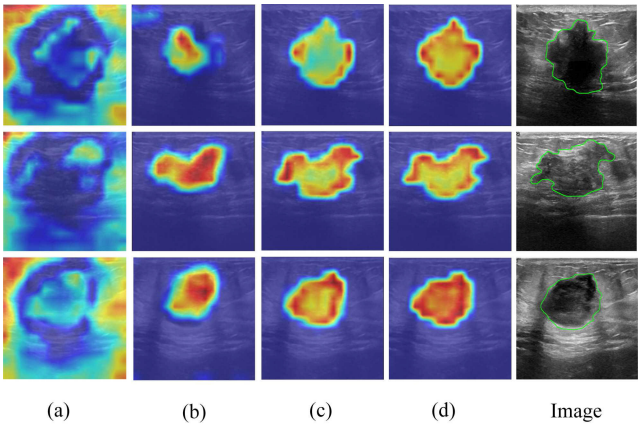


Fig. 3. Image illustrates the heatmaps of some convolution layers within the CAFF module. (a) Heatmap of the 1×1 convolutional layer (1×1 convolution + BN + ReLU). (b) and (c) Heatmaps of the 3×3 convolutional layer. (d) Heatmap of the whole module.

B. Feature Extraction Module

The self-attention method [37], [38] has been widely used to calculate the context relationship, which can better represent the features of convolutional layer learning and the relationship between spatial regions. However, they can only provide limited context information for outputting weighted feature maps of the global context. We propose FEM to replace the skip connection part of the traditional U-Net. Learning different scale features by convolutional blocks with different kernel sizes can effectively integrate local spatial information.

In fact, the low-level features f_1–f_3 contain more local features. In this way, we facilitate a more effective fusion of local features and partial spatial information, thereby enhancing the sensitivity of the model to the object. Therefore, the FEM takes the local feature map $f_i, i \in \{1, 2, 3\}$ on the CNN side as input and generates a weighted feature map containing the context relationship between pixels from the perspective of local space. First, we employ different convolution kernels to extract more detailed features from the original input features. The original features can effectively establish the dependencies between features at the current size latitude through different convolution kernels. Furthermore, compared with the self-attention mechanism, we add the residual connection, which can retain the key details of the original feature.

Fig. 4 illustrates our proposed FEM. For the input feature map (H, W, and C represent the height, width, and channel dimension, respectively), we use convolution kernels with convolution kernels of 1×1 and 3×3 to change the channels. Then, we define the features after applying 1×1 convolution as M and L, and the features after applying 3×3 convolution as N. Subsequently, M and L are reshaped, and N is reshaped and transposed, respectively. The product of N and L after the operation generates a size mapping of $HW \times HW$, and the mapping is normalized to obtain an attention map A. Then, we perform matrix multiplication on the attention map A and the reshaped M, and the result is reshaped into

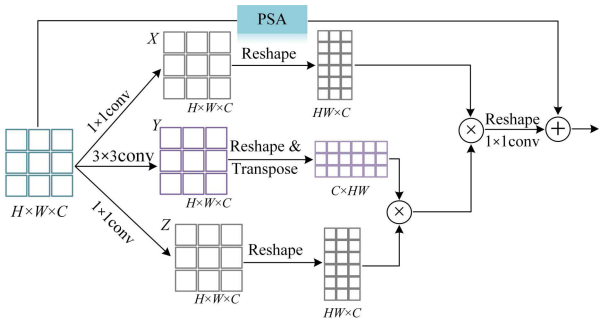


Fig. 4. FEM. PSA is polarized self-attention, the 1 × 1 CBR (1 × 1 convolutional layer, BN, and ReLU).

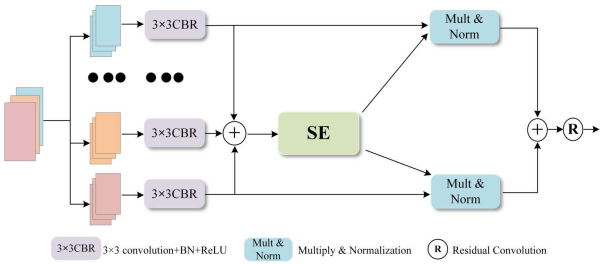


Fig. 5. MFA module. SE stands for squeeze-and-excitation modules.

$C \times H \times W$. After the 1 × 1 convolution operation, it is further added to the input feature map of the polarized self-attention (PSA) module [39] to generate a weighted feature map. It is worth noting that we utilize the PSA module to process the residual data stream. The intention is to mitigate the noise in the encoded shallow features, thereby enhancing the representation of the original features

$$D_i = \text{conv}(\text{Reshape}(M \cdot A)) + \text{PSA}(F_i). \tag{4}$$

C. MFA Module

Low-level features serve as a crucial foundation for enhancing high-level features. They have high resolution and help locate small objects. The lower resolution of higher level features facilitates the discovery of large objects. It is critical to merge the features from different levels to achieve accurate and robust segmentation.

However, direct add or concatenation operations may produce more redundant information. Inspired by SKNet [45], we propose to use the MFA module to adapt to different acceptance domains, not just relying on activation functions or channel additions. The main goal is to compress and aggregate three or more linear layers to calculate the corresponding weight coefficients for each input feature layer. MFA utilizes the group convolution to implement an adaptive weight calculation method for different feature layers, guaranteeing a smoother fusion. This adaptive fusion process uses the complementary information between different feature layers to ensure that the generated guidance information plays a greater role in the model training process.

Fig. 5 shows the details of the MFA module. First, we upsample or downsample the feature maps to make them have the same size. Then, according to the number of channels, the input feature maps are divided into several groups of subfeature maps. We perform the 3 × 3 convolution operation on each sub-feature map and then perform an add operation on the features of these branches. We employ the squeeze-and-excitation (SE) [40] module to assign different weights to different positions in the image from the perspective of the channel domain to explicitly model the interdependence between channels and improve the feature representation in the channel dimension. We multiply the SE features by the features from the 3 × 3 convolution operation and the features obtained from the convolution kernel are reweighted. Finally, the dimension is adjusted through a residual convolution block to obtain the final output feature map.

IV. Experiments

A. Datasets

We evaluate the proposed method on three public breast ultrasound datasets, including BUSI [41], BUS [42], and STU [43]. The BUSI dataset was created by AI-Dhabiani et al. [41], who collected data on 600 female patients at Baheya Hospital. It includes 780 masses, 647 abnormal cases (210 malignant and 437 benign masses), and 133 normal cases. In the experiment, we use abnormal samples for training, validating, and testing. It is worth noting that we divide the BUSI dataset into three datasets: BUSI-benign, BUSI-malignant, and BUSI-fusion. The BUS dataset is collected by Yap et al. [42], including 163 BUSIs. The third dataset STU provided by Zhuang et al. [43]. The STU dataset comprises 42 BUS images, each with an average size of 128 × 128 pixels, sourced from the Imaging Department of the First Affiliated Hospital of Shantou University. Since the STU dataset contains too few images, it is only used as external validation data to evaluate the generalization performance of the segmentation network.

B. Implementation Details and Evaluation Indicators

All experiments are implemented in the PyTorch framework, and the development environment is Ubuntu 20.04, python 3.9, and PyTorch 1.11. The model is trained on a single NVIDIA GeForceGTX 3090 GPU. During training, we choose the AdamW optimizer to update the network parameter, and the initial learning rate is 1e−4. The training period of the model is set to 150. In the 70th, 100th, and 120th periods, the learning rate gradually decreased by 0.5. In our experiment, all images, including training, validating, and testing images, are uniformly adjusted to 256 × 256. To avoid overfitting and improve the generalization ability, we also carried out several data enhancement methods, including horizontal flipping, vertical flipping, translation, and random rotation.

To verify the superiority and structural validity, we selected six widely used segmentation metrics to evaluate the accuracy of BUSI segmentation, including dice similarity coefficient (Dice), union intersection (IoU), precision (Pre), accuracy (ACC), sensitivity (SE), and specificity (Sp), for comprehensive evaluation. Mathematically, these metrics can be expressed as follows:

$$\text{Dice} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{5}$$

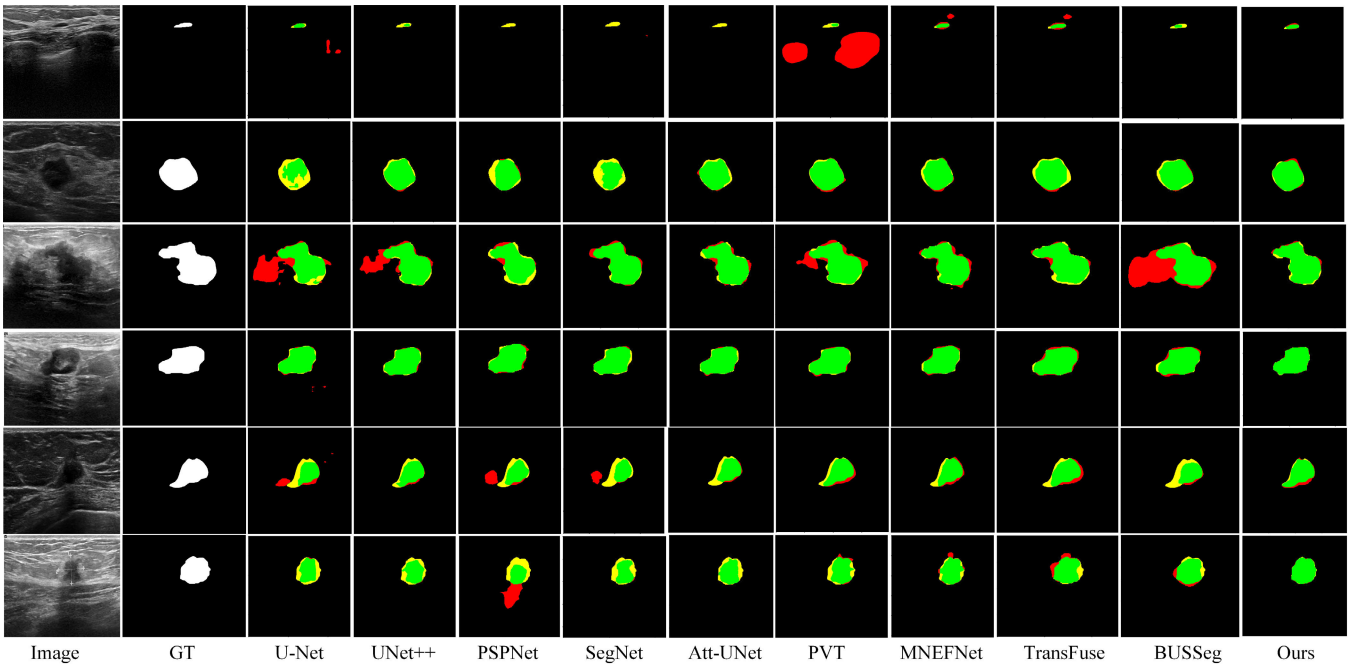


Fig. 6. Compare our model with the results of various SOTA methods on the full dataset. The green, red, and yellow areas represent the true positive, false positive, and false negative, respectively.

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{6}$$

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Acc} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

$$\text{SE} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{10}$$

where TP and TN represent the number of correctly segmented breast lesion pixels and background pixels, respectively. FP indicator is incorrectly segmented as the background pixel of the breast lesion pixel. FN is the breast lesion pixels that are incorrectly predicted as background pixels.

C. Comparison Experiments

To evaluate the superiority of the proposed method, we compare the method with nine other state-of-the-art (SOTA) methods, including U-Net, U-Net++, AAU-Net, PSPNet, SegNet, PVT, TransFuse, MNEF-Net, and BUSSeg [44]. For a fair comparison, we implemented all competitors in the same experimental environment. To validate the generalization ability of the method, we employ the BUSI-fusion dataset training weight to test and evaluate the effectiveness of each method on the STU dataset. To ensure a fair comparison, we conduct a fivefold cross validation to evaluate the test performance of all methods. It can be seen from Tables I–III that CMFF-Net has a relative advantage over other SOTA models in terms of feature modeling capabilities.

Compared with the traditional convolution, due to the advantages of adding location information and relying on a self-attention mechanism, the transformer can effectively capture the global context information. We can see from Tables I–III that the transformer-based methods achieve better performance in breast mass segmentation and generalization. CMFF-Net integrates the strengths of both CNNs and transformers, go a step further in segmentation performance compared to PVT, which relies solely on a pure transformer encoder. Moreover, as illustrated in Fig. 6, the similarity between the object and the background of the ultrasonic image, some error information is segmented together with the object to a certain extent. Compared with others, the method obtains the best segmentation results, in which our segmented results are closest to ground truth (GT).

Furthermore, we compared the mean and standard deviation (SD) of Dice, IoU, Pre, ACC, SE, and SP values with the most advanced methods. To test for statistical significance differences, we conduct paired t-tests between our method with other methods based on the Dice and IoU metrics, with a P_value less than 0.05 indicating significantly differences between CMFF-Net and the comparison method. Tables I–III show the detailed comparison results for the five datasets. Based on the quantitative results, we can see that the model has achieved the best results in the evaluation indicators. Notably, the Dice, IoU, Pre, ACC, SE, and SP indexes on the BUSI-malignant dataset reached 76.80%, 67.10%, 74.65%, 93.32%, 84.46%, and 94.75%, respectively. The two main indicators of Dice and IoU, compared with the second-ranked method, CMFF-Net increases Dice by 2.61% and IoU by 2.43% on the BUS dataset. For Dice and IoU on the BUSI-benign dataset, CMFF-Net improves by 0.53% and 0.29% compared with the suboptimal method. Compared with the suboptimal method on the BUSI-fusion dataset, the Dice and IoU indexes of the method are improved by 0.45% and 0.84%,

TABLE I

STATISTICAL COMPARISON WITH SOTA METHODS ON THE BUS AND BUSI-MALIGNANT DATASETS. WE CONDUCT FIVEFOLD CROSS VALIDATION AND REPORT THEIR AVERAGE RESULTS (MEAN ± SD)

Methods	Year	BUS						BUSI-malignant					
		Dice(%)	IoU(%)	Pre(%)	Acc(%)	SE(%)	SP(%)	Dice(%)	IoU(%)	Pre(%)	Acc(%)	SE(%)	SP(%)
U-Net	2015	61.10±5.89	51.38±4.43	79.09±11.74	97.44±1.38	64.11±12.45	99.26±0.42	62.54±4.09	49.69±3.66	69.06±5.22	90.72±0.92	63.63±4.39	95.80±0.81
PSPNet	2016	53.86±6.30	43.32±5.89	79.16±12.13	97.49±1.30	52.62±10.78	99.41±0.24	66.41±4.40	54.76±4.24	75.22±3.86	92.41±1.91	66.24±6.67	96.18±1.22
SegNet	2017	60.60±5.44	50.03±5.84	74.79±8.96	97.71±1.46	59.02±10.16	99.55±0.15	66.90±3.40	55.09±3.10	**79.35±4.59**	92.60±1.36	62.70±4.83	97.35±1.09
UNet++	2018	70.05±4.37	61.37±3.97	86.78±6.43	98.17±1.24	69.82±6.99	99.51±0.21	70.99±4.16	60.01±3.7	76.24±3.60	92.85±0.69	73.58±7.05	96.05±0.96
PVT	2021	77.34±10.67	67.45±11.99	71.11±15.13	97.89±1.14	93.12±4.28	98.33±1.24	74.75±4.53	64.18±5.21	70.53±6.80	92.43±1.09	**86.02±4.78**	93.24±1.66
TransFuse	2021	80.10±6.30	70.39±7.54	77.01±10.56	98.06±1.02	90.63±4.28	98.75±0.76	75.24±4.52	64.47±4.30	73.17±6.58	92.95±1.25	82.08±5.28	94.46±1.88
AAU-Net	2022	68.65±8.04	60.46±7.45	**89.05±3.47**	98.24±1.27	68.24±9.52	**99.72±0.06**	69.47±4.99	59.04±4.57	78.76±2.87	93.29±0.77	67.67±5.55	**97.37±0.86**
MNEF-Net	2023	82.49±3.3	73.34±3.85	77.93±5.59	98.43±0.96	92.61±2.87	98.91±0.31	73.06±3.05	62.27±2.63	67.96±2.72	91.82±0.54	85.54±4.80	92.67±1.53
BUSSeg	2023	78.54±6.69	69.83±6.86	56.97±14.11	98.15±0.87	82.19±3.04	98.96±0.62	71.32±2.29	59.86±2.21	52.36±4.08	92.43±1.82	80.41±4.77	94.00±2.63
Ours	—	**85.10±2.77**	**75.77±3.09**	81.15±2.82	**98.78±0.87**	**93.51±2.10**	99.31±0.11	**76.80±2.72**	**67.10±1.95**	74.65±1.74	**93.32±0.48**	84.46±5.10	94.75±0.73
P_value		<0.05(Dice), <0.05(IoU)						<0.05(Dice), <0.05(IoU)					

TABLE II

STATISTICAL COMPARISON WITH SOTA METHODS ON THE BUSI-BENIGN AND BUSI-FUSION DATASETS. WE CONDUCT FIVEFOLD CROSS VALIDATION AND REPORT THEIR AVERAGE RESULTS (MEAN ± SD)

Methods	Year	BUSI-benign						BUSI-fusion					
		Dice(%)	IoU(%)	Pre(%)	Acc(%)	SE(%)	SP(%)	Dice(%)	IoU(%)	Pre(%)	Acc(%)	SE(%)	SP(%)
U-Net	2015	66.78±7.42	57.67±7.51	73.51±8.18	96.38±0.99	71.10±9.29	98.54±0.90	67.47±2.90	57.40±3.07	77.04±3.36	94.46±1.00	68.59±5.05	98.31±0.39
PSPNet	2016	67.98±7.25	58.66±5.88	76.62±5.15	96.6±0.72	68.99±12.39	98.45±0.72	73.90±2.37	64.17±2.37	78.67±3.35	95.77±0.24	75.87±1.35	97.94±0.28
SegNet	2017	67.91±10.55	59.27±10.36	83.6±3.52	96.56±1.41	66.17±12.24	99.10±0.34	71.57±2.40	62.24±3.01	81.91±2.58	95.61±0.68	71.78±3.07	**98.60±0.28**
UNet++	2018	71.84±6.87	64.61±6.63	82.91±3.77	96.83±0.84	73.74±7.20	98.64±0.45	76.52±1.63	67.64±2.13	80.97±2.31	95.78±0.53	79.04±4.07	97.98±0.62
PVT	2021	80.47±4.22	73.21±4.52	78.59±4.25	97.03±0.90	85.87±6.57	97.92±0.71	81.24±2.75	72.68±3.48	78.99±4.22	96.41±0.59	**88.28±3.68**	97.33±0.74
TransFuse	2021	79.47±3.78	71.51±4.08	78.39±4.57	96.80±0.70	85.85±3.66	97.63±3.15	81.47±1.69	73.06±2.14	81.87±2.50	96.36±0.51	85.48±2.15	97.51±0.43
AAU-Net	2022	72.03±7.84	65.40±7.20	**84.61±3.00**	97.00±0.81	73.92±9.32	**98.78±0.48**	75.67±3.84	67.48±4.05	82.86±1.77	95.86±0.64	78.28±3.00	98.20±0.30
MNEF-Net	2023	79.47±2.50	71.65±2.07	78.14±1.31	96.89±0.48	85.63±5.77	97.75±0.84	80.74±1.93	72.29±1.99	78.61±2.93	96.04±0.87	87.80±3.55	96.92±0.85
BUSSeg	2023	74.70±6.43	67.14±6.43	52.17±5.49	96.98±0.62	78.78±7.60	98.41±0.48	78.46±2.15	69.51±2.96	65.51±4.73	96.19±0.69	83.15±1.57	97.60±0.38
Ours	—	**81.00±3.13**	**73.50±3.69**	79.49±2.57	**97.29±0.48**	**86.38±3.78**	98.29±0.17	**81.92±1.92**	**73.90±2.30**	**83.70±2.42**	**96.48±0.53**	84.57±2.83	98.01±0.38
P_value		<0.05(Dice), >0.05(IoU)						<0.05(Dice), <0.05(IoU)					

TABLE III

STATISTICAL COMPARISON WITH SOTA METHODS ON THE STU DATASET (MEAN ± SD)

Methods	Year	Param(M)	STU					
			Dice(%)	IoU(%)	Pre(%)	Acc(%)	SE(%)	SP(%)
U-Net	2015	31	69.17±6.90	56.70±7.52	78.55±16.56	93.77±2.29	70.38±12.99	97.67±1.92
PSPNet	2016	46	78.43±1.23	66.98±1.55	79.97±4.10	95.76±0.22	82.64±2.59	97.39±0.65
SegNet	2015	29	73.89±5.22	62.14±4.81	**85.30±4.43**	95.05±0.48	71.26±8.49	**98.47±0.57**
UNet++	2018	47	80.31±3.23	70.66±3.39	83.78±2.36	96.12±0.55	82.34±5.18	98.17±0.45
PVT	2021	25	87.09±0.64	78.07±1.03	82.37±2.22	97.29±0.23	**94.70±1.08**	97.64±0.38
TransFuse	2021	26	86.83±0.66	77.71±1.07	82.33±1.73	97.16±0.25	94.21±0.50	97.43±0.37
AAU-Net	2022	35	81.10±1.65	70.80±2.10	79.41±1.03	96.04±0.42	88.02±1.71	97.58±0.18
MNEF-Net	2023	58	82.56±7.44	72.04±9.38	77.96±7.21	95.81±1.48	91.58±8.97	96.55±0.78
BUSSeg	2023	26	85.61±4.38	75.99±5.89	74.61±8.71	96.84±5.77	90.24±4.78	97.43±1.06
Ours	—	37	**87.30±0.74**	**78.30±1.20**	83.17±2.08	**97.30±0.22**	94.12±0.96	97.57±0.38
P_value			<0.05(Dice), >0.05(IoU)					

respectively. In addition, compared with TransFuse using the attention mechanism, CMFF-Net is improved by 5.0% and 5.38% on the BUS dataset on Dice and IoU, respectively. On the BUSI-malignant dataset, the improvements are 1.56% for Dice and 2.63% for IoU, respectively. Our method achieves higher segmentation accuracy. As shown in Table III, the

TABLE IV
STATISTICAL COMPARISON OF ABLATION STUDIES ON BUSI-MALIGNANT, BUS, BUSI-BENIGN, AND BUSI-FUSION DATASETS. WE CONDUCT FIVEFOLD CROSS VALIDATION AND REPORT THEIR AVERAGE RESULTS (MEAN ± SD)

Group	Methods	BUS		BUSI-malignant		BUSI-benign		BUSI-fusion	
		Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)
a	Baseline	75.87±5.44	64.23±5.73	74.63±4.09	64.09±3.26	73.18±**2.03**	63.69±**2.30**	75.41±3.16	65.71±3.29
b	Baseline+FEM	80.70±5.70	70.47±6.41	76.57±3.72	66.79±3.47	79.23±3.80	71.52±4.14	80.47±**1.65**	71.79±**2.02**
	Baseline+MFA	82.09±5.31	72.68±5.63	76.60±4.01	67.16±3.33	79.25±3.62	71.43±3.71	80.32±2.07	72.07±2.62
	Baseline+CAFF	76.36±5.96	64.93±6.61	75.22±5.03	64.95±4.22	76.13±3.33	66.49±3.30	79.55±2.45	70.54±2.72
c	Baseline+FEM+MFA	81.17±4.60	71.47±4.88	76.68±4.21	**67.44**±3.29	80.06±4.08	72.38±4.18	80.92±2.06	72.64±2.43
	Baseline+FEM+CAFF	80.63±4.26	70.80±4.27	76.54±3.19	66.54±2.44	79.84±3.47	71.93±3.73	80.73±2.22	72.14±2.76
	Baseline+CAFF+MFA	81.75±5.94	72.26±6.06	76.65±4.40	66.34±3.47	79.74±4.50	72.43±4.76	81.13±2.12	72.82±2.65
d	Baseline+CAFF+MFA+FEM	**85.05±3.05**	**75.77±3.45**	**76.86±2.72**	67.10±**1.95**	**80.98**±3.50	**73.50**±4.13	**81.92**±2.14	**73.91**±2.55

generalization of our method has also yielded favorable outcomes. For the key indicators Dice and IoU, CMFF-Net achieves 0.21% and 0.23% increase in Dice and IoU compared with the suboptimal approach.

Regarding the statistical significance of the BUSI-malignant, BUSI-fusion, and BUS datasets, compared with the second-best method, our method has P_value of less than 0.05 on the Dice and IoU, demonstrating the reliability of significance. On the BUSI-benign dataset, the P_value of our method IoU is exceeds 0.05, indicating that the performance on this index is comparable to that of PVT. Nevertheless, for Dice, our approach is superior to competitors. Overall, the experimental results affirm that CMFF-Net delivers better performance than other segmentation networks on ultrasound public datasets.

Fig. 6 illustrates the visual segmentation results of different segmentation methods on the five datasets. Given the similar gray distribution, blurred edge, and irregular tumor morphology, accurate segmentation of ultrasound images is challenging. The selected images present at least one of the above challenges. From the third and fourth lines of Fig. 6, it is evident that our method excels in edge segmentation, delivering better results. This further shows that the method is superior to other methods in accurately identifying tumors in similar organ tissues and providing good segmentation results for irregular objects. In addition, the overall visual segmentation results of Fig. 6 indicate that CMFF-Net can achieve better segmentation results in segmenting tumors of different sizes. This also proves the potential of the method to improve the accuracy of breast tumor segmentation in clinical environments, especially in solving specific challenges posed by ultrasound images.

D. Ablation Study

To evaluate the effectiveness of different components in the proposed method, we conducted an ablation study to verify the effectiveness of the modules CAFF, FEM, and MFA. Initially, we removed the MFA module from CMFF-Net. In the skip connection part, we use the traditional direct addition instead of our FEM module. In addition, we use the concatenate operation instead of the dual-branch fusion CAFF module, relying on ResNet34 and P2T dual backbone as the baseline network. Subsequently, we incrementally reintroduce these components to the baseline. The performance of the baseline model with the addition of each module and the benchmark model combined with two modules is compared with the BUSI-malignant, BUS, BUSI-benign, and BUSI-fusion datasets.

To more clearly highlight the effectiveness of the designed modules in handling different tasks of the model, we divide the ablation experiments into four groups. In groups b and c, we present the experimental outcomes for the individual and the combined modules, respectively. The bold font indicates the best performance among all experimental groups, while the underlined font indicates the best performance within each group. The comparative results are tabulated in Table IV, and the segmentation outcomes are illustrated in Fig. 7.

We also perform attention heatmap visualization for each module of the model, as detailed in Fig. 8. It can be seen from Table IV and Fig. 7 that CMFF-Net is significantly better than the baseline in segmentation results and combination with other modules. The combination of the baseline and each module demonstrates the individual effectiveness of these modules and their collective contribution to improved segmentation accuracy. To enhance the clarity of the ablation experiment, we designed two bar graphs as a visual representation of the two key metrics of Dice and IoU on the BUSI-malignant, BUS, BUSI-benign, and BUSI-fusion datasets, as shown in Figs. 9 and 10.

1) Effectiveness of CAFF: Specifically, the visual comparison of the ablation experiment is presented in Fig. 7, which contains several typical challenging cases with blurred edges and similar appearance of the lesion and nonlesion areas. CAFF is used to fuse the local features extracted by CNN and the global information from the transformer. As indicated in the second line of Table IV, for the Dice, the combination of baseline + CAFF increased by 0.59% and

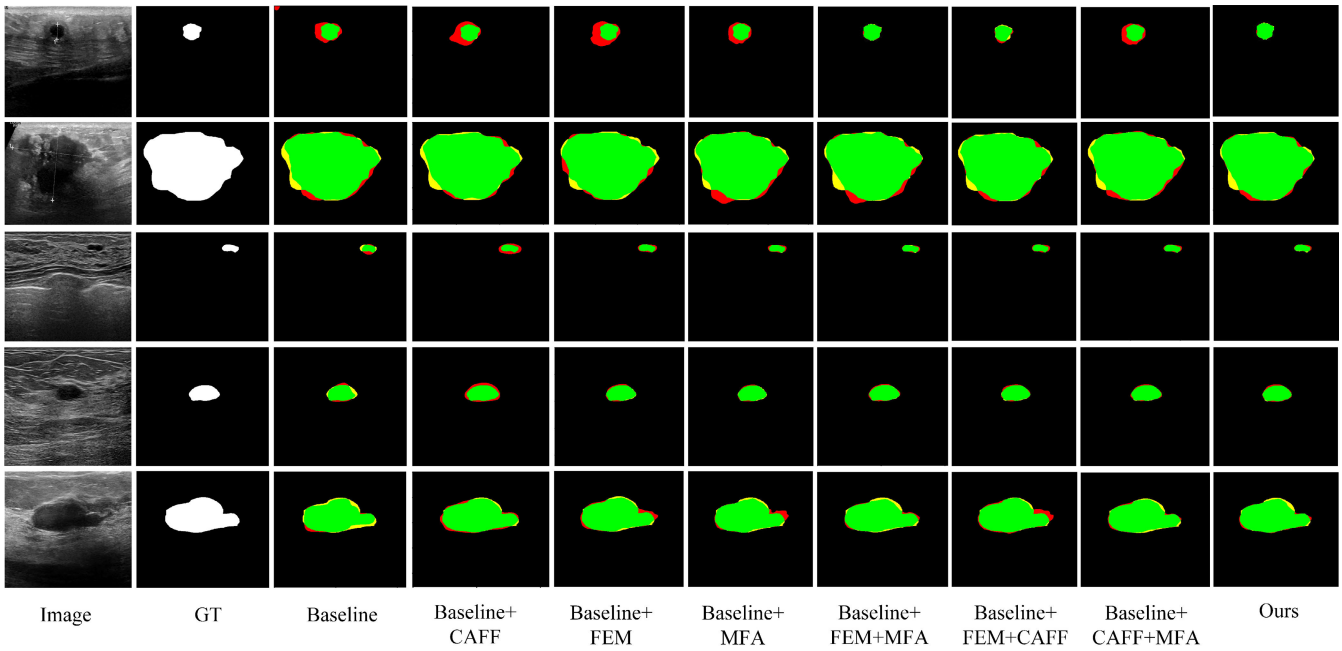


Fig. 7. Segmentation results of ablation experiments. Green, red, and yellow regions represent the true positive, false positive, and false negative, respectively.

0.49% on the BUSI-malignant and BUS datasets, respectively, and the IoU increased by 0.86% and 0.7%, respectively. We can observe from the visualization results shown in Fig. 7 that the baseline + CAFF has more true positive than the baseline, and the edge contour of the lesion can also be better positioned. Furthermore, on the BUSI-benign and BUSI-fusion datasets, Dice is increased by 2.95% and 4.14%, and IoU is increased by 2.8% and 4.83%, respectively. The attention maps in Figs. 3 and 8 demonstrate that the CAFF module effectively suppresses and eliminates noise while capturing a richer global context, clearly indicating the effectiveness of CAFF components. In addition, it can better integrate CNN local detail features and transformer long-term dependencies by incorporating the CAFF module. This integration compensates for the diluted spatial information during the decoding upsampling, thus improving performance.

2) Effectiveness of FEM: To highlight the local details of the object, we use the FEM instead of the traditional skip connection to enhance the feature representation on the CNN side. As evidenced by the BUSI-benign and BUSI-fusion datasets in Table IV, the baseline + FEM approach has yielded significant improvements over the baseline, with Dice increasing by 6.05% and 5.06%, and IoU increased by 7.83% and 6.08%, respectively. It can also be seen from Fig. 7 that baseline + FEM performs better in locating lesion edges than baseline, further indicating that FEM represent CNN local spatial features more effectively than traditional skip connections. The attention heatmaps in Fig. 8 indicate that FEM_1 and FEM_2 excel in the extraction of both local and global features, albeit with a higher noise content. FEM_3 assigns greater emphasis to the object area; yet, it still falls short of achieving precise focus on the object.

3) Effectiveness of MFA: In the decoder stage, we deploy the MFA module to establish the connection between high-level and low-level features, effectively preserving the low-level feature representation. At the same time, the adaptive weight calculation method of MFA for different feature layers ensures smoother fusion. Adding MFA to the baseline, it can be seen from Table IV that it significantly improves performance. The Dice and IoU increased by 6.22% and 8.45% on the BUS dataset, respectively. The attention map reveals the gradual concentration of attention from MFA_1 to MFA_2 and then to MFA_3, and the gradual enhancement of attention to boundaries. As depicted in the first line of Fig. 7, MFA can focus on the image edge region and effectively eliminate the redundant information in the image edge.

4) Effectiveness of Module Combination: In group c, the combination of baseline + CAFF + MFA and baseline + CAFF + FEM effectively utilizes the advantages of the two modules. (It is emphasized that in the baseline + CAFF + FEM, we use the fusion method of addition operation instead of the MFA module). The Dice on the BUSI-malignant dataset is 2.02% and 1.91% higher than that of the baseline, and the IoU coefficients are 2.25% and 2.45% higher than the baseline, which exceeds the baseline + CAFF. On the BUS dataset, the Dice is 5.95% and 4.83% higher than the baseline, and the IoU is 8.03% and 6.57% higher than the baseline. Compared with baseline + CAFF, and baseline + MFA, the Dice of baseline + CAFF + MFA increased by 3.61% and 0.49% on the BUSI-benign dataset. On the BUSI-fusion dataset, Dice is improved by 1.58% and 0.83%. For baseline + FEM, and baseline + MFA, the Dice of baseline + FEM + MFA is increased by 0.83% and 0.81% on the BUSI-benign dataset. On the BUSI-fusion dataset, Dice is improved by 0.45% and 0.6%. Both are better than the combination of baseline and individual modules.

In addition, in Table IV, we can see that for Dice and IoU, the baseline + CAFF + FEM is 0.32% and 0.56% lower than our method on the BUSI-malignant dataset, respectively. It indicates that the use of simple concatenate operations for

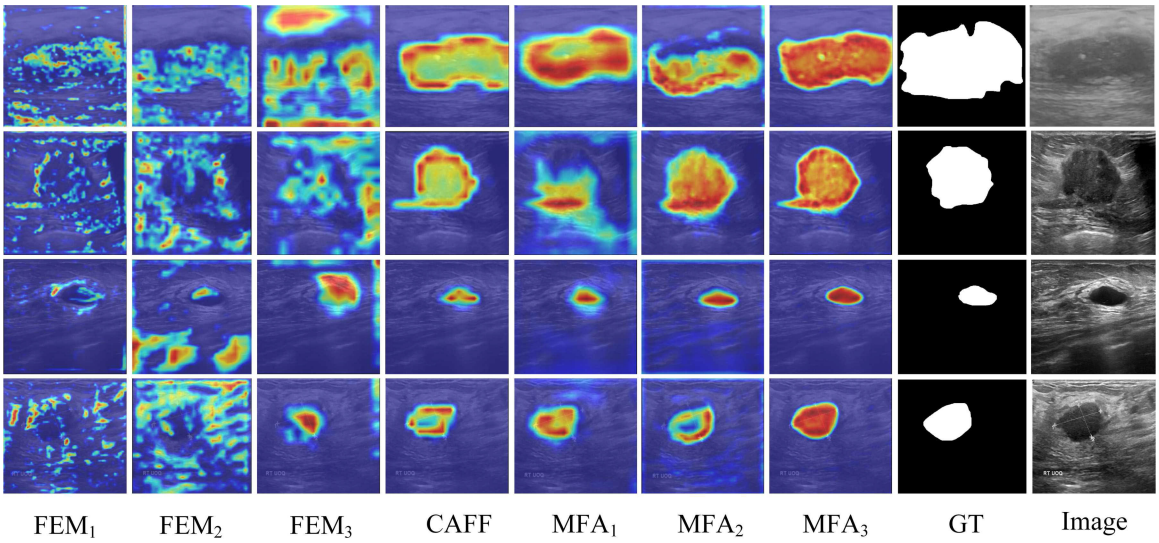


Fig. 8. Attention heatmaps of FEM_i, CAFF, and MFA_i modules in the model, where darker colors represent the higher attention.

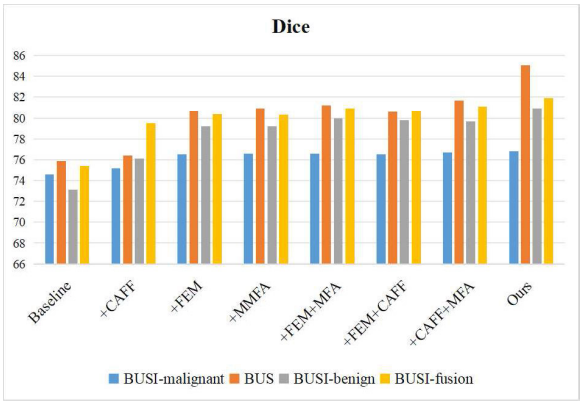


Fig. 9. Bar graph shows the ablation experiment of CMFF-Net, with Dice metrics on the BUSI-malignant, BUS, BUSI-benign, and BUSI-fusion datasets.

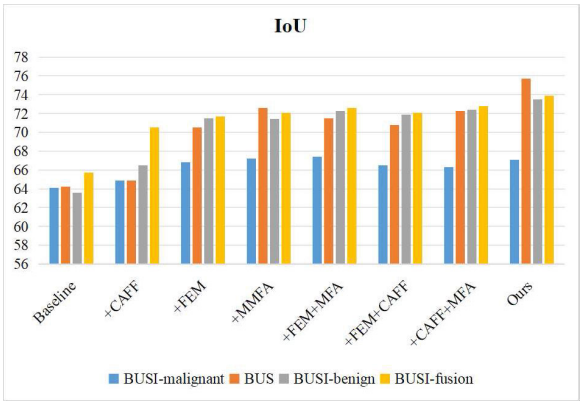


Fig. 10. Bar graph shows the ablation experiment of CMFF-Net, with IoU metrics on the BUSI-malignant, BUS, BUSI-benign, and BUSI-fusion datasets.

fusion does not achieve good results. Through the baseline + FEM + MFA, we can see the effectiveness of our MFA fusion method. On the BUS dataset, the baseline + CAFF + FEM, baseline + CAFF + MFA, and baseline + FEM + MFA decrease by 4.42%, 3.30%, and 3.88%, respectively, compared with our method on Dice. In the ablation experiment in group b, we can observe that FEM is the most stable in the BUSI-fusion task and has the highest Dice metrics. In contrast, the MFA module performs best on BUSI-malignant and BUS datasets. It is worth noting that due to the challenging processing of ultrasound datasets, the features extracted from module combinations may be redundant, which may adversely affect model performance. Especially in the BUSI-malignant dataset, we find that Dice and IoU metrics of baseline + FEM + CAFF are lower than baseline + FEM. However, as depicted in the fifth line of Fig. 7, the integration of the MFA module allows the model to effectively eliminate redundant information, thus improving the accuracy of segmentation results. The results of ablation experiments in group c further confirm the effectiveness of FEM and MFA modules in helping models understand the gap between different tasks. At the same time, the reduction in the SD of the ablation experiment in group c indicates that the model is more stable and robust, allowing it to achieve optimal performance on different tasks. Furthermore, the results of the experiment in group c further prove that the CAFF module extracts global information, which plays a critical role in the information interaction between modules.

It can be observed from Fig. 8 that when the similarity between the object and the background boundary is strong, each module can locate the object at different stages and pay good attention to the lesion area. It fully illustrates the validity and applicability of each module in the model.

V. Computational Efficiency

In this section, we compare the complexity of the proposed model with other methods, including the number of trainable parameters (Params) and floating-point operations (GFLOPs). CNNs enhance their feature extraction capability by superimposing layers. The deeper the network the larger the parameters, which leads to increased computational expenditure. U-Net++ utilizes a purely convolutional

TABLE V
COMPARISON RESULTS OF THE NUMBER OF PARAMETERS AND GFLOPS BETWEEN OUR MODEL AND OTHER SOTA MODELS

Methods	Year	Type	Params(M)	GFLOPs(G)
U-Net	2015	CNN	31	48
PSPNet	2016	CNN	46	34
SegNet	2017	CNN	29	40
UNet++	2018	CNN	47	152
AAU-Net	2022	CNN	35	66
PVT	2021	Transformer	**25**	**5**
TransFuse	2021	Hybrid	26	9
MNEF-Net	2023	Hybrid	59	9
BUSSeg	2023	Hybrid	26	**5**
CMFF-Net	Ours	Hybrid	37	8

architecture, which results in an extensive parameter count and a high level of GFLOPs. As illustrated in Table V, the MNEF-Net boasts 59M parameters, which is approximately 1.56 times the parameter count of our model. With 37M parameters and 8G GFLOPs, the segmentation performance of our approach is significantly better than that of the pure CNN architecture. BUSSeg and PVT have smaller parameters, but their segmentation performance is not as good as ours. Therefore, our hybrid structure can guarantee good results while maintaining reasonable computational costs, the model achieves an optimal trade-off between performance and model complexity in terms of parameters and GFLOPs.

VI. DISCUSSION

In this work, we propose a novel breast tumor segmentation model, named CMFF-Net. The results of the experimental indicators from Tables I–III and the visual segmentation results can be seen from Fig. 6. The existing methods have shortcomings in concurrently capturing fine-grained details and global contextual information. U-Net and its variants use pure convolution architecture, which is inherently less effective in encoding global features. Although AAU-Net attempts to address this issue by incorporating channel and spatial self-attention mechanisms to encapsulate global information, its performance in segmenting lesion areas with diverse shapes and high object-background similarity remains less than ideal. The hybrid architecture of TransFuse, MNEF-Net, and BUSSeg is adopted to give play to the ability of global features of the transformer and local features of CNN, and relatively good segmentation results are obtained.

However, they lack the performance of visual tasks in the case of blurred edges, variable tumor morphology, and large lesion distribution areas. It can be seen from the results of the ablation study in Table IV that our model shows its advantages on both datasets mentioned. We design a CAFF module to improve the interaction between CNN local features and transformer global features. CAFF can selectively emphasize important features and supplement rich global information to each decoder layer. Through the ablation study in Table IV, it can be concluded that the CAFF module has better feature capture ability, especially for malignant datasets. In addition, we use the FEM to replace the traditional skip connection, which can better refine the features of the local edge and alleviate the problem of feature loss. We further verified the effectiveness of the FEM in Table IV ablation study. In particular, we deploy the MFA module to highlight the difference between the background and the object so that the network can focus on the edge and enhance the feature representation.

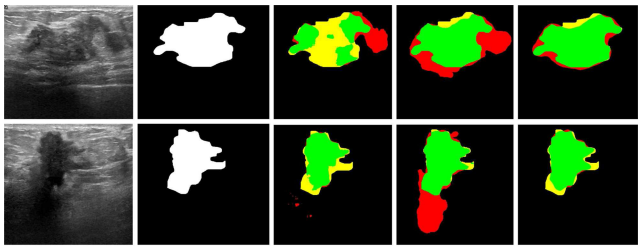


Fig. 11. Failure cases for competitors and our method. Green, red, and yellow regions represent the true positive, false positive, and false negative, respectively.

Finally, we conducted a comparative study with the SOTA method on five experimental datasets. The method achieves the best results on the six evaluation indicators for BUSIs. Our generalization has also achieved certain advantages. However, similar to other SOTA segmentation methods, our method may still not be able to deal with too much noise or too complex brightness distribution, as shown in Fig. 11. Nevertheless, even in the case of failure, CMFF-Net is still superior to other existing methods, especially in edge processing and has fewer error prediction areas.

VII. CONCLUSION

In this article, we propose a segmentation network named CMFF-Net. We introduce an FEM to refine local features on the CNN side, selectively emphasize important features and suppress irrelevant features, and establish remote dependencies. In addition, we design CAFF to fuse the features of CNN and transformer. The CAFF module uses high-level features as a priori knowledge to direct the attention of low-level features to the correct location of the object. In particular, we construct an MFA module to adaptively compute weights using group convolution to assign more weights to critical regions of the object, further improving the overall performance. We conduct experiments on three public datasets, including BUS, BUSI (BUSI-benign, BUSI-malignant, and BUSI-fusion), and STU datasets to evaluate CMFF-Net performance and compare it with the SOTA models. The results demonstrate that the proposed method achieves a good trade-off in the model complexity and segmentation performance.

In the field of medical image processing, due to the limited number of images in the dataset, we use various data enhancement methods to avoid potential overfitting risks. These methods include random rotation and random flipping. However, ultrasound segmentation of breast malignant lesions segmentation of images is still a challenging task. Given that the breast ultrasound dataset is small, but the breast cancer image data imaging ways are diverse. Therefore, we will consider the use of multimodal ideas to further achieve higher segmentation accuracy with low computational cost.

REFERENCES

[1] A. G. Waks and E. P. Winer, "Breast cancer treatment: A review," *J. Amer. Med. Assoc.*, vol. 321, no. 3, pp. 288–300, 2019.

[2] H. D. Cheng, X. J. Shi, R. Min, L. M. Hu, X. P. Cai, and H. N. Du, "Approaches for automated detection and classification of masses in mammograms," *Pattern Recognit.*, vol. 39, no. 4, pp. 646–668, Apr. 2006, doi: 10.1016/j.patcog.2005.07.006.

[3] H. D. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern Recognit.*, vol. 43, no. 1, pp. 299–317, Jan. 2010.

[4] M. Xian, Y. Zhang, H. D. Cheng, F. Xu, B. Zhang, and J. Ding, "Automatic breast ultrasound image segmentation: A survey," *Pattern Recognit.*, vol. 79, pp. 340–355, Jul. 2018, doi: 10.1016/j.patcog.2018.02.012.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[7] R. Almajalid, J. Shan, Y. Du, and M. Zhang, "Development of a deeplearning-based method for breast ultrasound image segmentation," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, 2018, pp. 1103–1108.

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[9] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2019.

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[11] L. C. Chen, G. Papandreou, and I. Kokkinos, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Jun. 2017.

[12] C. Xue et al., "Global guidance network for breast lesion segmentation in ultrasound images," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101989.

[13] X. Qu, Y. Shi, Y. Hou, and J. Jiang, "An attention-supervised full-resolution residual network for the segmentation of breast ultrasound images," *Med. Phys.*, vol. 47, no. 11, pp. 5702–5714, Nov. 2020.

[14] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[15] G. Chen, L. Li, Y. Dai, J. Zhang, and M. H. Yap, "AAU-Net: An adaptive attention U-Net for breast lesions segmentation in ultrasound images," *IEEE Trans. Med. Imag.*, vol. 42, no. 5, pp. 1289–1300, 2022.

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

[17] C. F. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, Aug. 2021, pp. 357–366.

[18] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 10012–10022.

[19] T. Huang, L. Huang, S. You, F. Wang, C. Qian, and C. Xu, "LightViT: Towards light-weight convolution-free vision transformers," 2022, *arXiv:2207.05557*.

[20] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2021, pp. 7262–7272.

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection withtransformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[22] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-PVT: Polyp segmentation with pyramid vision transformers," 2021, *arXiv:2108.06932*.

[23] G. Liu, J. Wang, D. Liu, and B. Chang, "A multiscale nonlocal feature extraction network for breast lesion segmentation in ultrasound images," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.

[24] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*, Cham, Switzerland: Springer, 2021, pp. 14–24.

[25] Y. H. Wu, Y. Liu, X. Zhan, M. M. Cheng, A. K. Davison, and R. Marti, "P2T: Pyramid pooling transformer for scene understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12760–12771, Aug. 2022.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 16, Jun. 2016, pp. 770–778.

[27] M. H. Yap et al., "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 4, pp. 1218–1226, Jul. 2017.

[28] Y. Hu et al., "Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model," *Med. Phys.*, vol. 46, no. 1, pp. 215–228, Jan. 2019.

[29] X. Cao, H. Chen, Y. Li, Y. Peng, S. Wang, and L. Cheng, "Dilated densely connected U-Net with uncertainty focus loss for 3D ABUS mass segmentation," *Comput. Methods Programs Biomed.*, vol. 209, Sep. 2021, Art. no. 106313.

[30] C. Xue et al., "Global guidance network for breast lesion segmentation in ultrasound images," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101989.

[31] R. Huang et al., "Boundary-rendering network for breast lesion segmentation in ultrasound images," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102478.

[32] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jgou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, vol. 39, no. 12, pp. 10347–10357.

[33] L. Yuan et al., "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2021, pp. 558–567.

[34] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.

[35] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.

[36] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 2017, pp. 2881–2890.

[37] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[38] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Oct. 2019, pp. 3146–3154.

[39] H. Liu, F. Liu, X. Fan, and D. Huang, "Polarized self-attention: Towards high-quality pixel-wise regression," 2021, *arXiv:2107.00782*.

[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

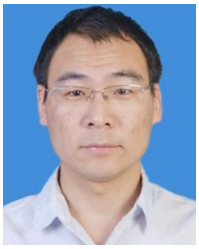
[41] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, Feb. 2020, Art. no. 104863.

[42] M. H. Yap et al., "Breast ultrasound region of interest detection and lesion localisation," *Artif. Intell. Med.*, vol. 107, Jul. 2020, Art. no. 101880, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365719306670.

[43] Z. Zhuang, N. Li, A. N. Joseph Raj, V. G. V. Mahesh, and S. Qiu, "An RDAU-NET model for lesion segmentation in breast ultrasound images," *PLoS One*, vol. 14, no. 8, 2019, Art. no. e0221535.

[44] H. Wu, X. Huang, X. Guo, Z. Wen, and J. Qin, "Cross-image dependency modelling for breast ultrasound segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 6, pp. 1619–1631, Jun. 2023.

[45] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.



Guoqi Liu received the Ph.D. degree from the School of Electronic and Information Engineering, South China University of Technology (SCUT), Guangzhou, China, in 2013.

He is currently an Associate Professor with the College of Computer and Information Engineering, Henan Normal University, Xinxiang, China. His research interests include image segmentation, machine learning, and partial differential equation.

Yanan Zhou is currently pursuing the master's degree with the Department of Computer Science and Information Engineering, Henan Normal University, Xinxiang, China, under the supervision of Prof. Guoqi Liu.

Her research interests include image segmentation and machine learning.

Jiajia Wang is currently pursuing the master's degree with the Department of Computer Science and Information Engineering, Henan Normal University, Xinxiang, China, under the supervision of Prof. Guoqi Liu.

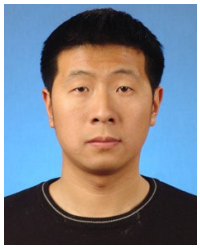
His research interests include image segmentation and machine learning.

Zongyu Chen is currently pursuing the master's degree with the Department of Computer Science and Information Engineering, Henan Normal University, Xinxiang, China, under the supervision of Prof. Guoqi Liu.

His research interests include image segmentation and machine learning.

Dong Liu (Member, IEEE) received the B.S. and M.S. degrees in computer science from Zhengzhou University, Zhengzhou, China, in 1998 and 2004, respectively, and the Ph.D. degree in computer science from Tianjin University, Tianjin, China, in 2013.

He is currently a Professor with the College of Computer and Information Engineering, Henan Normal University, Xinxiang, China. His research interests include educational data mining and complex network analysis.



Baofang Chang received the B.Sc. degree in information and computing science, the M.S. degree in computer software and theory, and the Ph.D. degree in applied mathematics from Lanzhou University, Lanzhou, China, in 2005, 2007, and 2011, respectively.

He is currently working with the College of Computer and information Engineering, Henan Normal University, Xinxiang, China. His research interests include optimization algorithm, machine learning, and edge computing.