HAU-Net: Hybrid CNN-transformer for breast ultrasound image segmentation

Article · January 2024

DOI: 10.1016/j.bspc.2023.105427

7 authors, including:

Huaikun Zhang
Lanzhou University
5 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Jing Lian
Lanzhou Jiaotong University
46 PUBLICATIONS **514** CITATIONS
SEE PROFILE

Yi Zetong
Lanzhou University
2 PUBLICATIONS **33** CITATIONS
SEE PROFILE

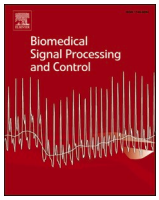
Ruichao Wu
Lanzhou University
4 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Biomedical Signal Processing and Control 87 (2024) 105427

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



HAU-Net: Hybrid CNN-transformer for breast ultrasound image segmentation

Huaikun Zhang [a], Jing Lian [b], Zetong Yi [a], Ruichao Wu [a], Xiangyu Lu [a], Pei Ma [a], Yide Ma [a,*]

[a] School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu, China
[b] School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, Gansu, China

ARTICLE INFO

Keywords:
Convolutional Neural Network
Transformer
Breast lesions segmentation
Ultrasound image

ABSTRACT

Breast cancer is a significant health concern that remains one of the leading causes of mortality in women worldwide. Convolutional Neural Networks (CNNs) have been shown to be effective in ultrasound breast image segmentation. Yet, because of the lack of long-distance dependence, the segmentation performance of CNNs is limited in addressing challenges typical of segmentation of ultrasound breast lesions, such as similar intensity distributions, the presence of irregular objects, and blurred boundaries. In order to overcome these issues, several studies have combined transformers and CNNs, to compensate for the shortcomings of CNNs with the ability of transformers to exploit long-distance dependence. Most of these studies limited themselves to rigidly plug transformer blocks into the CNN, lacking consistency in the process of feature extraction and therefore leading to poor performances in segmenting challenging medical images. In this paper, we propose HAU-Net(hierarchical attention-guided U-Net), a hybrid CNN-transformer framework that benefits from both the long-range dependency of transformers and the local detail representation of CNNs. To incorporate global context information, we introduce a L-G transformer block nested into the skip connections of the U shape architecture network. In addition, to further improve the segmentation performance, we added a cross attention block (CAB) module on the decoder side to allow different layers to interact. Extensive experimental results on three public datasets indicate that the proposed HAU-Net can achieve better performance than other state-of-the-art methods for breast lesions segmentation, with Dice coefficient of 83.11% for BUSI, 88.73% for UDIAT, and 89.48% for BLUI respectively.

1. Introduction

Breast cancer is characterized by high incidence and represents a great threat to women's health worldwide. In 2022, in the United States, breast cancer cases accounted for the highest proportion (31%) of all new cases of cancer in females, with a fatality rate of 15%, second only to that of lung cancer [1]. Early detection of symptoms and diagnostic treatment is therefore essential to provide timely care and improve survival rates [2]. Ultrasound imaging is widely applied in the early screening of breast cancer due to its convenience, low cost and non-radioactivity [3]. Automated breast ultrasound image segmentation has emerged as a promising technology that can potentially aid radiologists in enhancing the accuracy of breast cancer diagnosis. Accurate segmentation of lesions from breast ultrasound (BUS) images could help doctors characterize and locate tumors, enhancing the precision and speed of diagnosis [4]. Yet, due to similar intensity distributions, blurred boundaries, and irregular tumor morphology, it is challenging to accurately segment the lesion area. Several researchers are currently committed to designing new accurate and efficient segmentation methods.

Convolutional Neural Networks (CNNs) have been successful in image processing thanks to their powerful nonlinear expression ability. In 2015, Long et al. proposed FCN [5] which was the first method to use the CNNs for image semantic segmentation. Subsequently, U-Net [6] achieved excellent results in the field of medical image segmentation by leveraging its skip connections combined with an encoder-decoder network structure. Yap et al. [7] first applied the FCN variant network for breast lesion segmentation and outperformed traditional methods. In 2018, Almajalid et al. [8] developed a segmentation framework for breast ultrasound images based on U-Net which applied pre-processing techniques including contrast enhancement and speckle reduction to improve image quality.

* Corresponding author.
E-mail address: ydma@lzu.edu.cn (Y. Ma).

https://doi.org/10.1016/j.bspc.2023.105427
Received 24 April 2023; Received in revised form 26 July 2023; Accepted 12 September 2023
Available online 21 September 2023

Subsequently, a series of segmentation networks based on the encoder-decoder architecture have emerged. The U-shaped architecture based on U-Net has become the most widely used network structure in the medical image segmentation field. However, the traditional U-Net lacks higher-level contextual information and is restricted in its ability to capture deeper and wider semantic information due to the limited receptive field. Most of the new CNN-based segmentation models improve in two ways: by increasing the receptive field [9,10,15,16] and adding attention mechanisms [14,20]. Dilated convolution is a common approach to enlarge the receptive field. By setting different dilation factors and combining them, multi-scale feature maps with different receptive fields can be obtained. Attention mechanisms generally include spatial attention, channel attention, or a combination of the two. The aim is to enhance feature representation by adjusting weights to focus on crucial information. Although CNNs have been shown to be well in the field of breast tumor image segmentation, however, due to limitations of convolution itself, CNNs have difficulties in capturing long-range dependencies, causing poor segmentation or missed detection of irregular breast masses.

On the other hand, as transformers achieved state-of-the-art performance in the NLP domain [21], researchers have started to use transformers in computer vision tasks such as image classification [22,23,27,28], image segmentation [25,26], object detecting [24], and achieved remarkable results. Compared with CNN-based methods, transformers mainly leverage a Multi-head Self-Attention (MSA) mechanism to replace the convolution operation, which can effectively build long-range dependencies between a sequence of tokens and capture the global context [29]. A large challenge of transformers is that they need pre-training on large datasets to generalize and their quadratic complexity. In order to address this, SWIN-VIT [27] proposed the window based multi-head self-attention (W-MSA) and shifted window based multi-head self-attention (SW-MSA). W-MSA first partitions the feature map into multiple non-overlapping windows, then performs self-attention computation independently in each window. SW-MSA enables information interaction across different windows by a shifted window partitioning approach.

Due to the excellent ability of transformers to extract context semantic information, many researchers have applied them to the field of medical image segmentation making impressive progress in segmentation tasks of different human organs and lesions. TransUNet [30] was the first attempt of transformers in medical image segmentation which employs a hybrid CNN-Transformer architecture to leverage both the detailed high-resolution spatial information from CNN features and the global context encoded by transformers. After that, many studies followed TransUNet to build a hybrid CNN-Transformer architecture. A common approach is to embed the transformer block into the process of feature extraction to replace the high-level convolution modules. Yet, this practice results in lack of consistency in the process of feature extraction. CNNs and transformer modules can only perform respectively in the low level and high level, therefore cannot be fused effectively. In addition, they lack the information about interaction between features at multiple scales.

In order to overcome these challenges, we designed a hybrid CNN-transformer framework that effectively leverages the long-range dependency of transformers and the local detail representation feature of CNNs to perform accurate segmentation of ultrasound breast mass images. We followed a traditional U-shape architecture that involves an encoder, a decoder and several skip connections. A ResNet [39] backbone was developed as the encoder to extract the hierarchical features, and the decoder consisted of residual blocks and upsampling layers. Between the encoder and the decoder, we embedded a L-G transformer block to replace the skip connection, aiming to further refine the feature map extracted by the CNN and to achieve long-distance dependence. Compared to the naive vision transformer, the proposed L-G transformer Block could reduce the computational complexity while realizing competitive segmentation performance. Following the ResNet structure, there are four residual layers. L-G transformer Blocks will be added at the output of layer1 to 3 and subsequently concatenated with the corresponding up-sampling layers. In addition, we introduced a Cross Attention Block (CAB) to capture global dependency among different layers of features.

The motivation behind our research is twofold. Firstly, traditional CNN-based methods for ultrasound image segmentation often struggle to capture long-range dependencies in breast ultrasound images, particularly for tumors with complex and irregular shapes, diffuse boundaries, or intricate structures. Secondly, While transformers excel at modeling long-range dependencies, directly replacing CNNs or inserting transformers only at the last encoding layers presents challenges. Fully replacing CNNs may hinder capturing crucial local detailed information necessary for accurate segmentation. On the other hand, simply inserting transformers at the deepest layer is insufficient to capture long-range information from multi-scale features and it may potentially disrupt the original encoder's consistency. This can result in suboptimal segmentation performance.

Our proposed HAU-Net addresses these challenges by synergistically integrating the advantages of CNNs and transformers. By fusing the local detail representation feature of CNNs and the long-range dependency of transformers, HAU-Net aims to achieve accurate and efficient segmentation of breast ultrasound mass images.

To summarize, our contributions can be summarized as follows:

- We proposed HAU-Net for breast ultrasound image segmentation, which is a hierarchical hybrid CNN-transformer framework that efficiently fuses the long-range dependency of transformers with the local detail feature representation of CNNs.
- We proposed a Local-Global transformer block that includes both a local transformer module and a global transformer module. This novel design reduces computational complexity while maintaining competitive segmentation performance. The local transformer module captures local context information efficiently, while the global transformer module enables long-range dependency modeling, improving the network's ability to segment breast masses accurately.
- To capture global context information among multi-scale features from different layers, we introduced a Cross Attention Block (CAB). This module enhances feature integration and refines segmentation results, further improving the overall performance of HAU-Net.
- In addition, we obtained extensive experimental results on three public ultrasound breast datasets demonstrating the superiority of our method on breast tumors segmentation compared to existing methods.

2. Related work

2.1. CNN-based breast mass segmentation methods

In recent years, many researchers have proposed novel techniques for breast tumor segmentation using deep networks. For instance, Hu et al. [9] introduced dilated convolution in deeper network layers to increase the number of receptive fields in breast tumor segmentation. Byra et al. [13] developed a selective kernel U-Net that adjusts the network's receptive field via an attention mechanism and combines feature maps extracted with dilated and conventional convolutions. Xue et al. [12] proposed a global guidance block to aggregate non-local features and also introduced a breast lesion area boundary detection module for high-quality segmentation results with precise boundaries. Huang et al. [18] not only used multi-level feature fusion to obtain more semantic information for breast ultrasound image segmentation, but also introduced a boundary selection module to enable the network to automatically focus on the edge regions without additional supervision. It also equipped with a GCN-based boundary rendering module to convert the tuberculosis boundary into graph data for further feature

attribute mining of the entire contour. Lou et al. [17] introduced a new solution to overcome semantic gaps in U-shaped semantic segmentation networks, which result from the lack of compatibility between encoder and decoder features. Their approach, called the multi-level context refinement network (MCR-Net), includes two context refinement blocks: the inverted residual pyramid block (IRPB) and the context-aware fusion block (CFB). These blocks aim to improve the contextual relationships between the encoder and decoder features. Tang et al. [11] presented a feature pyramid non-local network (FPNN) that fuses multilevel features, considering long-range dependencies by combining the non-local module and the feature pyramid network. Chen et al. [14] introduced a hybrid adaptive attention module consisting of a channel self-attention module and a spatial self-attention module to replace the traditional convolution operation in their U-Net network.

2.2. Transformers in medical image segmentation

There have been several recent developments in medical image segmentation using a combination of convolutional neural networks (CNNs) and transformers. For example, LeViT-UNet [32] combines the U-Net architecture with a LeViT transformer module for fast and accurate segmentation. TransFuse [35] uses transformers and CNNs in a parallel style to capture both global dependencies and low-level spatial details. UNETR [33] introduced a novel architecture, dubbed as U-Net Transformers (UNETR), that utilizes a transformer as the encoder to learn sequence representations of the input volume and effectively capture the global multi-scale information, while also following the "U-shaped" network design for the encoder and decoder. Swin-unet [31] is a transformer-based network for medical image segmentation, using the Swin Transformer for the encoder, bottleneck, and decoder. MedT [34] operates on both the entire image and image patches to learn local and global features. The gated Axial-Attention model extends the existing CNNs architecture by introducing additional control mechanisms in the self-attention module. HiFormer [37] combines a CNN-based encoder with the Swin Transformer module to efficiently bridge CNNs and transformers for medical image segmentation. He et al. [38] proposed a cloud-based medical image segmentation method based on multi-feature extraction and interactive fusion. This approach leverages cloud computing to handle vast amounts of medical images and combines Transformers and CNN to extract global and local features, respectively. Xu et al. [39] introduced a hybrid feature extraction network that integrates CNN and Transformers, leveraging their respective advantages in feature extraction to enhance medical image segmentation performance. Zhu et al. [40] proposed a novel deep learning-based brain tumor segmentation method by jointly utilizing deep semantics and edge information in multimodal MRI. Specifically, they present a semantic segmentation module based on an improved Swin Transformer by introducing the shifted patch tokenization strategy for better training. X-Net [41] utilized convolutional networks and transformers to form dual encoding branches for feature extraction, simultaneously obtaining local and global features, thereby achieving superior results in medical image segmentation. Ma et al. [45] developed the ATFE-Net, which combines an axial Transformer with a feature enhancement-based CNN for ultrasound breast mass segmentation. In comparison to these transformer-based models, our approach inherits both the strengths of convolution in encoding precise spatial information and the long-distance dependence from the transformer.

3. Method

In this section, we first introduce the overall framework of HAU-Net and then discuss the proposed components and modules of the network.

3.1. Overview

The overall architecture of the HAU-Net is depicted in Fig. 1: it follows the classic U-shape structure of U-Net [6] which consists of three parts (encoder, decoder and skip-connections). Our aim was for our model to effectively extract local features of images while having the ability to learn long-range semantic information. We used resnet34 [42] as the backbone to extract local features and details of the feature maps on the encoder side, getting four residual block layers with different

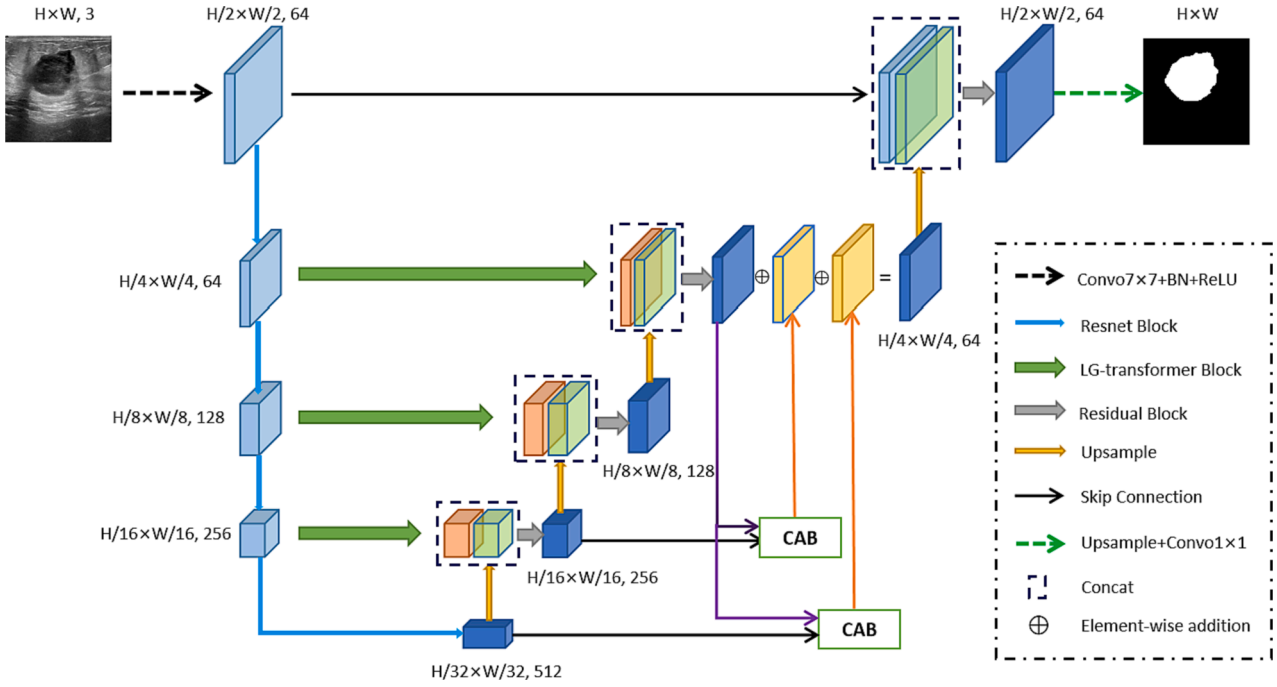


Fig. 1. The architecture of the proposed HAU-Net which is composed of encoder, decoder and skip connections. The skip connections of layer1 to layer3 are replaced by the proposed L-G transformer Blocks which are shown as the green arrows. Two CABs are utilized on the decoder side to obtain cross attention between different layers.

dimensions: layer 1 to 4. Every step in the decoder is composed of an upsampling of the feature map from the previous layer, a concatenation with the corresponding feature map from the encoder side, and a residual block which consisted of numerous combinations of Batch Normalization (BN), ReLU, and convolution layers. In order to capture long-range dependency of the feature maps, we nested the LG-transformer Block into the skip connections of layers 1 to 3. The output of the LG-transformer Block was then concatenated with the output of the corresponding up-sampling layers on the decoder side. Results then passed a residual block and up-sample to the next layer. We also proposed two CAB modules on the decoder side to obtain cross attention between different layers. In the following sections, we will detail the LG-transformer Block and CAB module.

3.2. L-G transformer block

The L-G transformer Block consisted of three main modules: a Local-Global Multi-head Self-Attention (L-G MSA), a Multi-Layer Perceptron (MLP) and a Squeeze-and-Excitation block (SE Block) (Fig. 2a), aiming to distill long-range context information from the output of the encoder.

3.2.1. Local multi-head Self-Attention (LMSA)

The proposed LMSA was inspired from the windows-based multi-head self-attention (WMSA) of SWIN-VIT [27]. Given an input feature map $X \in R^{H\times W\times C}$, we first divide it into $\frac{H\times W}{S^2}$ non-overlapping windows where S denotes the window size, and then applied multi-head self-attention within each window. The LMSA can be described as:

$$X_L = LMSA(X) = softmax\left(\frac{Q_X K_X^T}{\sqrt{d}}\right)V_X \tag{1}$$

where Q_X, K_X, V_X have been projected from X and denote the query, key and value matrices, and d denotes the dimension of the Q_X, K_X, V_X.

The LMSA could obtain the local dependence information of the feature map at a small computational cost. Yet, the lack of information interaction between windows led to the absence of the long-range dependencies. To address this issue, we proposed the Global MSA, which could add global dependencies and large receptive fields at a minimal computational cost.

3.2.2. Global multi-head self-attention (GMSA)

The GMSA and the LMSA were run in parallel. After partitioning X into windows, we obtained N window feature maps where $N=\frac{H\times W}{S^2}$. For assembling global information, we built global tokens from each window through the pyramid pooling illustrated in Fig. 2(b), inspired by [49]. First we applied 3×3, 2×2, 1×1 adaptive avg pooling on each window, then we flattened and concatenated the windows into a global token $\in R^{14\times C}$. After that we obtained N global tokens then we projected them to key and value matrices. We subsequently propagated the global context to each window by attention computation adopting feature maps partitioned by windows as query.

$$X_G = GMSA(X, GT) = softmax\left(\frac{Q_X K_{GT}^T}{\sqrt{d}}\right)V_{GT} \tag{2}$$

where Q_X represents the query matrices projected from the feature map, K_{GT} and V_{GT} are the key and value matrices projected from the global token. The final local–global combined feature map was computed through element-wise addition on X_L and X_G:

$$X_{LG} = X_L + X_G \tag{3}$$

3.2.3. Window size setting

We adopted a dynamic setup to determine the window size. For example, given an image $\in R^{224\times 224\times 3}$, the L-G transformer Block would be added at layers 1–3 of the network in which the output feature size is 56×56, 28×28, 14×14 respectively. The corresponding window size was set to 14, 7 and 7, realizing dynamical feature extraction ability depending on the size of the feature map.

3.2.4. Feed-Forward network (FFN)

In the FFN, LayerNorm (LN) was applied before the L-G MSA and the MLP blocks. In addition, based on what presented in [43], we added a Squeeze and Excitation (SE) block behind the output of the MLP layer. The SE block learns the channel weights through global spatial information, increasing the sensitivity of the effective feature maps, and suppressing irrelevant features of maps [44]. The FFN process can be expressed as:

$$\widehat{X} = LGMSA(LN(X)) + X \tag{4}$$

$$X_{out} = SE(MLP(LN(\widehat{X}))) + \widehat{X} \tag{5}$$

Note that the computation cost and complexities of our L-G transformer Block are very limited because the LMSA is computed with a small window and the size of the global token in the GMSA is much smaller than image and window size. We used the L-G transformer Block in the first to third layers to replace the skip connections, as a compromise between computational complexity and performance.

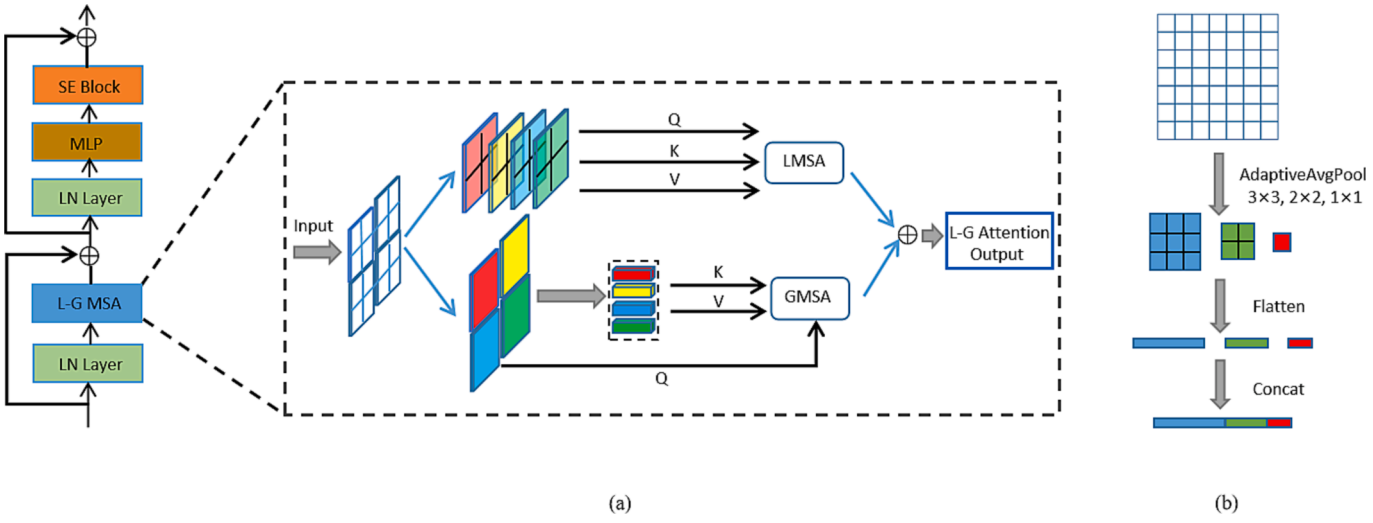


Fig. 2. (a)Schematic illustration of the proposed L-G transformer Block. (b)Pyramid pooling operation for obtaining the global token.

3.3. Cross attention block (CAB)

The CAB consisted of two MSA modules, and it was added to establish information interaction between different layers through attention computation. Directly computing the attention of two feature maps would be very computationally expensive, we therefore utilized a global token refined from the feature map as a medium to aggregate multi-scale global information and compute the attention, as illustrated in Fig. 3.

Given two feature maps $F_1 \in R^{H_1 \times W_1 \times C_1}$, and $F_2 \in R^{H_2 \times W_2 \times C_2}$, we first compress F_1 to a tensor of size $1 \times C_1$ through a Global Average Pooling (GAP) operation, and then projected it to a global token (GT) of size $1 \times C_2$ to match the channel dimension of F_2. A MSA block followed to aggregate long-range context between the GT and F_2 obtaining a Mixed Global Token (MGT) in which the query was projected from the GT, while the key and value were projected from the F_2.

$$MGT = MSA(GT, F_2) = softmax\left(\frac{Q_{GT}K_{F_2}^T}{\sqrt{d}}\right)V_{F_2} \tag{6}$$

We subsequently changed the dimension of MGT to C_1 by a linear transformation, and then applied another MSA block to combine information of F_1 and MGT. The CAB process can be expressed as:

$$F_{out} = MSA(F_1, (MGT)') = softmax\left(\frac{Q_{F_1}K_{(MGT)'}^T}{\sqrt{d}}\right)V_{(MGT)'} \tag{7}$$

where $(\mathrm{MGT})'$ is the result mapped from MGT, and F_{out} represents the final output of the CAB. As shown in Fig. 3, we utilized two CABs on the decoder side, one for layer 1 and layer 3, and another for layer 1 and layer 4, aiming to bring long-range dependencies across multi-scale layers.

4. Experiments and results

4.1. Datasets

The proposed method was evaluated on three public breast ultrasound datasets, BUSI [46], UDIAT [47] and BLUI [48].

The BUSI dataset was built by Al-Dhabyani et al., who collected data from 600 female patients at the Baheya Hospital. It comprises 780 images, of which 647 abnormal cases (210 malignant and 437 benign masses) and 133 normal cases. In our experiments, we used the abnormal samples for training and validation to verify the mass segmentation performance.

The UDIAT dataset was constructed by Yap et al. using a Siemens ACUSON Sequoia C512 system. The dataset contains 163 breast ultrasound images, of which 53 cancerous images and 110 benign images.

The BLUI dataset was provided by Abbasian Ardakani et al. which consists of 123 and 109 ultrasound images of malignant and benign breast lesions. All images are acquired from consecutive patients who presented to a cancer imaging center.

4.2. Implementation details

All experiments were implemented within the PyTorch framework and the models were trained on a single NVIDIA GeForce RTX2070 GPU with 8 GB graphic memory. We utilized Adam as the optimizer with a cosine annealing learning rate scheduler, setting the initial learning rate to 0.0001 and the minimum learning rate to 0.00001. The training epoch was set to 100, and the batch size to 16.

As part of our experiments, all images were resized to 224×224 before being fed into the model. The number of heads of multi-head self-attention used in different L-G transformer Blocks is [4,8,16], with corresponding window sizes of [14,7,7]. Several data augmentation approaches with horizontal flipping, vertical flipping, mirroring, transposition, and random rotation were adopted to avoid over-fitting and to enhance the generalization ability of the models. In addition, we used the weights pre-trained on ImageNet for the Resnet34 to initialize the parameters.

4.3. Loss function

To effectively address the challenge of extremely unbalanced class distribution and ensure smoother gradient descent [50], we adopted a hybrid segmentation loss function that comprises binary cross-entropy (BCE) loss and Dice loss. This combination allows us to benefit from the unique characteristics of both loss functions, achieving more robust and accurate segmentation results. Formally, the hybrid loss function is mathematically defined as follows:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{8}$$

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} y_i p_i}{\sum_{i=1}^{N}(y_i + p_i)} \tag{9}$$

$$L = \frac{L_{BCE} + L_{Dice}}{2} \tag{10}$$

Where N denotes the total number of pixels in the input image. $y_i \in \{0, 1\}$ represents the ground-truth label of the i-th pixel, with 1 indicating the positive class and 0 referring to the negative class. $p_i \in [0, 1]$ indicates the probability that pixel i is predicted to belong to the positive class.

4.4. Evaluation metrics

In order to conduct fair and systematic comparisons, the dice score

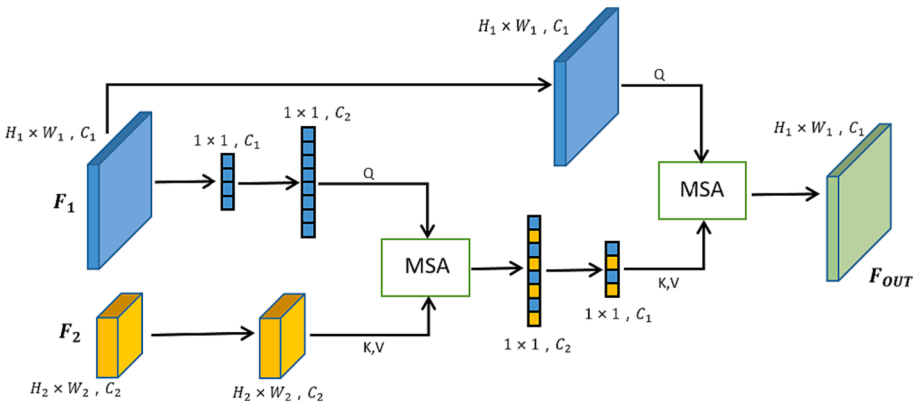


Fig. 3. The structure of our proposed Cross Attention Block (CAB) module.

(Dice) and a 95% Hausdorff Distance (HD) were used as main quantitative indicators. Other auxiliary metrics, such as Intersection over Union (IoU), Accuracy (Acc), Specificity (Spe), and Precision (Pre), were also calculated for a comprehensive assessment. Mathematically, these metrics can be expressed as follows:

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

$$HD = max(h(pred, gt), h(gt, pred)) \tag{12}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{13}$$

$$Acc = \frac{TP + FN}{TP + TN + FP + FN} \tag{14}$$

$$Spe = \frac{TN}{TN + FP} \tag{15}$$

$$Pre = \frac{TP}{TP + FP} \tag{16}$$

where TP, TN, FP, FN are the true positive, true negative, false positive, and false negative, and $h(A, B) = \max_{a \in A}\left\{\min_{b \in B}\left\|a - b\right\|\right\}$ (where $\| \bullet \|$ indicates the euclidean distance between two pixels).

4.5. Comparison study

To evaluate the effectiveness of the method proposed in this paper, we first compared it with state-of-the-art deep learning methods for breast ultrasound images segmentation on the BUSI, UDIAT and BLUI datasets respectively. Subsequently, to validate the generalization ability of our approach, we merged the three datasets into one and evaluated the effectiveness of each method on this combined dataset. Furthermore, following the approach in [51], to further evaluate the robustness and generalization ability of our method to specific features in ultrasound datasets, we partitioned the BUSI dataset based on benign and malignant tumors, as well as relative tumor sizes, and performed evaluation analyses on different methods. To ensure fair comparisons, we adopted five-fold cross-validation in all the experiments mentioned above. Moreover, to test for statistical significance differences, we conducted paired t-tests between our method and the second-best method based on the Dice and HD metrics, with P_value less than 0.05 indicating significant differences between our method and the comparison method.

4.5.1. Comparison with state-of-the-art methods

We first compared HAU-Net with state-of-the-art deep learning methods for breast ultrasound images segmentation on three public breast ultrasound datasets, BUSI, UDIAT and BLUI. The methods we compared included four CNN-based methods (U-Net(2015) [6], R34 DeepLabv3(2018) [16], CE-Net(2019) [10], MCRNet(2021) [17]) and five transformer-based methods (TransUNet(2021)[30], SwinUnet (2021) [31], FAT-Net(2022) [36], HiFormer(2022) [37], and AFTE-Net (2022) [45]).

The results of the comparisons for the three datasets are presented in Table 1, Table 2 and Table 3. Based on the quantitative results, our HAU-Net achieved the best results on most evaluation metrics. For the two main metrics of Dice and HD, HAU-Net achieved 83.11% and 10.67 on the BUSI dataset, with an increase of 0.65% in Dice and a decrease of 10.56% in HD compared to the second ranked method. On the UDIAT dataset, HAU-Net outperforms all competitors on all six metrics with a 1.17% increase in Dice and a 28.2% decrease in HD compared to the second-ranked method. As shown in Table 3, HAU-Net also performed well on the BLUI dataset and achieved the best results in all metrics except for ACC. In terms of the two key metrics, Dice and HD, HAU-Net achieved a 0.26% increase in Dice and a 5.78% decrease in HD compared to the second best method.

Regarding statistical significance, on the BUSI and UDIAT datasets, our method exhibited P_value less than 0.05 for both Dice and HD metrics compared to the second-best method, demonstrating significant superiority. On the BLUI dataset, our method's P_value for Dice was greater than 0.05, indicating comparable performance to TransUNet in this metric. However, for HD, our method significantly outperformed the competitor. Overall, The experimental results indicate that HAU-Net achieved better performance than other segmentation networks on three public ultrasound datasets and exhibited stable performance.

Fig. 4 shows the visual segmentation results achieved by different segmentation methods on three breast ultrasound image datasets. The first three rows correspond to the BUSI dataset, followed by the UDIAT dataset in the next three rows, and the final three rows show images from the BLUI dataset. Accurately segmenting ultrasound images is challenging due to the similar gray-scale distribution, blurred borders, and irregular tumor morphology. The selected images presented at least one of the aforementioned challenges.

The first three rows of images corresponded to the BUSI dataset, where the network may easily mistake the target due to the similar gray distribution. In the first row, Fat-Net, Deeplabv3 and AFTE-Net provided better segmentation results, in the second row, HiFormer and AFTE-Net accurately segmented the target, while in the third row, SwinUnet and Fat-Net could correctly locate the target, whereas the other methods had misjudgment and large area segmentation errors. Our method achieved accurate target localization and segmentation in all three images. Images in the fourth to the sixth row have the problems of blurred boundaries and irregular shapes, and our method outperformed the others, segmenting the target more accurately and being the closest to the real label in the target contour. Especially in the image corresponding to the fifth row, due to issues with similar grayscale intensities and blurred boundaries, the segmentation results of most methods tend to either overestimate or underestimate the tumor region, making it challenging to accurately delineate the tumor contours. However, our proposed method demonstrates the ability to more precisely segment the target, which is also reflected in our significantly superior HD metric

Table 1
Comparison results (mean ± std) of the proposed method on the BUSI dataset.

Method	Dice	Hd95	Iou	Acc	Spe	Pre
U-Net	72.41 ± 4.43	26.01 ± 3.88	63.39 ± 0.77	95.18 ± 0.56	97.56 ± 0.62	74.08 ± 1.16
R34 DeepLabv3	81.55 ± 2.07	13.08 ± 2.51	73.25 ± 2.47	96.47 ± 0.28	97.96 ± 0.38	82.59 ± 2.75
MCRNet	82.29 ± 1.92	12.31 ± 1.73	74.21 ± 2.20	96.48 ± 0.43	97.61 ± 0.58	81.96 ± 2.41
CE-Net	82.38 ± 1.64	12.02 ± 1.56	74.42 ± 1.88	96.65 ± 0.34	97.94 ± 0.37	83.04 ± 1.75
TransUNet	80.84 ± 2.06	15.45 ± 3.05	72.79 ± 2.49	95.87 ± 0.82	97.19 ± 1.06	81.46 ± 3.23
SwinUnet	80.19 ± 2.00	14.29 ± 2.62	71.65 ± 2.41	96.03 ± 0.62	97.37 ± 0.55	80.39 ± 2.06
FAT-Net	81.48 ± 1.99	13.81 ± 1.96	73.47 ± 2.13	96.36 ± 0.36	97.93 ± 0.33	82.55 ± 2.23
ATFE-Net	81.93 ± 2.76	11.93 ± 2.61	73.21 ± 3.07	96.50 ± 0.61	97.76 ± 0.57	80.68 ± 4.14
HiFormer	82.57 ± 2.50	12.13 ± 2.13	74.43 ± 2.73	**96.93 ± 0.34**	98.35 ± 2.48	84.38 ± 1.86
HAU-Net	**83.11 ± 2.07**	**10.67 ± 2.44**	**75.26 ± 2.08**	96.80 ± 0.16	**98.52 ± 0.24**	**86.08 ± 2.52**
P_value	<0.05(Dice), <0.05(Hd95)					

Table 2
The segmentation results (mean ± std) of different competing methods on the UDIAT dataset.

Method	Dice	Hd95	Iou	Acc	Spe	Pre
U-Net	80.58 ± 5.82	12.26 ± 4.96	71.53 ± 7.08	98.48 ± 0.76	99.11 ± 0.53	77.42 ± 6.37
R34 DeepLabv3	86.68 ± 2.47	5.21 ± 3.46	78.14 ± 3.08	98.81 ± 0.39	99.37 ± 0.27	85.82 ± 3.78
MCRNet	87.68 ± 1.73	5.43 ± 2.87	79.72 ± 1.99	98.86 ± 0.33	99.48 ± 0.29	86.44 ± 2.35
CE-Net	87.70 ± 1.96	5.08 ± 3.74	79.70 ± 2.33	98.96 ± 0.31	99.47 ± 0.13	86.41 ± 3.05
TransUNet	87.28 ± 2.71	6.11 ± 3.84	79.30 ± 2.99	98.86 ± 0.36	99.37 ± 0.14	85.84 ± 2.11
SwinUnet	85.32 ± 1.20	5.49 ± 1.52	76.11 ± 1.67	98.76 ± 0.34	99.33 ± 0.23	84.70 ± 2.18
FAT-Net	87.35 ± 3.27	5.68 ± 3.94	79.36 ± 4.03	98.94 ± 0.44	99.37 ± 0.21	85.39 ± 3.71
ATFE-Net	87.75 ± 1.84	5.88 ± 2.76	79.41 ± 2.15	98.98 ± 0.33	99.54 ± 0.20	86.38 ± 3.32
HiFormer	87.25 ± 2.03	5.07 ± 3.36	79.07 ± 2.28	98.97 ± 0.40	99.36 ± 0.20	85.84 ± 1.63
HAU-Net	**88.73 ± 2.11**	**3.64 ± 2.26**	**81.22 ± 2.30**	**99.03 ± 0.32**	**99.60 ± 0.12**	**88.68 ± 2.25**
P_value			<0.05(Dice), <0.05(Hd95)			

Table 3
Comparison results(mean ± std) of the proposed method on the BLUI dataset.

Method	Dice	Hd95	Iou	Acc	Spe	Pre
U-Net	84.45 ± 2.60	11.82 ± 2.23	75.61 ± 2.88	95.76 ± 0.44	97.49 ± 0.71	84.52 ± 1.63
R34 DeepLabv3	87.96 ± 1.57	6.19 ± 1.37	80.02 ± 2.03	96.90 ± 0.09	97.87 ± 0.48	87.79 ± 1.57
MCRNet	87.54 ± 1.82	7.67 ± 1.44	79.81 ± 2.17	96.64 ± 0.30	97.50 ± 0.35	86.32 ± 2.92
CE-Net	88.49 ± 1.21	5.97 ± 1.08	80.96 ± 1.44	96.71 ± 0.33	98.11 ± 0.55	89.77 ± 1.79
TransUNet	89.24 ± 1.29	5.71 ± 0.95	81.87 ± 1.70	**97.01 ± 0.32**	97.59 ± 0.33	88.18 ± 2.36
SwinUnet	88.52 ± 1.24	5.82 ± 1.01	80.53 ± 1.66	96.95 ± 0.37	97.95 ± 0.45	88.38 ± 1.60
FAT-Net	87.71 ± 1.59	8.08 ± 1.71	79.70 ± 2.14	96.76 ± 0.30	97.90 ± 0.53	87.42 ± 1.45
ATFE-Net	87.78 ± 1.26	6.88 ± 2.03	80.75 ± 2.05	96.77 ± 0.59	97.72 ± 0.98	88.75 ± 2.05
HiFormer	88.34 ± 1.16	7.23 ± 0.79	80.61 ± 1.44	96.93 ± 0.32	97.92 ± 0.56	88.37 ± 2.17
HAU-Net	**89.48 ± 0.44**	**5.38 ± 0.66**	**82.12 ± 0.85**	96.96 ± 0.42	**98.17 ± 0.29**	**89.93 ± 1.15**
P_value			>0.05(Dice), <0.05(Hd95)			

compared to other competitors in the comparative experiments. The last three rows of images were from the BLUI dataset, which was relatively less challenging to segment, and most methods performed well. However, our method outperformed the others in the segmentation details.

The visual segmentation results indicate that our proposed method outperformed the other methods in accurately identifying tumors from similar organ tissues and providing excellent segmentation results for irregular targets. These findings demonstrate the potential of our method in improving the accuracy of breast tumor segmentation in clinical settings, particularly in addressing the specific challenges posed by ultrasound images.

4.5.2. Generalization performance and robustness analysis

Table 4 presents the comparison results on the merged dataset, where our method achieved the best results in all six metrics. Regarding statistical significance, our method showed comparable performance to Hiformer in terms of Dice (P_value greater than 0.05), indicating similar capabilities in Dice-based segmentation. However, for HD, P_value less than 0.05 indicated that our model outperformed other methods significantly. Overall, the performance on the merged dataset demonstrated that our method has clear advantages in terms of generalization ability compared to other methods.

Regarding the experiments on the generalization ability to specific features in the ultrasound dataset, we divided the BUSI dataset into benign and malignant tumors as well as different relative tumor sizes. In the first experiment, we divided the BUSI dataset into benign and malignant groups. In the second experiment, we partitioned the BUSI dataset into three groups based on relative tumor size: small (less than 5% relative tumor size), medium (greater than 5% and less than 20% relative tumor size), and large (greater than 20% relative tumor size). Subsequently, we conducted five-fold validation on the divided datasets and selected the top five methods based on previous experimental performance for comparison. For simplicity and clarity, only the Dice and HD metrics were used in this experiment.

Table 5 presents the comparison results based on tumor benign and malignant characteristics, where each method performed significantly better on the benign dataset compared to the malignant dataset. This discrepancy is attributed to the fact that malignant tumors often have irregular shapes and non-smooth edges, making their accurate segmentation more challenging for deep learning models. Our method achieved the best results for both benign and malignant tumor segmentation, with a particularly noticeable advantage in segmenting malignant tumors. This indicates that our approach outperforms other methods in accurately delineating irregular tumor boundaries and effectively overcoming the issue of edge ambiguity.

Table 6 displays the performance of different methods in segmenting tumors of different sizes. In the case of small-sized tumors (less than 5% relative tumor size), our method ranked second in Dice score, slightly lower than CENET, but ranked first in HD. For medium and large-sized tumors, our method ranked first in both Dice and HD metrics. Particularly, in segmenting tumors larger than 20% relative size, our method exhibited an 2.2% improvement in Dice coefficient compared to the second-ranked method, with a 13.2% decline in HD, significantly outperforming other competitive methods. This experiment underscores the efficacy of our model in achieving favorable segmentation results for tumors of various sizes through the effective fusion of CNNs and transformers.

By conducting the paired *t*-test between our method and the second-best method based on the Dice and HD metrics, both p-values were found to be less than 0.05, thereby demonstrating that our experimental results exhibit statistical significance in terms of differences between groups.

4.6. Ablation study

4.6.1. Impact of the proposed modules

To evaluate the performance of the proposed modules, we conducted ablation experiments on BUSI. We firstly removed the L-G transformer block and the Cross Attention Block (CAB) from the HAU-Net and set this up as the baseline network. We then added various components to the baseline and compared it with Hiformer, which was the second-best performing method in the comparison experiment. The results of the

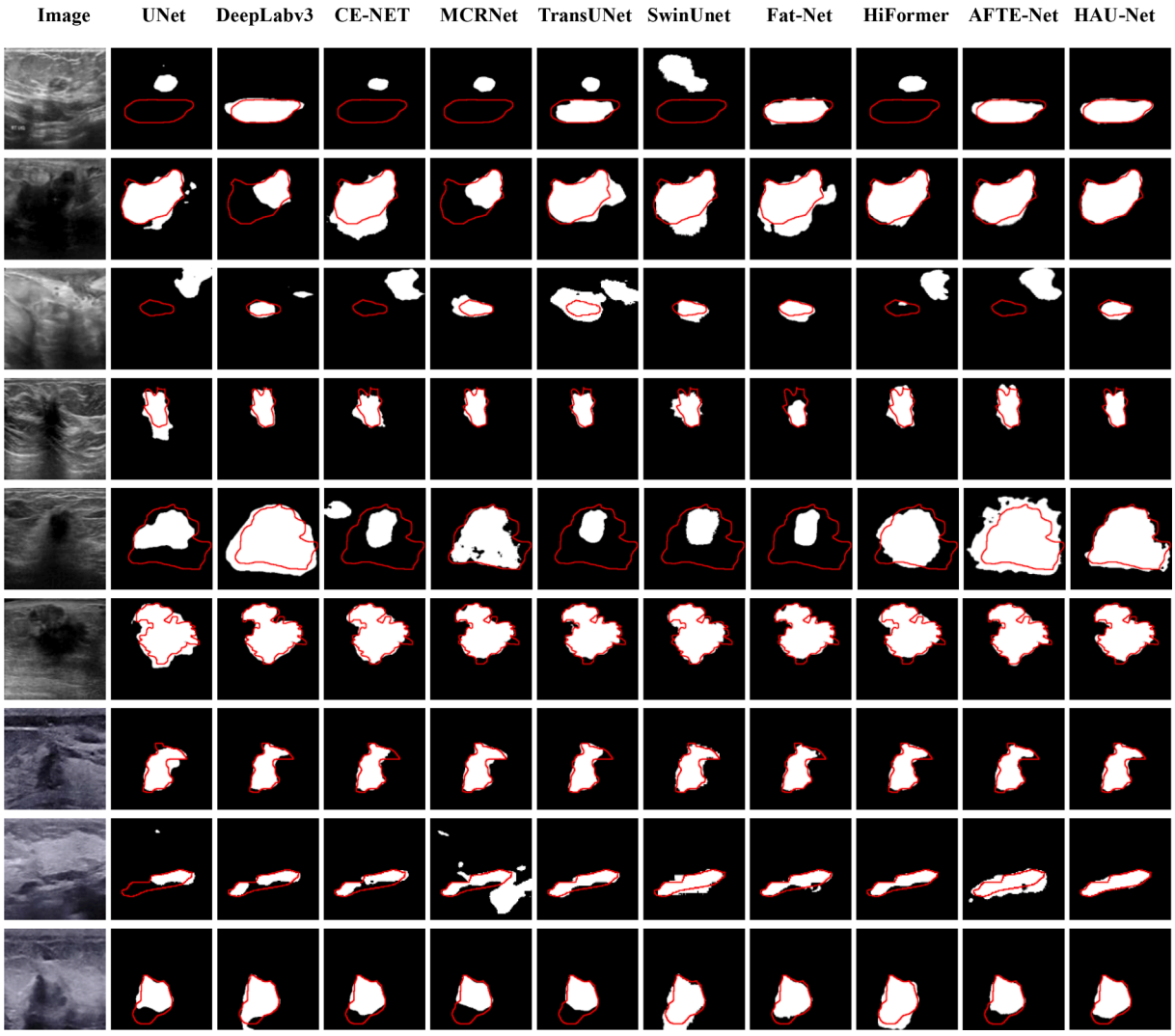


Fig. 4. Segmentation results of different methods on BUSI, UDIAT and BLUI. The images of the first three rows come from the BUSI dataset, the middle three rows come from the UDIAT dataset while the last three rows come from the BLUI dataset. From left to right are images of Input, U-net, R34 DeepLabv3, CE-Net, MCRNet, TransUNet, SwinUnet, Fat-Net, HiFormer, AFTE-Net, and Ours. The red curve is the boundary of the breast lesion.

Table 4
Comparison results (mean ± std) of the proposed method on the merged dataset.

Method	Dice	Hd95	Iou	Acc	Spe	Pre
U-Net	73.59 ± 7.02	21.44 ± 5.18	65.37 ± 7.13	96.59 ± 0.31	98.36 ± 0.38	74.40 ± 9.39
R34 DeepLabv3	82.32 ± 2.65	11.64 ± 1.01	74.26 ± 3.28	97.16 ± 0.59	98.33 ± 0.26	82.14 ± 5.47
MCRNet	82.27 ± 3.20	12.40 ± 2.39	74.66 ± 3.50	97.11 ± 0.39	98.20 ± 0.23	81.96 ± 4.50
CE-Net	82.20 ± 2.58	11.38 ± 1.92	74.70 ± 2.78	97.20 ± 0.67	98.35 ± 0.33	82.77 ± 4.89
TransUNet	81.87 ± 3.70	12.94 ± 2.80	74.16 ± 3.69	96.99 ± 0.45	98.11 ± 0.21	81.87 ± 3.70
SwinUnet	80.83 ± 2.98	12.41 ± 1.80	72.49 ± 3.13	96.94 ± 0.59	98.16 ± 0.35	81.00 ± 3.81
FAT-Net	81.81 ± 3.49	11.89 ± 2.04	74.21 ± 3.88	97.25 ± 0.56	98.52 ± 0.17	82.54 ± 4.39
ATFE-Net	82.55 ± 3.23	11.13 ± 1.94	74.71 ± 3.50	97.42 ± 0.59	98.50 ± 0.28	82.60 ± 3.97
HiFormer	82.82 ± 2.56	11.49 ± 2.02	75.01 ± 2.81	97.23 ± 0.53	98.43 ± 0.11	83.11 ± 4.95
HAU-Net	**82.85 ± 2.64**	**9.98 ± 1.26**	**75.15 ± 2.58**	**97.55 ± 0.62**	**98.68 ± 0.50**	**83.97 ± 3.37**
P_value			>0.05(Dice), <0.05(Hd95)			

ablation experiments, as presented in Table 7, indicate that the modules designed in our study play significant roles in enhancing the network's performance. Specifically, the L-G transformer block improved the DICE metric by 1.20% and reduced the HD metric by 12.21% from the baseline, while the CAB module further enhanced these metrics by 0.87% and 9.50%, respectively. Compared with HiFormer, the baseline model exhibited inferior performance in all metrics. However, after incorporating the L-G transformer block, the performance of our model became comparable to that of HiFormer. Furthermore, with the addition of the CAB module, our model outperformed HiFormer in all metrics. The results of our ablation experiments confirm the significance of the proposed modules for enhancing the network's performance in breast ultrasound image segmentation.

4.6.2. *Ablation on different adaptive avg pooling combination types*

The L-G transformer block was introduced in section 3.2, to obtain

Table 5
The segmentation results (mean ± std) of benign and malignant lesions in BUSI by different methods.

Method	Benign		Malignant	
	Dice	Hd95	Dice	Hd95
MCRNet	82.36 ± 2.44	13.60 ± 4.26	76.27 ± 3.65	20.06 ± 6.95
CE-Net	83.79 ± 2.26	10.65 ± 2.75	76.69 ± 2.38	21.05 ± 4.45
TransUNet	83.10 ± 1.83	12.68 ± 0.99	76.68 ± 3.72	21.36 ± 7.99
ATFE-Net	83.42 ± 0.96	10.52 ± 1.26	76.95 ± 3.79	20.71 ± 6.98
HiFormer	84.09 ± 1.74	8.90 ± 1.43	76.86 ± 3.56	21.71 ± 6.48
HAU-Net	**84.55 ± 1.39**	**8.05 ± 1.13**	**77.61 ± 2.34**	**17.41 ± 6.57**
P_value	<0.05(Dice), <0.05(Hd95)			

Table 6
The segmentation results (mean ± std) of distinct relative tumor size in BUSI by different methods.

Method	Small		Medium		Large	
	Dice	Hd95	Dice	Hd95	Dice	Hd95
MCRNet	82.27 ± 1.37	8.33 ± 0.99	83.04 ± 4.89	11.43 ± 3.78	80.67 ± 4.09	14.42 ± 5.34
CE-Net	**83.26 ± 0.56**	8.38 ± 1.02	83.05 ± 3.78	11.63 ± 3.30	81.71 ± 4.48	15.16 ± 5.18
TransUNet	82.20 ± 2.51	8.40 ± 1.10	82.96 ± 3.53	12.90 ± 4.21	81.42 ± 2.83	15.26 ± 3.81
ATFE-Net	82.55 ± 2.81	8.69 ± 2.33	82.32 ± 4.05	11.89 ± 2.90	81.36 ± 3.78	15.13 ± 4.32
HiFormer	82.74 ± 0.97	9.09 ± 2.47	82.93 ± 3.60	12.64 ± 3.32	83.06 ± 5.60	14.22 ± 4.59
HAU-Net	82.98 ± 1.87	**7.34 ± 1.23**	**83.31 ± 3.92**	**11.03 ± 3.34**	**84.90 ± 4.43**	**11.92 ± 4.17**
P-values	<0.05(Dice), <0.05(Hd95)					

Table 7
Ablation study on different component combinations of HAU-Net on the BUSI dataset.

Method	Dice	Hd95	Iou	Acc	Spe	Pre
HiFormer	82.57 ± 2.50	12.13 ± 2.13	74.43 ± 2.73	96.93 ± 0.34	98.35 ± 2.48	84.38 ± 1.86
Baseline	81.41 ± 2.17	13.43 ± 2.89	73.53 ± 2.29	96.43 ± 0.53	98.11 ± 0.51	83.10 ± 2.26
Baseline + L-G transformer Block	82.39 ± 2.35	11.79 ± 2.67	74.52 ± 2.41	96.71 ± 0.45	98.18 ± 0.49	83.97 ± 2.35
HAU-Net (Baseline + L-G transformer Block + CAB)	83.11 ± 2.07	10.67 ± 2.44	75.26 ± 2.08	96.80 ± 0.16	98.52 ± 0.24	86.08 ± 2.52

the global tensor we utilized the adaptive avg pooling which consisted of three adaptive avg pooling with different sizes on the windows partitioned feature map, the pooling sizes were 3 × 3, 2 × 2and1 × 1. In the ablation study, we investigated the impact of different pooling combinations. As shown in Fig. 5, we compared two adaptive avg pooling combination types, 3 × 3+1 × 1(Type B) and 2 × 2 + 1 × 1(Type C), with the adaptive avg pooling combination approach (Type A) proposed in this paper, and performed five-fold cross-validation experiments on BUSI and UDIAT.

As in Table 8, the network segmentation performance gradually decreases together with the computation cost from Type A to Type C. Although the numbers of parameters decrease significantly from Type A to Type C, the corresponding decrease in computational power required was negligible. Based on this and on comprehensive consideration about the network performance, we believe type A is the optimal choice.

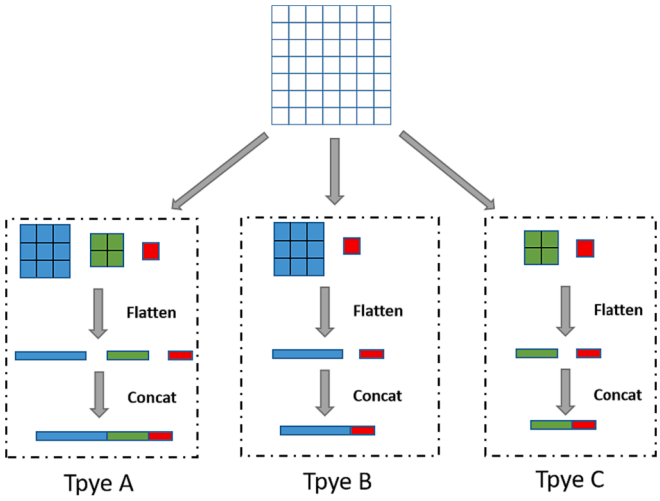


Fig. 5. Three adaptive avg pooling combination types.

Table 8
Ablation study on different adaptive avg pooling combination types of HAU-Net on BUSI and UDIAT respectively.

	Method	Dice	Hd95	Params (M)	GFLOPs
BUSI	HAU-Net(Type A)	83.11 ± 2.07	10.67 ± 2.44	76.65	10.76
	HAU-Net(Type B)	82.66 ± 2.08	11.64 ± 2.22	51.87	10.58
	HAU-Net(Type C)	82.17 ± 2.39	13.06 ± 2.18	32.51	10.43
UDIAT	HAU-Net(Type A)	88.73 ± 2.11	3.64 ± 2.26	76.65	10.76
	HAU-Net(Type B)	88.51 ± 1.64	3.87 ± 1.93	51.87	10.58
	HAU-Net(Type C)	88.03 ± 2.22	4.92 ± 4.31	32.51	10.43

5. Discussion and conclusion

In this study, we present a novel breast tumor segmentation model, the Hierarchical attention-guided U-Net (HAU-Net), which combines the strengths of transformers and convolutional neural networks (CNNs) to accurately detect breast lesions in ultrasound images. Our approach is unique in that we replace the traditional skip connection with a L-G transformer block, which allows for long-range dependency modeling, while maintaining network integrity and consistency during the feature extraction process. In addition, we introduce the Cross Attention Block (CAB) to improve information interaction among multi-size feature layers. The CAB selectively emphasizes important features and suppresses irrelevant ones, which leads to better feature representation and higher segmentation accuracy.

We conducted experiments on three publicly available datasets, BUSI, UDIAT and BLUI, to evaluate the performance of HAU-Net. The results show that our method outperforms state-of-the-art methods in terms of Dice coefficient, Hausdorff distance, IoU, Accuracy, Specificity, and Precision. Subsequently, the experimental results on the merged dataset demonstrate that our method continues to exhibit the most outstanding performance in terms of generalization compared to other competing methods. Additionally, to further validate the robustness of our approach, we partitioned the BUSI dataset based on tumor case characteristics and relative tumor sizes, conducting comparative experiments on these partitioned datasets. The experimental results indicate that our method achieves favorable outcomes on all partitioned datasets, particularly showing a significant advantage over other

competing methods in segmenting malignant tumors and large-sized tumors. This also illustrates the effectiveness of combining CNN and transformer in our approach, enabling more precise segmentation of large-sized and irregular tumor contours.

Despite the promising results, our method still faces some limitations, Firstly, it has difficulties in handling small, irregular targets and blurred edges, especially when the pixel intensity of target is very close to the background. Secondly, as a traditional supervised segmentation method, our study relies on sufficient manually labeled samples for training, which requires at least hundreds of labeled samples, but the available data in practical applications may be very scarce. In future work, we plan to improve our HAU-Net model by incorporating a region-based attention mechanism in the encoder. This will enable the model to focus more on local regions of interest, better capturing important features while ignoring irrelevant or noisy ones. We also aim to propose self-supervised or semi-supervised model training to reduce the dependence on manually labeled samples, and verify the effectiveness of this method in real clinical diagnosis.

Overall, our proposed HAU-Net model shows great potential in accurate detection of breast lesions in medical images, and we believe it can have significant impact in clinical practice.

CRediT authorship contribution statement

Huaikun Zhang: Conceptualization, Writing – original draft, Methodology, Data curation. **Jing Lian:** Resources, Funding acquisition, Methodology. **Zetong Yi:** Data curation. **Ruichao Wu:** Data curation. **Xiangyu Lu:** Visualization. **Pei Ma:** Visualization. **Yide Ma:** Supervision, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This work is jointly supported by the National Natural Science Foundation of China (No.62061023 and No.61961037) and the Natural Science Foundation of Gansu Province (No.18JR3RA288 and No.21JR7RA345).

References

[1] R.L. Siegel, K.D. Miller, H.E. Fuchs, et al., Cancer statistics, CA: Cancer J. Clin. 72 (1) (2022), https://doi.org/10.3322/caac.21708.
[2] H.D. Cheng, X.J. Shi, R. Min, et al., Approaches for automated detection and classification of masses in mammograms, Pattern Recogn. 39 (4) (2006) 646–668, https://doi.org/10.1016/j.patcog.2005.07.006.
[3] H.D. Cheng, J. Shan, W. Ju, et al., Automated breast cancer detection and classification using ultrasound images: A survey, Pattern Recogn. 43 (1) (2010) 299–317, https://doi.org/10.1016/j.patcog.2009.05.012.
[4] M. Xian, Y. Zhang, H.D. Cheng, et al., Automatic breast ultrasound image segmentation: A survey, Pattern Recogn. 79 (2018) 340–355, https://doi.org/10.1016/j.patcog.2018.02.012.
[5] J. Long, E. Shelhamer, and T. Darrell, 2015. Fully convolutional networks for semantic segmentation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 3431–3440, doi: 10.1109/TPAMI.2016.2572683.
[6] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional net works for biomedical image segmentation, in: Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Berlin, Germany: Springer, 2015, pp. 234–241, doi: 10.48550/arXiv.1505.04597.
[7] M.H. Yap, G. Pons, J. Martí, et al., Automated breast ultrasound lesions detection using convolutional neural networks, IEEE J. Biomed. Health Informat. 22 (4) (2017) 1218–1226, https://doi.org/10.1109/JBHI.2017.2731873.
[8] R. Almajalid, J. Shan, Y. Du, et al., Development of a deeplearning-based method for breast ultrasound image segmentation, in: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, pp. 1103–1108, doi: 10.1109/ICMLA.2018.00179.
[9] Y. Hu, Y. Guo, Y. Wang, et al., Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model, Med. Phys. 46 (2019) 215–228, https://doi.org/10.1002/mp.13268.
[10] Z. Gu, J. Cheng, H. Fu, et al., CE-Net: Context encoder network for 2D medical image segmentation, IEEE Trans. Med. Imaging 38 (10) (2019) 2281–2292, https://doi.org/10.1109/TMI.2019.2903562.
[11] P. Tang, X. Yan, Y. Nan, et al., Feature pyramid non-local network with transform modal ensemble learning for breast tumor segmentation in ultrasound images, IEEE Trans. Ultrasonics Ferroelect. Freq. Control. 68 (12) (2021) 3549–3559, https://doi.org/10.1109/TUFFC.2021.3098308.
[12] C. Xue, L. Zhu, H. Fu, et al., Global guidance network for breast lesion segmentation in ultrasound images, Med. Image Anal. 70 (2021), 101989, https://doi.org/10.1016/j.media.2021.101989.
[13] M. Byra, P. Jarosik, A. Szubert, et al., Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network, Biomed. Signal Process. Control 62 (2020) 1–10, https://doi.org/10.1016/j.bspc.2020.102027.
[14] G. Chen, Y. Dai, J. Zhang, et al., AAU-net: An adaptive attention U-net for breast lesions segmentation in ultrasound images, IEEE Trans. Med. Imaging 9968268 (2022), https://doi.org/10.1109/TMI.2022.3226268.
[15] Y. Yan, Y. Liu, Y. Wu, et al., Accurate segmentation of breast tumors using AE U-net with HDC model in ultrasound images, Biomed. Signal Process. Control 72 (2022), 103299, https://doi.org/10.1016/j.bspc.2021.103299.
[16] L. Chen, Y. Zhu, G. Papandreou, et al., Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818, doi: 10.48550/arXiv.1802.0261.
[17] M. Lou, J. Meng, Y. Qi, et al., MCRNet: Multi-level context refinement network for semantic segmentation in breast ultrasound imaging, Neurocomputing 470 (2021) 154–169, https://doi.org/10.1016/j.neucom.2021.10.102.
[18] R. Huang, M. Lin, H. Dou, et al., Boundary-rendering network for breast lesion segmentation in ultrasound images, Med. Image Anal. 102478 (2022), https://doi.org/10.1016/j.media.2022.102478.
[20] H. Lee, J. Park, J.Y. Hwang, Channel attention module with multiscale grid average pooling for breast cancer segmentation in an ultrasound image, IEEE Trans Ultrason. Ferroelectr. Freq. Control. 67 (2020) 1344–1353, https://doi.org/10.1109/TUFFC.2020.2972573.
[21] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, Adv. Neural Inf. Proces. Syst. 30 (2017) 5998–6008, https://doi.org/10.5555/3295222.3295349.
[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: Transformers for image recognition at scale, Int. Conf. Learn. Represent. (2020), https://doi.org/10.48550/arXiv.2010.11929.
[23] C.F. Chen, Q. Fan, and R. Panda, 2021. Crossvit: Cross-attention multiscale vision transformer for image classification, arXiv preprint arXiv:2103.14899, doi: 10.48550/arXiv.2103.14899.
[24] N. Carion, F. Massa, G. Synnaeve, et al., End-to-end object detection with transformers, Eur. Conf. Comput. Vis. (2020) 213–229, https://doi.org/10.48550/arXiv.2005.12872.
[25] R. Strudel, R. Garcia, I. Laptev, et al., Segmenter: Transformer for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 7262–7272, doi: 10.1109/ICCV48922.2021.00717.
[26] E. Xie, W. Wang, Z. Yu, et al., Segformer: Simple and effificient design for semantic segmentation with transformers. arXiv preprint arXiv:2105.15203, 2021, doi: 10.48550/arXiv.2105.15203.
[27] Z. Liu, Y. Lin, Y. Cao, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 10012–10022, doi: 10.1109/ICCV48922.2021.00986.
[28] T. Huang, L. Huang, S. You, et al., LightViT: Towards Light-Weight Convolution-Free Vision Transformers, arXiv preprint arXiv:2207.05557, 2022, doi: 10.48550/arXiv.2207.05557.
[29] Y. Yang, L. Jiao, X. Liu, et al., Transformers Meet Visual Learning Understanding: A Comprehensive Review, arXiv preprint arXiv:2203.12944, 2022, doi: 10.48550/arXiv.2203.12944.
[30] J. Chen, Y. Lu, Q. Yu, et al., Transunet: Transformers make strong encoders for medical image segmentation, arXiv preprint arXiv:2102.04306, 2021, doi: 10.48550/arXiv.2102.04306.
[31] H. Cao, Y. Wang, J. Chen, et al., Swin-unet: Unet-like pure transformer for medical image segmentation, arXiv preprint arXiv:2105.05537, 2021, doi: 10.48550/arXiv.2105.05537.
[32] G. Xu, X. Wu, X. Zhang, et al., Levit-unet: Make faster encoders with transformer for medical image segmentation, arXiv preprint arXiv: 2107.08623, 2021, doi: 10.48550/arXiv.2107.08623.
[33] A. Hatamizadeh, Y. Tang, V. Nath, et al., netr: Transformers for 3d medical image segmentation, arXiv preprint arXiv:2103.10504, 2021, doi: 10.48550/arXiv.2103.10504.
[34] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, et al., Medical transformer: Gated axial-attention for medical image segmentation, arXiv preprint arXiv:2102.10662, 2021.
[35] Y. Zhang, H. Liu, and Q. Hu, TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation, arXiv preprint arXiv:2102.08005, 2021, doi: 10.48550/arXiv.2102.08005.

[36] H. Wu, S. Chen, G. Chen, et al., Fat-net: Feature adaptive transformers for automated skin lesion segmentation, Med. Image Anal. 76 (2022), 102327, https://doi.org/10.1016/j.media.2021.102327.
[37] M. Heidari, A. Kazerouni, and M. Soltany, et al., HiFormer: Hierarchical Multi-scale Representations Using Transformers for Medical Image Segmentation, in: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6191-6201, doi: 10.1109/WACV56688.2023.00614.
[38] X. He, G. Qi, Z. Zhu, et al., Medical image segmentation method based on multi-feature interaction and fusion over cloud computing, Simul. Model. Pract. Theory 126 (2023), 102769, https://doi.org/10.1016/j.simpat.2023.102769.
[39] Y. Xu, X. He, G. Xu, et al., A medical image segmentation method based on multi-dimensional statistical features, Front. Neurosci. 16 (2022) 1009581, https://doi.org/10.3389/fnins.2022.1009581.
[40] Z. Zhu, X. He, G. Qi, et al., Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI, Informat. Fus. 91 (2023) 376–387, https://doi.org/10.1016/j.inffus.2022.10.022.
[41] Y. Li, Z. Wang, L. Yin, et al., X-Net: a dual encoding–decoding method in medical image segmentation, Vis. Comput. 39 (2023) 2223–2233, https://doi.org/10.1007/s00371-021-02328-7.
[42] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[43] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, Proc. of CVPR (2018) 7132–7141, https://doi.org/10.1109/CVPR.2018.00745.
[44] D. Jha, P.H. Smedsrud, M.A. Riegler, et al., Resunet++: An advanced architecture for medical image segmentation, Proc. of IEEE ISM. (2019) 225–230, https://doi.org/10.1109/ISM46123.2019.00049.
[45] Z. Ma, Y. Qi, C. Xu, et al., ATFE-Net: Axial transformer and feature enhancement-based CNN for ultrasound breast mass segmentation, Comput. Biol. Med. 153 (2023), 106533, https://doi.org/10.1016/j.compbiomed.2022.106533.
[46] W. Al-Dhabyani, M. Gomaa, H. Khaled, et al., Dataset of breast ultrasound images, Data Brief 28 (2020), 104863, https://doi.org/10.1016/j.dib.2019.104863.
[47] M.H. Yap, G. Pons, J. Marti, et al., Automated breast ultrasound lesions detection using convolutional neural networks, IEEE J. Biomed. Heal. Informatics. 22 (2017) 1218–1226, https://doi.org/10.1109/JBHI.2017.2731873.
[48] A. Abbasian Ardakani, A. Mohammadi, M. Mirza-Aghazadeh-Attari, et al., An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies, Comput. Biol. Med. 152 (2023) (2022), 106438, https://doi.org/10.1016/j.compbiomed.106438.
[49] K. He, X. Zhang, S. Ren, et al., Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE T PATTERN ANAL. 37 (9) (2015) 1904–1916, https://doi.org/10.1109/TPAMI.2015.2389824.
[50] C. H Sudre, W. Li, T. Vercauteren, et al., Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2017,pp. 240–248, doi: 10.1007/978-3-319-67558-9_28.
[51] W. Gómez-Flores, W.C.D.A. Pereira, A comparative study of pre-trained convolutional neural networks for semantic segmentation of breast tumors in ultrasound, Comput. Biol. Med. 126 (2020), 104036, https://doi.org/10.1016/j.compbiomed.2020.104036.