Computers in Biology and Medicine 155 (2023) 106629

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

# HCTNet: A hybrid CNN-transformer network for breast ultrasound image segmentation

Qiqi He, Qiuju Yang *, Minghao Xie

*School of Physics and Information Technology, Shaanxi Normal University, Xi'an, China*

ARTICLE INFO

*Keywords:*
Breast lesion segmentation
Convolutional neural network
Transformer
Deep learning
Ultrasound imaging

ABSTRACT

Automatic breast ultrasound image segmentation helps radiologists to improve the accuracy of breast cancer diagnosis. In recent years, the convolutional neural networks (CNNs) have achieved great success in medical image analysis. However, it exhibits limitations in modeling long-range relations, which is unfavorable for ultrasound images with speckle noise and shadows, resulting in decreased accuracy of breast lesion segmentation. Transformer can obtain sufficient global information, but it is deficient in acquiring local details and needs to be pre-trained on large-scale datasets. In this paper, we propose a Hybrid CNN-Transformer network (HCTNet) for boosting the breast lesion segmentation in ultrasound images. In the encoder of HCTNet, Transformer Encoder Blocks (TEBlocks) are designed to learn the global contextual information, which are combined with CNNs to extract features. In the decoder of HCTNet, a Spatial-wise Cross Attention (SCA) module is developed based on the spatial attention mechanism, which reduces the semantic discrepancy with the encoder. Moreover, residual connection is used between decoder blocks to make the generated features more discriminative by aggregating contextual feature maps at different semantic scales. Extensive experiments on three public breast ultrasound datasets demonstrate that HCTNet outperforms other medical image segmentation methods and the recent semantic segmentation methods on breast ultrasound lesion segmentation.

## 1. Introduction

Female breast cancer has now surpassed lung cancer as the leading cause of cancer incidence worldwide in 2020 and is also the fifth leading cause of cancer death worldwide [1]. Ultrasound imaging has been widely used in the diagnosis of breast masses due to its unique advantages, such as safety, affordability and efficiency. Accurate segmentation of breast lesions from ultrasound images is an important step in computer-aided diagnosis (CAD), which helps the diagnosis and treatment of breast cancer and thus effectively reduces mortality. However, segmentation of breast ultrasound lesions remains a challenging task due to the low quality of ultrasound imaging caused by speckle noise and strong shadows (as shown in Fig. 1(a)), as well as the irregularities of breast lesions, such as the varying tumor shapes and sizes among patients [2].

The convolutional neural networks (CNNs) are an end-to-end deep learning method. It can extract deep features, and has made amazing progress in dealing with breast ultrasound image segmentation [3–6]. Among these networks, the symmetric encoder–decoder architecture like U-Net [8] is the mainstream model architecture. Because of the simple and excellent performance of U-shaped structure, more and more various U-shaped networks are proposed, including UNet++ [9], Attention Unet [34] and FPN [10]. However, these methods are based on CNNs by stacking convolutional kernels to gradually acquire the large receptive fields and integrate the global contextual information, which has some limitations. First, the deep stacking of convolutional operations increases the network parameters, leading to computational inefficiencies and difficulties in network optimization [7]. Second, the local operation of convolution makes the network under- or over-segmented when segmenting breast ultrasound lesions. As shown in Fig. 1(d)–(f), CNNs such as Unet [8], Unet++ [9], and FPN [10] segment lesions by sliding the convolutional kernels, which would lead to some normal tissues with similar appearance to lesions being misidentified, resulting in unsatisfactory segmentation.

There are many normal pixels in the ultrasound image that are distant from but similar in appearance to the breast lesions. Combining these pixels can provide long-term non-local features for the segmentation of breast ultrasound lesions, allowing the network to learn

* Corresponding author.
*E-mail address:* yangqiuju@snnu.edu.cn (Q. Yang).

https://doi.org/10.1016/j.compbiomed.2023.106629
Received 18 July 2022; Received in revised form 11 January 2023; Accepted 4 February 2023
Available online 9 February 2023

discriminative features [2]. Transformer is undoubtedly an ideal way to this task. Recently, transformer-based medical image segmentation has been intensively developed. By using self-attention, Transformer can directly model long-range dependencies to compensate the deficiency of CNNs in dealing with long-range dependencies, and thus improve the network performance [11]. However, transformer-based models need to be trained on large-scale datasets to perform well because they lack some of the inductive biases inherent to CNNs [14] and may have difficulty learning the position encoding of the images when applied to breast ultrasound medical tasks with small datasets [12]; and the computational cost required to apply Transformer directly to the high-resolution original images is very expensive.

CNNs have the advantage of extracting local features but suffer from a lack of direct modeling of global information. In contrast, based on the self-attention mechanism, Transformer can learn the relationship between global pixel points but are not as powerful as CNNs in representing local details [15]. Therefore, some researchers have started to explore how to properly combine convolution and self-attention to build an optimal medical segmentation network [19–21]. Gao et al. [19] proposed UTNet, in which CNNs and Transformer are alternately applied to encoder and decoder subnets at different resolutions to improve the segmentation performance. nnFormer [20] was proposed for volumetric image segmentation by exploiting the combination of interleaved convolution and self-attention operations, as well as introducing local and global volume-based self-attention mechanism. Xu et al. [21] integrated rich local features and global contextual information at different scales and applied self-attention mechanisms to multi-scale feature maps for the segmentation of breast ultrasound lesions.

In this work, we propose a Hybrid CNNs-Transformer network (HCTNet) for breast ultrasound image segmentation. Specifically, a Transformer Encoder Block (TEBlock) is designed in the encoder to capture the long-range dependencies. The encoder uses a hybrid stem where convolution and Transformer are interleaved to give full play to their strengths. A Spatial-wise Cross Attention (SCA) module is developed in the decoder to reduce the problem of semantic discrepancy between the encoder and decoder subnets. HCTNet was extensively compared with seven state-of-the-art segmentation methods on three public dataset of breast ultrasound lesions. The contributions of this paper are mainly in the following three aspects.

- First, this paper proposes a segmentation network HCTNet for breast ultrasound image segmentation, which integrates the advantages of CNNs and Transformer and achieves better performance than state-of-the-art segmentation networks on three public datasets.
- Second, we propose a TEBlock in the encoder to compute the interactions between pixels in ultrasound images, compensating for the lack of global information captured by CNNs.
- Third, we develop a SCA module to reduce the semantic discrepancy between the encoder and decoder subnets by fusing the spatial attention maps. Additionally, residual connections are utilized between decoder blocks to enhance the position information of lesions in breast ultrasound images.

## 2. Related works

In the past decades, the work related to breast ultrasound medical image processing has developed significantly in the wave of deep learning. Here, we briefly review the CNN-based, Transformer-based, and CNN-Transformer fusion approaches in breast ultrasound medical image segmentation and relative position coding in Transformer.

### 2.1. CNN-based methods

In recent years, with the continuous development of deep learning, various classical CNN models have been developed, such as FCN [32], Unet [8], Unet++ [9], FPN [10] and DeepLabv3+ [33]. Based on these models, amazing progress has been made related to breast ultrasound images. Shareef et al. [3] utilized row-column-wise kernels to adapt the breast anatomy and fused contextual information at different scales in two encoder branches for segmenting small breast tumors. Hu et al. [4] proposed to combine a dilated convolutional network with a phase-based active contour model to automatically segment breast lesion regions. Yap et al. [5] used CNN methods for breast ultrasound lesion detection and comparatively investigated three CNN-based methods. Zhu et al. [6] developed a second-order subregion network for breast lesion segmentation by utilizing the second-order statistics of multiple feature subregions.

### 2.2. Transformers-based methods

The core idea of Transformer is to apply self-attention mechanisms to capture long-term dependencies. Mo et al. [38] proposed the HoVer-Trans model to use Transformer for breast cancer diagnosis in breast ultrasound images. Cao et al. [13] proposed an Unet-like pure Transformer Swin-Unet, which used hierarchical Swin Transformer with shifted windows as the encoder for medical image segmentation. However, Transformer requires pre-training on large datasets and the high computational complexity results in the substantial overhead of training and inference. Therefore, few works utilize only Transformers, most of which are CNN-Transformer fusion methods.

### 2.3. CNN and transformer fusion methods

The fusion of CNNs and Transformer has become an important research direction. Compared with CNNs, Transformer relaxes the local inductive bias by encoding image features, making it more capable of handling non-local interactions [20], which compensates for the deficiency of CNNs in handling long-range dependencies. Chen et al. [16] proposed TransUnet to explore the potential of Transformer in medical image segmentation for the first time. They used CNNs as feature extractors and utilized Transformer to extract global contextual information from the feature maps of CNNs. Wang et al. [37] proposed TFNet, using Transformer to fuse the features extracted by CNNs, and achieved better performance than CNNs. Yao et al. [17] first used convolutional operations to obtain the feature map and then used Transformer to encode the image patches to obtain global contextual information. In the above fusion methods, Transformer is applied on top of the

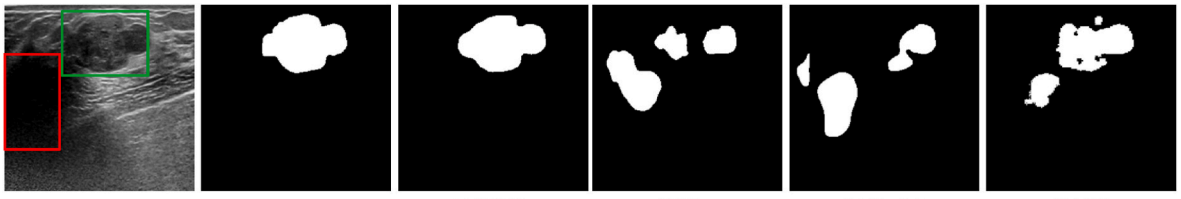


**Fig. 1.** Illustration of segmentation results with different approaches. (a) Input image; (b) Ground truth; (c)–(f) are the segmentation results with HCTNet, Unet [8], Unet++ [9], and FPN [10], respectively. The green box in (a) shows the area of the breast lesion, and the red box shows the ultrasound shadow similar to the lesion.

low-resolution feature maps from the CNN, which does not fully exploit the advantages of Transformer. Methods like UTNet [19] and nnFormer [20] improve this fusion strategy by employing a hybrid stem that intertwines convolution and self-attention, taking full advantage of their strengths. A similar approach is formulated in the encoder of our network. However, to further reduce the complexity of the model and the difference between the semantic features of the encoder and the decoder, instead of using the CNNs-Transformer fusion approach in the decoder, we design the SCA module to recover the spatial features of the whole image.

### 2.4. Relative positional encoding

Combining position information in an explicit representation is an essential consideration in Transformer. The standard self-attention module discards position information completely and is perturbatively equivalent, limiting visual task expressiveness [23]. Therefore, Shaw et al. [24] proposed an extension of self-attention for merging the relative position information of sequences to improve machine translation performance. Bello et al. [23] introduced a new 2D relative self-attention mechanism for image classification training by independently adding relative height and width information to achieve 2D relative self-attention. Ramachandran et al. [25] verify whether self-attention can be a useful independent layer using a self-attention module with 2D relative position embedding.

## 3. Materials and methods

### 3.1. Datasets

We use three breast ultrasound image datasets to evaluate the effectiveness of the proposed network. The first one is the dataset BUSI [28] from Baheya Womenś Hospital for Early Detection and Treatment of Cancer (Cairo, Egypt). It was acquired by the LOGIQ E9 ultrasound system and the LOGIQ E9 Agile ultrasound system. BUSI collected 780 images from 600 female patients aged 25–75 years, with an average image size of 500 × 500 pixels, including 437 benign cases, 210 malignant masses, and 133 normal cases.

The second dataset is BUS [29] from the UDIAT Diagnostic center of Parc Taulí Corporation in Sabadell, where images were collected using a Siemens ACUSON Sequoia C512 system with a 17L5HD linear array transducer (8.5 MHz). BUS collected 163 images from different women with an average image size of 760 × 570 pixels, including 53 images of malignant cases and 110 images of benign cases.

The last dataset is Dataset B [30] from the Cancer Center of Sun Yat-sen University, which was acquired by the HDI 5000 SonoCT System (Philips Medical Systems) with an L12-5 50 mm Broadband Linear Array at the imaging frequency of 7.1 MHz. Dataset B collected 320 images from patients with a mean age of 46.6 ± 14.2 years, with 160 cases of benign and 160 cases of malignant masses, and the image size was normalized to 128 × 128 pixels.

Our task is to segment lesions in breast ultrasound images, so we remove the normal cases without breast lesions in the BUSI dataset. Moreover, five-folder cross-validation is applied to evaluate different segmentation methods on the three aforementioned datasets.

### 3.2. Segmentation method

Fig. 2 illustrates the architecture of HCTNet. HCTNet takes a breast ultrasound image as the input and produces a segmented mask in an end-to-end manner. Specifically, HCTNet gradually extracts features of different resolutions from shallow to deep layers in the encoder. For the feature maps with the same resolution, we first extract the original local features using CNNs and then learn the long-range dependencies by TEBlock. We can obtain the global contextual relationships of the feature maps at different scales in multiple stages of the encoder. In addition, CNNs can complement some inductive biases lacking in Transformer, which Transformer need to learn by pre-training with large-scale datasets [14]. In the decoder, we upsample the feature maps from deep to shallow at different scales. The SCA modules fuse the feature maps from the encoder and decoder to reduce the semantic differences during upsampling. To further optimize the segmentation results, we adopt residual connection between decoder blocks to refine the details of the lesions. Finally, HCTNet produces the prediction map as the segmentation result. The following subsections will describe the details of Transformer Encoder Block (TEBlock) and SCA module in HCTNet.

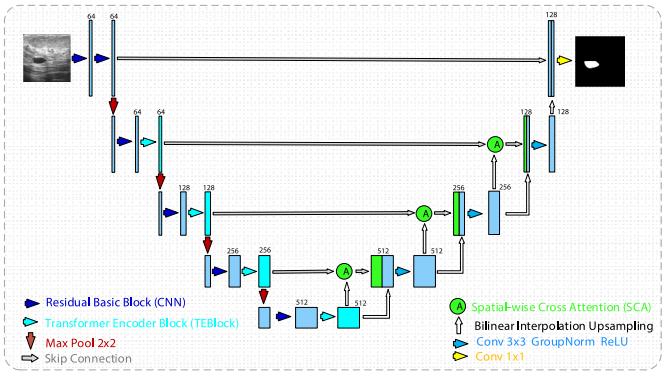


**Fig. 2.** The schematic diagram of HCTNet. The encoder extracts global features of breast ultrasound images at different scales by CNNs and TEBlocks. The SCA module is used in the decoder to reduce the semantic differences with the encoder. The details of TEBlock and SCA are shown in Figs. 3 and 4, respectively.

#### 3.2.1. Transformer Encoder Block

Ultrasound images inherently have speckle noise and shadow areas similar to the lesions. Convolution cannot directly learn long-range dependencies, resulting in CNNs that often contain non-lesion regions or lose part of lesion regions when extracting local spatial feature information. To address this problem of CNNs, we compensate for the deficiencies of convolution by introducing TEBlocks at multiple stages of the encoder to learn global contextual information. As shown in Fig. 2, the TEBlock explores the contextual relationships between pixel points in the feature maps. It takes the feature maps processed by the current convolutional layer as input and outputs the global contextual feature maps.

Fig. 3 shows a schematic illustration of TEBlock. TEBlock is constructed based on multi-head self-attention (MHSA), which allows the model to jointly attend to information from different representation subspaces at different positions [11]. TEBlock takes the convolutional feature map $X \in \mathbb{R}^{C\times H\times W}$ as input, where $H$, $W$ are the spatial height, width and $C$ is the number of channels. To avoid the excessive computational resources consumed when MHSA is applied directly to the feature map $X$, we first squeeze the channel $C$ with $1 \times 1$ convolution to obtain $X' \in \mathbb{R}^{C'\times H\times W}$, then input $X'$ into MHSA to obtain the attention-weighted feature map, and finally restore the channel of the feature map to $C$ by a $1 \times 1$ convolution. With MHSA, interdependencies are established among the pixels of the feature map $X'$, which helps HCTNet to distinguish the lesion and non-lesion regions. The residual connection [22] is applied after the second $1 \times 1$ convolution, which adds the previous feature map $X$ to the feature map processed by MHSA to accelerate the model optimization and reduce the learning difficulty of the attention-weighted feature map. The output $Y \in \mathbb{R}^{C\times H\times W}$:

$$Y = \sigma\{f_{conv}[Attention(\sigma(f_{conv}(X))) + X]\}, \quad (1)$$

where $f_{conv}$ denotes $1 \times 1$ convolution, $\sigma$ represents the ReLU activation function, and *Attention* denotes MHSA operations, respectively.

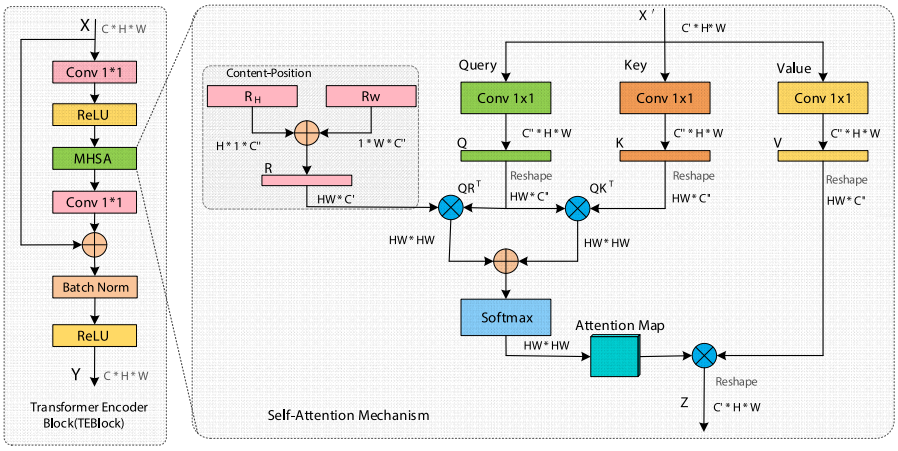


**Fig. 3.** The schematic diagram of TEBlock. MHSA takes a single self-attention mechanism as an example, where $\oplus$ and $\otimes$ denote matrix addition and matrix multiplication, respectively.

MHSA is the core of TEBlock, and we use four heads in this paper. For ease of presentation, Fig. 3 takes a single self-attention mechanism as an example. The input $X^{'}$ is projected into the embedding space of query (denoted Q), key (denoted K), and value (denoted V) using three $1 \times 1$ convolutions, respectively, where $\mathrm{Q}, \mathrm{K}, \mathrm{V} \in \mathbb{R}^{C'' \times H \times W}$ and $C''$ denotes the dimension of the channel at each head embedding. Then $\mathrm{Q}, \mathrm{K}, \mathrm{V}$ are reshaped and transposed to obtain a matrix of size $HW \times C''$. Each vector $q_i$ in matrix Q is dotted with the transpose matrix of $K$ and normalized by a Softmax function to get the contextual aggregation matrix or similarity matrix. Specifically, the contextual aggregation matrix of the $i$-th query vector (i.e., $q_i$) can be expressed as:

$$\mathscr{P}_{i} = softmax\left(\frac{q_i K^T}{\sqrt{C''}}\right), \tag{2}$$

where $\mathscr{P}_{i} \in \mathbb{R}^{1 \times HW}$. The outputs of all vectors in matrix Q then form a similarity matrix $\mathscr{P} \in \mathbb{R}^{HW \times HW}$. Using $\mathscr{P}$ as a weight to collect global contextual information from $V$, which represents the feature map information. The output of self-attention can be expressed as follows:

$$Attention(Q, K, V) = \underbrace{softmax\left(\frac{QK^T}{\sqrt{C'}}\right)}_{\mathcal{P}} V. \tag{3}$$

The attention mechanism mentioned above also requires special consideration of the relative position changes between pixel points, which is perturbation equivariant if the self-attention does not incorporate an explicit representation of the position information [23], which leads to ineffective modeling of highly structured image contents. Based on previous work [23–25], we adopt the 2D relative position encodings suitable for visual tasks [23]. By adding relative height and width information to the self-attention independently, the TEBlock takes into account the content information of the feature map and the relative distances between features at different positions, thus enabling an effectively associate between content information and positions. Therefore, we use relative position encoding for the feature map before the Softmax to represent the relative position relationship between the pixel points of the feature map. The pairwise attention logits between pixel $i$ and pixel $j$ can be expressed as:

$$l_{i,j} = \frac{q_i^T}{\sqrt{c'}}\left(k_j + r^W_{j_x - i_x} + r^H_{j_y - i_y}\right), \tag{4}$$

where $q_i$ represents the query vector ($i$-th row of Q) for pixel $i$, the $k_j$ is the key vector for pixel $j$, $r^W_{j_x - i_x}$ and $r^H_{j_y - i_y}$ are the learned embeddings for relative width $j_x - i_x$ and relative height $j_y - i_y$, respectively. The 2D relative position self-attention can be expressed as:

$$\mathrm{Z} = \mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = softmax\left(\frac{QK^T + S^{rel}_H + S^{rel}_W}{\sqrt{C'}}\right) V, \tag{5}$$

where Z is the output of MHSA in TEBlock, and $S^{rel}_H, S^{rel}_W \in \mathbb{R}^{HW \times HW}$ are matrics of relative position logits along height and width dimensions that satisfy $S^{rel}_H[i,j] = q_i^T r^H_{j_y - i_y}, S^{rel}_W[i,j] = q_i^T r^W_{j_x - i_x}$, respectively.

### 3.2.2. Spatial-wise Cross Attention

It is important to process the spatial information at the pixel level in breast ultrasound image segmentation. The problem of semantic discrepancies between the global feature maps obtained by the encoder at different scales and the feature maps from the decoder often makes the segmentation of breast ultrasound lesions unsatisfactory. We develop an SCA module to solve this problem. The SCA module squeezes the feature maps from the encoder and decoder along the channel dimension to obtain a fused attention-weight map and spatially excites it with the feature maps from the encoder, which we consider mitigates the problem of semantic discrepancies and can more clearly represent the position information of the lesion. In contrast to the existing spatial attention mechanisms such as Attention Gating (AG) in Attention Unet [34], the SCA module fuses the attention maps from the encoder and decoder respectively, which not only enhances the features of the region of interest, but also makes the final fused attention map alleviate the semantic discrepancy problem between the encoder and decoder.

As shown in Fig. 4, the input $E \in \mathbb{R}^{C \times h \times w}$ and $D \in \mathbb{R}^{C \times h \times w}$ are feature maps from the encoder and decoder, respectively. Both are squeezed by $1 \times 1$ convolution to generate the projection maps $p_{E(i,j)}, p_{D(i,j)} \in \mathbb{R}^{h \times w}$, where $E(i,j)$ and $D(i,j)$ corresponds to the spatial location $(i,j)$ with $i \in \{1, 2, \cdots, h\}$ and $j \in \{1, 2, \cdots, w\}$. These projection maps are rescaled to [0,1] by Sigmoid function to get the attention map $A_E$, $A_D \in \mathbb{R}^{1 \times h \times w}$. $A_E(i,j)$ and $A_D(i,j)$ correspond to the relative importance of the spatial location $(i,j)$ for the given feature maps $E$ and $D$, respectively. To reduce the semantic differences between the encoder and decoder, we add $A_E$ and $A_D$ to get the attention map $M \in \mathbb{R}^{h \times w}$, which integrates the relative importance of the information in feature maps E and D:

$$M = \left[\underbrace{\sigma_1(W_E E)}_{A_E} + \underbrace{\sigma_1(W_D D)}_{A_D}\right] \Big/ 2.0, \tag{6}$$

where $W_E, W_D$ are $1 \times 1$ convolution, and $\sigma_1$ is the Sigmoid activation function. $M$ is used to obtain the weighted feature map $T \in \mathbb{R}^{c \times h \times w}$ from the feature map $E$ with global contextual information. Therefore, the output feature map $T$ of the SCA module emphasizes the position in-

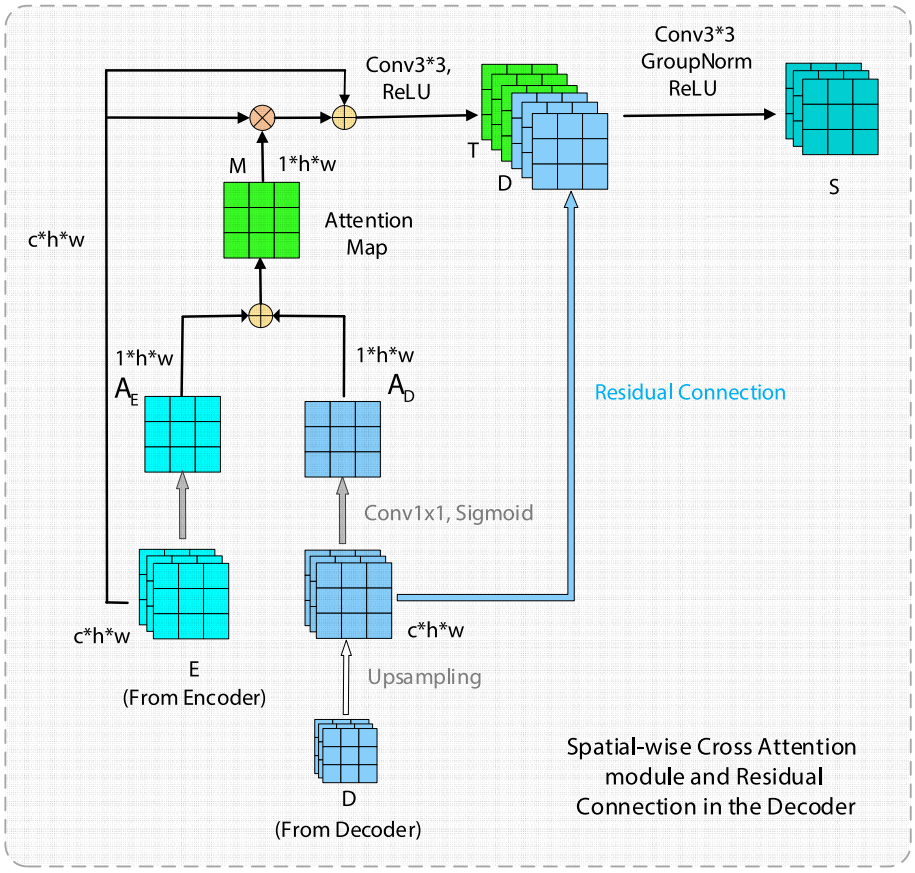


**Fig. 4.** The schematic illustration of the details of SCA module and the residual connection in the decoder. The input $E$ and $D$ of the SCA module are the feature maps from the encoder and decoder, respectively, and the output feature map is $T$. A residual connection is used after the upsampling of $D$ to obtain the final output feature map $S$ of HCTNet.

formation of the lesions based on the global context information:

$$T = \sigma_2[W_T(M \times E + E)], \tag{7}$$

where $\sigma_2(x) = \max(0, x)$ denotes the ReLU activation function and $W_T$ denotes the $3 \times 3$ convolution operation.

To further complement the details of lesions in the feature map $T$ and reduce noise disturbance, a residual connection is used after the upsampling of D. The merging of feature maps T and D at different semantic scales is achieved using the residual connection for multi-scale fusion prediction. After convolution, we obtained the output feature maps of the decoder block $S \in \mathbb{R}^{c \times h \times w}$:

$$S = \sigma_2[W_Z(Concate(T, D))], \tag{8}$$

where $W_Z$ denotes the $3 \times 3$ convolution operation.

## 4. Experiments and results

### 4.1. Implementation details

To speed up the training process, we initialize the parameters of the feature extraction network using ResNet18 [22] pre-trained on ImageNet while other parameters are initialized using Pytorch default method (kaiming_uniform [26]). HCTNet is trained by Adam optimizer with a mini-batch size of 4 and an epoch of 80. We use an initial learning rate of 0.0001 and use the ReduceLROnPlateau method to adjust it. In our experiments, we used the Dice loss [39] to train the model. All experiments are conducted using Pytorch on a NVIDIA GeForce GTX 1060 with 6 GB RAM.

### 4.2. Evaluation metrics

Six widely-used segmentation metrics are employed for quantitative comparison of different methods, including Dice coefficient (denoted as Dice), Jaccard index (denoted as Jaccard), Recall, Precision, Accuracy, and Hausdorff distance (denoted as HD).

### 4.3. Ablation study

The ablation study experiments in this section are conducted on the BUSI [28] dataset. We will show the effectiveness of the principal components of our network, including TEBlock, SCA, and residual connection in the decoder (RC-decoder).

Table 1 shows the comparison results of our method with different components. The first row represents the baseline, whose encoder is ResNet18 and decoder is Unet. By comparing the results of SCA (2nd row) and baseline, it proves that the SCA module has the advantage of reducing semantic differences, which helps capture the positional correlation of breast lesion segmentation. By comparing the results of TEBlock (3rd row) and baseline, we can conclude that modeling the long-range dependencies can achieve superior performance in segmenting breast lesions from ultrasound images. Comparing the 2nd row and the 4th row, it is apparent that SCA + RC-decoder have better results than SCA, showing that adding residual connection can further improve the breast lesion segmentation performance. Additionally, the combination of the TEBlocks and the SCA modules has superior segmentation results over only using TEBlocks, demonstrating that SCA modules can effectively solve the problem of semantic differences between the encoder and decoder sub-networks. Finally, HCTNet with all components has the best segmentation results in Dice, Jaccard, Recall, and

**Table 1**
Metric results of different components on the BUSI dataset.

| TEBlock | SCA | RC-decoder | Dice % | Jaccard % | Recall % | Precision % | Accuracy % |
|---|---|---|---|---|---|---|---|
| | | | 79.24(± 2.07) | 68.03(± 2.62) | 71.84(± 2.52) | 82.89(± 1.30) | 96.18(± 0.47) |
| | ✓ | | 80.81(± 0.92) | 69.34(± 1.17) | 80.62(± 1.41) | 83.04(± 1.17) | 96.20(± 0.36) |
| ✓ | | | 80.90(± 2.04) | 69.31(± 2.84) | 80.59(± 2.52) | 83.05(± 1.66) | 96.26(± 0.50) |
| | ✓ | ✓ | 81.30(± 1.49) | 70.12(± 2.02) | 80.35(± 1.84) | **84.37(± 2.59)** | 96.21(± 0.54) |
| ✓ | ✓ | | 81.72(± 1.38) | 70.66(± 1.82) | 82.11(± 1.94) | 82.84(± 1.39) | 96.28(± 0.70) |
| ✓ | ✓ | ✓ | **82.00(± 1.89)** | **71.84(± 2.52)** | **82.14(± 2.03)** | 83.24(± 1.86) | **96.94(± 0.48)** |

Accuracy. The suboptimal Precision may be due to the fact that the combination of the SCA module with RC-decoder is a simpler and lighter network than HCTNet, thus improving the Precision.

Fig. 5 visually compares the segmentation results produced by the baseline, baseline + SCA, baseline + TEBlock, and HCTNet on two examples. From Fig. 5, we can see that baseline + TEBlock performs better than baseline and baseline + SCA, which demonstrates that TEBlock can learn the long-range dependencies to boost the breast lesion segmentation performance. However, Transformer-based TEBlock has deficiencies in local (detail) processing, as shown in Fig. 5(e), the lesion region boundaries still affect the segmentation results. HCTNet (TEBlock + SCA + RC-decoder) can make a comparatively accurate segmentation performance.

## 5. Comparison with the state-of-the-arts

We compared HCTNet against several deep-learning-based segmentation methods, including U-Net [8], Unet++ [9], FCN [32], FPN [10], DeepLabV3+ [33], Attention U-net [34], TransUnet [16], and ultrasound breast lesions segmentation methods TFNet [37]. For TFNet, the quantitative performance is directly copied from the corresponding papers since the same dataset is used. For the other models, we obtained segmentation results by exploiting their public implementations and fine-tuning the training parameters.

Table 2, Table 3 and Table 4 report the segmentation results of different methods on the BUSI dataset, BUS dataset and Dataset B, respectively. Specifically, Table 2 quantitatively compares the average result (% ± standard deviation) of the six metrics of HCTNet with the other eight segmentation networks and presents the number of parameters for each model. As shown in Table 2, HCTNet improves the second-best method by 0.82%, 1.92%, 0.66% and 0.99% in Dice, Jaccard, Accuracy and Recall, respectively, and has the smallest HD value. TransUnet also applies the self-attention mechanism to the feature maps and integrates local features and global contextual information, thus obtaining quite good results on the BUSI and BUS datasets, as shown in Tables 2 and 3. Meanwhile, compared to Unet, HCTNet only increased the number of parameters by 4.94 M, but achieved better results. In Table 3, HCTNet improves 0.17%, 0.5%, 0.55% and 1.38% over the second-best method in Dice, Jaccard, Recall, and Precision. Also, in Table 4, HCTNet outperforms all competitors on all six metrics. The experimental results show that HCTNet achieves superior performance against other segmentation networks on all three datasets.

Fig. 6 visualizes the segmentation results of different methods on three breast ultrasound image datasets. Rows I - IV are the representative segmentation results on BUSI, and rows V - VI and row VII are the selected segmentation results on BUS and Dataset B, respectively. For the breast ultrasound images in the BUSI dataset shown in rows I - IV, the images mostly contain shaded or normal regions similar to lesions. The compared CNNs methods lack the learning of global context when processing local information, so they tend to ignore some details of the breast lesion region or incorporate other non-lesion regions into their predicted segmentation results. Although TransUnet can capture long-range dependencies, it uses the CNN as the backbone for feature extraction and applies Transformer to low-resolution feature maps, so the problem of over- or under-segmentation still exists, making the segmentation results unsatisfactory. For the breast ultrasound images in the BUS dataset shown in rows V - VI, the lesion regions are well-defined but the tumor size varies widely, and the segmentation needs to be combined with the dependency with distant pixels for a more accurate segmentation. The CNNs lack this capability and therefore misidentify some of the normal regions, generating over-segmentation and failing to accurately segment the lesion regions. For the breast ultrasound images in the Dataset B dataset shown in the last row, the lesion is large and clear contoured, so all methods achieve good segmentation results.

In general, the lack of learning long-range dependencies in CNNs leads to the segmentation results of these networks (Unet, Unet++, FPN, FCN8s, Deeplabv3+, Attention Unet) including some non-lesioned regions or lose some breast lesioned regions, which reduces the performance of segmentation. One major characteristic of TransUNet is to treat CNNs as main bodies, on top of which Transformer are further applied to capture long-term dependencies. However, this may lead to a problem that the advantages of Transformer are not fully exploited.

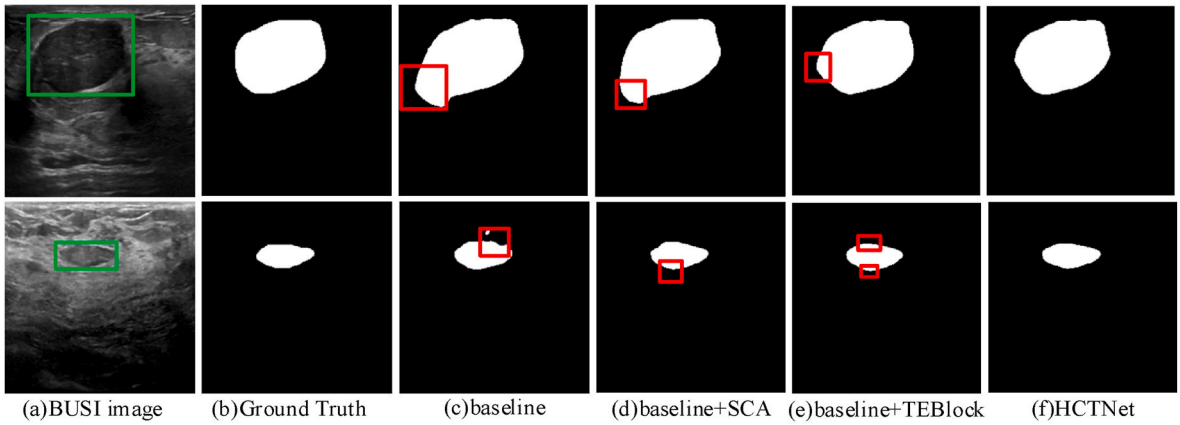


**Fig. 5.** Visualization of ablation experimental results. (a) Input images; (b) Ground truth; (c–f) indicate the segmentation results of baseline, baseline + SCA, baseline + TEBlock, and HCTNet, respectively. The green box indicates the position of the lesions in the ultrasound image, and the red box shows the mis-segmented areas.

**Table 2**
Quantitative comparison of the number of parameters of different methods and their segmentation performance on BUSI dataset. The best results are shown in bold.

| Method | Dice % | Accuracy % | Jaccard % | Recall % | Precision % | HD | Parameters |
|---|---|---|---|---|---|---|---|
| Unet [8] | 73.58(± 1.02) | 95.22(± 0.29) | 59.88(± 1.31) | 73.88(±1.91) | 76.87(±2.53) | 52.76(±2.43) | 17.266 M |
| Unet++ [9] | 76.40(± 2.52) | 95.39(± 0.86) | 63.34(± 3.29) | 77.51(±2.97) | 78.47(±4.70) | 43.90(±3.28) | 9.163 M |
| FCN8s [32] | 77.56(± 2.24) | 95.96(± 0.47) | 65.27(± 2.87) | 74.76(±1.84) | 83.51(±2.58) | 40.45(±3.44) | 11.171 M |
| FPN [10] | 78.92(± 2.75) | 96.11(± 0.56) | 67.15(± 3.42) | 75.31(±4.84) | **85.98(±1.70)** | 44.79(±1.67) | 12.492 M |
| DeepLabV3+ [33] | 80.22(± 3.40) | 96.19(± 0.80) | 68.97(± 3.99) | 79.03(±2.67) | 83.93(±3.81) | 38.78(±4.50) | 40.341 M |
| AttentionUnet [34] | 75.51(± 2.66) | 95.49(± 0.49) | 62.65(± 2.74) | 73.68(±4.65) | 81.18(±2.22) | 43.60(±3.64) | 34.877 M |
| TransUnet [16] | 81.18(± 2.27) | 96.28(± 0.51) | 69.92(± 3.03) | 81.85(±3.89) | 82.51(±3.01) | 35.26(±3.37) | 44.00 M |
| TFNet [37] | 77.3(± 5.5) | – | 63.0(± 7.0) | 79.5(± 11.1) | 75.5(± 2.3) | – | 33.89 M |
| HCTNet | **82.00(± 1.89)** | **96.94(± 0.48)** | **71.84(± 2.52)** | **82.14(±2.03)** | 83.24(±1.86) | **34.55(±3.71)** | 22.204 M |

**Table 3**
Quantitative comparison of segmentation performance using different methods on BUS dataset. The best results are shown in bold.

| Method | Dice % | Accuracy % | Jaccard % | Recall % | Precision % | HD |
|---|---|---|---|---|---|---|
| Unet [8] | 77.47(± 5.74) | 97.98(± 0.47) | 65.00(± 6.52) | 73.95(± 7.20) | 85.24(± 1.82) | 41.47(± 7.83) |
| Unet++ [9] | 79.00(± 2.58) | 98.07(± 0.43) | 66.04(± 3.40) | 73.63(± 3.18) | 87.12(± 2.43) | 35.94(± 5.53) |
| FCN8s [32] | 78.31(± 5.01) | 98.06(± 0.40) | 65.20(± 6.39) | 73.97(± 5.11) | 85.25(± 1.67) | 23.50(± 4.88) |
| FPN [10] | 83.32(± 4.59) | 98.38(± 0.39) | 72.34(± 6.15) | 80.51(± 4.53) | 84.90(± 3.36) | 50.56(± 9.61) |
| DeepLabV3+ [33] | 81.41(± 5.18) | 98.32(± 0.28) | 69.68(± 6.92) | 79.19(± 6.76) | 86.04(± 2.46) | 24.04(± 6.72) |
| AttentionUnet [34] | 79.39(± 3.39) | 98.10(± 0.48) | 67.09(± 4.09) | 76.02(± 3.96) | 85.57(± 2.83) | 32.21(± 6.83) |
| TransUnet [16] | 83.96(± 3.26) | **98.60(± 0.46)** | 73.33(±4.22) | 82.64(± 2.78) | 84.57(± 4.00) | **21.14(± 8.79)** |
| HCTNet | **84.13(± 2.02)** | 98.49(± 0.36) | **73.83(± 2.78)** | **83.19(± 3.12)** | **88.50(± 3.06)** | 21.66(± 5.52) |

**Table 4**
Quantitative comparison of segmentation performance using different methods on Dataset B. The best results are shown in bold.

| Method | Dice % | Accuracy % | Jaccard % | Recall % | Precision % | HD |
|---|---|---|---|---|---|---|
| Unet [8] | 95.32(± 0.22) | 95.57(± 0.20) | 91.08(± 0.41) | 95.65(± 0.35) | 95.03(± 0.35) | 19.56(± 2.13) |
| Unet++ [9] | 95.27(± 0.30) | 95.54(± 0.26) | 91.00(± 0.55) | 95.45(± 0.75) | 95.15(± 0.73) | 21.50(± 3.00) |
| FCN8s [32] | 96.60(± 1.05) | 96.83(± 0.96) | 93.46(± 1.94) | 96.21(± 1.07) | 97.02(± 1.04) | 19.56(± 1.90) |
| FPN [10] | 96.95(± 1.09) | 97.12(± 1.05) | 94.11(± 2.04) | 96.87(± 0.99) | 97.06(± 1.18) | 21.02(± 1.74) |
| DeepLabV3+ [33] | 94.09(± 0.31) | 94.46(± 0.28) | 88.88(± 0.54) | 94.32(± 0.68) | 93.99(± 1.01) | 19.89(± 1.73) |
| AttentionUnet [34] | 95.42(± 0.26) | 95.69(± 0.17) | 91.25(± 0.48) | 95.63(± 0.45) | 95.25(± 0.45) | 20.29(± 1.18) |
| TransUnet [16] | 94.83(± 0.35) | 95.12(± 0.29) | 90.18(± 0.63) | 95.54(± 1.11) | 94.20(± 1.01) | 19.65(± 2.68) |
| HCTNet | **97.23(± 1.07)** | **97.41(± 0.95)** | **94.63(± 2.00)** | **97.33(± 1.02)** | **97.14(± 1.10)** | **19.35(± 2.42)** |

Although TFNet uses Transformer to fuse multi-scale features, the decoding process completed by two-step upsampling has defects in recovering the image detail. In contrast, HCTNet uses a hybrid stem where convolution and self-attention are interleaved to give full play to their strengths. Therefore, HCTNet outperforms TFNet in terms of segmentation performance and model size. Table 2, Table 3, and Table 4 show that the experimental results obtained by the same network on three different public datasets differ significantly. We speculate that this may be due to the different ultrasound image acquisition devices and their different processing methods of the images leading to the influence and bias on the results. The BUSI dataset is the most challenging because the lesions in this dataset are highly variable in size, some lesion boundaries are irregular, and the effect of noise is more evident, yet the segmentation results of HCTNet in this dataset are still good.

## 6. Discussions

Our study shows that bridging CNN and Transformer is very useful for ultrasonic breast tumor segmentation. The segmentation results on three public datasets show that HCTNet performs best for all breast tumor segmentation tasks of different difficulties, indicating the good robustness of our method. However, there are still some limitations in this work. First, we did not use post-processing methods such as CRF [35] to further optimize the segmentation results. Second, if the image to be segmented is very complex, it will lead to incorrect segmentation results. For example, as shown in Fig. 7(I), when there are too many lesion-like shadows in the ultrasound image, the segmentation results are not satisfactory. Another example is shown Fig. 7(II), where the lesion edges are too confusing and blurred. Although HCTNet can segment the lesion region roughly correctly, it is inaccurate on boundary details. In future work, we will analyze these problems in depth.

Furthermore, we believe that the effectiveness of segmentation can be further improved by the multi-functionalization of the model, such as adding classification, detection, and identification of benign and malignant lesion regions to the segmentation network. In clinical practice, it will also be more beneficial to doctors for the diagnosis and subsequent treatment of the disease.

## 7. Conclusion

In this paper, we propose a Hybrid CNN-Transformer Network (HCTNet) for breast ultrasound image segmentation. The encoder of HCTNet combines the inductive bias of CNNs in modeling spatial correlations and the power of Transformer in modeling long-range dependencies, which makes up for the inefficiency of CNNs in acquiring global contextual information and complements the short-comings of Transformer in developing local details. In the decoder part, we propose the SCA module to fuse the feature maps in the encoder and decoder subnetworks to address the inconsistency of semantic information. Moreover, we use residual connection for the fusion of semantic information at different scales, which can reduce the noise in ultrasound images and the loss of details caused by direct upsampling, and obtain the accurate position information of lesion regions. The performance of HCTNet was evaluated on three public datasets, showing that HCTNet has good robustness and generalization. Furthermore, compared to Unet, HCTNet adds very few parameters and it remains lightweight, which is very

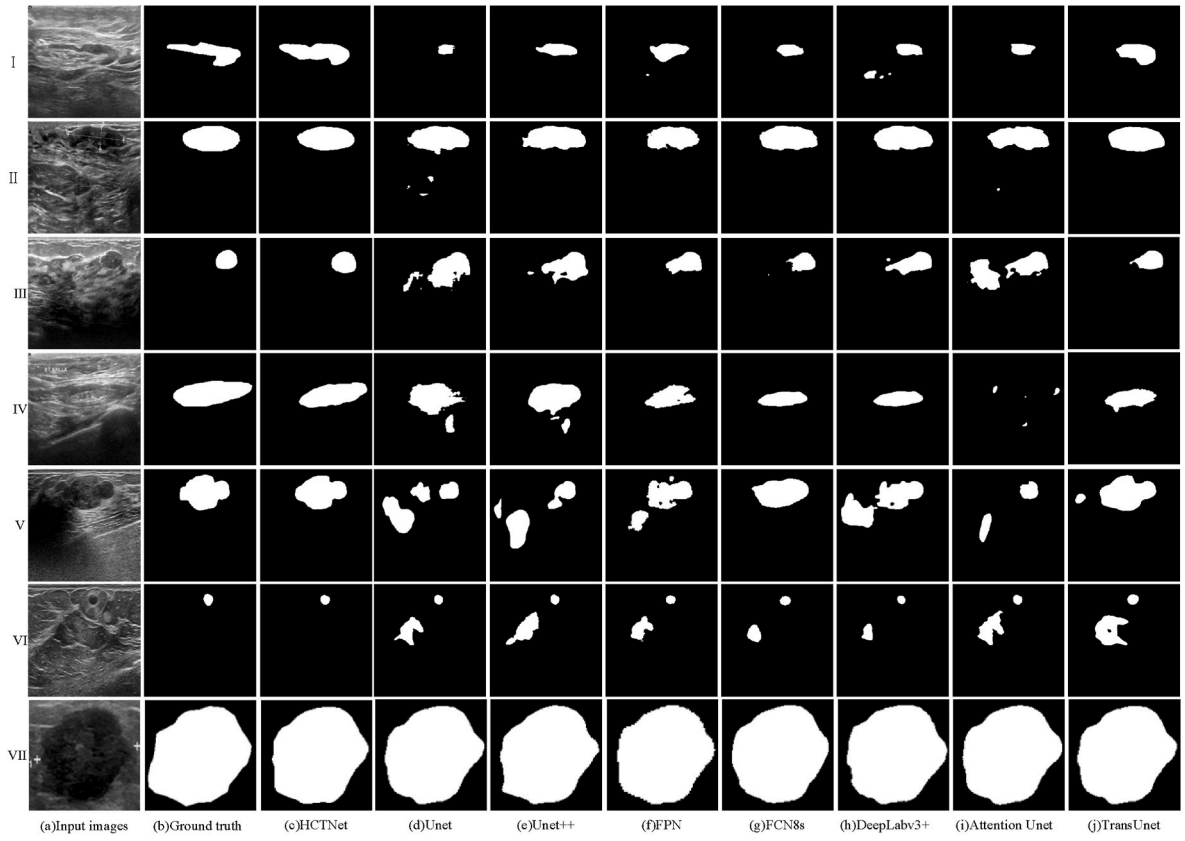


**Fig. 6.** Visualization of the segmentation results of breast lesions generated by different segmentation methods. (a) Input images; (b) Ground truth; (c)–(j) are the segmentation results of breast lesions obtained by HCTNet, Unet, Unet++, FPN, FCN8s, Deeplabv3+, Attention Unet, and TransUnet, respectively.

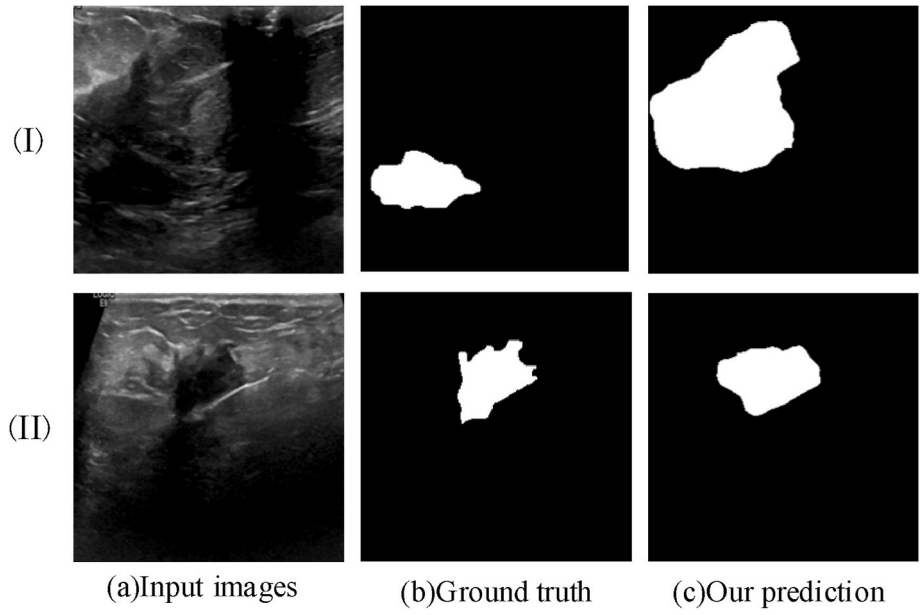


**Fig. 7.** Visualization of the incorrect segmentation results. (a) The input breast ultrasound image; (b) Ground truth of the breast ultrasound lesion; (c) segmentation results obtained by HCTNet.

important for computer-aided diagnosis systems.

## Declaration of competing interest

All authors declare that there are no conflict of interests, we do not have any possible conflicts of interest.

## Acknowledgments

This work was supported by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2023-JC-YB-228, and the Open Fund of State Key Laboratory of Loess and Quaternary Geology under Grant SKLLQGZR2201.

## References


[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA A Cancer J. Clin. 71 (3) (2021) 209–249, https://doi.org/10.3322/caac.21660.

[2] C. Xue, L. Zhu, H. Fu, X. Hu, X. Li, H. Zhang, P. Heng, 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.

[3] B. Shareef, A. Vakanski, M. Xian, P.E. Freer, ESTAN: Enhanced Small Tumor-Aware Network for Breast Ultrasound Image Segmentation, 2020, 12894, https://doi.org/10.48550/arXiv.2009.12894. ArXiv, abs/2009.

[4] Y. Hu, Y. Guo, Y. Wang, J. Yu, J. Li, S. Zhou, C. Chang, Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model, Med. Phys. 46 (1) (2019) 215–228, https://doi.org/10.1002/mp.13268.

[5] M.H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, R. Marti, Automated breast ultrasound lesions detection using convolutional neural networks, IEEE j. biomed. health inf. 22 (4) (2017) 1218–1226, https://doi.org/10.1109/JBHI.2017.2731873.

[6] L. Zhu, R. Chen, H. Fu, C. Xie, L. Wang, L. Wan, P.A. Heng, A second-order subregion pooling network for breast lesion segmentation in ultrasound, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2020, October, pp. 160–170, https://doi.org/10.1007/978-3-030-59725-2_16.

[7] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803, https://doi.org/10.1109/CVPR.2018.00813.

[8] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2015, October, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

[9] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imag. 39 (6) (2019) 1856–1867, https://doi.org/10.1109/TMI.2019.2959609.

[10] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125, https://doi.org/10.1109/CVPR.2017.106.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017), https://doi.org/10.48550/arXiv.1706.03762.

[12] J.M.J. Valanarasu, P. Oza, I. Hacihaliloglu, V.M. Patel, Medical transformer: gated axial-attention for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2021, September, pp. 36–46, https://doi.org/10.1007/978-3-030-87193-2_4.

[13] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like Pure Transformer for Medical Image Segmentation, 2021, https://doi.org/10.48550/arXiv.2105.05537 arXiv preprint arXiv:2105.05537.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, N. Houlsby, An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020, https://doi.org/10.48550/arXiv.2010.11929 arXiv preprint arXiv: 2010.11929.

[15] H. Wang, X. Chen, T. Zhang, Z. Xu, J. Li, CCTNet: coupled CNN and transformer network for crop segmentation of remote sensing images, Rem. Sens. 14 (2022) 1956, https://doi.org/10.3390/rs14091956.

[16] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, 2021, https://doi.org/10.48550/arXiv.2102.04306. ArXiv, abs/2102.4306.

[17] Y. Chang, M. Hu, G. Zhai, X. Zhang, TransClaw U-Net: Claw U-Net with Transformers for Medical Image Segmentation, 2021, https://doi.org/10.48550/arXiv.2107.05188. ArXiv, abs/2107.5188.

[19] Y. Gao, M. Zhou, D.N. Metaxas, UTNet: a hybrid transformer architecture for medical image segmentation, MICCAI (2021), https://doi.org/10.1007/978-3-030-87199-4_6.

[20] H.Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, Y. Yu, nnFormer: Interleaved Transformer for Volumetric Segmentation, 2021, https://doi.org/10.48550/arXiv.2109.03201 arXiv preprint arXiv:2109.03201.

[21] M. Xu, K. Huang, Q. Chen, X. Qi, Mssa-net: multi-scale self-attention network for breast ultrasound image segmentation, in: 2021 IEEE 18th International Symposium on Biomedical Imaging, ISBI), 2021, pp. 827–831, https://doi.org/10.1109/ISBI48211.2021.9433899.

[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90, doi: 10.1109/CVPR.2016.90.

[23] I. Bello, B. Zoph, A. Vaswani, J. Shlens, Q.V. Le, Attention augmented convolutional networks, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3285–3294, https://doi.org/10.1109/ICCV.2019.00338.

[24] P. Shaw, J. Uszkoreit, A. Vaswani, Self-attention with relative position representations, arXiv preprint arXiv:1803.02155 (2018), https://doi.org/10.48550/arXiv.1803.02155.

[25] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, Stand-alone self-attention in vision models, Adv. Neural Inf. Process. Syst. 32 (2019), https://doi.org/10.48550/arXiv.1906.05909.

[26] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on ImageNet classification, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026–1034, https://doi.org/10.1109/ICCV.2015.123.

[28] W. Al-Dhabyani, M.M. Gomaa, H. Khaled, A.A. Fahmy, Dataset of breast ultrasound images, Data Brief 28 (2020), https://doi.org/10.1016/j.dib.2019.104863.

[29] M.H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, A.K. Davison, R. Marti, Automated breast ultrasound lesions detection using convolutional neural networks, IEEE j. biomed. health inf. (2017), https://doi.org/10.1109/JBHI.2017.2731873 doi: 10.1109/JBHI.2017.2731873.

[30] Q. Huang, Y. Huang, Y. Luo, F. Yuan, X. Li, Segmentation of breast ultrasound image with semantic classification of superpixels, Med. Image Anal. 61 (2020), 101657, https://doi.org/10.1016/j.media.2020.101657.

[32] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440, https://doi.org/10.1109/CVPR.2015.7298965.

[33] L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818, https://doi.org/10.1007/978-3-030-01234-2_49.

[34] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, D. Rueckert, Attention U-Net: Learning where to Look for the Pancreas, 2018, https://doi.org/10.48550/arXiv.1804.03999 arXiv preprint arXiv:1804.03999.

[35] P. Krähenbühl, V. Koltun, Efficient inference in fully connected crfs with Gaussian edge potentials, Adv. Neural Inf. Process. Syst. 24 (2011), https://doi.org/10.48550/arXiv.1210.5644.

[37] T. Wang, Z. Lai, H. Kong, TFNet: transformer fusion network for ultrasound image segmentation, in: Asian Conference on Pattern Recognition, Springer, Cham, 2022, pp. 314–325, https://doi.org/10.1007/978-3-031-02375-0_23.

[38] Yu Mo, et al., HoVer-Trans: Anatomy-Aware HoVer-Transformer for ROI-free Breast Cancer Diagnosis in Ultrasound Images, 2022, https://doi.org/10.48550/arXiv.2205.08390. ArXiv abs/2205.08390 n. pag.

[39] Fausto Milletarì, et al., V-net: fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), 2016, pp. 565–571, https://doi.org/10.1109/3DV.2016.79.