Expert Systems With Applications 213 (2023) 119024

Contents lists available at ScienceDirect

Expert Systems With Applications

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

CSwin-PNet: A CNN-Swin Transformer combined pyramid network for breast lesion segmentation in ultrasound images ☆

Haonan Yang[1], Dapeng Yang *

State Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin 150001, China
Artificial Intelligence Laboratory, Harbin Institute of Technology, Harbin 150001, China

ARTICLE INFO

Keywords:
Breast lesion segmentation
Ultrasound image
Swin transformer
Pyramid network
Feature fusion

ABSTRACT

Currently, the automatic segmentation of breast tumors based on breast ultrasound (BUS) images is still a challenging task. Most lesion segmentation methods are implemented based on a convolutional neural network (CNN), which has limitations in establishing long-range dependencies and obtaining global context information. Recently, transformer-based models have been widely used in computer vision tasks to build long-range contextual information due to their powerful self-attention mechanism, and their effect is better than that of a traditional CNN. In this paper, a CNN and a Swin Transformer are linked as a feature extraction backbone to build a pyramid structure network for feature encoding and decoding. First, we design an interactive channel attention (ICA) module using channel-wise attention to emphasize important feature regions. Second, we develop a supplementary feature fusion (SFF) module based on the gating mechanism. The SFF module can supplement the features during feature fusion and improve the performance of breast lesion segmentation. Finally, we adopt a boundary detection (BD) module to pay additional attention to the boundary information of breast lesions to improve the boundary quality in the segmentation results. Experimental results show that our network outperforms state-of-the-art image segmentation methods on breast ultrasound lesion segmentation.

1. Introduction

Breast cancer is a terrible disease that affects the lives and health of women worldwide and is one of the most common causes of death among women. According to statistics reported by the American Cancer Society (2021), an estimated 43,600 breast cancer deaths occurred in 2021 (Siegel et al., 2021). An early diagnosis of breast cancer is critical for improved survival (Bleicher et al., 2015). Ultrasound imaging, as a mature technology, is a noninvasive, nonradiative and low-cost imaging method that is widely used in the clinical diagnosis of breast tumors (Xian et al., 2018). However, the low contrast, high noise, and high similarity between tissues in ultrasound images make it difficult to identify the complete lesion tissue, even for ultrasound experts. There are differences in the observations of different experts. Computer-aided diagnosis (CAD) systems can assist radiologists in interpretation and diagnosis (Bai et al., 2021; Jalalian et al., 2017; Samulski et al., 2010; Yanase & Triantaphyllou, 2019). In this case, a CAD system based on breast ultrasound (BUS) images can assist physicians in detecting lesions, improve diagnostic accuracy and reduce subjectivity to a certain extent. A CAD system uses machine learning and computer vision techniques to extract morphological and textural features from ultrasound images to separate lesion areas from the background for lesion segmentation (Yassin et al., 2018; Zhu et al., 2021). CAD systems have been proven to be efficient and accurate in achieving lesion segmentation (Horsch et al., 2001; Moon et al., 2020).

In recent years, convolutional neural networks (CNNs) have been widely used in medical imaging fields, including detection, classification and semantic segmentation, and have achieved outstanding performance (Rai et al., 2019; Thiyagarajan & Murukesh, 2020). The most representative networks are the fully convolutional network (FCN) (Shelhamer et al., 2017) and U-Net (Ronneberger et al., 2015). Yap et al. (2019) developed a segmentation model for breast lesions in BUS images based on FCN. Hu, Guo et al. (2019) developed a fully convolutional network for breast tumor segmentation based on FCN using fused dilated convolutions. Ghosh et al. (2020) proposed an improved U-Net model by incorporating global features for the automatic segmentation of BUS images. However, ultrasound images are characterized by low

☆ Funding: This work was supported in part by the National Natural Science Foundation of China Grant #52075114, Interdisciplinary Research Foundation of HIT (IR2021218), and Postdoctoral Scientific Research Development Fun (LBH-W18058) to D. Yang.
* Correspondence to: Technology Innovation Building K404, Harbin Institute of Technology, Harbin 150001, China.
E-mail addresses: haonan_yang@hit.edu.cn (H. Yang), yangdapeng@hit.edu.cn (D. Yang).
[1] Technology Innovation Building K404, Harbin Institute of Technology, Harbin 150001, China.

https://doi.org/10.1016/j.eswa.2022.119024
Received 22 June 2022; Received in revised form 30 July 2022; Accepted 9 October 2022
Available online 17 October 2022

contrast, high noise, blurred edges, and variable shapes and locations of breast lesions, and a simple CNN may not perform well. An attention mechanism (Mnih et al., 2014), as an effective means to improve the performance of CNNs, improves feature extraction through the sensitive control of local features, the most representative of which is Attention U-Net (Oktay et al., 2018). Tong et al. (2021) proposed an improved U-Net model based on a mixed attention loss function and Attention U-Net. Zhuang et al. (2019) proposed the residual attention gate network (RDAU-Net), which utilizes residual units to enhance edge information and improve segmentation performance. Saliency maps can be used to highlight visually salient regions or objects in an image. These maps can strengthen the network's attention to the region of interest and help to improve the segmentation performance of the network (Ramadan et al., 2020; Vakanski et al., 2020). For example, Ning et al. (2022) proposed a saliency-guided morphology-aware U-Net (SMU-Net) model that utilizes salient foreground and background images. SMU-Net integrates shape-aware, edge-aware and position-aware blocks to improve the network's ability to learn the morphological information of breast lesions. In addition, adding prior knowledge to the network also helps feature segmentation (Xi et al., 2017).

Ultrasound images have low contrast, and many pixels do not contain diseased tissue that has a similar appearance to the diseased tissue pixels. These nonlocal features can be learned by capturing global background information to improve the ability of the network to discriminate features. Previous studies have proposed the use of dilated convolution to expand the receptive field (Chen et al., 2018; Li et al., 2021). However, it fails to capture contextual information from a global perspective and capturing long-distance dependency information is helpful for distinguishing features. Xue et al. (2021) proposed a global guidance network (GG-Net) for breast tumor segmentation. They used multilayer CNN information as guiding information to learn long-range nonlocal dependencies in space and channels, thereby enhancing the network learning ability. Recently, the transformer architecture (Vaswani et al., 2017), originally applied to natural language processing (NLP) tasks, has received much attention in the computer vision field. The multiheaded self-attention (MSA) mechanism in the transformer architecture that is used for the construction of global relationships is applicable to pixel-based CV tasks. Carion et al. (2020) established the first end-to-end object detection model based on the transformer model. Dosovitskiy et al. (2020) used a transformer to replace a CNN, established a transformer-based image recognition model called Vision Transformer (ViT), and achieved comparable performance to other state-of-the-art methods that use convolutional technology. Wang et al. (2021) built a pyramid ViT (PVT) model to extract multiscale feature maps. The PVT model reduces the computational complexity to a certain extent, but its complexity is still quadratic with the image size. Liu, Lin et al. (2021) proposed a Swin Transformer using a moving window strategy, which greatly reduces the computational complexity and enables cross-window information exchange with advanced performance. The transformer architecture has shown great potential for wide use in downstream computer vision tasks (classification, detection and segmentation) (Liu, Zhang et al., 2021). Due to the successful application of transformers in computer vision tasks, some studies have explored the possibility of transformers in medical image processing (Chen, Liu et al., 2021; Hatamizadeh et al., 2022; Valanarasu et al., 2021). Cao et al. (2021) replaced the convolutional encoding and decoding operations in U-Net with a Swin Transformer module and established Swin-UNet. Lin et al. (2022) constructed a Swin-UNet-like architecture by using parallel dual-Swin Transformer modules and proposed a transformer-based feature interaction fusion module. Wang et al. (2022) used transformers to replace skip connections in U-Net to obtain an improved U-Net model. Zhang et al. (2021) proposed a ViT-based TransFuse network and tried to fuse a transformer and a CNN to achieve feature extraction. Chen, Lu et al. (2021) proposed TransUNet, which uses a CNN to extract features and then feeds them into a transformer for long-range dependency modeling. To alleviate the insufficiency of CNNs in global modeling, inspired by TransUNet, this paper proposes a novel segmentation network for breast lesion segmentation in BUS images. The Swin Transformer module is used to assist the CNN in global feature extraction, and a feature pyramid network (FPN) structure (Lin et al., 2017) is constructed to achieve multiscale feature fusion. In summary, our main contributions are as follows:

1. We build a residual Swin Transformer block (RSTB) based on a Swin Transformer to globally model the features extracted by the CNN.
2. We design an interactive channel attention (ICA) module based on a channel attention mechanism. It utilizes the supervised encoder features to output the features of each layer of the encoder to focus on tumor-related regions and assigns large weights to these feature channels.
3. We propose a supplementary feature fusion (SFF) module based on a gating mechanism. The encoding process is made to selectively receive features from the encoder to supplement the feature information of the encoder.
4. We add a boundary detection (BD) module in the middle layer of the decoder to identify the boundary map of breast lesions and obtain lesion features with high-quality boundaries.

2. Methodology

The novel breast lesion segmentation network proposed in this paper is shown in Fig. 1. Overall, the network takes BUS images as input and produces segmentation results of glandular lesions in an end-to-end manner. The shallow layer of the CNN pays more attention to the texture and structure information, which is not conducive to the transformer capturing the global information, and it introduces additional computation (Gao et al., 2021). Therefore, our network uses a CNN to process shallow structural information and a transformer to extract deep global features to build the pyramid network. In the encoder stage, a residual module with the GEUL activation function (Fig. 2(a)) is used to process shallow structural information to obtain features with different spatial resolutions. The convolution operation with stride = 2 is used to replace the maxpool operation to prevent the loss of image features during the downsampling process. The output image size after each feature extraction step is 1, 1/2, and 1/4 of the input image size. The output features of the three residual modules are denoted as E_i, $i \in \{1,2,3\}$. Then, the feature map output by the last residual block is concatenated into an RSTB through a 1×1 convolution for deep feature extraction. To better extract the context information in the deep features, this paper uses a Swin Transformer instead of a standard transformer, and it is described in detail in Section 2.1. In the decoder stage, a residual module with the ReLU activation function (Fig. 2(b)) is used for feature decoding. The output feature map of each layer of the decoder is denoted as D_i, $i \in \{1,2,3,4\}$. To utilize the feature information of different layers in the encoding stage, ICA and SFF are introduced. The convolutional layer is enhanced to pay attention to the lesion area through the channel attention of ICA. The feature complementation of the decoding subnetwork is realized by using SFF. Subsequently, BD is introduced in the decoding intermediate layer to capture the breast lesion contours and obtain segmentation features with high-quality boundary information.

2.1. Residual Swin Transformer block

In ViT, a standard transformer block includes MSA, a multilayer perceptron (MLP) and a layer norm (LN), as shown in Fig. 3(a). The output of standard transformer z^l is:

$$\begin{aligned} \hat{z}^l &= MSA(LN(z^{l-1})) + z^{l-1}, \\ z^l &= MLP(LN(z^l)) + \hat{z}^l. \end{aligned} \tag{1}$$

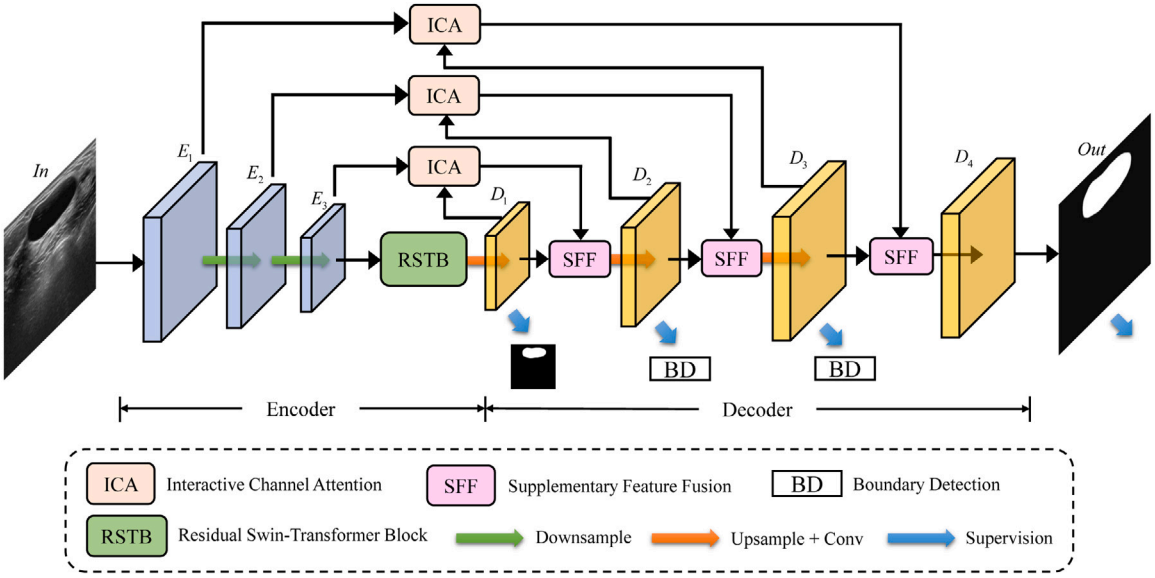

Fig. 1. The architecture of the proposed breast lesion segmentation network (CSwin-PNet) in this work.

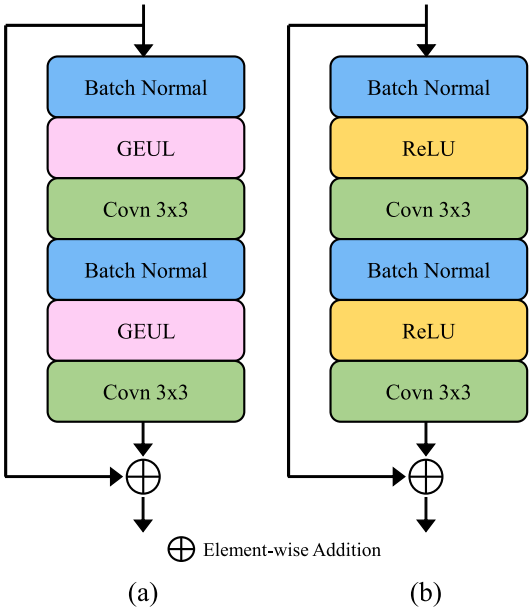

Fig. 2. Residual convolution block.

The standard transformer block uses MSA to compute the relationship between each marker. The computational complexity of MSA is equal to the fourth power of the number of markers, which makes it unsuitable for dense prediction and high-resolution image tasks. In contrast to ViT, a Swin Transformer is a hierarchical transformer and uses an efficient self-attention strategy. The normal MSA approach is replaced by the regular window-based MSA (W-MSA) approach and the shifted window MSA (SW-MSA) approach, as shown in Fig. 3(b). The W-MSA strategy places some patches inside each window, and each window is independent of the others and contains M × M patches (default M = 7). The SW-MSA approach handles the interaction of information between windows and allows cross-window connections. They are denoted as follows:

$$\begin{aligned} \hat{z}^{l} &= W-MSA(LN(z^{l-1}))+z^{l-1}, \\ z^{l} &= MLP(LN(\hat{z}^{l}))+\hat{z}^{l}, \\ \hat{z}^{l+1} &= SW-MSA(LN(z^{l}))+z^{l}, \\ z^{l+1} &= MLP(LN(\hat{z}^{l+1}))+\hat{z}^{l+1}. \end{aligned} \tag{2}$$

The interaction between information is enhanced by the alternate execution of W-MSA and SW-MSA. However, this approach limits the attention to the local window, which to a certain extent weakens the global modeling capability of the transformer. Hence, we design an RSTB (see Fig. 4(a)) to further enhance information exchange. We use a Swin Transformer with a "tiny" configuration pretrained on ImageNet. It has been shown that pretrained transformer models based on ImageNet can also be well applied to feature extraction of medical images and can perform better than pretrained models based on medical images (Brandt et al., 2021; Matsoukas et al., 2022). In addition, we use the feature information output from each Swin Transformer layer to achieve complementarity between the different layers. The process is as follows: (1) The output features of the last layer and the middle layer are fed into a 1 × 1 convolution operation and unsampled to spatially match the output features of the top layer. (2) The output feature map of each RSTB layer is used in the summation processing method for fusion. The fusion result is used as the output of the RSTB.

2.2. Interactive channel attention module

Features in different channels have different semantic information, but not all semantic objects need to be detected. We need to effectively highlight the object to be detected in the feature channel and suppress the irrelevant information in the feature channel to alleviate the interference of non-salient objects. We design an ICA module to focus on the channel features of interest; (see Fig. 4(b)). The channel feature of supervision can better highlight the object to be detected. We assign large weights to the channel feature to focus on the lesion regions in the output feature channels of the convolutional layer. According to ECA-Net (Wang et al., 2020), it is shown that the dimensional decay in the two FC layers after GAP in SENet (Hu, Shen et al., 2019) affects the weight learning of channel attention. We use a single linear layer

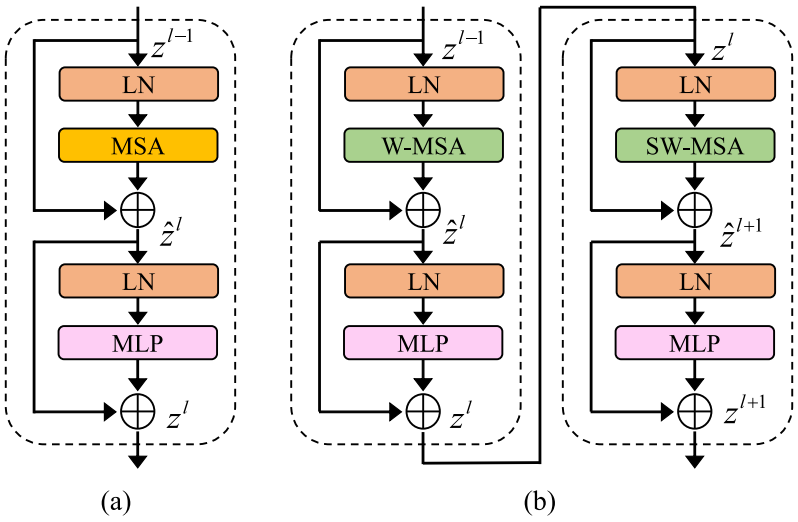

Fig. 3. Transformer blocks: (a) Standard transformer; (b) Swin Transformer.

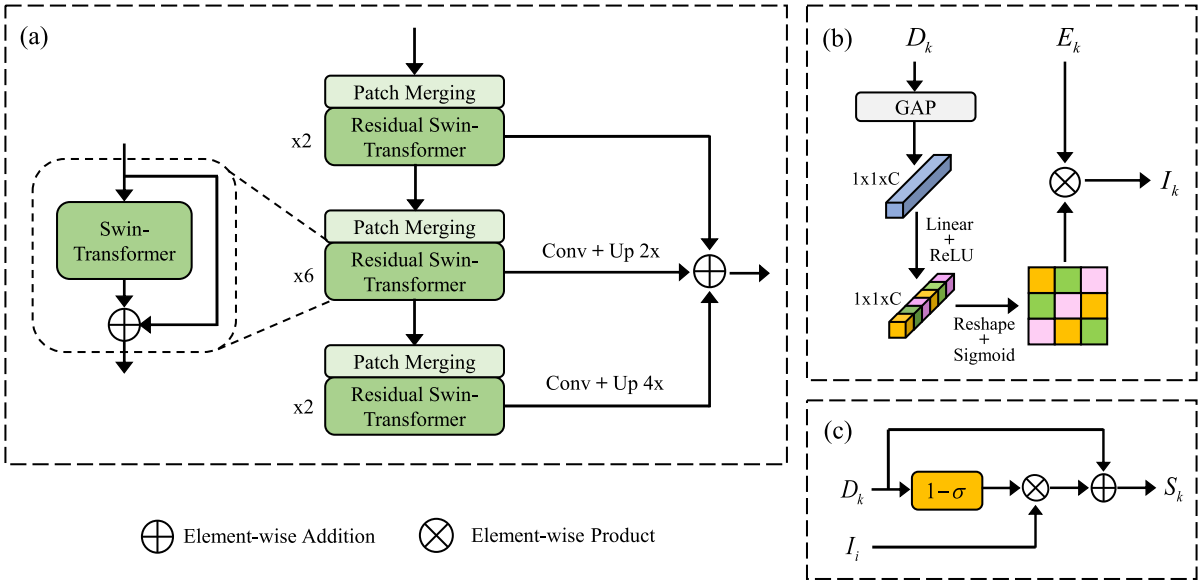

Fig. 4. Structure of (a) RSTB, (b) ICA, and (c) SFF.

instead of the two FC layers. The process can be detailed as follows:

$$G(\mathrm{X}) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathrm{X}^{c}(i,j) \tag{3}$$

$$F_k = \sigma(\delta(L \cdot G(D_k))) \tag{4}$$

$$I_k = F_k \otimes E_k \tag{5}$$

where $\mathrm{G(X)} \in \mathrm{R}^{\mathrm{C}\times 1\times 1}$ is the global average pooling (GAP) layer. $L \in \mathrm{R}^{\mathrm{C}\times\mathrm{C}}$ is denoted as the weights of the linear layer. $\delta(\cdot)$ and $\sigma(\cdot)$ indicate the ReLU operation and sigmoid activation function, respectively. The features D_k, $k \in \{1,2,3\}$ after supervision in the decoder and the input features E_k, $k \in \{1,2,3\}$ in the convolutional layer of the encoder are used as inputs. After the input D_k goes through the GAP layer, we use a linear layer and a sigmoid function to build the channel attention map. Eq. (3) represents the importance of each channel. The channel attention map is matched to the input E_k space by a reshape operation. Ultimately, the output of ICA is generated by multiplying the channels of the weights F_k and the input features E_k. When information flows through ICA, important features are emphasized, while unnecessary features are suppressed.

2.3. Supplementary feature fusion module

The core of the design of the SFF module is to supplement the missing content of the current task by introducing useful features; (see Fig. 4(c)). During our previous research, we found that due to the characteristics of low contrast, noise, and blurred boundaries in ultrasound images, the foreground and background are complexly mixed. Some areas that are in the foreground unfortunately have relatively weak responses. Regions with underestimated responses may be filtered. We supplement weakly responsive regions with information from other tasks and give higher attention to features with weak responses to guide the network to detect underestimated regions. This process is beneficial to segmentation task execution. Otherwise, erroneous recognition results will inevitably be produced. Specifically, we first use the sigmoid function to obtain a weight matrix W_1. Then, we obtain an inverse attention weight matrix W_2 by subtracting the weight matrix W_1 from a weight matrix E, where all elements are 1. After that, ICA output is multiplied by W_2 to obtain the inverse attention feature. We eventually add the reverse attention features to D_k to re-evaluate the lesion area.

$$S_k = D_k + (1 - \sigma(D_k)) \otimes I_k \tag{6}$$

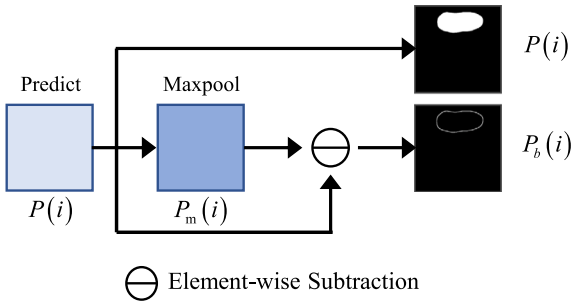

Fig. 5. Structure of the BD.

where $\sigma(\cdot)$ denotes the sigmoid activation function, and $1-\sigma(\cdot)$ is used to generate the reverse mask. The SFF module is implemented based on the gating mechanism without additional parameters. Useful information can be supplemented by focusing attention on the correct location, and useless information can be suppressed. The module can obtain a rich representation via the aggregation of features.

2.4. Boundary detection module

After SFF, the region of interest of the feature becomes larger than before. To ensure the integrity of the feature structure, suppress irrelevant feature regions, and enhance the fuzzy boundary feature extractions of breast lesions, we use a BD module, as shown in Fig. 5. While predicting the shape and structure features of breast lesions, the segmentation results are enhanced by additional boundary prediction. Specifically, we first obtain the breast lesion segmentation map $P(i)$ by using a 1×1 convolutional layer with an output channel of 1. Then, we use a maxpool operation (Feng et al., 2019) to expand the pixel range to obtain the feature map $P_m(i)$ and subtract $P(i)$ from $P_m(i)$ to obtain the boundary map B of the breast lesion. The boundary mask is obtained from the breast lesion mask by the same operation.

2.5. Hybrid loss function

We design a hybrid loss function to efficiently train our proposed model. It includes the loss of the RSTB module, denoted by L_S; the BD loss, denoted by L_B; and the loss of the final segmentation result of the network denoted by L_O. The total loss, denoted by L, is a weighted combination of these three loss functions:

$$L = \lambda_1 L_S + \lambda_2 L_B + \lambda_3 L_O \tag{7}$$

where λ_1, λ_2 and λ_3 denote weight coefficients and are experimentally set to 0.5, 0.5 and 1. We introduce the Dice loss L_{dice} and binary cross-entropy loss L_{bce} to define L_O and L_S. L_* denotes L_O and L_S:

$$L_* = \mu_1 L_{dice} + \mu_2 L_{bce} \tag{8}$$

where μ_1 and μ_2 denote weight coefficients, and their values are set as 1 and 1. L_{dice} and L_{bce} are calculated as:

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

$$L_{bce} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1-y_i)\log(1-p_i)) \tag{10}$$

where N is the number of pixels. p_i and y_i denote the predicted value and the ground truth, respectively. L_B mainly consists of the loss for segmented shape structure prediction L_{shape} and boundary prediction $L_{boundary}$:

$$L_B = \alpha L_{shape} + \beta L_{boundary} \tag{11}$$

where α and β denote weight coefficients, and their values are set as 1 and 10. L_{shape} is defined in Eq. (8). $L_{boundary}$ is the mean square error (MSE) and computed as:

$$L_{boundary} = \sum_{i=1}^{N_p}((P_b)_i - (Y_b)_i)^2 \tag{12}$$

where N_p is the number of pixels. P_b and Y_b denote the predicted value and the ground truth, respectively.

3. Experiments

3.1. Datasets

We used two different datasets to evaluate the performance of our model. Dataset 1, provided by UDIAT Diagnostic Center of the Parc TaulíCorporation, Sabadell (Spain) (Yap et al., 2018), contains 163 ultrasound images (110 benign and 53 malignant) obtained with a Siemens ACUSON Sequoia C512 system 17L5 HD linear array transducer (8.5 MHZ). Dataset 2, provided by Baheya Hospital, Cairo (Egypt) (Al-Dhabyani et al., 2019), contains 780 ultrasound images (437 benign, 210 malignant, and 133 normal) obtained with the LOGIQ E9 ultrasound system and LOGIQ E9 Agile ultrasound system. Since the main purpose of breast lesion segmentation is to evaluate lesions and identify the distribution of lesions, normal cases without masks in dataset 2 were removed.

3.2. Evaluation metrics

Five common metrics, namely, the Dice coefficient, Jaccard index (IoU), precision, recall and F1-score, are used to evaluate the performance of the segmentation models. We use the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) rates to compute these metrics:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{13}$$

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

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$F1 - score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{17}$$

In addition, the Hausdorff distance (HD) 95% and average surface distance (ASD) are used to assess the performance of the segmentation models.

3.3. Implementation details

All experiments are performed on the same data partition based on 5-fold cross-validation. Four folds (80% of images) are used for training, and one-fold (20% of images) is used for testing. During training, 20% of the training data are used for validation. All images are resized to a fixed size of 224×224. We use the Adam optimizer to train the network. We explored different learning rates, batch sizes and epochs with the Adam optimizer. The best results are obtained by setting the learning rate, batch size and maximum epoch number as 0.0001, 4, and 200, respectively. Random rotation and horizontal flip operations are used to perform data enhancement on the training set. All experiments are implemented on the PyTorch v1.10 library and a single NVIDIA GeForce RTX 3080 10-GB GPU.

Table 1
Results of ablation experiments with our proposed method.

	Methods	IoU (%)	Dice (%)	Precision (%)	Recall (%)	F1-score (%)
Dataset1	baseline	75.67 ± 2.18	84.24 ± 2.23	86.27 ± 2.33	85.86 ± 2.02	86.04 ± 1.65
	baseline+ICA	77.14 ± 1.47	85.76 ± 1.41	86.60 ± 1.95	86.88 ± 1.71	87.12 ± 0.96
	baseline+SFF	77.06 ± 1.63	85.58 ± 1.48	85.85 ± 2.08	87.58 ± 2.19	86.68 ± 1.29
	baseline+ICA+SFF	78.25 ± 1.41	86.63 ± 1.36	88.09 ± 2.51	87.51 ± 2.07	87.78 ± 1.80
	baseline+ICA+SFF+BD (our)	**78.61 ± 1.23**	**87.25 ± 1.19**	**88.61 ± 1.34**	**88.10 ± 1.61**	**88.33 ± 0.74**
Dataset2	baseline	72.64 ± 1.67	81.15 ± 1.59	83.93 ± 1.84	82.56 ± 2.38	83.21 ± 1.16
	baseline+ICA	73.59 ± 1.25	82.13 ± 1.19	84.14 ± 2.15	84.12 ± 2.13	84.09 ± 1.07
	baseline+SFF	74.02 ± 1.18	82.46 ± 1.14	84.62 ± 1.92	84.51 ± 1.25	84.74 ± 0.97
	baseline+ICA+SFF	74.55 ± 1.21	83.15 ± 1.23	85.05 ± 0.97	84.98 ± 2.34	85.00 ± 1.04
	baseline+ICA+SFF+BD (our)	**75.11 ± 1.13**	**83.68 ± 1.14**	**85.71 ± 1.51**	**85.87 ± 1.34**	**85.78 ± 1.09**

3.4. Ablation study

In this section, we aim to verify the effectiveness of the ICA, SFF and BD modules in the network. We conducted ablation experiments on dataset 1 and dataset 2. The baseline is constructed by removing the ICA, SFF and BD modules from our network. Table 1 shows the results of comparing our approach with the models containing different components. Our base network (baseline) has good segmentation performance. The IoU and Dice scores of the baseline are 75.67% and 84.24% on dataset 1 and 72.64% and 81.15% on dataset 2, respectively.

First, we investigate the effectiveness of the ICA module. We add the ICA module to the base network and call this model baseline+ICA. Compared to the baseline, it has improved in terms of IoU and Dice by 1.47% and 1.52% on dataset 1, respectively, and by 0.95% and 0.98% on dataset 2. The ICA module uses the features output by the RSTB to guide the channel features of the CNN in the encoding layer to be passed to the decoding layer. This can highlight the objects that need to be detected in the feature channel, suppress irrelevant information in the feature channel, and pass effective information to other tasks.

Second, we evaluate the performance of the SFF module. We add the SFF module to the base network and call this model baseline+SFF. The results show that the IoU and Dice scores are improved by 1.39% and 1.34% on dataset 1, respectively, and by 1.38% and 1.31% on dataset 2, respectively. The SFF module implements feature complementation by introducing the output features of each layer of the encoder and the output features of each layer of the decoder. It is difficult to accurately extract effective information from the simple encoding output because of the complex mixture of foreground and background. Giving high attention to weak features through the SFF module guides the network to re-evaluate the features, which helps to achieve better segmentation performance.

Finally, the proposed model (baseline+ICA+SEF+BD) is formed by fusing the ICA and SFF modules into the base network and adding BD supervision at the decoding layer. Compared with the baseline model, the IoU and Dice scores on dataset 1 are improved by 2.94% and 3.01%, respectively, and on dataset 2 by 2.47% and 2.53%, respectively. The fusion of the ICA, SFF and BD modules allows for more efficient interaction between features at different scales, which leads to a more efficient fusion of feature representations for multiscale branching and helps achieve better segmentation performance.

We also compare the feature fusion methods in the RSTB module. The residual structure compensates for the global modeling ability of the Swin Transformer and improves its segmentation ability, as shown in Table 2. Table 3 shows the comparison of RSTB feature aggregation methods. Compared with the feature connection method, the feature summation method performs better and is more efficient.

Table 2
Comparisons of our method with and without a residual structure in RSTB.

Methods	Dataset1		Dataset2	
	IoU (%)	Dice (%)	IoU (%)	Dice (%)
Our_w/oR	76.90 ± 1.51	85.33 ± 1.47	73.50 ± 1.66	82.12 ± 1.53
Our	**78.61 ± 1.23**	**87.25 ± 1.19**	**75.11 ± 1.13**	**83.68 ± 1.14**

Table 3
Comparisons of feature aggregation methods in RSTB.

Aggr.	Dataset1		Dataset2	
	IoU (%)	Dice (%)	IoU (%)	Dice (%)
Concat	77.03 ± 1.90	85.71 ± 1.84	73.83 ± 1.66	82.41 ± 1.48
Sum	**78.61 ± 1.23**	**87.25 ± 1.19**	**75.11 ± 1.13**	**83.68 ± 1.14**

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

To validate the effectiveness of our proposed method, we conducted comparison experiments with deep learning image segmentation methods and state-of-the-art transformer-based biomedical image segmentation methods. The competing methods include the following: U-Net (Ronneberger et al., 2015), Attention U-Net(AU-Net) (Oktay et al., 2018), U-Net++ (Zhou et al., 2018), FPN (Lin et al., 2017), ViT (Dosovitskiy et al., 2020), TransUNet (Chen, Lu et al., 2021), and Swin-UNet (Cao et al., 2021). All methods were evaluated both quantitatively and qualitatively. Tables 4 and 5 show the quantitative evaluation results of all methods on dataset 1 and dataset 2, respectively. Compared with other segmentation methods, our method has larger IoU, Dice, precision, recall and F1-score values, as well as smaller HD and ASD values. In Table 4, the average IoU and Dice scores of our method are 78.61% and 87.25%, respectively, which are 3.98% and 4.04% higher than those of the second-best Swin-UNet model. The HD and ASD values of our proposed method are 9.42 and 2.94, which are 4.26 and 1.17 smaller than those of Swin-UNet. Similarly, in Table 5, the average IoU and Dice scores of our method are 75.11% and 83.68%, respectively, which are 1.94% and 2.23% higher than those of the second-best Swin-UNet model. The HD and ASD values are 16.97 and 5.6, which are 1.9 and 0.91 smaller than those of Swin-UNet. This shows that our segmentation network can segment breast lesions from ultrasound images more accurately than all competitors, and identifies edge information more accurately. Additionally, these segmentation performance metrics on dataset 1 are better than the segmentation performance metrics on dataset 2. The reason for this is that the BUS image quality of dataset 1 is better than that of dataset 2, which results in better segmentation performance.

We also qualitatively compare the segmentation results of different methods. Figs. 6 and 7 show a visual comparison of the segmentation results of different methods on dataset 1 and dataset 2. The transformer-based methods outperform the traditional CNN methods. From the comparison of ViT and TransUNet, the effect of TransUNet is better than that of ViT. We argue that although transformers pay more attention to global information, using transformers too early is not conducive to shallow feature extraction. Although Swin-UNet has a good segmentation effect, using only a transformer as the backbone for feature extraction is not conducive to the success of feature extraction. Obviously, the segmentation result of Swin-UNet is quite different, and

Table 4
Comparing our method (CSwin-PNet) with the state-of-the-art methods on dataset 1.

Methods	IoU (%)	Dice (%)	Precision (%)	Recall (%)	F1-score (%)	HD	ASD
U-Net	68.25 ± 4.36	77.83 ± 3.78	81.83 ± 3.76	79.16 ± 2.92	80.46 ± 3.20	24.35 ± 3.93	7.51 ± 1.62
AU-Net	68.78 ± 3.33	78.28 ± 3.45	80.21 ± 3.51	83.32 ± 3.15	81.71 ± 3.05	23.53 ± 3.32	6.73 ± 1.22
U-Net++	69.56 ± 3.28	79.23 ± 3.06	83.06 ± 3.77	81.36 ± 3.55	82.11 ± 2.14	22.87 ± 2.39	6.26 ± 1.11
FPN	74.12 ± 2.29	83.22 ± 2.07	85.61 ± 2.72	84.23 ± 2.24	84.90 ± 2.26	13.40 ± 3.39	4.32 ± 1.16
ViT	71.45 ± 3.24	80.52 ± 3.52	85.05 ± 4.74	80.21 ± 4.93	82.44 ± 3.46	15.05 ± 3.02	5.32 ± 1.47
TransUNet	73.13 ± 3.31	82.00 ± 2.31	86.52 ± 4.39	81.14 ± 3.45	83.65 ± 2.51	14.36 ± 4.48	4.56 ± 1.46
Swin-UNet	74.63 ± 1.67	83.21 ± 1.67	87.82 ± 2.60	82.42 ± 1.37	85.02 ± 1.69	13.68 ± 1.81	4.11 ± 0.76
CSwin-PNet (our)	**78.61 ± 1.23**	**87.25 ± 1.19**	**88.61 ± 1.34**	**88.10 ± 1.61**	**88.33 ± 0.74**	**9.42 ± 1.33**	**2.94 ± 0.41**

Table 5
Comparing our method (CSwin-PNet) with the state-of-the-art methods on dataset 2.

Methods	IoU (%)	Dice (%)	Precision (%)	Recall (%)	F1-score (%)	HD	ASD
U-Net	65.36 ± 1.81	74.35 ± 1.72	78.3 ± 1.98	77.86 ± 2.24	78.05 ± 1.36	30.85 ± 4.24	9.81 ± 1.05
AU-Net	66.35 ± 1.50	75.22 ± 1.55	77.94 ± 1.78	79.43 ± 1.93	78.67 ± 1.55	27.76 ± 4.19	9.27 ± 0.82
U-Net++	67.20 ± 1.67	75.80 ± 1.26	79.87 ± 1.71	77.90 ± 2.51	78.87 ± 1.99	26.96 ± 2.55	9.22 ± 1.17
FPN	72.32 ± 1.36	80.78 ± 1.21	83.50 ± 1.32	82.54 ± 1.15	83.01 ± 0.66	20.51 ± 1.69	6.93 ± 0.73
ViT	70.43 ± 0.81	79.47 ± 0.73	81.48 ± 1.91	82.16 ± 2.11	81.78 ± 0.62	20.47 ± 1.22	6.89 ± 0.71
TransUNet	72.40 ± 1.86	80.97 ± 1.67	82.28 ± 2.28	84.26 ± 2.11	83.23 ± 1.84	19.73 ± 2.05	6.87 ± 0.79
Swin-UNet	73.17 ± 0.71	81.45 ± 0.62	83.49 ± 1.36	83.36 ± 1.60	83.40 ± 0.31	18.87 ± 1.57	6.51 ± 0.67
CSwin-PNet (our)	**75.11 ± 1.13**	**83.68 ± 1.14**	**85.71 ± 1.51**	**85.87 ± 1.34**	**85.78 ± 1.09**	**16.97 ± 2.00**	**5.60 ± 0.70**

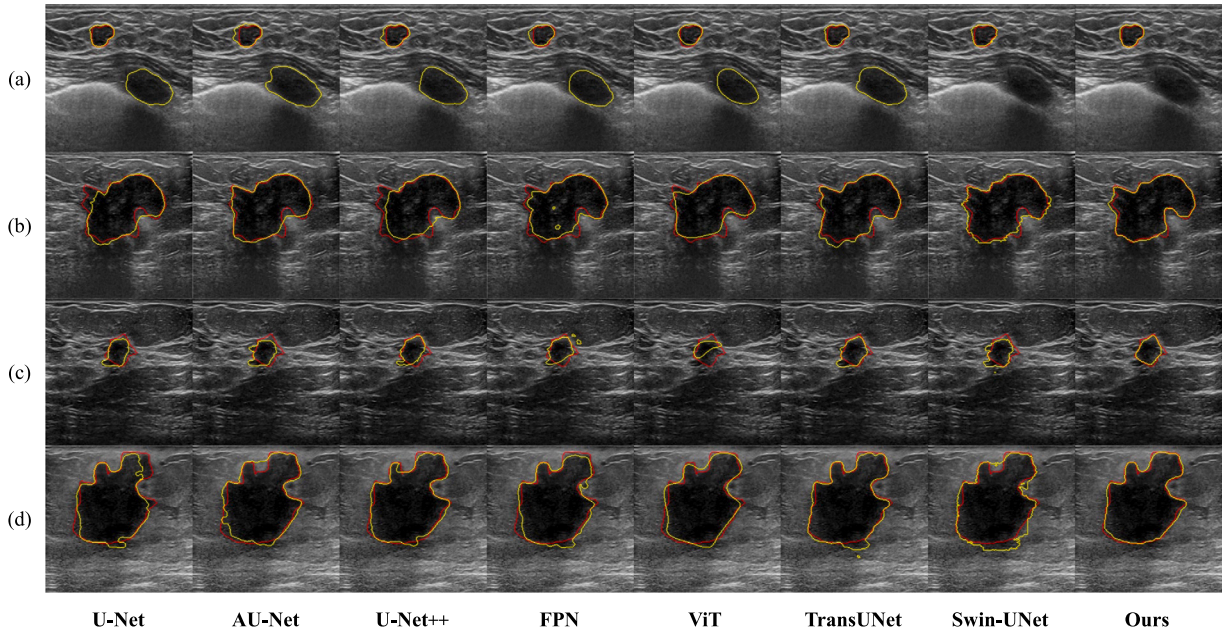

Fig. 6. Visual comparison of segmentation results for different methods on dataset 1. (a)–(d) represent four different breast ultrasound images. Each column represents the segmentation results for each method. (Col. 1) U-Net, (Col. 2) AU-Net, (Col. 3) U-Net++, (Col. 4) PFN, (Col. 5) ViT, (Col. 6) TransUNet, (Col. 7) Swin-UNet, and (Col. 8) our method. The red and yellow contours represent the ground truth and the segmentation results, respectively.

the boundary is not smooth on dataset 2. Overall, our method has the best segmentation results. Our proposed method can accurately identify lesion areas when processing BUS images with similar contrast and nonbreast lesions. When dealing with relatively blurry images, our proposed method can better identify lesions and handle lesion boundary information.

4. Discussion

In clinical diagnosis, the general purpose of breast lesion segmentation is to evaluate lesions, track changes in lesions, identify the distribution and severity of lesions, and assist doctors in diagnosis. The performance of the segmentation network has a significant impact on lesion segmentation. We combine the respective advantages of a CNN and a transformer to design a segmentation network with spatial and channel guidance. The semantic features of breast lesions in BUS images are first extracted using a CNN. Then, we take advantage of the Swin Transformer's ability to capture global contextual information to establish long-range dependencies on the semantic features of breast lesions and improve spatial feature learning. Channel attention is adopted to guide the improvement in channel features, and an edge detection module is used to learn edge feature information to further improve the segmentation performance of the network. Compared with state-of-the-art methods, our method achieves better performance in breast lesion segmentation.

Although the segmentation performance of the proposed method is strong, the segmentation accuracy of some BUS images is limited. Fig. 8 shows some failed cases. These cases neglect some regions of breast lesions on the ultrasound images. This indicates that it is difficult to accurately segment the lesion tissue when the lesion boundary is not clear and the lesion region is heterogeneous in intensity. As with other segmentation methods, when the target breast lesion area has

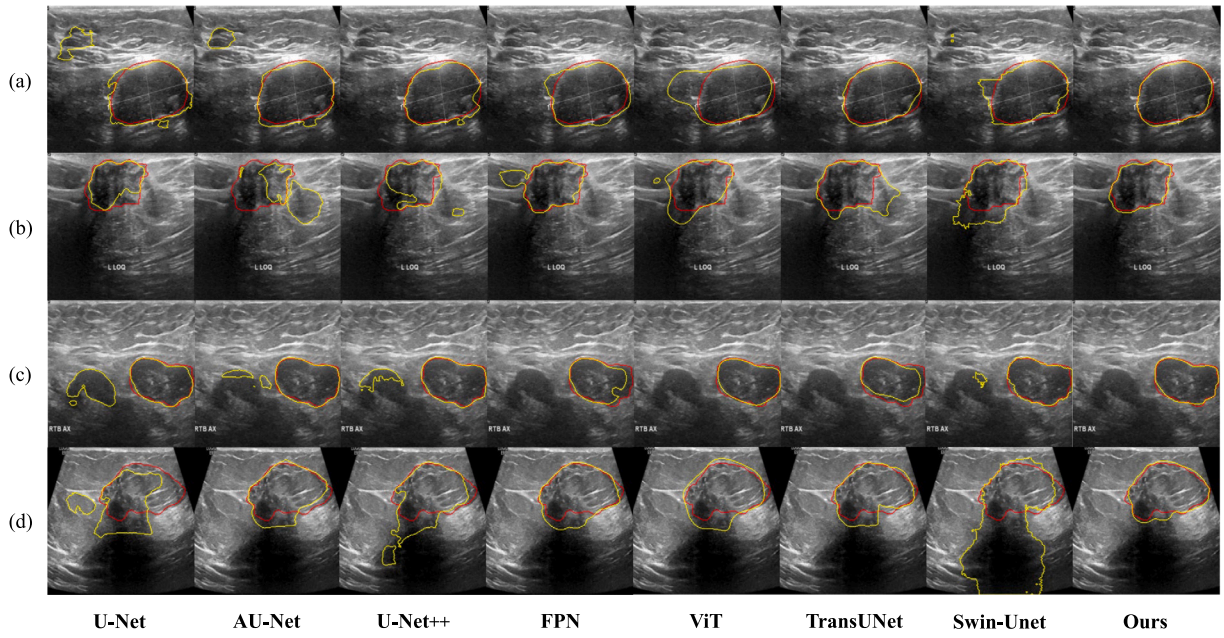

Fig. 7. Visual comparison of segmentation results for different methods on dataset 2. (a)–(d) represent four different breast ultrasound images. Each column represents the segmentation results for each method. (Col. 1) U-Net, (Col. 2) AU-Net, (Col. 3) U-Net++, (Col. 4) PFN, (Col. 5) ViT, (Col. 6) TransUNet, (Col. 7) Swin-UNet, and (Col. 8) our method. The red and yellow contours represent the ground truth and the segmentation results, respectively.

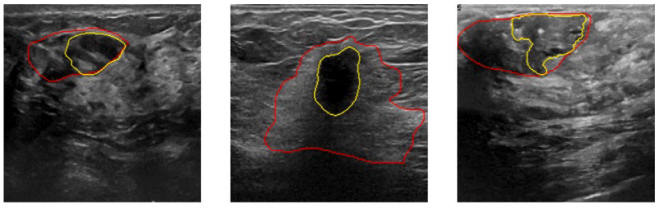

Fig. 8. Some inaccurate segmentation results of the proposed method. (a)–(c) represent three different breast ultrasound images from the two different datasets. The red and yellow contours represent the ground truth and the segmentation results, respectively.

insignificant intensity values or unclear boundaries, the network often cannot fully detect the breast lesion area. In the future, we will explore more effective feature extraction modules. These include embedding CNNs into transformers or building more powerful transformer variants to enhance the perception and extraction of breast lesion regions and boundary information.

In general, transformer-based networks outperform U-Net and variant networks, and FPN has competitive performance. The U-Net skip link operation can supplement the features from the encoder stage into the decoder to compensate for the image detail information during decoding. From the segmentation results of U-Net and AU-Net, it can be found that the attention mechanism can encourage the network to pay attention to more information and improve the segmentation performance of the network. A transformer, as a multiheaded self-attention mechanism, can capture global contextual information and establish long-range dependencies on a target to extract more powerful features. TransUNet and Swin-UNet achieve competitive performance through the transformer mechanism and multiscale fusion strategy. Our proposed method utilizes multiscale features and multitask information to achieve feature complementarity to obtain strong semantic features. Our proposed method obtains strong semantic features through a combination of a CNN and a transformer using multiscale features and feature complementation of multitask information. We adopt an additional edge-aware unit for the network to facilitate breast lesion boundary recognition.

5. Conclusions

In this paper, we propose a combined CNN-Swin Transformer model for breast lesion segmentation in ultrasound images. We construct an RSTB based on the powerful self-attention mechanism of the Swin Transformer. The backbone of our network is established by using the CNN for feature localization and the RSTB for global feature extraction. Through the feature pyramid structure design, the ICA module and SFF module are applied to effectively fuse the multiscale features of the encoder and complement the decoder features. The BD module predicts breast lesion boundaries to help improve segmentation performance. We evaluate our network on two different ultrasound image datasets by comparing it against state-of-the-art methods, and the experimental results show that our network can segment breast lesions more accurately than all competitors.

CRediT authorship contribution statement

Haonan Yang: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Data curation, Writing – original draft. **Dapeng Yang:** Validation, Resources, Writing – review & editing, Supervision.

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.

Formatting of funding sources

Funding: This work was supported in part by the National Natural Science Foundation of China Grant #52075114, Interdisciplinary Research Foundation of HIT (IR2021218), and Postdoctoral Scientific Research Development Fun (LBH-W18058) to D. Yang.

References

Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2019). Dataset of breast ultrasound images. *Data in Brief*, *28*, Article 104863, URL: https://doi.org/10.1016/j.dib.2019.104863.

Bai, J., Posner, R., Wang, T., Yang, C., & Nabavi, S. (2021). Applying deep learning in digital breast tomosynthesis for automatic breast cancer detection: A review. *Medical Image Analysis*, *71*, Article 102049, URL: https://doi.org/10.1016/j.media.2021.102049.

Bleicher, R. J., Ruth, K., Sigurdson, E. R., Beck, J. R., Ross, E., Wong, Y.-N., Patel, S. A., Boraas, M., Chang, E. I., Topham, N. S., & Egleston, B. L. (2015). Time to surgery and breast cancer survival in the United States. *JAMA Oncology*, *2*(3), 1–10, URL: https://doi.org/10.1001/jamaoncol.2015.4508.

Brandt, I. v. d., Fok, F., Mulders, B., Vanschoren, J., & Cheplygina, V. (2021). Cats, not CAT scans: a study of dataset similarity in transfer learning for 2D medical image classification. arXiv e-prints. arXiv:2107.05940.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2021). Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv e-prints. arXiv:2105.05537.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. arXiv e-prints. arXiv:2005.12872.

Chen, B., Liu, Y., Zhang, Z., Lu, G., & Zhang, D. (2021). TransAttUnet: Multi-level attention-guided U-Net with transformer for medical image segmentation. arXiv e-prints. arXiv:2107.05274.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arXiv e-prints. arXiv:2102.04306.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv e-prints. arXiv:1802.02611.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv e-prints. arXiv:2010.11929.

Feng, M., Lu, H., & Ding, E. (2019). Attentive feedback network for boundary-aware salient object detection. In *2019 IEEE/CVF conference on computer vision and pattern recognition* CVPR, (pp. 1623–1632). URL: https://doi.org/10.1109/CVPR.2019.00172.

Gao, Y., Zhou, M., & Metaxas, D. N. (2021). UTNet: a hybrid transformer architecture for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention – MICCAI 2021* (pp. 61–71). Cham: Springer, URL: https://doi.org/10.1007/978-3-030-87199-4_6.

Ghosh, D., Kumar, A., Ghosal, P., Chowdhury, T., Sadhu, A., & Nandi, D. (2020). Breast lesion segmentation in ultrasound images using deep convolutional neural networks. In *2020 IEEE Calcutta conference* CALCON, (pp. 318–322). URL: https://doi.org/10.1109/CALCON49167.2020.9106568.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., & Xu, D. (2022). Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. arXiv e-prints. arXiv:2201.01266.

Horsch, K., Giger, M. L., Venta, L. A., & Vyborny, C. J. (2001). Automatic segmentation of breast lesions on ultrasound. *Medical Physics*, *28*(8), 1652–1659, URL: https://doi.org/10.1118/1.1386426.

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

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2019). Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(8), 2011–2023, URL: https://doi.org/10.1109/tpami.2019.2913372.

Jalalian, A., Mashohor, S., Mahmud, R., Karasfi, B., Saripan, M. I. B., & Ramli, A. R. B. (2017). Foundation and methodologies in computer-aided diagnosis systems for breast cancer detection. *EXCLI Journal*, *16*, 113–137, URL: https://doi.org/10.17179/excli2016-701.

Li, J., Cheng, L., Xia, T., Ni, H., & Li, J. (2021). Multi-scale fusion U-net for the segmentation of breast lesions. *IEEE Access*, *9*, 137125–137139, URL: https://doi.org/10.1109/ACCESS.2021.3117578.

Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., & Zhang, D. (2022). DS-TransUNet: Dual swin transformer U-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–15, URL: https://doi.org/10.1109/TIM.2022.3178991.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *2017 IEEE conference on computer vision and pattern recognition* CVPR, (pp. 936–944). URL: https://doi.org/10.1109/CVPR.2017.106.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision* ICCV, (pp. 9992–10002). URL: https://doi.org/10.1109/ICCV48922.2021.00986.

Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., & He, Z. (2021). A survey of visual transformers. arXiv e-prints. arXiv:2111.06091.

Matsoukas, C., Haslum, J. F., Sorkhei, M., Söderberg, M., & Smith, K. (2022). What makes transfer learning work for medical images: Feature reuse & other factors. arXiv e-prints. arXiv:2203.01825.

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent models of visual attention. arXiv e-prints. arXiv:1406.6247.

Moon, W. K., Lee, Y.-W., Ke, H.-H., Lee, S. H., Huang, C.-S., & Chang, R.-F. (2020). Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Computer Methods and Programs in Biomedicine*, *190*, Article 105361, URL: https://doi.org/10.1016/j.cmpb.2020.105361.

Ning, Z., Zhong, S., Feng, Q., Chen, W., Zhang, Y., & Ning, Z. (2022). SMU-net: Saliency-guided morphology-aware U-net for breast lesion segmentation in ultrasound image. *IEEE Transactions on Medical Imaging*, *41*(2), 476–490, URL: https://doi.org/10.1109/TMI.2021.3116087.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). Attention U-net: Learning where to look for the pancreas. arXiv e-prints. arXiv:1804.03999.

Rai, M., Datta, P., & Ansari, R. (2019). An introduction to deep learning techniques in ultrasound image modality. In *2019 2nd international conference on power energy, environment and intelligent control* PEEIC, (pp. 293–298). URL: https://doi.org/10.1109/PEEIC47157.2019.8976806.

Ramadan, H., Lachqar, C., & Tairi, H. (2020). Saliency-guided automatic detection and segmentation of tumor in breast ultrasound images. *Biomedical Signal Processing and Control*, *60*, Article 101945, URL: https://doi.org/10.1016/j.bspc.2020.101945.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention – MICCAI 2015* (pp. 234–241). Cham: Springer, URL: https://doi.org/10.1007/978-3-319-24574-4_28.

Samulski, M., Hupse, R., Boetes, C., Mus, R. D. M., Heeten, G. J. d., & Karssemeijer, N. (2010). Using computer-aided detection in mammography as a decision support. *European Radiology*, *20*(10), 2323–2330, URL: https://doi.org/10.1007/s00330-010-1821-8.

Shelhamer, E., Long, J., & Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 640–651, URL: https://doi.org/10.1109/TPAMI.2016.2572683.

Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2021). Cancer statistics, 2021. *CA: A Cancer Journal for Clinicians*, *71*(1), 7–33, URL: https://doi.org/10.3322/caac.21654.

Thiyagarajan, A., & Murukesh, C. (2020). A survey on deep learning architectures and frameworks for cancer detection in medical images analysis. *International Journal of Innovative Technology and Exploring Engineering*, *9*(11), 28–34, URL: https://doi.org/10.35940/ijitee.K7654.0991120.

Tong, Y., Liu, Y., Zhao, M., Meng, L., & Zhang, J. (2021). Improved U-net MALF model for lesion segmentation in breast ultrasound images. *Biomedical Signal Processing and Control*, *68*, Article 102721, URL: https://doi.org/10.1016/j.bspc.2021.102721.

Vakanski, A., Xian, M., & Freer, P. E. (2020). Attention-enriched deep learning model for breast tumor segmentation in ultrasound images. *Ultrasound in Medicine & Biology*, *46*(10), 2819–2833, URL: https://doi.org/10.1016/j.ultrasmedbio.2020.06.015.

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. arXiv e-prints. arXiv:2102.10662.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. arXiv e-prints. arXiv:1706.03762.

Wang, H., Cao, P., Wang, J., & Zaiane, O. R. (2022). UCTransNet: Rethinking the skip connections in U-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 36* (pp. 2441–2449). URL: https://doi.org/10.1609/aaai.v36i3.20144.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF conference on computer vision and pattern recognition* CVPR, (pp. 11531–11539). URL: https://doi.org/10.1109/CVPR42600.2020.01155.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *2021 IEEE/CVF international conference on computer vision* ICCV, (pp. 548–558). URL: https://doi.org/10.1109/ICCV48922.2021.00061.

Xi, X., Shi, H., Han, L., Wang, T., Ding, H. Y., Zhang, G., Tang, Y., & Yin, Y. (2017). Breast tumor segmentation with prior knowledge learning. *Neurocomputing*, *237*, 145–157, URL: https://doi.org/10.1016/j.neucom.2016.09.067.

Xian, M., Zhang, Y., Cheng, H., Xu, F., Zhang, B., & Ding, J. (2018). Automatic breast ultrasound image segmentation: A survey. *Pattern Recognition*, *79*, 340–355, URL: https://doi.org/10.1016/j.patcog.2018.02.012.

Xue, C., Zhu, L., Fu, H., Hu, X., Li, X., Zhang, H., & Heng, P.-A. (2021). Global guidance network for breast lesion segmentation in ultrasound images. *Medical Image Analysis*, *70*, Article 101989, URL: https://doi.org/10.1016/j.media.2021.101989.

Yanase, J., & Triantaphyllou, E. (2019). A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications*, *138*, Article 112821, URL: https://doi.org/10.1016/j.eswa.2019.112821.

Yap, M. H., Goyal, M., Osman, F. M., Martí, R., Denton, E., Juette, A., & Zwiggelaar, R. (2019). Breast ultrasound lesions recognition: end-to-end deep learning approaches. *Journal of Medical Imaging*, *6*(1), Article 011007, URL: https://doi.org/10.1117/1.JMI.6.1.011007.

Yap, M. H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A. K., & Marti, R. (2018). Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*, *22*(4), 1218–1226, URL: https://doi.org/10.1109/JBHI.2017.2731873.

Yassin, N. I., Omran, S., Houby, E. M. E., & Allam, H. (2018). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer Methods and Programs in Biomedicine*, *156*, 25–45, URL: https://doi.org/10.1016/j.cmpb.2017.12.012.

Zhang, Y., Liu, H., & Hu, Q. (2021). Transfuse: Fusing transformers and cnns for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention – MICCAI 2021* (pp. 14–24). Cham: Springer, URL: https://doi.org/10.1007/978-3-030-87193-2_2.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 3–11). Cham: Springer, URL: https://doi.org/10.1007/978-3-030-00889-5_1.

Zhu, Y.-C., AlZoubi, A., Jassim, S., Jiang, Q., Zhang, Y., Wang, Y.-B., Ye, X.-D., & DU, H. (2021). A generic deep learning framework to classify thyroid and breast lesions in ultrasound images. *Ultrasonics*, *110*, Article 106300, URL: https://doi.org/10.1016/j.ultras.2020.106300.

Zhuang, Z., Li, N., Raj, A. N. J., Mahesh, V. G. V., & Qiu, S. (2019). An RDAU-NET model for lesion segmentation in breast ultrasound images. *PLoS ONE*, *14*(8), Article e0221535, URL: https://doi.org/10.1371/journal.pone.0221535.