Original Article

# Dilated transformer: residual axial attention for breast ultrasound image segmentation

**Xiaoyan Shen[1], Liangyu Wang[1], Yu Zhao[1], Ruibo Liu[1], Wei Qian[1], He Ma[1,2]**

[1]The College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China; [2]Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Northeastern University, Shenyang, China

*Contributions:* (I) Conception and design: X Shen, H Ma; (II) Administrative support: H Ma, W Qian; (III) Provision of study materials or patients: H Ma, W Qian; (IV) Collection and assembly of data: X Shen, L Wang, Y Zhao; (V) Data analysis and interpretation: X Shen, L Wang, R Liu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*Correspondence to:* He Ma. The College of Medicine and Biological Information Engineering, Northeastern University, 195 Chuangxin Road, Hunnan District, Shenyang 110819, China. Email: mahe@bmie.neu.edu.cn.

**Background:** The segmentation of breast ultrasound (US) images has been a challenging task, mainly due to limited data and the inherent image characteristics involved, such as low contrast and speckle noise. Although convolutional neural network-based (CNN-based) methods have made significant progress over the past decade, they lack the ability to model long-range interactions. Recently, the transformer method has been successfully applied to the tasks of computer vision. It has a strong ability to capture distant interactions. However, most transformer-based methods with excellent performance rely on pre-training on large datasets, making it infeasible to directly apply them to medical images analysis, especially that of breast US images with limited high-quality labels. Therefore, it is of great significance to find a robust and efficient transformer-based method for use on small breast US image datasets.

**Methods:** We developed a dilated transformer (DT) method which mainly uses the proposed residual axial attention layers to build encoder blocks and the introduced dilation module (DM) to further increase the receptive field. We evaluated the proposed method on 2 breast US image datasets using the 5-fold cross-validation method. Dataset A was a public dataset with 562 images, while dataset B was a private dataset with 878 images. Ground truth (GT) was delineated by 2 radiologists with more than 5 years of experience. The evaluation was followed by related ablation experiments.

**Results:** The DT was found to be comparable with the state-of-the-art (SOTA) CNN-based method and outperformed the related transformer-based method, medical transformer (MT), on both datasets. Especially on dataset B, the DT outperformed the MT on metrics of Jaccard index (JI) and Dice similarity coefficient (DSC) by 2.67% and 4.68%, respectively. Meanwhile, when compared with Unet, the DT improved JI and DSC by 4.89% and 4.66%, respectively. Moreover, the results of the ablation experiments showed that each add-on part of the DT is important and contributes to the segmentation accuracy.

**Conclusions:** The proposed transformer-based method could achieve advanced segmentation performance on different small breast US image datasets without pretraining.

**Keywords:** Breast ultrasound (US); tumor segmentation; transformer; residual; axial attention

Submitted Jan 14, 2022. Accepted for publication Jun 15, 2022.
doi: 10.21037/qims-22-33
**View this article at:** https://dx.doi.org/10.21037/qims-22-33

## Introduction

Image segmentation is one of the essential tasks of computer-aided diagnosis (CAD) systems which are developed to help doctors make reliable diagnostic decisions swiftly, especially during the early screening and diagnosis of breast cancer using breast ultrasound (US) (1-3). It requires finding actual tumor boundaries contained in the US images which are generated in real-time during the process of inspection, since only accurate tumor segmentation can provide reliable and complete auxiliary information for subsequent screening and diagnosis (4). However, breast US images are associated with inherent issues, such as speckle noise and low contrast. These issues are prone to yielding false positives during the process of image segmentation (5-7). This presents a challenge to the popularization of CAD systems. Therefore, innovating a robust image segmentation method is crucial to reducing false positives and achieving efficient segmentation of breast US images.

Use of a convolutional neural networks (CNN) has become the mainstream method due to its ability to automatically extract hierarchical feature representation of images, and it has developed rapidly in the field of computer vision (8). Since 2013, the automated segmentation methods applied to CAD systems have mainly been deep learning methods based on CNN. In the field of medical image segmentation, Unet (9) was the first and is the most classical CNN-based segmentation method. Many improved versions of Unet have subsequently been proposed and reported to perform excellently on medical image segmentation. For example, Unet++ (10) concatenates the 4 layers of Unet together and aggregates the feature maps with different scales into the decoder, which improves the segmentation accuracy, and ResUnet (11) replaces each submodule of Unet with the form of residual connection. Some researchers have focused on improving the main network structure of Unet and others have added attention modules or information extraction modules which can help control the importance of the input feature by changing the weights of the input variable. For example, Attention Unet (12) introduced a local spatial attention module by adding an attention module to each skip connection, and SE-Net (13) introduced a channel attention mechanism to increase the network awareness of different channels of an image. Moreover, considering that the spatial attention module only focuses on the target region and the channel attention mechanism only focuses on the local information of each channel, researchers have proposed a mechanism based on the mixed domain, such as convolutional block attention module (CBAM) (14) and residual attention learning (15). However, although CNN-based deep learning methods have made significant progress in the field of medical image segmentation (16,17), a bottleneck remains.

The CNN-based methods perform weakly on modeling long-range dependencies of an image. They extract feature maps by using convolution kernels of different sizes to perform convolution operations on a target image; one kernel of each convolution layer only captures the corresponding local correlations, but global interactions information is of great importance for medical images (18). Especially for breast US image segmentation—due to the existence of speckle noise and low contrast—the lack of long-range dependency learning can easily lead to false positives in the segmentation results. Many researchers have tried to improve the receptive field of CNN by proposing methods such as the pyramid scene parsing network (19) and atrous convolution (20), but capturing long-range interactions is still a challenge due to CNN's poor scaling properties with respect to large receptive fields (21). Since the use of self-attention has succeeded in addressing the problem of long-range dependencies in the field of natural language processing (NLP) (22), CNN-based methods have begun to evolve tremendously. Global attention layers have been added to the existing convolution networks, such as SE-Net and spatially-aware attention mechanism (23). Some researchers (24,25) have used self-attention layers to replace the convolution layers of ResUnet. Since the proposal of the pretrained Vision Transformer (ViT) (26) model, which regards the inputting images patches with their position embedding information as "word vector" of NLP models, the popularity of transformer-based segmentation methods has exploded (27-29). Of particular note is the swim transformer, which has successfully applied a transformer to the image semantic segmentation task (30).

However, most transformer-based methods require pretraining on large datasets to achieve satisfactory performance, yet large datasets and labels with high quality images are very expensive to obtain in the field of medical image segmentation. Some researchers have made related attempts to improve the adaptability of transformer-based methods to small datasets (16,18,27). The most typical work is medical transformer (MT) which was proposed by Valanarasu *et al.* (18). They proposed a gated axial attention layer which is used for building multihead attention modules and used a transformer-based encoder. Meanwhile,

 

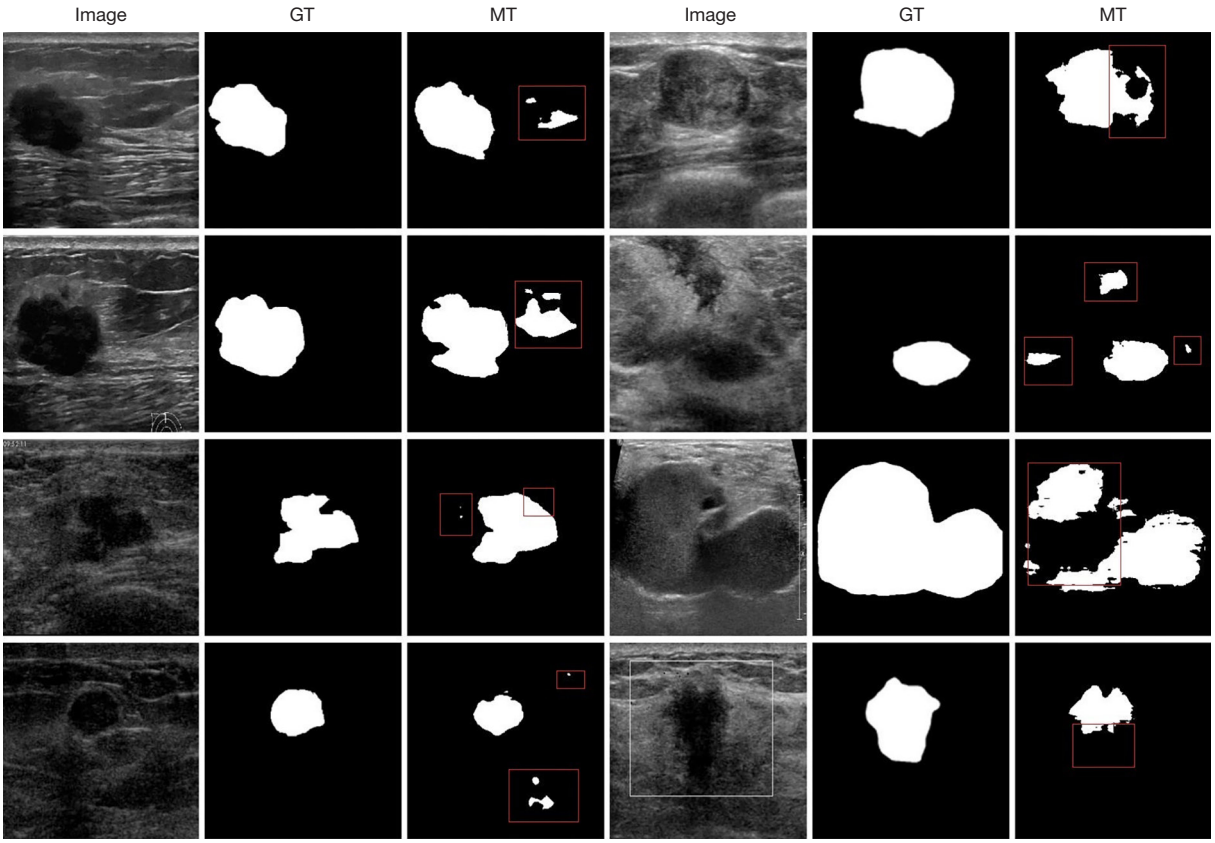


**Figure 1** Segmentation examples of breast US images obtained by MT (images in the third column and the sixth column). GT, ground truth; MT, medical transformer; US, ultrasound.

they proposed a LoGo training strategy and achieved good segmentation performance on 3 medical image datasets without pretraining. However, MT has been found not to perform particularly well on breast US images with low contrast. As shown in *Figure 1*, false positives as well as incomplete segmentation results (highlighted with a red rectangle) are generated by MT, due to its lack of special attention to the region of interest (ROI).

Therefore, to achieve robust and efficient segmentation performance on small breast US image datasets, we propose the dilated transformer (DT), which uses the proposed residual axial attention layer as the building layer of a multihead attention block. It has a strong ability to model long-range interactions of images and achieve efficient segmentation of breast US images. The main contributions are as follows: (I) a residual axial attention mechanism is proposed to help the model capture more detailed local information, (II) a dilated-convolution module (DM) is introduced to the end of the encoder pipeline to extract more global interactions information from a larger receptive field, and (III) a transformer-based architecture is proposed to achieve robust and efficient segmentation of breast US images through use of the proposed residual axial-attention layer as the main building block of the encoder and the introduced DM as the bridge between the end of the encoder and the head of the decoder. We present the following article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-22-33/rc).

## Methods

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013), and was approved by the Biological and Medical Ethics Committee of Northeastern University. Individual consent for this retrospective analysis was waived. As shown in *Figure 2*, the DT has an encoder-decoder architecture. Its input is a feature map obtained by a convolution block with 3 convolution layers followed by batch normalization (BN)

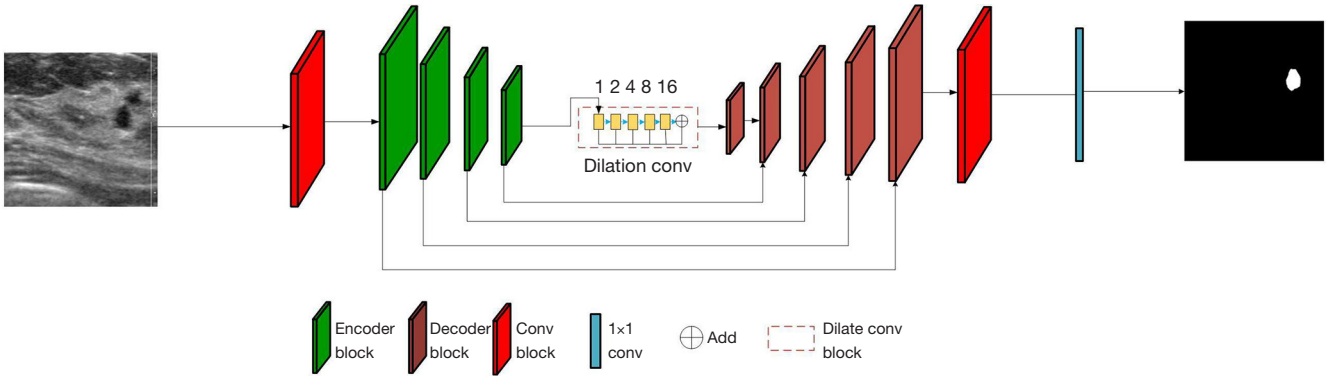


**Figure 2** Architecture of the proposed DT model. DT, dilated transformer.

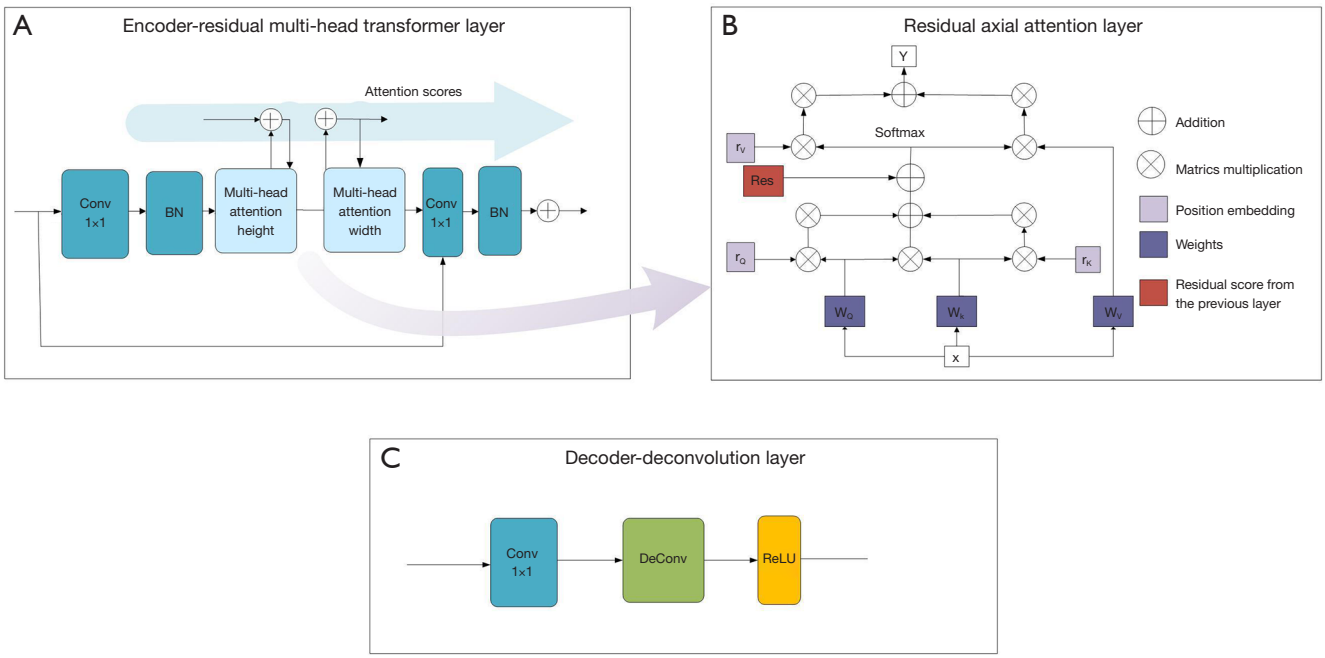


**Figure 3** Illustration of the structure of encoder and decoder blocks used in the DT model. (A) Encoder block—the residual axial transformer layer. (B) Residual axial attention layer, which is the primary building of both height and width multi-head attention blocks found in the residual axial transformer layer. (C) Decoder block—deconvolution layer. BN, batch normalization; ReLU, rectified linear unit; DT, dilated transformer.

and rectified linear unit (ReLU), respectively. In the encoder pipeline, 4 residual transformer-based encoder blocks of different sizes are designed. As shown in *Figure 3A*, in each encoder block, a 1×1 convolution layer is first used to change the size of input feature maps in the bottleneck of each encoder block. Then, after BN, these feature maps are fed into 2 residual multihead attention layers. They are the height-axial-based multihead attention layer, which operates along the height and width-axial-based multihead attention layer, which works along the width. Each is composed of 8 residual axial attention heads, as shown in *Figure 3B*. At the end of the encoder pipeline, another 1×1 convolution layer is used to produce attention maps that are passed to DM to capture information from a larger receptive field after normalization. Meanwhile, softmax attention scores are regarded as residual information to be passed over

 

all attention layers. As indicated by the arrow in *Figure 3A*, softmax attention scores from the previous layer are first given as an additional input to the height-axial-based multihead attention layer. Then, the new softmax attention scores are passed to the width-axial-based multihead attention layer. In the decoder pipeline, which is shown as *Figure 3C*, 5 decoder blocks are used for upsampling. Each has a convolution layer, an upsampling layer, and ReLU activation. After another convolution block and a stand-alone 1×1 convolution layer, final segmentation results can be obtained

### *The proposed residual axial attention*

#### Standard self-attention

Originally, self-attention was proposed to model the relationships of sequences in the field of NLP (31,32). They have different forms, such as additive attention (33) and dot-product attention. Due to its characteristic of being implemented in a space-efficient manner by use of highly optimized matrix-operation code and the ability to avoid degrade explosion by multiplying $1/\sqrt{d_k}$, the scaled dot-product attention introduced by Vaswani *et al.* (22) became the most popular. Standard attention is defined in Eq. [1]: if we assume a feature map with height of $H$, the width of $W$ and channels of $C_{in}$ is denoted as $x \in \mathbb{R}^{C_{in}\times H\times W} \in \mathbb{R}^{C_{in}\times H\times W}$:

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad [1]$$

where $Q$, $K$, and $V$ are 3 matrices obtained by packing the queries $q$, keys of dimension of $d_k$, and values of dimension of $d_v$, respectively.

#### Axial attention

First, researchers customarily applied a self-attention mechanism to augment the outputs of CNN-based networks as an additional module (34,35). Later, spatial convolution layers were entirely replaced with stacked attention layers (19,36), obtaining promising results at a great computational cost. To improve the computation efficiency, Wang *et al.* (37) proposed stand-alone axial attention by factorizing the common self-attention into two 1D attention mechanisms which operate along height-axial and width-axial, respectively. In addition, the position term plays a great role in capturing shapes or boundaries of the ROI from a global receptive. Especially for medical images, it is crucial to insert position terms when using self-attention to deal with image segmentation tasks (37). Taking width-axial attention for an example, we assume that pixels with location $(I, j)$ in feature map have the vectors of $q_{ij}$, $k_{ij}$, and $v_{ij}$; self-attention can also be given as follows:

$$y_{ij} = \sum_{h=1}^{H}\sum_{w=1}^{W}\text{softmax}\left(q_{ij}^T k_{hw}\right)v_{hw} \quad [2]$$

where $y_{ij} \in \mathbb{R}^{C_{out}\times H\times W} y_{ij} \in \mathbb{R}^{C_{out}\times H\times W}$ denotes the output of a self-attention layer. Therefore, the width-axial attention with position embedding can be defined as follows:

$$y_{ij} = \sum_{w=1}^{W}\text{softmax}\left(q_{ij}^T k_{iw} + q_{ij}^T r_{iw}^q + k_{iw}^T r_{iw}^k\right)\left(v_{iw} + r_{iw}^v\right) \quad [3]$$

where $r^q$, $r^k$, and $r^v$ denote the position term inserted into the vector of $q$, $k$, and $v$, when traversing point by point along the direction of width, respectively. Similarly, height-axial attention with position encoding can be given as follows:

$$y_{ij} = \sum_{h=1}^{H}\text{softmax}\left(q_{ij}^T k_{ih} + q_{ij}^T r_{ih}^q + k_{ih}^T r_{ih}^k\right)\left(v_{ih} + r_{ih}^v\right) \quad [4]$$

The width-axial attention layer and height-axial attention layer are set to work consecutively to capture global interactions. They both work alone in their directions. The final results are obtained by contacting the total information together.

#### Residual axial attention

Inspired by the use of residual units which have succeeded in helping Unet achieve quite good performance in the field of medical image segmentation, we sought to apply the residual idea to axial attention layers to further improve the segmentation performance of self-attention mechanism on small breast US image datasets. Similar to the residual attention layer proposed by He *et al.* (38), we proposed residual axial attention by creating a "direct" path to transport the raw attention scores from the previous axial attention layer to the next axial attention layer in the whole transformer network. Following the principle that height- and width-axial attention are originally obtained in a separate way (37), we added height- and width-residual axial attention scores to the axial attention mechanism, respectively, and concatenated the outputs of the multi-head on both height-layer and width-axial layer together as the new output of the present layer, as shown in *Figure 3A*. We provide a more detailed introduction using the corresponding equations expressions.

We assume that residual attention score from the previous attention layer is denoted as Res' and it can be

 

calculated by Eq. [1]. New residual attention score can be obtained as follows:

$$\text{Res} = \text{Res}^{'} + \frac{Q^{'}k^{T}}{\sqrt{d_k}} \quad [5]$$

Therefore, the proposed residual axial-attention can be defined as follows:

$$RAA\left(Q^{'},K^{'},V^{'},\text{Res}^{'}\right) = \text{softmax}\left(\text{Res}^{'} + \frac{Q^{'}K^{T}}{\sqrt{d_k}}\right)V^{'} \quad [6]$$

In addition, it is noted that in each axial-attention layer, 8 attention heads are included. We can obtain the output of each multi-head attention layer as follows:

$$Mh = Concat\left(h_1, h_2, ..., h_h\right)W^{o} \quad [7]$$

where $W^{o}$ is a matrix which is used to transform the concatenation of the outputs of the heads, and $h_h$ denotes the output of the "height-axial" attention head, which could be calculated according to Eq. [1].

Therefore, the final output of the multi-head axial-attention layer can be obtained as follows:

$$O = Concat\left(Mh_h, Mh_w\right) \quad [8]$$

where $Mh_h$ and $Mh_w$ denote the output of height-axial multihead layers and width-axial multihead layers, respectively.

In terms of how to propagate residual attention scores, we give a brief introduction. First, the outputs of the residual axial-attention layer along both height-axial and width-axial are obtained sequentially by using Eq. [6]. Meanwhile, the new residual attention scores, which are defined as Eq. [5], are retained and passed to the next attention layer. Then, we can obtain the output of each multihead axial-attention layer by Eq. [8]. Finally, we can obtain the output of the residual multihead axial attention block by concatenating the output of the height-axial residual multihead layer and width-axial residual multihead layer.

### *Dilated convolution module*

We note that the information of some interactions may be missed when using axial-attention layers. For example, the intimate information between pixels about the direction of diagonal is neglected when axial attention is only used in the direction of height and width. However, any global interactions information is important for medical image segmentation. As shown in *Figure 1*, we assume that false positives in the segmentation results of MT are caused by the lack of enough global information obtained by the model to some degree. Therefore, in order to capture more global interactions information, we introduced a dilated convolution module which is often used in the architecture of Unet (39-41) to improve the receptive field, and the dilation operation can be described as follows:

$$z\left(x,y\right) = \alpha\left\{\sum_{ij} f\left(s + i \times r, y + j \times r\right) \times g\left(i,j\right) + \beta\right\} \quad [9]$$

where $\alpha$ denotes the activate function, $\beta$ is a biased unit, and $r$ is a changeable parameter used to determine how big the receptive field of convolution kernel is. In terms of the size of the receptive field of convolution kernel, we give its calculation formula Eq. [10]. Referring to Eq. [10], when $r$=2, the size of receptive field is increased to 5×5.

$$N = \left\{\left(k_s + 1\right) \times \left(r - 1\right) + k_s\right\}^2 \quad [10]$$

where $k_s$ denotes the size of a convolution kernel and $N$ represents the size of the receptive field.

As shown in *Figure 2*, a dilated module is inserted into the end of the encoder pipeline including five 3×3 convolution kernels with dilation parameters r of 1, 2, 4, 8, and 16, respectively. According to Eq. [10], we can determine that the sizes of the receptive field are N=3×3, 7×7, 15×15, 31×31, and 63×63, respectively. Note that the output feature maps can contain more information due to the wide receptive field of the dilated module.

## Results

### *Datasets and evaluation metrics*

We used 2 breast US image datasets to evaluate our proposed method. Dataset A was a public dataset from Zhang *et al.* (42) consisting of 562 breast US images, each of which contains only 1 tumor which mostly displays a relatively regular size and is located in the center of the image. According to the demonstrations of Zhang *et al.* (42), images from dataset A were of multiple resolutions including 550×357, 555×490, 546×360, and 600×480. They were collected from different US devices, including GE VIVID 7 (GE Healthcare, Milwaukee, WI, USA), Hitachi EUB-6500 (Hitachi Medical Systems, Tokyo, Japan), Philips iU22 (Philips Healthcare, Amsterdam, The

 

Netherlands), and Siemens ACUSON S2000 (Siemens Healthineers, Erlangen, Germany). Four experienced radiologists participated in the generating work of ground truth (GT). More details about dataset A and how to obtain the final GT are provided in a previous paper (42). Different from dataset A, dataset B was a private breast US image dataset including 878 breast US images with the resolutions of 775×580, 1,024×768, and 850×649. They mostly contained tumors of different sizes and uncertain locations. With the approval of the local biological and medical Ethics Committee of the Northeastern University, Shenyang, China, we collected them from different US devices, including GE LOGIQ E9 (GE Healthcare) and Philips EPIQ 5 (Philips Healthcare), in a local hospital. The exclusion criteria were contraindications to undergoing breast US screening, dot-enhancements distributed in multiple quadrants, and non-mass enhancement lesions (none were excluded). We concealed the information of all patients and invited 2 experienced radiologists to ascertain the GT. One radiologist was responsible for sketching contours, and the other was responsible for the examination.

In order to compare our proposed method with state-of-the-art (SOTA) segmentation methods fairly and subjectively, the metrics used in previous papers (42,43) were adopted. The metrics included accuracy (ACC), true-positive ratio (TPR), false-positive ratio (FPR), Jaccard index (JI), Dice similarity coefficient (DSC), area error ratio (AER), Hausdorff error (HE), and mean absolute error (MAE), respectively. As previously illustrated (38), the segmentation performance is indicated as quite good when JI and DSC are both large, and AER, HE, and MAE are all small.

### *Implementation details*

We used PyTorch (Meta AI, New York, NY, USA) to implement the proposed method on 1 Nvidia GeForce GTX 1080ti GPU (Nvidia, Santa Clara, CA, USA), and set the learning rate, batch size, and epoch number to 0.001, 4, and 400, respectively. All of the input images were resized to 256×256. Identical with MT (18), we also used the same data preprocessing method which had been given a detailed introduction in the supplement file of MT (18), as well as the binary cross-entropy loss function to measure the error between GT and predictions.

In similarity with the testing method used in the paper (42), for both dataset A and dataset B segmentation, we used the 5-fold cross-validation experiment method to test our proposed method, as well as MT which is the most similar to the DT. Meanwhile, in order to demonstrate that transformer-based methods have a better ability to model the distant dependencies than do convolution-based methods, we also compared our proposed method with Unet (9) and Unet++ (10), which are typical convolution-based medical image segmentation models with excellent performance. The average result of the 5-fold tests was taken as the final result for each metric, which decreased the possibility of unstable performance of the model and guaranteed persuasiveness to a great degree. In addition, we used the Mann-Whitney U test to statistically analyze these results. A 95% confidence level was used for obtaining the corresponding confidence interval (95% CI).

Additionally, for dataset A segmentation, we also compared the DT with some fully automated segmentation methods (44-46) which have achieved SOTA on dataset A. The salient attention Unet (SAUnet) (43) was the most recent method to achieve SOTA performance on dataset A. It is also an improved CNN-based method and incorporates radiologists' visual attention. To compare the DT with it subjectively, we took the test result obtained by the model with the best performance in the 5-fold test.

### *Dataset A*

Quantitative results of comparison experiments on dataset A are shown in *Table 1*. It shows that the DT achieves the highest TPR of 0.887, highest JI of 0.813, and highest DSC of 0.889. Meanwhile, the DT has the lowest FPR of 0.104, lowest AER of 0.224, and lowest HE of 38.890. Compared with MT (18), the DT improves TPR, JI, and DSC by 1.3%, 2.4%, and 1.6%, respectively. Especially, on metrics of FPR, AER, HE, and MAE, DT outperforms MT by 1.7%, 2.7%, 25.21, and 15.53, respectively. Compared with typical medical images segmentation methods such as Unet and Unet++, DT outperforms them on all metrics. Especially on the metrics of FPR, HE, and MAE, DT reduces them significantly by 5.8%, 15.02, and 10.12, respectively, which indicates that DT has a strong ability to model global interactions. Compared with fully automated segmentation methods which have performed SOTA on dataset A (42), the DT improves TPR, JI, and DSC by 7.7%, 9.3%, and 5.9%, respectively and reduces FPR, AER, and HE by 5.5%, 1.4%, and 10.33, respectively, although DT is second only to the SOTA on metric of MAE at a disadvantage of 1.76.

In addition, *Tables 2,3* list the standard error of mean on all metrics for each method and the 95% CI and P values,

 

**Table 1** Quantitative results of comparison experiments on dataset A

| Baseline | Method | TPR | FPR | JI | DSC | AER | HE | MAE |
|---|---|---|---|---|---|---|---|---|
| Transformer-based | MT | 0.874 | 0.121 | 0.789 | 0.873 | 0.251 | 64.101 | 30.011 |
| | DT | 0.887 | 0.104 | 0.813 | 0.889 | 0.224 | 38.890 | 14.484 |
| Convolution-based | Unet | 0.843 | 0.228 | 0.754 | 0.840 | 0.378 | 53.909 | 24.601 |
| | Unet++ | 0.849 | 0.162 | 0.759 | 0.849 | 0.310 | 83.893 | 42.834 |
| Others | (44) | 0.810 | 0.159 | 0.720 | 0.830 | 0.362 | 49.221 | 12.721 |
| | (45) | 0.809 | 1.063 | 0.592 | 0.701 | 1.251 | 107.610 | 26.619 |
| | (46) | 0.674 | 0.180 | 0.612 | 0.710 | 0.510 | 69.202 | 21.306 |

TPR, true-positive ratio; FPR, false-positive ratio; JI, Jaccard index; DSC, Dice similarity coefficient; AER, area error ratio; HE, Hausdorff error; MAE, mean absolute error; MT, medical transformer; DT, dilated transformer.

**Table 2** Standard error analysis of the comparison experiment on dataset A (± standard error)

| Method | TPR (±) | FPR (±) | JI (±) | DSC (±) | AER (±) | HE (±) | MAE (±) |
|---|---|---|---|---|---|---|---|
| MT | 0.005 | 0.011 | 0.006 | 0.004 | 0.012 | 2.445 | 1.193 |
| DT | 0.005 | 0.008 | 0.005 | 0.004 | 0.009 | 2.032 | 1.0541 |
| Unet | 0.007 | 0.031 | 0.008 | 0.007 | 0.032 | 2.551 | 1.566 |
| Unet++ | 0.007 | 0.018 | 0.007 | 0.006 | 0.019 | 3.067 | 1.939 |

TPR, true-positive ratio; FPR, false-positive ratio; JI, Jaccard index; DSC, Dice similarity coefficient; AER, area error ratio; HE, Hausdorff error; MAE, mean absolute error; MT, medical transformer; DT, dilated transformer.

**Table 3** CI and P value analysis in the comparison experiment on dataset A

| Metrics | 95% CI | | | | P value | | |
|---|---|---|---|---|---|---|---|
| | MT | DT | Unet | Unet++ | DT *vs.* MT | DT *vs.* Unet | DT *vs.* Unet++ |
| TPR | (0.862, 0.881) | (0.877, 0.895) | (0.829, 0.859) | (0.829, 0.859) | 0.006 | 0.021 | 0.097 |
| FPR | (0.096, 0.138) | (0.088, 0.121) | (0.168, 0.290) | (0.122, 0.194) | 0.110 | 0.733 | 0.088 |
| JI | (0.786, 0.808) | (0.803, 0.822) | (0.730, 0.762) | (0.747, 0.776) | 0.070 | 1.681e-6 | 1.294e-6 |
| DSC | (0.872, 0.889) | (0.884, 0.898) | (0.823, 0.850) | (0.839, 0.863) | 0.070 | 1.681e-6 | 1.291e-6 |
| AER | (0.223, 0.268) | (0.201, 0.237) | (0.322, 0.448) | (0.268, 0.342) | 0.111 | 2.394e-6 | 9.843e-7 |
| HE | (45.940, 55.521) | (34.961, 42.933) | (48.914, 58.911) | (77.870, 89.891) | 0.005 | 0.001 | 6.929e-21 |
| MAE | (27.361, 32.040) | (12.440, 16.572) | (21.532, 27.671) | (39.020, 46.621) | 0.008 | 3.038e-5 | 8.201e-22 |

CI, confidence interval; MT, medical transformer; DT, dilated transformer; TPR, true-positive ratio; FPR, false-positive ratio; JI, Jaccard index; DSC, Dice similarity coefficient; AER, area error ratio; HE, Hausdorff error; MAE, mean absolute error.

respectively. From *Table 3*, we can observe that the results in *Table 1* all are contained in the corresponding CI, and 3 of 7 of P values of DT *vs.* MT, 6 out of 7 of P values of DT *vs.* Unet, and 5 out of 7 of P values of DT *vs.* Unet++ are less than 0.05. Therefore, we conclude that the DT outperformed other methods significantly and statistically on dataset A.

*Table 4* lists the quantitative results obtained by comparing the DT with SAUnet. It shows that DT outperforms SAUnet slightly on all metrics of TPR, FPR, DSC, and ACC, as well as JI. Therefore, from the perspective of quantitative evaluation metrics, DT has achieved SOTA segmentation performance on dataset A,

 

**Table 4** Quantitative result obtained by comparing DT with SAUnet on dataset A

| Method | TPR | FPR | JI | DSC | ACC |
|---|---|---|---|---|---|
| SAUnet | 0.899 | 0.106 | 0.825 | 0.896 | 0.978 |
| DT | 0.903 | 0.102 | 0.825 | 0.901 | 0.980 |

The JI value shown in this table are after rounding, and the exact value of JI for DT is 0.034 percentage points higher than that of SAUnet. DT, dilated transformer; SAUnet, salient attention Unet; TPR, true-positive ratio; FPR, false-positive ratio; JI, Jaccard index; DSC, Dice similarity coefficient; ACC, accuracy.

especially in reducing false positives, which can also be observed intuitively from *Figure 4*.

As shown in *Figure 4*, the segmentation results obtained by both the DT and SAUnet have relatively few false positives, indicating that both methods perform well in distinguishing precise boundaries of the ROI. In addition, the DT has a better ability to refine segmentation. It could capture more detailed information than SAUnet. Taking the image in the fourth row as an example, the DT could distinguish the invaginated part at the top of the tumor, but SAUnet could not. Similarly, the DT could identify the raised feature which is located at the 7 o'clock position of the tumor in the second row, but SAUnet could not. Especially for the image in the fifth row, due to the influence of the glandular tissue with high echo characteristics, SAUnet was unable to predict well at the 5 o'clock position although the DT could. This is mainly on account of the self-attention mechanism used in the main building blocks. Therefore, MT also has a better ability to model long-range dependencies than convolution-based methods such as Unet and Unet++. Taking images in the second and fifth row as examples, compared with Unet and Unet++, MT reduces false positives significantly. Due to the interference caused by low contrast or subcutaneous fat tissue or duct tissue with hypoechoic characteristics and the insufficient ability to grasp global dependencies, MT performs relatively weakly in distinguishing the actual boundary of the ROI and also caused a few false positives (such as images in the first, fourth, seventh, and eighth row), while the DT could better identify the tumor location precisely by capturing more global information than could the MT. Therefore, the DT has a strong ability to model distant interactions and capture more features of an ROI, with enhanced attention given to tumor boundaries.

### *Dataset B*

*Table 5* lists the quantitative results of comparison experiments on dataset B. It shows that the DT achieves the highest TPR of 0.798, highest JI of 0.662, highest DSC of 0.769, and the lowest HE of 83.500 and lowest MAE of 35.741, which indicates that the DT has the best segmentation performance. Due to the self-attention mechanism used in the main building blocks of the encoder, when compared with convolution-based methods, DT and MT both have a stronger ability to model the distant dependencies and could identify the actual location of the ROI with the lowest false positives. Especially, on the metrics of FPR and AER, MT outperformed Unet by 9.3% and 8.2%, respectively. On the metrics of TPR, JI, DSC, HE, and MAE, the DT outperformed Unet by 5.0%, 4.9%, 4.6%, 10.88, and 16.27, respectively. When compared with MT, the DT improved TPR, JI, and DSC by 5.90%, 2.67%, and 4.68%, respectively. This indicates that the DT has a better ability to capture more and detailed information of the ROI than does MT. In addition, MT achieved the lowest FPR of 0.303 and the lowest AER of 0.565 although it has disadvantages in metrics of TPR, JI, and DSC, which indicates that MT lacks the ability to learn enough detailed information on the ROI, especially in the precision of boundary identification.

The results of statistical analysis on the test results of dataset B are shown in *Tables 6*,7, respectively. We can observe that the results in *Table 5* are all contained in the corresponding CIs, and 5 of 7 of the P values of the DT *vs.* MT, 6 out of 7 of the P values of the DT *vs.* Unet, and 6 out of 7 of the P values of the DT *vs.* Unet++ are less than 0.05. Therefore, we conclude that the DT outperformed other methods significantly and statistically on dataset B. It can also be observed from *Figure 5* intuitively.

In *Figure 5*, from left to right, the sample images, GT, and segmentation results obtained by MT, DT, Unet, and Unet++ are shown respectively. Compared with GT, it is easy for us to observe that the segmentation results of MT (images in the third column) are incomplete. Moreover, they locate the correct position of the ROI with few false positives, which indicates that although MT could find the location of ROI, it could not capture enough detailed

 

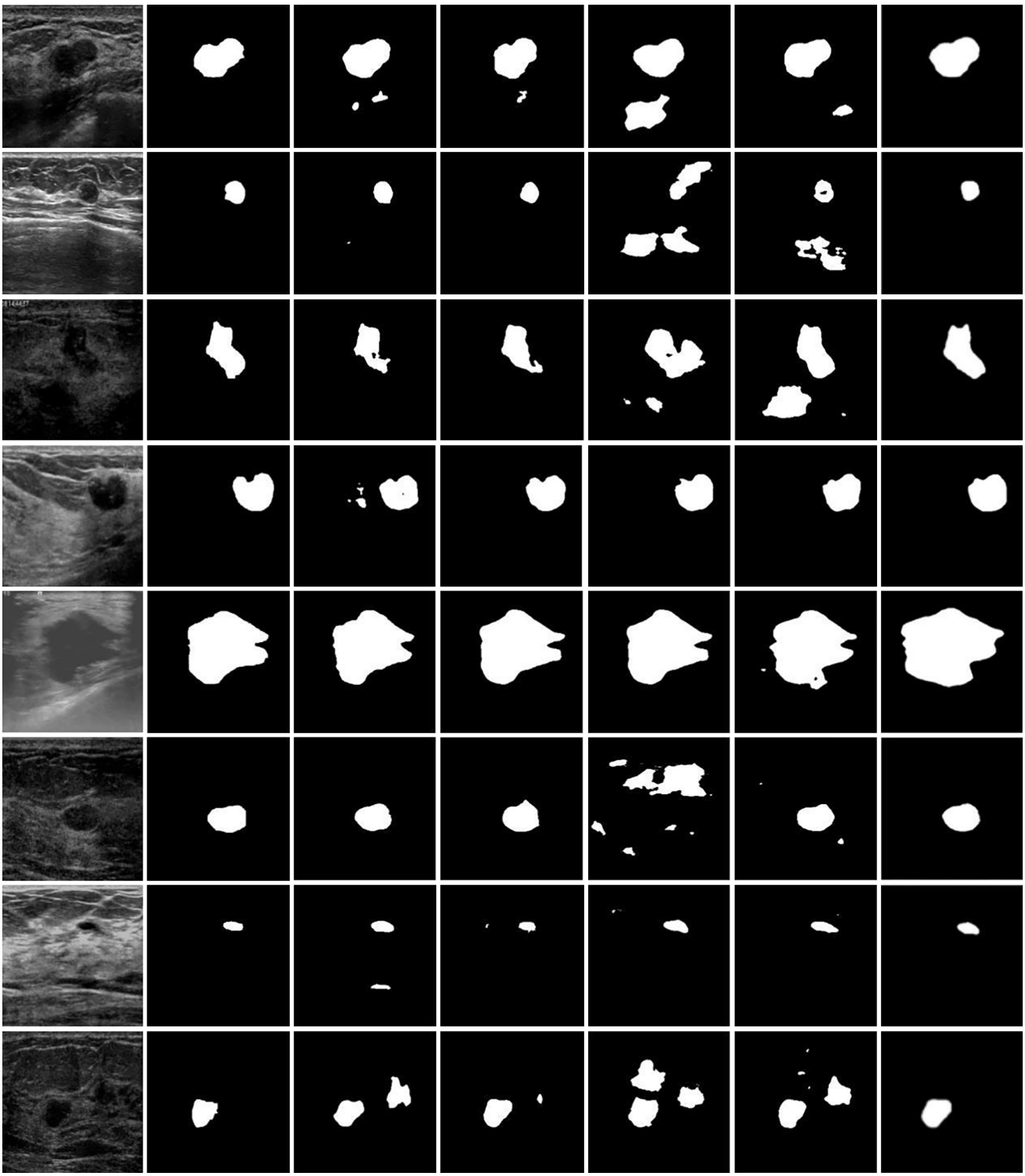

**Figure 4** Segmentation results obtained by different segmentation methods on dataset A. From left to right, they are sample images, GT, segmentation results obtained by MT, DT, Unet, Unet++ and SAUnet, respectively. GT, ground truth; MT, medical transformer; DT, dilated transformer; SAUnet, salient attention Unet.

**Table 5** Quantitative results of comparison experiments on dataset B

| Baseline | Method | TPR | FPR | JI | DSC | AER | HE | MAE |
|---|---|---|---|---|---|---|---|---|
| Transformer-based | MT | 0.739 | 0.303 | 0.635 | 0.723 | 0.565 | 89.181 | 48.922 |
| | DT | 0.798 | 0.386 | 0.662 | 0.769 | 0.588 | 83.500 | 35.741 |
| Convolution-based | Unet | 0.748 | 0.396 | 0.613 | 0.723 | 0.647 | 94.381 | 52.013 |
| | Unet++ | 0.726 | 0.428 | 0.587 | 0.704 | 0.702 | 105.480 | 65.180 |

TPR, true-positive ratio; FPR, false-positive ratio; JI, Jaccard index; DSC, Dice similarity coefficient; AER, area error ratio; HE, Hausdorff error; MAE, mean absolute error; MT, medical transformer; DT, dilated transformer.

**Table 6** Standard error analysis of the comparison experiment on dataset B (± standard error)

| Method | TPR (±) | FPR (±) | JI (±) | DSC (±) | AER (±) | HE (±) | MAE (±) |
|---|---|---|---|---|---|---|---|
| MT | 0.008 | 0.027 | 0.007 | 0.007 | 0.029 | 1.902 | 1.361 |
| DT | 0.008 | 0.035 | 0.008 | 0.007 | 0.036 | 1.985 | 1.181 |
| Unet | 0.009 | 0.033 | 0.008 | 0.008 | 0.034 | 1.944 | 1.360 |
| Unet++ | 0.009 | 0.035 | 0.008 | 0.008 | 0.037 | 1.939 | 1.492 |

TPR, true-positive ratio; FPR, false-positive ratio; JI, Jaccard index; DSC, Dice similarity coefficient; AER, area error ratio; HE, Hausdorff error; MAE, mean absolute error; MT, medical transformer; DT, dilated transformer.

**Table 7** CI and P value analysis in the comparison experiment on dataset B

| Metrics | 95% CI | | | | P value | | |
|---|---|---|---|---|---|---|---|
| | MT | DT | Unet | Unet++ | DT *vs.* MT | DT *vs.* Unet | DT *vs.* Unet++ |
| TPR | (0.721, 0.753) | (0.782, 0.812) | (0.731, 0.766) | (0.709, 0.743) | 2.912e-12 | 1.490e-5 | 1.750e-13 |
| FPR | (0.242, 0.350) | (0.318, 0.453) | (0.332, 0.460) | (0.360, 0.496) | 3.310e-8 | 0.847 | 0.103 |
| JI | (0.732, 0.761) | (0.647, 0.676) | (0.597, 0.630) | (0.570, 0.602) | 0.004 | 5.021e-5 | 1.221e-13 |
| DSC | (0.732, 0.761) | (0.755, 0.782) | (0.706, 0.739) | (0.688, 0.720) | 0.004 | 5.021e-5 | 1.231e-13 |
| AER | (0.502, 0.616) | (0.518, 0.659) | (0.581, 0.714) | (0.631, 0.774) | 0.087 | 1.852e-4 | 2.922e-12 |
| HE | (84.172, 91.631) | (79.612, 87.393) | (90.571, 598.190) | (101.721, 109.313) | 0.060 | 6.050e-5 | 4.790e-15 |
| MAE | (45.671, 51.001) | (33.431, 38.061) | (49.371, 54.701) | (62.301, 68.160) | 7.152e-13 | 2.992e-22 | 6.460e-56 |

CI, confidence interval; MT, medical transformer; DT, dilated transformer; TPR, true-positive ratio; FPR, false-positive ratio; JI, Jaccard index; DSC, Dice similarity coefficient; AER, area error ratio; HE, Hausdorff error; MAE, mean absolute error.

information. In addition, in *Figure 5*, we can see that there are many false positives in the segmentation results of Unet and Unet++, yet MT and the DT could avoid these false positives to a certain degree. Taking the image in the seventh row as an example, fat tissue with hypoechoic characteristics can easily be mistaken by Unet and Unet++ as a tumor region, but the DT could identify the actual ROI, which is shown in the fourth column and seventh row. This indicates that transformer-based methods are superior to convolution-based methods in modeling global interactions and reducing false positives of segmentation results.

## Discussion

Due to the strong ability to model global interactions information, transformer-based methods have become increasingly popular in the area of computer vision as well as medical image analysis. However, limited datasets in the field of medical images restricts the development of

 

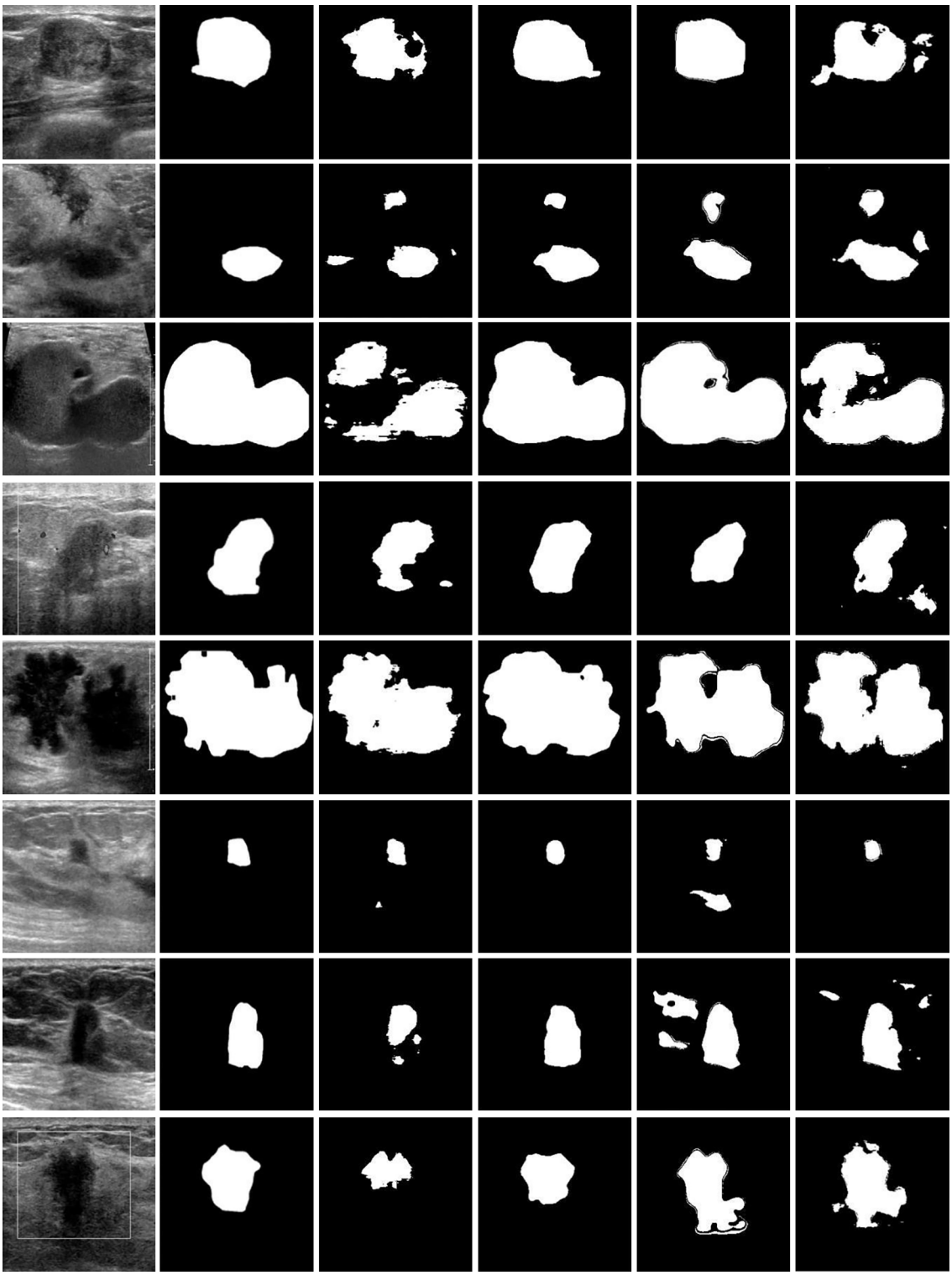

**Figure 5** Segmentation results obtained by different segmentation methods on dataset B. From left to right are original images, GT, segmentation results obtained by MT, the DT, Unet and Unet++, respectively. GT, ground truth; MT, medical transformer; DT, dilated transformer.

**Table 8** Study on the setting of the number of encoder blocks

| Block number | A | | | B | | |
|---|---|---|---|---|---|---|
| | JI | DSC | FPR | JI | DSC | FPR |
| 2 | 0.815 | 0.894 | 0.109 | 0.677 | 0.780 | 0.384 |
| 3 | 0.816 | 0.894 | 0.145 | 0.664 | 0.786 | 0.072 |
| 4 | 0.825 | 0.901 | 0.102 | 0.702 | 0.803 | 0.456 |
| 5 | 0.786 | 0.872 | 0.140 | 0.681 | 0.787 | 0.289 |

JI, Jaccard index; DSC, Dice similarity coefficient; FPR, false-positive ratio.

transformers, because most transformer-based methods rely heavily on pretraining on large datasets. Therefore, we proposed the DT, which is based on the proposed residual axial-attention mechanism. The DT could perform well on small breast US image datasets without pretraining, which is of significance to the popularity of CADs. Of note, the DT significantly outperformed other related methods on dataset B when compared to dataset A. This indicates that the DT has a more stable and better generalization performance since we found it more challenging to accurate locate the tumors in dataset B than in dataset A. The method most closely related to the DT is the MT (18). Compared with the MT, the DT has only 1 branch of encoder-decoder architecture. We set 4 encoder blocks in the encoder path (in terms of the number of encoder blocks we set, more detailed demonstration is given in the following subsection). In addition, a dilation module (DM) was added to the end of the encoder pipeline to improve the receptive field and compensate for the loss of global interactions information caused by the process of changing 2D attention to axial attention with only 1 dimension. In addition, we added a path used to propagate residual information between axial-attention layers to capture more detailed local information. To give an illustration of the critical role that each add-on part plays, we conducted the following ablation study on both dataset A and dataset B. Three typical metrics of JI, DSC, and FPR were taken as the evaluation metrics.

### *Setting of the number of encoder blocks*

To verify that the DT network with four encoder blocks has a better performance, we conducted the following study. The number of encoder blocks was set to 2, 3, 4, and 5, respectively. Other network components were kept the same as those of the DT. During the training period, batch size, epoch, and learning rates were set to 4, 400, and 0.001, respectively. After training the DT with 2, 3, 4, and 5 encoder blocks on both dataset A and dataset B, 4 models which were denoted as $DT_2$, $DT_3$, $DT_4$, and $DT_5$ were obtained on dataset A and dataset B, respectively. After evaluation on dataset A and dataset B, respectively, the quantitative results were obtained (*Table 8*).

From *Table 8*, we can observe that $DT_4$ achieves the highest JI and DSC both on dataset A and dataset B, and it outperforms $DT_5$ on the metrics of JI and DSC by 3.94% and 2.87%, respectively. This indicates that the DT with 4 encoder blocks has the best ability to capture enough details to find a complete tumor region and that the performance would not be necessarily improved if the number of encoder blocks were increased to 5. In dataset B, when the encoder blocks number was increased from 2 to 5, three evaluation metrics did not show a common trend of positive growth. However, when the encoder blocks number was increased from 2 to 4, JI and DSC generally had a positive growth trend. This indicates that adding encoder blocks could help to learn more features of the ROI and capture more global interactions when the number of encoder blocks is less than 5. We assume that the risk of overfitting that easily causes unstable prediction results also increases when adding encoder blocks. To further verify this assumption, we plotted 2 figures to illustrate how validation loss changes during the period of training on both dataset A and dataset B, respectively.

As shown in *Figures 6*,*7*, we can observe that when the encoder blocks number is set to 2 and 4, the validation loss curves (line in blue and red) all show a better and stable decline, while the curves in cyan and green show a sudden "rise" where there is an inflection point both on dataset A and dataset B. This indicates that when the number of encoder blocks is set to 3 and 5, the model is unstable and has a greater risk of over-fitting. However, when the encoder blocs number is 2 and 4, the model shows a relatively stable condition. Especially on dataset B, the curve in the red line converges faster than the curve

 

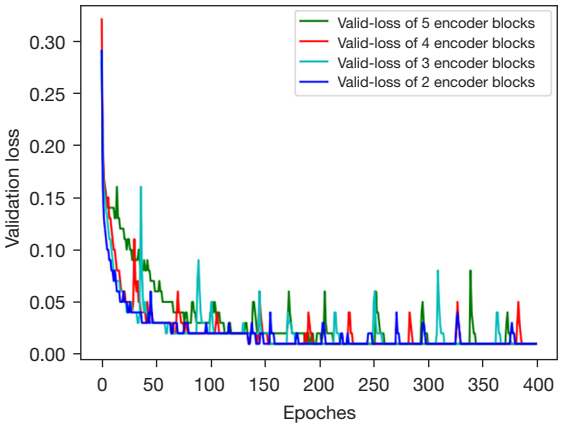


**Figure 6** Validation loss on dataset A

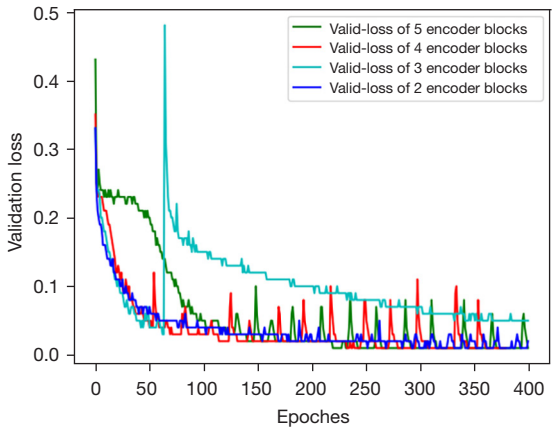


**Figure 7** Validation loss on dataset B.

**Table 9** Results of the ablation experiment

| DM | Res | A | | | B | | |
|---|---|---|---|---|---|---|---|
| | | JI | DSC | FPR | JI | DSC | FPR |
| × | × | 0.648 | 0.759 | 0.578 | 0.797 | 0.881 | 0.156 |
| √ | × | 0.687 | 0.789 | 0.357 | 0.814 | 0.893 | 0.107 |
| × | √ | 0.693 | 0.797 | 0.439 | 0.815 | 0.894 | 0.110 |
| √ | √ | 0.702 | 0.803 | 0.456 | 0.825 | 0.901 | 0.103 |

DM, dilation module; Res, residual path; JI, Jaccard index; DSC, Dice similarity coefficient; FPR, false-positive ratio.

in blue, which indicates that when the number of encoder blocks is set to 4, the model has a better adaptability and generalization performance on both datasets.

### *Contribution made by each of add-on parts*

To verify the contribution made by the dilated module and residual score path, we conducted the following ablation study on both dataset A and dataset B. As shown in *Table 9*, a check mark indicates that the module is used, and a cross mark indicates that it is not used. First, we can observe that the DT with both DM and Res has the highest JI of 0.702 and highest DSC of 0.803 on dataset A; and has the highest JI of 0.825, highest DSC of 0.901, and the lowest FPR of 0.103 on dataset B. This indicates that the proposed DT method can achieve the best performance with the help of DM and Res path together. Second, for both dataset A and dataset B segmentation, the DT with only DM or only Res both outperform the DT without DM and Res on the 3 metrics, which indicates that DM and Res both contribute to the excellent final performance. In particular, when only adding the DM module, the FPR is reduced by 22.06% and 4.91% on dataset A and dataset B, respectively. This indicates that DM plays a significant role in reducing false positives and helps the DT capture much global information by improving the receptive field. In addition, when only adding Res, JI and DSC are all improved by 4.50% and 3.72%, and 1.7% and 1.28% on dataset A and dataset B, respectively. This indicates that residual axial-attention information passed between layers matters and contributes to capturing more detailed local information. On the whole, the introduced DM and the proposed residual axial attention are both important and help the DT to achieve the best performance together.

### *Computational complexity*

The proposed DT method was implemented by PyTorch and evaluated on 2 breast US image datasets using 1 NVIDIA GeForce GTX 1080ti GPU. *Table 10* shows the results of analysis on the computation complexity which measures the amount of computing resources that a particular algorithm

 

**Table 10** Analysis of the computation complexity

| Method | Trainable parameters (M) | GFLOPs | Inference time (s/image) | Training time (h) |
|---|---|---|---|---|
| MT | 5.97 | 7.77 | 0.44 | 11.5 |
| DT | 16.37 | 9.03 | 0.13 | 7.25 |

Training time is the average of the training time spent on the 2 datasets. GFLOPs, Giga Floating point Operations Per Second; MT, medical transformer; DT, dilated transformer.

consumes when it runs. We can observe that the DT has a few more trainable parameters and Giga Floating point Operations Per Second (GFLOPs) than does the most related work, MT. This is mainly due to the introduced DM modules where multiple convolution operations are introduced. However, it is worth noting that the DT requires less training and inference time than does the MT, which indicates the superiority of the architecture and the practicality of the DT. Therefore, the DT has relatively good spatial and computational complexity and has both precise segmentation performance and high efficiency.

## Conclusions

This paper proposes a DT that uses the proposed residual axial attention layers to build up transformer encoders to improve the segmentation performance of transformer-based methods on small breast US images datasets. It was evaluated on 2 small breast US image datasets and was shown to effectively without pretraining, especially in identifying the accurate location and precise boundary of ROIs and reducing false positives. In addition, ablation experiments were conducted, and the results showed that the DT with 4 encoder blocks has a relatively stable performance and that both the introduced DM and the proposed residual axial-attention contribute to accurate segmentation with high efficiency.

## Acknowledgments

*Funding:* This research was supported in part by the National Natural Science Foundation of China (No. 61702087) and in part by the Fundamental Research Funds for the Central Universities under (No. N172008008).

## Footnote

*Reporting Checklist:* The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-22-33/rc

*Conflicts of Interest:* All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-22-33/coif). The authors have no conflicts of interest to declare.

*Ethical Statement:* The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Biological and Medical Ethics Committee of Northeastern University, Shenyang, China, and individual consent for this retrospective analysis was waived.



## References

1. Gonçalves VM, Delamaro ME, Nunes FLS. A systematic review on the evaluation and characteristics of computer-aided diagnosis systems. Revista Brasileira de Engenharia Biomédica 2014;30:355-83.
2. Tsochatzidis L, Zagoris K, Arikidis N, Karahaliou A, Costaridou L, Pratikakis I. Computer-aided diagnosis of mammographic masses based on a supervised content-based image retrieval approach. Pattern Recognit 2017;71:106-17.
3. Halalli, B, Makandar, A. Computer Aided Diagnosis -

 

Medical Image Analysis Techniques. In: Kuzmiak, CM. editor. Breast Imaging. London: IntechOpen, 2017. Available online: https://www.intechopen.com/chapters/56615
4. Stavros TA. Breast ultrasound. 2004.
5. Yu Y, Acton ST. Speckle reducing anisotropic diffusion. IEEE Trans Image Process 2002;11:1260-70.
6. Shan J. A fully automatic segmentation method for breast ultrasound images. Utah State University, 2011.
7. Hiremath P, Akkasaligar PT, Badiger S. Speckle Noise Reduction in Medical Ultrasound Images. In: Gunarathne G, editor. Advancements and Breakthroughs in Ultrasound Imaging. London: IntechOpen, 2013. Available online: https://www.intechopen.com/chapters/45101
8. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Bing S, Liu T, Wang X, Wang G, Cai J, Chen T. Recent advances in convolutional neural networks. Pattern Recognit 2018;77:354-77.
9. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. MICCAI 2015: International Conference on Medical image computing and computer-assisted intervention. Munich: Springer, 2015:234-41.
10. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2018) 2018;11045:3-11.
11. Zhang Z, Liu Q, Wang Y. Road extraction by deep residual u-net. IEEE Geosci Remote Sens Lett 2018;15:749-53.
12. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D. Attention gated networks: Learning to leverage salient regions in medical images. Med Image Anal 2019;53:197-207.
13. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. CPVR 2018: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018:7132-41.
14. Woo S, Park J, Lee JY, Kweon IS. Cbam: Convolutional block attention module. ECCV 2018: Proceedings of the European conference on computer vision. Munich: Springer, 2018:3-19.
15. Cai S, Tian Y, Lui H, Zeng H, Wu Y, Chen G. Dense-UNet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. Quant Imaging Med Surg 2020;10:1275-85.
16. Olveres J, González G, Torres F, Moreno-Tagle JC, Carbajal-Degante E, Valencia-Rodríguez A, Méndez-Sánchez N, Escalante-Ramírez B. What is new in computer vision and artificial intelligence in medical image analysis applications. Quant Imaging Med Surg 2021;11:3830-53.
17. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Tang X. Residual attention network for image classification. CPVR 2017: Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, Hawaii: IEEE, 2017:3156-64.
18. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: Gated axial-attention for medical image segmentation. MICCAI 2021: International Conference on Medical Image Computing and Computer-Assisted Intervention. Strasbourg: Springer, 2021:36-46.
19. Zhao H, Shi J, Qi X, Wang X, and Jia J. Pyramid scene parsing network. CVPR 2017: Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, Hawaii: Springer, 2017:2881-90.
20. Chen L, Papandreou G, Kokkinos L, Murphy K, and Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected crfs. ICLR 2015: International Conference on Learning Representations. California, USA, 2015.
21. Ramachandran P, Parmar N, Vaswani A, Bello I, Levskaya A, Shlens J. Stand-alone self-attention in vision models. NeurIPS 2019: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019:68-80.
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. NeurIPS 2017: Proceedings of the 31rd International Conference on Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates Inc., 2017.
23. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. CVPR 2016: Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2016:770-778.
24. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu P. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 2020;21:1-67.
25. Hu H, Gu J, Zhang Z, Dai JF, Wei Y. Relation networks for object detection. CVPR 2018: Proceedings ofthe IEEE conference on computer vision and pattern recognition. Salt Lake City, UT, USA: IEEE, 2018:3588-97.
26. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021: International Conference on Learning Representations. Vienna, Austria, 2021.

 

27. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jegou H. Training data-efficient image transformers & distillation through attention. ICML 2021: International Conference on Machine Learning. Virtual Event. PMLR, 2021: 10347-57.
28. Yuan L, Chen Y, Wang T, Yu W, Shi Y, Jiang Z, Tay FE, Feng J, and Yan S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. ICCV 2021: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, 2021:538-547.
29. Wang W, Xie E, Li X, Fan D, Song K, Liang D, Lu T, Luo P, Shao L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. ICCV 2021: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, 2021:568-78.
30. Liu Z, Lin YT, Cao Y, Hu H, Wei YX, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. ICCV 2021: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, 2021:10012-22.
31. Parikh AP, Tackstrom O, Das D, Uszkoreit J. A decomposable attention model for natural language inference. EMNLP 2016: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, TX, USA: The Association for Computational Linguistics, 2016:2249-55.
32. Paulus R, Xiong C, Socher R. A deep reinforced model for abstractive summarization. ICLR 2018: International Conference on Learning Representations. Vancouver, 2018.
33. Bahdanau D, Cho K, Bengio Y. Neural machine translation byjointly learning to align and translate. ICLR 2015: International Conference on Learning Representations. California, USA: 2015.
34. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. CVPR 2018: Proceedings ofthe IEEE conference on computer vision and pattern recognition. Salt Lake City, UT, USA: IEEE, 2018:7794-803.
35. Huang Z, Wang X, Huang L, Huang C, Wei YC, Liu W. Ccnet: Criss-cross attention for semantic segmentation. ICCV 2019: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, 2019:603-12.
36. Hu H, Zhang Z, Xie Z, Lin S. Local relation networks for image recognition. ICCV 2019: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, 2019:3464-73.
37. Wang HY, Zhu YK, Green B, Adam H, Yuille A, Chen LC. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. ECCV 2020: European Conference on Computer Vision. Online event: Springer, 2020:108-26.
38. He RN, Ravula A, Kanagal B, Ainslie J. Realformer: Transformer likes residual attention. ACL-IJCNLP 2021: Findings of The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Bangkok, 2021.
39. Chen Y, Guo Q, Liang X, Wang J, and Qian Y. Environmental sound classification with dilated convolutions. Appl Acoust 2019;148:123-32.
40. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. ICLR 2016: International Conference on Learning Representations. San Juan, 2016.
41. Zhuang Z, Li N, Joseph Raj AN, Mahesh VGV, Qiu S. An RDAU-NET model for lesion segmentation in breast ultrasound images. PLoS One 2019;14:e0221535.
42. Zhang Y, Xian M, Cheng HD, Shareef B, Ding J, Xu F, Huang K, Zhang B, Ning C, Wang Y. BUSIS: A Benchmark for Breast Ultrasound Image Segmentation. Healthcare (Basel) 2022;10:729.
43. Vakanski A, Xian M, Freer PE. Attention-Enriched Deep Learning Model for Breast Tumor Segmentation in Ultrasound Images. Ultrasound Med Biol 2020;46:2819-33.
44. Xian M, Zhang Y, Cheng H. Fully automatic segmentation of breast ultrasound images based on breast characteristics in space and frequency domains. Pattern Recognit 2015;48:485-97.
45. Torbati N, Ayatollahi A, Kermani A. An efficient neural network based method for medical image segmentation. Comput Biol Med 2014;44:76-87.
46. Shan J, Cheng HD, Wang Y. A novel segmentation method for breast ultrasound images based on neutrosophic l-means clustering. Med Phys 2012;39:5669-82.

**Cite this article as:** Shen X, Wang L, Zhao Y, Liu R, Qian W, Ma H. Dilated transformer: residual axial attention for breast ultrasound image segmentation. Quant Imaging Med Surg 2022;12(9):4512-4528. doi: 10.21037/qims-22-33

 