Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

BPAT-UNet: Boundary preserving assembled transformer UNet for ultrasound thyroid nodule segmentation

Hui Bi[a,b,c], Chengjie Cai[a], Jiawei Sun[b,d,e], Yibo Jiang[f], Gang Lu[g,h], Huazhong Shu[g,h], Xinye Ni[b,d,e,*]

[a] School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou, Jiangsu 213164, China
[b] The Affiliated Changzhou NO.2 People's Hospital of Nanjing Medical University, Changzhou, Jiangsu 213003, China
[c] Key Laboratory of Computer Network and Information Integration, Southeast University, Nanjing, Jiangsu 211096, China
[d] Jiangsu Province Engineering Research Center of Medical Physics, Changzhou, Jiangsu 213003, China
[e] Center of Medical Physics, Nanjing Medical University, Changzhou, Jiangsu 213003, China
[f] Changzhou Institute of Technology, Changzhou, Jiangsu 213032, China
[g] Laboratory of Image Science and Technology, Southeast University, Nanjing, Jiangsu 210096, China
[h] Centre de Recherche en Information Biomédicale Sino-Français, Rennes F-35000, France

ARTICLE INFO

Article history:
Received 13 December 2022
Revised 23 April 2023
Accepted 15 May 2023

Keywords:
Medical ultrasound image segmentation
Thyroid nodules segmentation
Computer-aided diagnosis and treatment
Transformer-based network
BPAT-UNet

ABSTRACT

Background and Objective: Accurate and efficient segmentation of thyroid nodules on ultrasound images is critical for computer-aided nodule diagnosis and treatment. For ultrasound images, Convolutional neural networks (CNNs) and Transformers, which are widely used in natural images, cannot obtain satisfactory segmentation results, because they either cannot obtain precise boundaries or segment small objects. Methods: To address these issues, we propose a novel Boundary-preserving assembly Transformer UNet (BPAT-UNet) for ultrasound thyroid nodule segmentation. In the proposed network, a Boundary point supervision module (BPSM), which adopts two novel self-attention pooling approaches, is designed to enhance boundary features and generate ideal boundary points through a novel method. Meanwhile, an Adaptive multi-scale feature fusion module (AMFFM) is constructed to fuse features and channel information at different scales. Finally, to fully integrate the characteristics of high-frequency local and low-frequency global, the Assembled transformer module (ATM) is placed at the bottleneck of the network. The correlation between deformable features and features-among computation is characterized by introducing them into the above two modules of AMFFM and ATM. As the design goal and eventually demonstrated, BPSM and ATM promote the proposed BPAT-UNet to further constrain boundaries, whereas AMFFM assists to detect small objects. Results: Compared to other classical segmentation networks, the proposed BPAT-UNet displays superior segmentation performance in visualization results and evaluation metrics. Significant improvement of segmentation accuracy was shown on the public thyroid dataset of TN3k with Dice similarity coefficient (DSC) of 81.64% and 95th percentage of the asymmetric Hausdorff distance (HD95) of 14.06, whereas those on our private dataset were with DSC of 85.63% and HD95 of 14.53, respectively. Conclusions: This paper presents a method for thyroid ultrasound image segmentation, which achieves high accuracy and meets the clinical requirements. Code is available at https://github.com/ccjcv/BPAT-UNet.

1. Introduction

Ultrasound imaging has become the preferred technology for the diagnosis of thyroid nodules due to its outstanding advantages of no radiation, low cost, and real-time performance [1,2]. However, ultrasound images have low contrast and a large amount of noise, resulting in blurred edges and unexpectedly varied boundaries of thyroid nodules.

With the rapid development of computer medical technology, the image segmentation algorithm based on deep learning has gradually become the mainstream method in the field of medical image segmentation. Convolutional neural networks (CNNs) shine brightly with their strong processing scale invariance and the ability to model the inductive bias of images [3]. CNNs with encoder-

[*] Corresponding authors.
E-mail addresses: bihui@cczu.edu.cn (H. Bi), caichengjie666@163.com (C. Cai), 921173049@qq.com (J. Sun), jiangyb@czust.edu.cn (Y. Jiang), lugang@seu.edu.cn (G. Lu), shu.list@seu.edu.cn (H. Shu), nxy@njmu.edu.cn (X. Ni).

https://doi.org/10.1016/j.cmpb.2023.107614
0169-2607/© 2023 Elsevier B.V. All rights reserved.

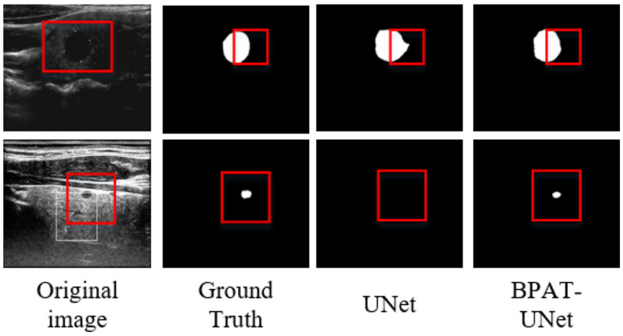


Fig. 1. Examples of the problems of ultrasound thyroid image segmentation, where the first row of images indicates the problem of inaccurate boundary, and the second row shows that small nodules are easily ignored. However, our BPAT-UNet addresses these two issues.The sample data derive from the TN3k dataset.

decoder structure show excellent segmentation performance, its encoder is used for feature extraction and is usually accompanied with multi-scale down-sampling, and the responsibility of the decoder is to restore the resolution of the image as much as possible [4,5].

Fully convolutional network (FCN) is the representative network of CNNs for image semantic segmentation [6]. Compared with FCN, UNet plays a greater role in the field of medical image processing where data are relatively scarce [7]. The key core is the skip connection that can transmit the low-level features to reserve local information. UNet-based networks have made significant progress in recent years. However, Fig. 1 shows that UNet architecture still suffers from two limitations in the task of thyroid ultrasound image segmentation.

(1) **Hard to deal with the disseminated boundary**. The first row of Fig. 1 shows a typical example of thyroid ultrasound image segmentation. UNet segmentation results fail to maintain accurate boundaries. The reason is that the gray value of the pixels on the edge of thyroid nodules is usually very close to the surrounding pixels, which can easily lead to errors in automatic segmentation. There will be a high probability that the border area will be missing if the segmentation methods cannot deal with fuzzy boundaries [8]. To solve the fuzzy boundary issue, researchers devoted themselves to proposing many methods to pay close attention to boundary information, such as manually adding parameters for post-processing [9], and adding boundary constraints into the network [10,11].

(2) **Lack of small object sensitivity**. The second row of Fig. 1 also shows why UNet segmentation results fail to recognize the relatively small thyroid nodules. The reason is that the UNet-based network uses convolution block for spatial representation that fails to model context dependencies and global information interaction [12]. In addition, traditional down-sampling structures can easily lead to smaller features being discarded [13].

To detect small targets better, multi-scale representation and contextual information have become the two pillars of mainstream use [14]. Multi-scale representation combines high- and low-level features to help demonstrate the small object in multiple resolutions [15,16]. Dilated convolution is proposed to enlarge the receptive field, however, the large dilated rates may ignore small objects [17]. Recently, the self-attention mechanism considering long-range dependencies was established [18,19].

1.1. Medical image segmentation based on CNN

In the early days of medical image segmentation, traditional medical image segmentation methods were based on thresholding, region, edge detection, clustering, and deformable models [21].

With the development of deep learning techniques, FCN is the mainstream approach for image segmentation [6]. UNet contains two concatenated paths of contracting and expanding. The contracting path is used to extract image features, capture contexts, and compress images into feature maps composed of features. The expanding path is adopted for precise localization, features detection, and prediction. In addition, the skip connections transmit low-frequency information from the lower-level layers to the higher-level layers [7]. As a result of the excellent performance of UNet for medical image segmentation, the corresponding variants continue to emerge. Res-UNet builds the powerful Resnet into a U-shaped network [22,23]. UNet++ chooses dense skip connections to capture fine details in the foreground [24].

To detect the small objects, Dense-UNet fuses small-scale features within and between slices in a densely connected manner to enhance context [25]. FactSeg improves the accuracy of small object segmentation by multi-scale fusion of two branches [26]. To obtain the boundaries of the tissues more precisely, some studies add shape constraints to the network. Lee explicitly introduces the boundary point detection module in CNNs, which uses multi-scale atrous convolution to generate boundary point prediction maps [11]. SAUNet also constructs a shaped flow and passes gradients to capture rich shape-related information [27].

1.2. Medical image segmentation based on CNNs combining self-attention

Although CNNs have achieved great success in the field of medical segmentation due to their strong local representation ability, they suffer from a lack of attention to global information based on feature maps.

Self-attention is better at capturing the internal correlation of features by calculating the interaction between image patches which solves the problem of long-distance dependence [28]. U-net transformer improves UNet by introducing self-attention and cross-attention [29]. AttnUNet proposes a novel Attention gate (AG) model for medical imaging that automatically learns to focus on target structures of varying shapes and sizes [18]. TransAttUNet introduces an adaptive weighted sum of self-attention and global spatial attention at the bottleneck of the U-shaped network [30]. UTNet introduces self-attention in UNet in the form of residual connections, which are beneficial to induce bias learning [31].

1.3. Medical image segmentation based on transformers

The transformer structure is first proposed for natural language processing [32]. In vision tasks, Vision transformer (ViT) is proposed for image classification [33]. Although replacing CNNs with Transformers for medical image segmentation requires further study, the transformer-based networks show impressive performance and accuracy in medical image processing [34,35]. As a serial scheme, Pyramid vision transformer v2 (PVT.v2) replaces block embeddings with overlapping convolutions as block embeddings [36]. This serial kind of combination cannot handle both high-frequency and low-frequency information at each layer simultaneously.

To solve this issue, global and local image transformer mixes convolutional and transformer features in a parallel manner [37]. To fully fuse the local and global features, the subsequent HiLo attention and inception mixer are proposed to fuse the information of the two frequencies to a certain extent [38,39].

For medical images, transformer-based methods combine long-range dependencies that helps multiple objects analysis and local information that help small objects. The MedT proposes a gated axial-attention model to extend the Local-Global strategy [40]. In particular, for ultrasound images, the nodules' shape is always irregular. Feature extraction based on local window attention that is similar to Swin transformer, and the convolution block leads to the neglect of local fine details [41].

In this paper, considering the unsolved problems and challenges of the above ultrasound image segmentation methods, we introduce more efficient feature mining based on Transformer, deformable convolution, and multi-pooling methods to handle irregular and relatively small thyroid nodes. In addition, shape constraints are considered to handle thyroid nodules with large shape variations. Although several improved methods already exist, developing an effective algorithm for thyroid clinical diagnosis and treatment remains a major challenge. To settle the aforementioned two issues in thyroid ultrasound segmentation simultaneously, we propose the Boundary preserving assembled transformer UNet (BPAT-UNet) in this paper. This design mainly focus on the blurring of thyroid image edges and the enhancement of localization of nodular regions according to global background features.

The main contributions of our work are as follows.

(1) The proposed Boundary point supervision module (BPSM), which preserves the thyroid boundary information, involves not only two kinds of local features, but also the relations between them.

(2) The adaptive multi-scale features fusion module (AMFFM) is adopted to enhance local features. The deformable convolutional and deformable attention blocks are alternately used can better capture the characteristics of the nodules.

(3) The Assembled transformer module (ATM) performs two self-attention routes to process global and local information for small nodules detection. The Extern attention (EA) block can learn the underlying correlations sample-among of the dataset.

(4) According to the quantitative results on the thyroid public dataset TN3k [20] and our private dataset, the proposed approach displays significant accuracy advantages compared with other state-of-the-art techniques.

The rest of the paper is organized as follows. In Section II, we describe the architecture of the proposed BPAT-UNet and its modules in detail for thyroid nodule segmentation. Section III presents our experimental results and compares the proposed method with other advanced segmentation methods. The effects of the key elements involved in our method are analyzed by a series of ablation studies. Conclusions and perspectives are drawn in Section IV.

2. Methods

In this section, Fig. 2 shows the overall architecture of our proposed BPAT-UNet, and we describe the three important modules of BPAT-UNet in the following subsection. (1) The purpose of the Boundary point supervision module (BPSM) is to maintain the bounding global receptive field of objects. (2) The Adaptive multi-scale feature fusion module (AMFFM) is designed to express attention to objects of smaller sizes and objects with large differences in shape. (3) The Assembled transformer module (ATM) is designed to fully integrate global and local information.

2.1. Overall structure design

The UNet framework mainly consists of an encoder, bottleneck, decoder, and skip connections, where both encoder and decoder are based on Double convolution blocks (DCB). The traditional UNet consists of multi-scale layers, where n indicates the n-th layer of the network, $n = 1, 2, 3, 4, 5$.

To effectively maintain the shape of the irregular nodule's boundary and small object features, we augment three important modules into the UNet framework. The BSPM is inserted into the $3rd$, $4th$ and $5th$ layers for boundary point feature supervision. The AMFFM is inserted into the $3rd$ and $4th$ layers for feature aggregation. The ATM is inserted into the bottleneck for samples-among adaptation.

For a given thyroid image $X \in R^{H\times W\times 3}$, the encoder performs multi-scale feature processing to obtain depth features. The output of the DCB on each layer is calculated by Eq. (1),

$$F_n^{DCB} \in R^{(H/2^{n-1})\times(H/2^{n-1})\times(2^{n-1})C} \tag{1}$$

where $C = 64$, which is halved by maximizing the pooling layer. Instead of down-sampling F_n^{DCB} directly, the convolution features F_n^{DCB} are fed to BPSM to generate the boundary-enhanced features F_n^{BPSM}, where $n = 3, 4, 5$. Subsequently, F_n^{BPSM} and F_{n+1}^{BPSM} are fed into AMFFM to fuse the multi-scale features F_n^{AMFFM}, where $n = 3, 4$. Besides, we augment a boundary points prediction based on F_n^{BPSM}, where $n = 3, 4, 5$. The deepest boundary-enhanced features F_n^{BPSM} are transferred into ATM to obtain a full mixture of global features and local features F_n^{ATM}, where $n = 5$.

2.2. Boundary points supervision module (BPSM)

The BSPM is proposed to supervise boundary point features with key points of the ground truth. Fig. 2 shows the location of BPSM in our entire model and Fig. 3 shows the structure of BPSM in detail.

It takes features F_n^{DCB} as input, where $F_n^{DCB} \in R^{w_n\times h_n\times c_n}$ $(n = 3, 4, 5)$, where w_n, h_n and c_n represents the length, width and channel number of BPSM input feature in the n-th layer, respectively. F_n^{BPSM} $(n = 3, 4, 5)$ is used to enhance features and generate boundary constrain features.

To extract more effective features for representing thyroid boundaries, we propose Stripe pooling self-attention (SPSA), including stripe pooling and self-attention calculation [36,42].

Notably, we redesign SPSA that adds stripe pooling before the calculation of key (K), value (V), and query (Q). The SPSA can be calculated by Eq. (2) and Eq. (3),

$$SPSA(Q, K, V) = Concat(head_0, \cdots, head_N)W^O \tag{2}$$

$$head_j = Attention(QW_j^Q, SP(K)W_j^K, SP(V)W_j^V) \tag{3}$$

where $W^O \in R^{C\times C}$, $W_j^Q \in R^{C\times d_{head}}$, $W_j^K \in R^{C\times d_{head}}$, and $W_j^V \in R^{C\times d_{head}}$ are the parameters of linear projection, respectively. N is the number of heads. is the operation of stripe pooling, which can be expressed by Eq. (4):

$$SP(\mathbf{x}) = Norm(Concat(reshape_1(\mathbf{x})\dot{W}^s, reshape_2(\mathbf{x})\dot{W}^s)) \tag{4}$$

Among them, given input feature $\mathbf{x} \in R^{H\times W\times C}$, pooling $\mathbf{x}$ in horizontally way to achieve, where $\mathbf{x}_h = reshape_1(\mathbf{x}$, where $\mathbf{x}_h \in R^{W\times 1\times C}$ and in vertically way to achieve $x_v = reshape_1(\mathbf{x})$, where $\mathbf{x}_v \in R^{H\times 1\times C}$. W^s is the parameter of linear projection. *Concat* means concatenating features channel-wise.

The redesigned SPSA is calculated by Eq. (5)

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_{head}}}\right)\dot{V} \tag{5}$$

In addition, the Pyramid pooling self-attention (PPSA) is introduced by considering variable receptive fields. Following a 1 × 1 convolution, the boundary key point feature can be generated [43]. PPSA is similar to the proposed SPSA. Instead of the reshape operation, PPSA pooling deals with the features in multiple scales directly.

Subsequently, SPSA and PPSA features are used for multiplication, and addition with the residual connection. With such two pooling self-attention used simultaneously, we can capture the

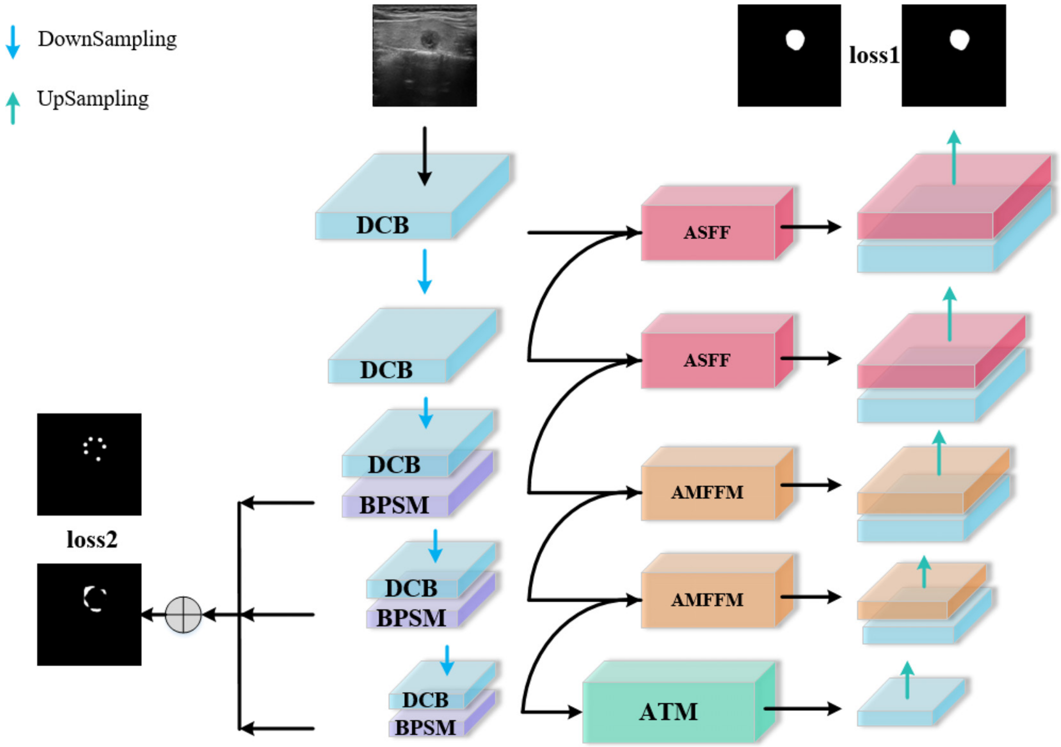


Fig. 2. Overall architecture of our proposed BPAT-UNet.

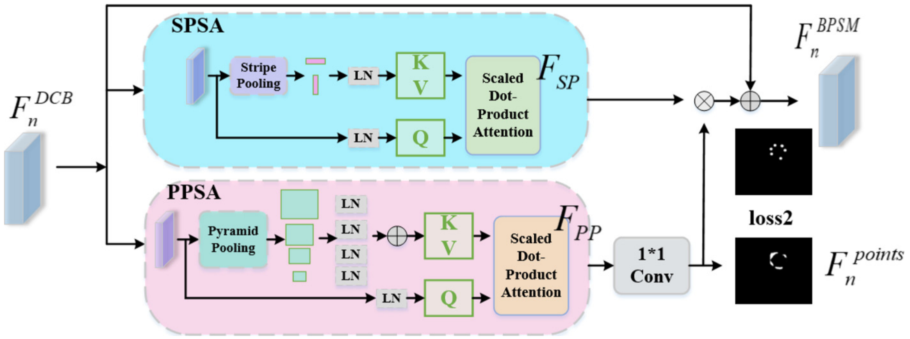


Fig. 3. Structure of the proposed Boundary Points Supervision Module in detail.

long-distance relationship and multi-scale features of isolated regions. The overall process can be formulated by Eq. (6):

$$F_n^{BPSM} = F \oplus (F_{SP} \otimes (conv_{1\times1})(F_{PP})) \tag{6}$$

where $\oplus$ means feature addition, $\otimes$ means matrix multiplication, and $conv_{1\times1}$ stands for convolution. F, F_{SP}, and F_{PP} represent the original input feature, SPSA feature, and PPSA feature, respectively.

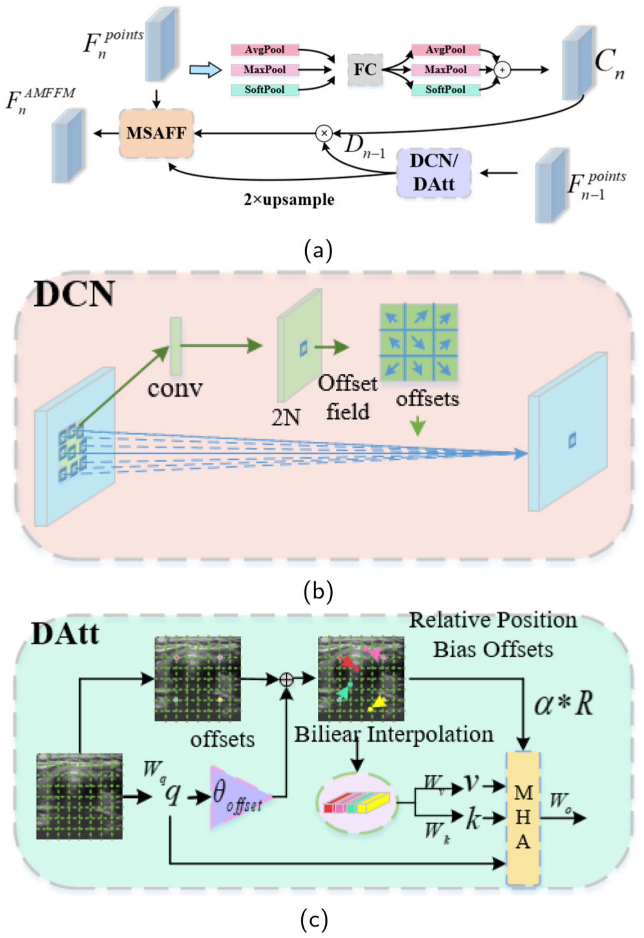
To better introduce boundary prior and improve the segmentation performance, we propose a boundary point supervision to accurately represent the thyroid contour's shape. The details of the boundary key points selection algorithm are described in as Table 1 shown. We design a boundary point generator to generate relatively ideal boundary key points. We remove the redundant points of the boundary contour to reduce the amount of computation for subsequent processing.

Table 1
Boundary key point selection algorithm.

```
**Algorithm1**: Boundary key point selection algorithm
Input: Ground truth segmentation map
Output:Boundary key points
$M$: boundary points by Canny
$P_m^D \leftarrow (x_1^D, y_1^D), (x_2^D, y_2^D), \cdots, (x_M^D, y_M^D)$
$N$: boundary key points by DP approximation algorithm from
$P_n^D \leftarrow (x_1^D, y_1^D), (x_2^D, y_2^D), \cdots, (x_N^D, y_N^D)$
for $i = 1, 2, 3, \ldots, N-1$ do
  if $ED(P_{i+1}^D, P_i^D) < 5$ then
    delete $P_i^D$
  end
end
Return $\overline{P}$
```

First, we choose the traditional edge detection Canny operator to generate the set of boundary points. Second, we use the DP approximation algorithm to extract the boundary key points considering that, concave parts are prone to appear in the boundary area, leading to the large difference in the thyroid nodules shape. By using the DP approximation algorithm, we achieve an accurate sampling of the edge points of the labeled images. However, some key points generated by the DP approximation algorithm still almost overlap, which may cause the model to focus too much on the relevant parts and ignore other parts. To solve this issue, we remove points that are too close to each other by setting a distance threshold, removing one of the pairs of points whose Euclidean distance (ED) is less than 5.

Third, we draw circles for the selected key points and set the circle radius of the edge points to 10. Fig. 3 shows the point ground



(a)

(b)

(c)

Fig. 4. Adaptive Multi-scale Feature Fusion Module. (a) The overall structure of the Adapted Multi-scale Feature Fusion Module. (b) Details of Deformable Convolutions. (c) Details of Our Deformable Attention.

truth results. It can be observed that the distance between the boundary joint points is suitable.

Furthermore, considering the amount of computation, the boundary point supervision module is only placed in the 3*rd*, 4*th* and 5*th* layers of the encoder.

2.3. Adaptive multi-scale feature fusion module (AMFFM)

AMFFM is proposed to capture smaller features placed at the 3*rd* and 4*th* top-to-bottom skip connections [24]. Fig. 2 shows the location of AMFFM in our entire model and Fig. 4 shows the structure of AMFFM.

AMFFM takes features from two adjacent layers in the encoder for fusion to excavate the relationship between channel dimensions [44]. Fig. 4(a) shows that deformable convolution and deformable attention are used to process the lower-scale features, whhereas the improved channel attention is used to process the features of the shallower layer. By introducing deformable convolution and deformable attention in different layers, it is suitable to objects with different shapes [45,46]. Fig. 4(b) and Fig. 4(c) show the structures of deformable convolution and deformable attention, respectively.

To prevent excessive computation and strengthen the long-range dependencies between deep semantic information, the deformable convolution is used in the shallow 3*rd* layer. Given the flexibility position of the deformable convolution, more accurate features of the thyroid are achieved. The deformable attention is located at the deep 4*th* layer. We can obtain the long-term dependencies features with low computational costs.

Considering the offset in deformable attention, an ideal value is often difficult to obtain by setting a constant predefined factor to prevent large offsets. Therefore, we propose the use of learnable parameters to control the offset value, which can achieve a more suitable and stable effect than the predefined offset method.

To highlight relatively important channels in features, we use three different types of pooling, namely, average pooling, max pooling, and soft pooling to enhance channel dependencies and build channel attention [47].

This channel attention process is represented by Eq. (7),

$$\begin{aligned} F_n^{poolCombine} = {} & sig[Relu(FC(AvgPool(F_n^{Points}))) \\ & + Relu(FC(MaxPool(F_n^{Points}))) \\ & + Relu(FC(SoftPool(F_n^{Points})))] \end{aligned} \tag{7}$$

where $n = 3, 4$. The FC represents a fully connected layer, sig means sigmoid function, $AvgPool$, $MaxPool$, and $SoftPool$ represent average pooling, max pooling, and soft pooling, respectively. Average pooling preserves the background information, max pooling preserves texture characteristics, and soft pooling maintains expressive features and is differentiable. Finally, we embed channel features into local/global variables to obtain refined features through multiplication and add original features in the form of residual connections.

In addition, considering that the information contained in different scale features is inconsistent, we adopt the Adaptive spatial feature fusion (ASFF), which adaptively fuses deep deformable features and shallow features. The multi-scale adaptive fusion method uses features of different scales as fusion weights, which effectively solves the problem that deep and shallow features that only contain rich semantic or location information are difficult to effectively fuse.

This process can be denoted by Eq. (8),

$$\begin{aligned} F_n^{AMFFM} = {} & F_n^{Points} \oplus (P_n^{poolCombine} \odot D_{n-1}) \\ & \hat{\oplus} upsample(D_{n-1})) \\ D_{n-1} = {} & \begin{cases} D_{n-1}^{conv} & n = 4 \\ D_{n-1}^{att} & n = 5 \end{cases} \end{aligned} \tag{8}$$

where $\odot$ stands for element-level multiplication, $\hat{\oplus}$ represents ASFF.

2.4. Assembled transformer module (ATM)

ATM is proposed for local and global feature fusion and it captures potential relationships between different samples. Fig. 2 shows the location of the ATM in our entire model and Fig. 5 shows the structure of the ATM.

The core purpose of two-path attention is to focus on both global and local information [38,48]. The module consists of two parts, namely, two routes self-attention and external attention. We pay more attention to the full fusion of global and local information while maintaining global dependency attention, including not only high-frequency information but also local details and low-frequency information.

For ultrasound thyroid segmentation, the two paths are redesigned to process global and local features in parallel. Features are fused according to channel partitioning are transferred to high-frequency and low-frequency attention modules. Local features are obtained by windowed attention and convolution, whereas global features are obtained by standard self-attention. This behavior effectively learns comprehensive features that contain high-frequency and low-frequency information in visual data.

Furthermore, we replace the Multi-layer perception (MLP) in transformer with External Attention, which can explore the interrelationships between different data samples. By replacing the

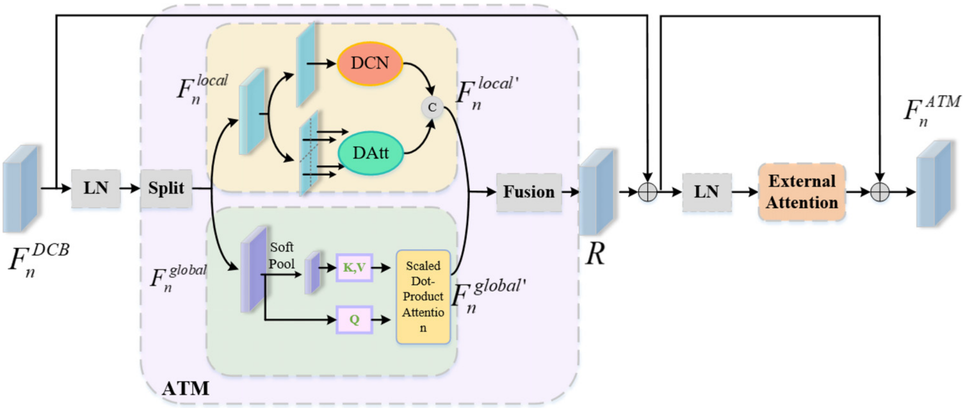


Fig. 5. Details of Assembled Transformer Module.

original transformer, better performance and lower computational complexity can be achieved.

Two Routes Self-attention: Given the input feature map $F_n^{DCB} \in R^{N\times\frac{C_5}{2}}$, $n = 5$, it splits into two parts with half channel as $F_n^{local} \in R^{N\times\frac{C_5}{2}}$, $F_n^{global} \in R^{N\times\frac{C_5}{2}}$, respectively.

They are calculated in two different ways. First, we propose a parallel architecture that mixes the deformable convolution and the windowed deformable attention to learn high-frequency and local components in detail. The purpose of introducing deformable feature calculation is to reduce the recognition problem of objects with large differences in shape.

Windowed deformable attention computes deformable attention by dividing the features into four equal-sized windows, each of which is $w_i \in R^{\frac{N}{4}\times\frac{C_5}{2}}$, $i = 1, 2, 3, 4$, and finally stitches the four features back together.

This process can be formulated by Eq. 9,

$$F_n^{local'} = D^{cn}(F_n^{local}) \oplus WD^{att}(F_n^{local}) \tag{9}$$

where $n = 5$, D^{cn} means deformable convolution and WD^{att} refers to window deformable attention.

In addition, for the learning of low-frequency components, K/V is calculated by feature pooling based on linear spatial-reduction attention [36].

For thyroid segmentation, we change the pooling method to soft-pool to achieve $\frac{K}{V}$ for greater strength of feature activation. Finally, local and global features are fused channel-wise.

Two Routes Self-attention can be formulated by Eq. 10,

$$R = Concat(F_n^{local'}, LSRA(F_n^{global})) \qquad (n = 5) \tag{10}$$

External Attention: Considering that External Attention can solve the problem of self-attention by ignoring potential correlations between samples, the second half of the module adds External Attention similar to the MLP structure.

2.5. Loss function

The Soft Dice loss is used to train our BPAT-UNet, which minimizes not only the difference between segmentation prediction (S_{pred}) and label image (S_{GT}), but also boundary key points prediction (S_{pred_point}) and boundary point labels (S_{GT}).

$$L_{total} = L_{seg} + L_{point} \tag{11}$$

$$L_{seg} = \varphi_{dice}(S_{pred}, S_{GT}) \tag{12}$$

$$L_{point} = \varphi_{dice}(S_{pred\_point}, S_{GT}) \tag{13}$$

where φ_{dice} represents the dice loss function.

3. Results

3.1. Datasets

To evaluate our model, we use two data sets: the public TN3k dataset and our own private thyroid dataset.

(1) **TN3k dataset**: The TN3k dataset contains 3493 ultrasound images collected from 2421 patients, all of which are grayscale. We chose 2303 images as the training set, leaving 576 images as the validation set, and an additional 614 images for testing the best model. They are resized to 256 × 256 and we performed data expansion such as normalization and random inversion.

(2) **Our own thyroid dataset**: Our own thyroid dataset consists of 328 ultrasound images collected at The Affiliated Changzhou No. 2 Peoples Hospital of Nanjing Medical University (No.2020_KY146-01). They were acquired by three commercial scanners from Philips Healthcare/ Best/ Netherlands, Siemens Healthineers/ Erlangen/ Germany, GE Healthcare/ Chicago/ USA.

We divide the datasets in a ratio of 8:1:1 for training, validation and testing. The delineation of the nodules was performed by three physicians with extensive clinical experience in the ultrasound department. We convert contours to contours and binary masks as ground truths. To remove patient privacy and other irrelevant information, we crop the images and labels to 512 × 512. They are also adjusted to 256 × 256 for normalization and random inversion.

3.2. Implementation details

The network BPAT-UNet was built with PyTorch 1.7.0, using the Adam optimizer with a weight decay of $1e-4$. The initial learning rate is set to $1e-4$, and the learning rate is decayed by a warm-up cosine annealing algorithm. All experiments are performed on Tesla V100 32G with batch size set to 16 and a maximum of 150 epochs. We use the weights of the TN3k dataset for transfer learning by considering the small number of private thyroid datasets to avoid over-fitting.

3.3. Qualitative evaluation

First, we evaluated the proposed method from the perspective of visual quality. We compared the visual quality between the proposed BPAT-UNet and EIGHT other advanced methods in the image semantic segmentation field, namely, UNet [7], FCN [6], Deeplabv3+Resnet50 [49], AttnUNet [18], SmaAt-UNet [50], UT-Net [31], UNet Transformer [29], and TransUNet [12]. UNet, FCN, and Deeplabv3+Resnet50 methods are based on traditional CNNs structure. Both AttnUNet and SmaAt-UNet are a combination of convo-

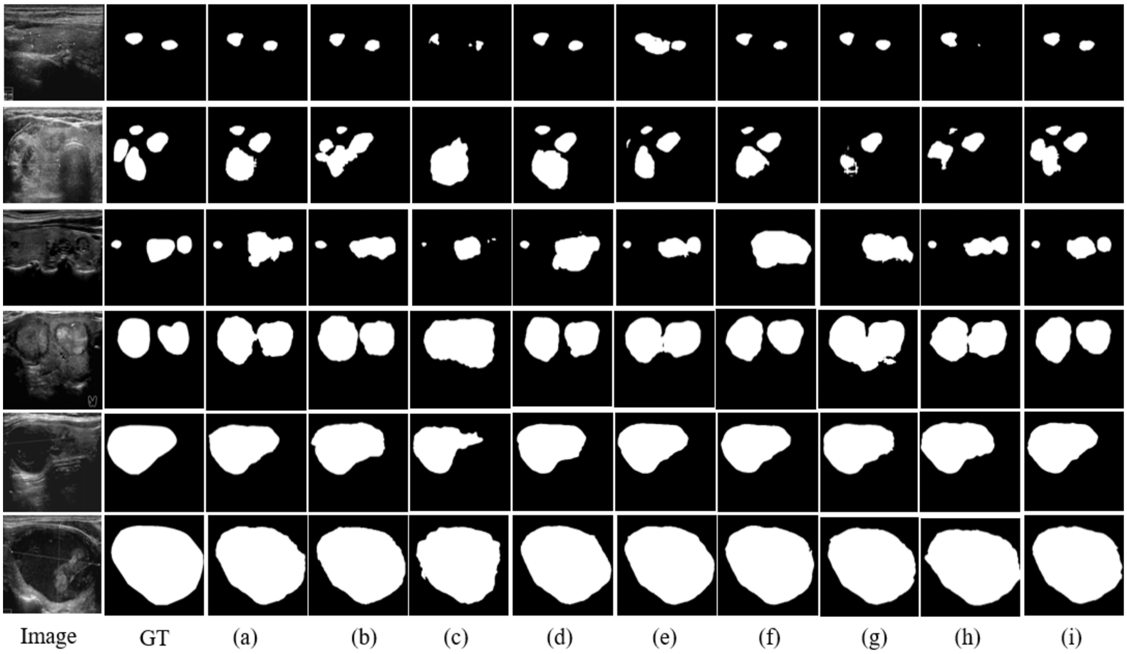


Fig. 6. Segmentation results of different existing ways and our method on the TN3k dataset. (a) UNet. (b) FCN (c) Deeplabv3-Resnet50. (d) AttnUNet. (e) SmaAt-UNet. (f) UT-Net. (g) UNet Transformer. (h)TransUNet (i). BPAT-UNet.

lution and attention mechanisms. UT-Net, UNet Transformer, and TransUNet are based on the Transformer structure. For these networks, different pre-training parameters are adopted. FCN uses the Imagenet pre-trained VGG backbone, deeplabv3+ uses the pre-trained Resnet50 backbone, and the rest of the network is trained from scratch without using any pre-trained weights.

3.3.1. Results on the TN3k dataset

Firstly, we evaluated these methods on the public TN3k dataset. Fig. 6 presents the segmentation results on five example images of the testing dataset obtained by the proposed BPAT-UNet and the EIGHT other methods aforementioned. The segmentation results provided by these methods are displayed on the third to the tenth columns, respectively. The first to sixth lines in Fig. 6 present the segmentation results of six patients. Obviously, the proposed BPAT-UNet is the most effective to detect thyroid nodules from ultrasound images visually. Our BPAT-UNet has a superior perception ability for small thyroid nodule objects compared with other models, from the segmentation effect of small objects in the first three rows of Fig. 6. Thus that our AMFFM compensates for the smaller features that are easily lost in downsampling. Alternatively, the proposed network can obtain this information by paying extra attention to shape differences through deformable computation, which can compare the differences between different models from the second and third rows in Fig. 6. For large thyroid objects(rows 4, 5, and 6 in Fig. 6), BPAT UNet also performs well, which is attributed to our BPSM maintaining the boundary area of the object through a point approach. It should be emphasized that for the diagnosis and treatment of thyroid nodules, the boundary area must be grasped, and BPAT UNet has rich experience in this area.

3.3.2. Results on the private dataset

Second, we evaluated these methods on our own private thyroid dataset. Fig. 7 presents the segmentation results on three sample images of the testing dataset obtained by the proposed BPAT-UNet and the EIGHT methods aforementioned. The segmentation results provided by these methods are displayed on the third to the tenth columns, respectively. The first to fifth rows of Fig. 7 present the segmentation results of five patients. Also, the proposed BPAT-UNet is the most effective in detecting thyroid nodules from ultrasound images visually.

Compare with other methods, it is found that the proposed BPAT-UNet can identify non-nodule regions from the thyroid ultrasound images. Similar to the segmentation performance of TN3k dataset, the segmentation prediction results show that for dense and small-scale objects, our BPAT-UNet performs significantly better than other models for thyroid ultrasound image segmentation. In addition, BPAT-UNet achieves the best performance in maintaining the shape of segmented objects compared to other methods, making more accurate judgments on the boundaries of thyroid nodule regions.

3.4. Quantitative evaluation

Furthermore, we evaluated the proposed method quantitatively. Dice Similarity Coefficient (DSC), Intersection Over Union (IoU), 95th percentage of the asymmetric Hausdorff distance (HD95), F1-Score, Accuracy, AUC, Recall, and Precision are used for evaluation. AUC stands for Area Under the ROC Curve. We also compare the proposed BPAT-UNet and EIGHT aforementioned methods.

They can be calculated as follows:

$$DSC = \frac{2 \cdot TP}{FP + 2 \cdot TP + FN} \tag{14}$$

$$IoU = \frac{TP}{FP + TP + FN} \tag{15}$$

$$HD95 = \max_{k95\%}[d(X,Y), d(Y,X)] \tag{16}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

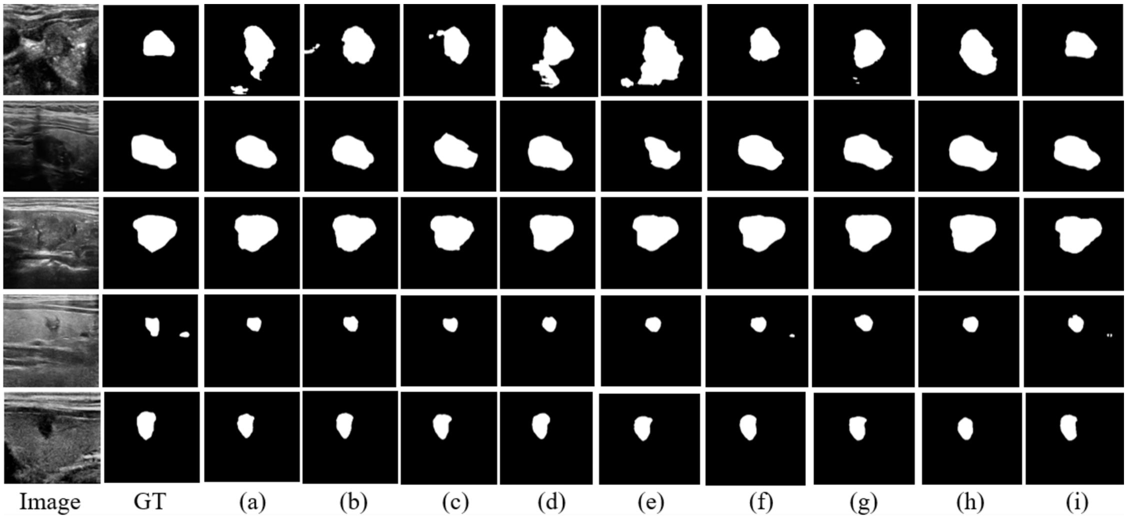


Fig. 7. Segmentation results of different existing ways and our method on our own dataset. (a) UNet. (b) FCN (c) Deeplabv3-Resnet50. (d) AttnUNet. (e) SmaAt-UNet. (f) UT-Net. (g) UNet Transformer. (h) TransUNet. (i) BPAT-UNet.

Table 2
Performance summury on the public TN3k dataset.

Model	F1(%)	Accuracy(%)	IoU(%)	DSC(%)	HD95	AUC(%)	Recall(%)	Precision(%)
Deeplab-Resnet50[49]	72.78	95.36	55.76	71.60	27.58	84.21	70.33	75.40
UNet[7]	81.62	96.62	67.77	80.79	21.51	91.59	85.07	78.44
FCN[6]	81.95	96.86	68.27	81.14	18.58	90.78	83.20	80.74
AttnUNet[18]	81.96	96.81	68.42	81.25	19.05	89.84	81.11	82.83
UNet Transformer[29]	81.26	96.63	67.79	80.80	19.30	90.35	82.42	80.14
UT-Net[31]	82.13	96.67	68.84	81.54	16.83	91.01	83.79	80.55
TransUNet[12]	81.34	96.59	68.04	80.98	18.81	90.26	82.35	80.35
SmaAt-UNet[50]	84.03	97.20	71.46	83.36	15.58	91.53	84.42	**83.65**
BPAT-UNet (ours)*	**84.23**	**97.22**	**71.87**	**83.64**	**14.06**	**92.03**	**85.57**	82.94

* The best results are highlighted in bold.

$$F1 - Score = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} \tag{20}$$

where TP, FP, TN, and FN mean true positive, false positive, true negative, and false negative, respectively. d represents the calculation of one-way Hausdorff distance of two sets.X and Y stand for prediction and ground truth, respectively.

3.4.1. Results on TN3k dataset

Table 2 and Fig. 8 present the experimental results in two different forms on the TN3K dataset. Quantitative results show that our proposed BPAT-UNet achieves superior segmentation performance on evaluation metrics compared with other schemes, scoring 71.87% in IoU, 83.64% in DSC, 14.06 in HD95, 84.23% in F1-Score, 97.22% in Accuracy, 92.03% in AUC, 85.57% in Recall and 82.94% in Precision. Compared to baseline method UNet, BPAT-UNet yields 4.1%, 2.85%, 7.45, 2.61%, 0.6%, 0.44%, 0.5%, and 4.5% improvement in IoU, DSC, HD95, F1-Score, Accuracy, AUC, Recall and Precision.

As a result of the numerous indicators, we mainly compare multiple methods for representative indicators such as DSC and HD95. For DSC, the proposed BPAT-UNet reaches the score of 83.64% which is an improvement of 2.85% compared with the score obtained by UNet (80.79%), which is an improvement of 12.04% compared with Deeplabv3-Resnet50 (71.60%), which is an improvement of 2.5% compared with FCN (81.14%), which is an improvement of 2.39% compared with AttnUnet (81.25%), which is an improvement of 2.84% compared with UNet Transformer (80.80%), which is an improvement of 2.1% compared with UT-Net (81.54%), which is an improvement of 2.66% compared with TransUnet (80.98%), which is an improvement of 0.28% compared with SmaAt-UNet(83.36%).

For HD95, the proposed BPAT-UNet reaches the score of 83.64% which is an improvement of 7.45 compared with the score obtained by UNet (21.51), which is an improvement of 13.52 compared with Deeplabv3-Resnet50 (27.58), which is an improvement of 4.52 compared with FCN (18.58), which is an improvement of 4.99 compared with AttnUnet (19.05), which is an improvement of 5.24 compared with UNet Transformer (19.30), which is an improvement of 2.77 compared with UT-Net (16.83), which is an improvement of 4.75 compared with TransUnet (18.81), which is an improvement of 1.52 compared with SmaAt-UNet (15.58).

From the DCS and HD95 results, we find that UNet achieves the ideal segmentation effect on the TN3k dataset as it integrates the features from the underlying space through a multi-scale structure and skip connection. FCN slightly outperforms UNet based on the pre-training weights. Although DeepLab V3+ has a well-designed atrous spatial pyramid pooling and decoder structure, it is not ideal for the thyroid TN3k dataset. In the scheme of adding the attention mechanism to convolution, AttnUNet achieves a high DSC score. By adding channel attention to convolution, SmaAt-UNet achieves the second-highest DSC score.

With the help of the introduction of the Transformer into the CNN, the problem that the convolutional receptive field has limited effect and is difficult to capture the long-range context is solved. UT-Net introduces residual self-attention to form a residual Transformer that slightly improves the DSC score. UNet Transformer adds self-attention and cross-attention to bottleneck and skip connection, respectively, increasing the global receptive field, and has achieved minimal performance improvement. TransUNet combines

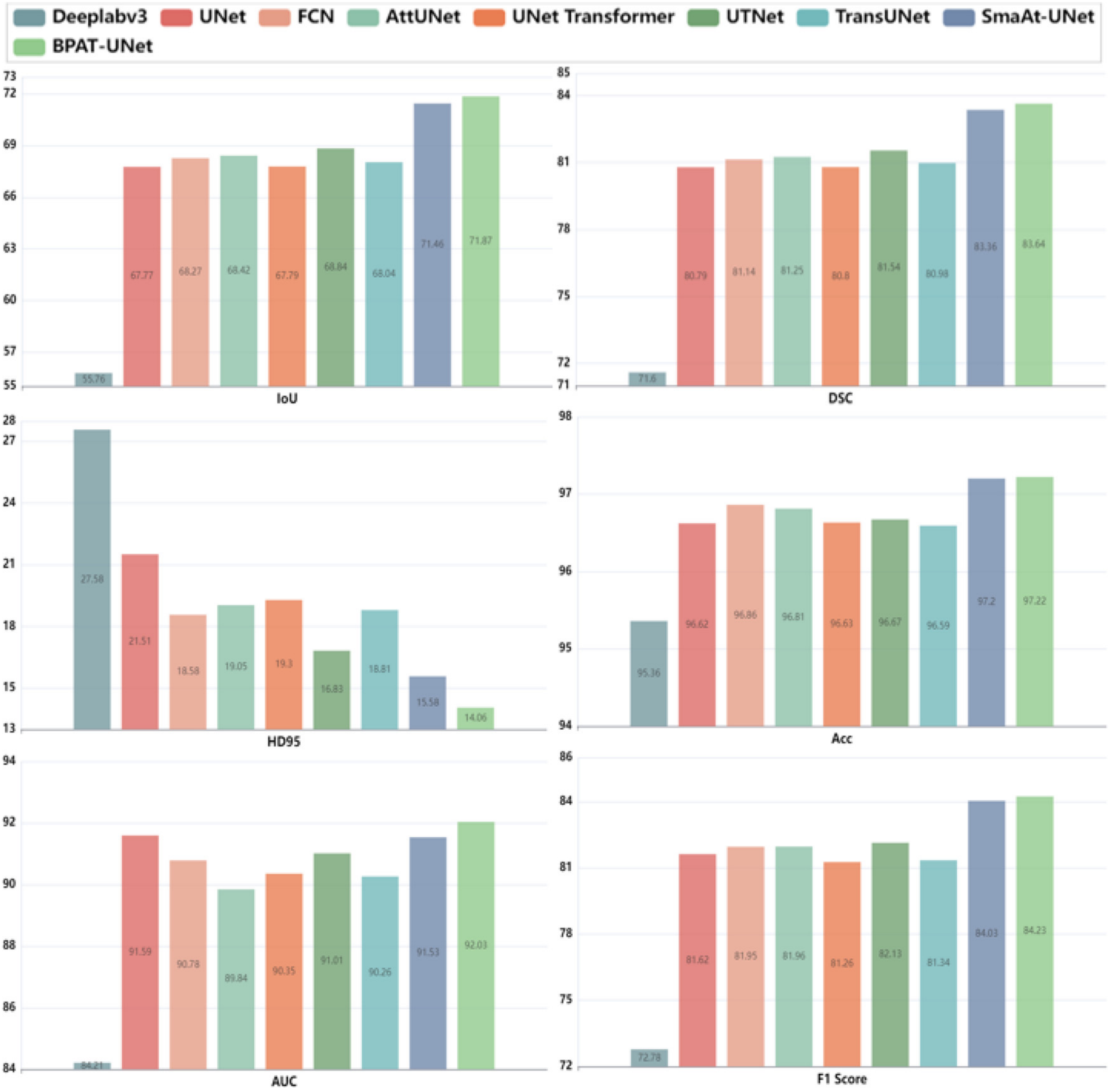


Fig. 8. Visualization of evaluation metrics of the compared methods and the proposed BPAT-UNet on the public TN3k dataset.

the convolutional layer and the Transformer layer in turn achieves good results. Similarly, our BPAT-UNet inherits Transformer's excellent global modeling capabilities and has excellent DSC metric results on the TN3k dataset. It is noted that our BPAT-UNet does not simply insert Transformers or self-attention like the Transformer network mentioned above. We focus on both global and local feature information, which may be the fundamental reason for obtaining superior DSC metrics. Considering the correlation between HD95 indicators and the shape and boundary of thyroid nodules, we compare them through HD95, which is the focus of this article. Compared with the existing segmentation methods, the proposed BPAT-UNet performs more prominently on HD95. This is due to the boundary point monitoring module we designed and the introduction of deformable computing operations. These features assist the model to remain comfortable in the face of thyroid with large shape differences and fuzzy boundaries, which is what other models lack.

3.4.2. Results on the private dataset

We also conduct experiments on a private thyroid dataset to further illustrate the effectiveness of our proposed method.

Given the small number of private datasets, overfitting may occur when training the network, so we use the pre-trained weights of the TN3k dataset for transfer learning.

Table 3 and Fig. 9 show that our proposed BPAT-UNet achieves the best segmentation performance on DSC compared with other schemes. Similar to the TN3k dataset, we also focus on analyzing DSC and HD95. For DSC aspect, the proposed BPAT-UNet reaches the score of 85.63% which is an improvement of 4.91% compared with the score obtained by UNet (80.72%), which is an improvement of 3.07% compared with FCN (82.56%), which is an improvement of 9.63% compared with Deeplabv3+Resnet50 (76.00%), which is an improvement of 2.64% compared with AttnUNet (82.99%), which is an improvement of 2.82% compared with SmaAt-UNet (82.81%), which is an improvement of 1.97% compared with UT-Net (83.66%), which is an improvement of 2.24% compared with UNet Transformer (83.39%), which is an improvement of 4.99% compared with TransUNet (80.64%).

For HD95, The proposed BPAT-UNet reaches the score of 14.53, which is an improvement of 12.81 compared with the score obtained by UNet (27.34), which is an improvement of 3.75 compared with FCN (18.28), which is an improvement of 10.21 compared with Deeplabv3+Resnet50 (24.74), which is an improvement of 3.79 compared with AttnUNet (18.32), which is an improvement

Table 3
Performance summary on the private thyroid ultrasound dataset.

Model	F1(%)	Accuracy(%)	IoU(%)	DSC(%)	HD95	AUC(%)	Recall(%)	Precision(%)
Deeplab-Resnet50[49]	76.20	96.22	61.30	76.00	24.74	88.92	80.57	72.28
UNet[7]	80.96	96.85	67.68	80.72	27.34	89.78	80.91	81.02
FCN[6]	81.80	97.52	70.30	82.56	18.28	88.58	77.90	86.11
AttnUNet[18]	82.24	97.06	70.92	82.99	18.32	89.82	80.74	83.79
UNet Transformer[29]	82.44	97.67	71.51	83.39	16.78	90.14	81.48	83.42
UT-Net[31]	83.97	97.00	71.91	83.66	15.35	91.57	85.00	82.96
TransUNet[12]	81.27	96.54	67.55	80.64	19.90	90.19	82.61	79.98
SmaAt-UNet[50]	82.61	97.14	70.67	8281	19.14	89.83	81.06	84.23
BPAT-UNet (ours)*	**85.81**	**97.80**	**74.87**	**85.63**	**14.53**	**92.19**	**85.33**	**86.30**

* The best results are highlighted in bold.

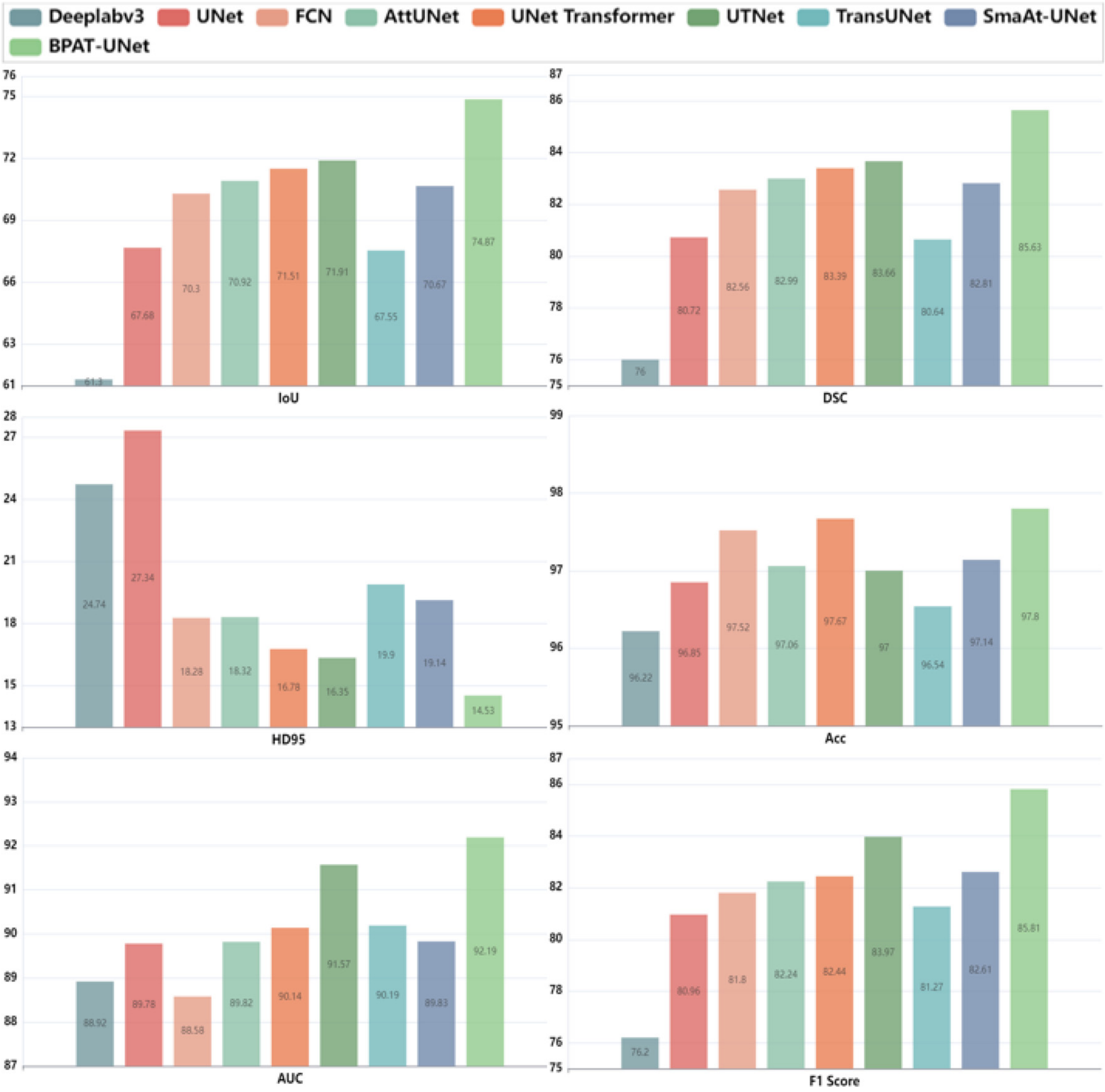


Fig. 9. Visualization of evaluation metrics of the compared methods and the proposed BPAT-UNet on the private thyroid ultrasound dataset.

of 4.61 compared with SmaAt-UNet (19.14), which is an improvement of 0.82 compared with UT-Net (15.35), which is an improvement of 2.25 compared with UNet Transformer (16.78), which is an improvement of 5.37 compared with TransUNet (19.90).

Our proposed BPAT-UNet ranks first in both DSC and HD95 metrics among various segmentation methods. By observing the segmentation performance of the aforementioned models on the two datasets, we find that the results are basically consistent, that is, models that perform well on TN3k also have ideal effects on our own thyroid datasets, and vice versa. The reason is that the image distribution of the two datasets is nearly similar so the segmentation results will not differ. As with the previous results based on public datasets, models with attention mechanisms on private datasets and models with a mixture of Transformer and CNN outperform convolutional models in metrics.

Notably, we also find that the experimental results on our thyroid dataset were generally better than those on the TN3k dataset because transfer learning can help the model converge better, and

Table 4
Ablation experiment of the proposed modules on the public TN3k dataset.

Model	F1(%)	Accuracy(%)	IoU(%)	DSC(%)	HD95	AUC(%)	Recall(%)	Precision(%)
UNet	81.62	96.62	67.77	80.79	21.51	91.59	85.07	78.44
UNet + BPSM (Point)	83.14	96.97	70.44	82.66	15.45	93.50	85.87	80.58
UNet + AMFFM	82.47	96.91	69.22	81.81	16.26	90.85	83.17	81.79
UNet + ATM	83.10	96.98	70.68	82.82	16.55	91.56	84.79	81.48
UNet + BPSM (Point) + AMFFM + ATM(ours)*	**84.23**	**97.22**	**71.87**	**83.64**	**14.06**	**92.03**	**85.57**	**82.94**

* The best results are highlighted in bold.

Table 5
Ablation experiment of the proposed modules on the private thyroid dataset.

Model	F1(%)	Accuracy(%)	IoU(%)	DSC(%)	HD95	AUC(%)	Recall(%)	Precision(%)
UNet	80.96	96.85	67.68	80.72	27.34	89.78	80.91	81.02
UNet + BPSM (Point)	83.30	97.40	70.47	82.68	18.99	90.70	82.74	83.87
UNet + AMFFM	82.84	97.45	70.29	82.56	18.61	90.24	81.69	84.03
UNet + ATM	82.40	97.60	71.15	83.15	18.64	89.74	80.48	84.41
UNet + BPSM (Point) + AMFFM + ATM(ours)*	**85.81**	**97.80**	**74.87**	**85.63**	**14.53**	**92.19**	**85.33**	**86.30**

* The best results are highlighted in bold.

Table 6
Performance on different BPSM locations and point types on the public TN3k dataset.

Model	F1(%)	IoU(%)	DSC(%)	HD95
UNet	81.62	67.77	80.79	21.51
UNet+BPSM (2)[1] (Point)	81.79	68.31	81.17	17.89
UNet+BPSM (5) (Point)	82.42	68.95	81.62	17.65
UNet+BPSM (3+4+5) (Only Canny)	82.41	69.31	81.87	17.92
UNet+BPSM (3+4+5) (Point)(ours)*	**83.14**	**70.44**	**82.66**	**15.45**

[1] The numbers in parentheses represent the number of layers corresponding to the encoder. * The best results are highlighted in bold.

the quality of our dataset is relatively high. The proposed BPAT-UNet shows its ability for small region features description. Meanwhile, our BPAT-UNet also shows significantly better ability in shape boundary preservation compared with other models. These two important performance improvements can be reflected in DSC and HD95.

3.5. Ablation study

To validate the role of each module in the proposed network, we use UNet as the baseline network and perform ablation studies on the TN3k dataset and our own thyroid dataset.

Table 4 and Table 5 show the different network structures by adding the proposed BPSM, AMFFM, and ATM to UNet, respectively. The segmentation performance on the TN3k dataset is significantly improved, such that DSC increases by 1.87%, 1.02%, and 0.82%, whereas HD95 decreases by 6.06, 5.25, and 4.96. Compare with UNet, our strategy improves DSC with 2.85% and HD95 with 7.45. The segmentation performance on our private dataset also improves such that DSC increases by 1.96%, 1.84%, and 2.43%, whereas HD95 decreases by 8.35, 8.73, and 8.7. Compare with UNet, our strategy improves DSC with 4.91% and HD95 with 12.81. The above improvements are sufficient to demonstrate the effectiveness of the modules we insert on the ultrasound thyroid dataset, and each module will be described in detail later.

3.5.1. Effect of boundary points supervision module

To validate the role of BPSM, we place BPSM in different layers for the ablation experiment. Table 6 and Fig. 10 show the various results on the TN3k dataset. We used three strategies for comparison. We place the BPSM module in the 2nd layer, 5th layer of the network, and key points selection without the DP algorithm. The performance is significantly improved, such that DSC increases by 0.38%, 0.83%, and 1.08%, whereas HD95 decreases by 3.62, 3.86, and 3.59. Our strategy improves DSC by 1.87% and HD95 by 6.06.

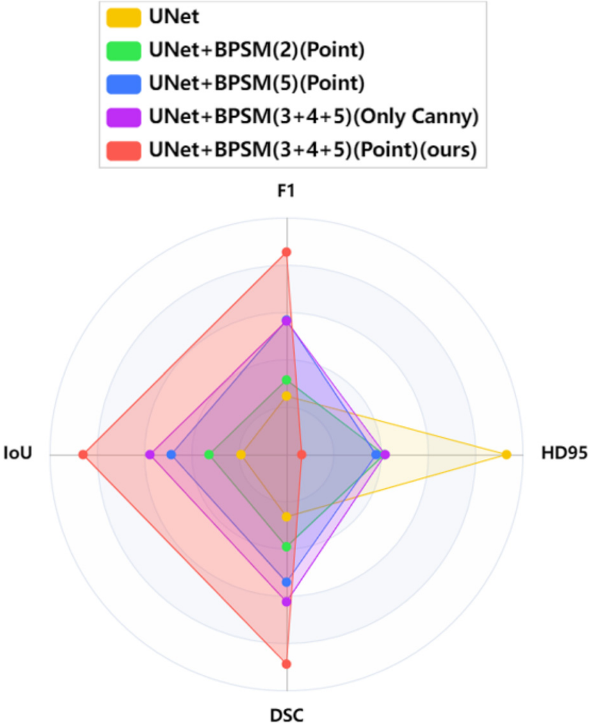


Fig. 10. Radar chart of different placement of our Boundary Points Supervision Module.

From the results, we investigate the effect of the location and number of BPSMs on the TN3k dataset. When we place BPSM in the 2nd layer of the encoder stage, the improvement effect is minimal, but the calculation amount increases considerably because the low-level features are difficult to accurately predict key points. The performance is significantly improved when boundary point features are merged at multiple scales. Therefore, BPSMs are finally placed at 3*rd*, 4*th*, and 5*th* layers.

When introducing BPSM supervised by edge key points in UNet, the segmentation accuracy is improved. Compared with the supervision introduction without boundary point map selection, the segmentation results are improved by 0.79% and 2.47 on DSC and HD95, respectively, which indicates that the boundary key point

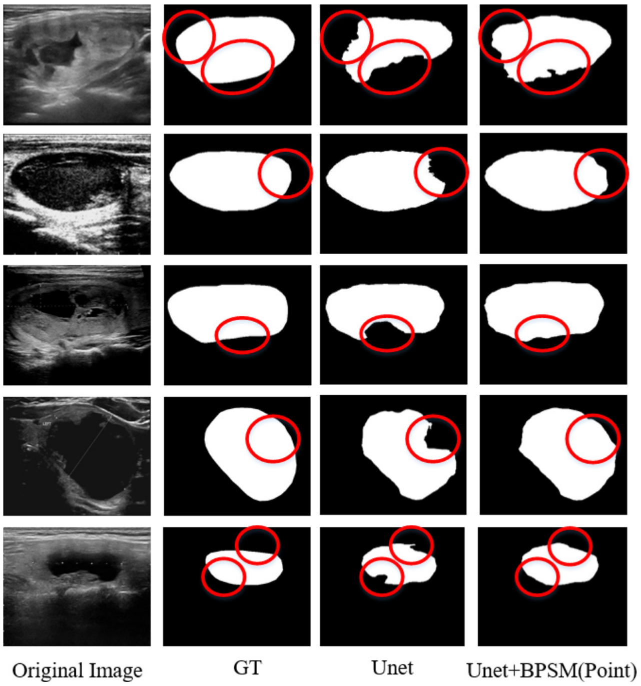


Fig. 11. Comparison between the results of our Boundary Points Supervision Module and the original model.

selection algorithm proves to be effective and effective boundary preservation is difficult to achieve solely through the canny operator.

In addition to evaluating the model by segmentation metrics, visual evaluation of the segmentation results is essential.We also evaluate the visual segmentation results for the analytical discussion of the proposed method. Fig. 11 visually shows the experimental results before and after adding BPSM. It is obvious that UNet often fails to preserve the shape of the boundaries. The proposed BPSM has the ability for ultrasound image segmentation when applied at multiple scales, it can handle the border pixel values almost the same between the boundary and the background.

3.5.2. Effect of adapted multi-scale feature fusion module

Table 4 and Table 5 also show that, on the TN3k dataset, when AMFFM is introduced, DSC and HD95 improve by 1.02% and 5.25, respectively. On the private thyroid dataset, the experimental results also show significant improvement, which DSC and HD95 improved by 1.84% and 8.73, respectively. Fig. 11 shows the experimental results of these models.

The first row of Fig. 12 shows that small objects are discarded using UNet directly. The third row shows that the proposed BPAT-UNet overcomes this issue and can segment relatively small nodules, which can confirm that our AMFFM can compensate for the problem of smalle feature loss, which can bring about serious problems such as missing targets and treatment difficulty during the medical diagnosis. Furthermore, the third row of Fig. 12 shows that objects with extremely irregular shapes can be segmented relatively accurately, which confirms the functionality of deformable computing.

3.5.3. Effect of assembled transformer module

Furthermore, we conducted experiments to compare different structures of ATM. Table 4 and Table 5 illustrate that DSC and HD95 improve by 2.03% and 4.96 on the TN3k dataset after introducing ATM into UNet. On the private dataset, segmentation results are improved by 4.91% and 12.81 on DSC and HD95, respectively.

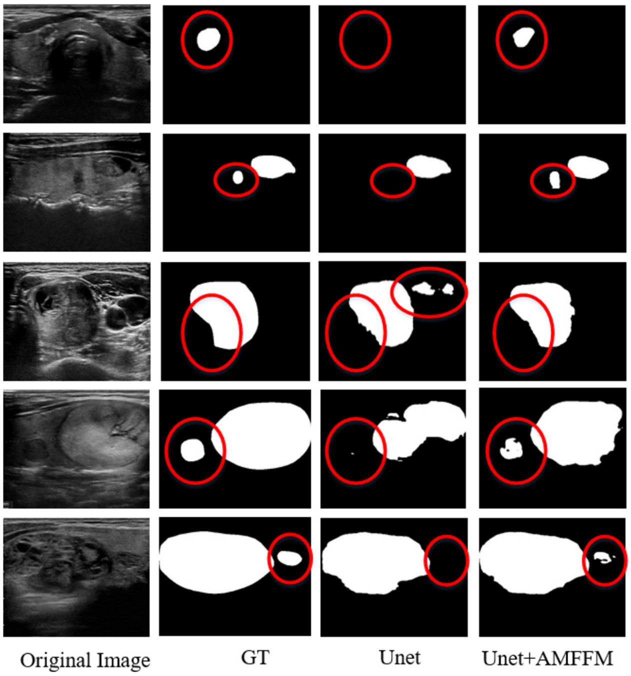


Fig. 12. Comparison between the results of our Adapted Multi-scale Feature Fusion Module and the original mode.

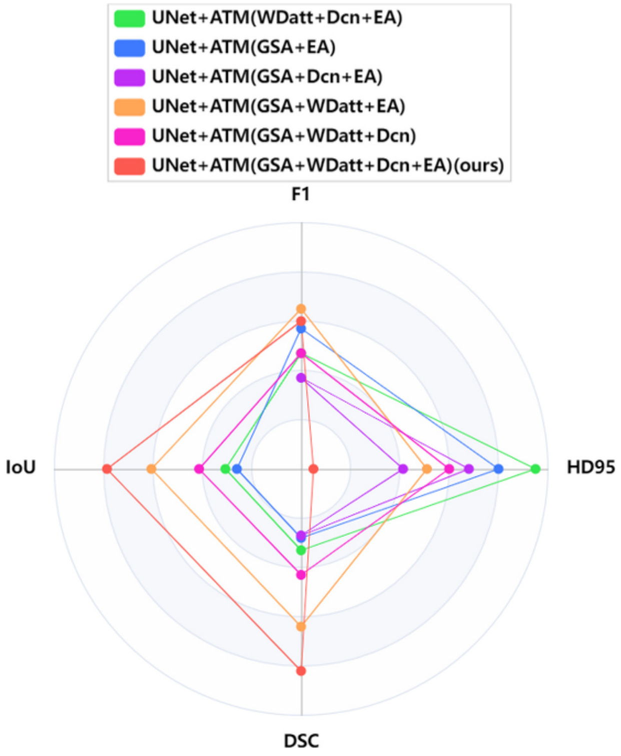


Fig. 13. Radar chart of our different ATM styles on the public TN3k dataset.

In addition, we also conducted experiments to compare different structures of ATM.

We separated ATM into three blocks the local features extraction block including WDatt and Dcn, the global feature extraction block, and EA block. Then, we combine these blocks for experiments. The segmentation results are listed in Table 7 and Fig. 13.

Compared with the UNet baseline, the results display significant improvement in both inserting local branches and global branches.

Table 7
Performance on different ATM styles on the public TN3k dataset.

Model	F1(%)	IoU(%)	DSC(%)	HD95
UNet	81.62	67.77	80.79	21.51
UNet+ATM (WDatt+Dcn+EA)[1]	82.97	69.96	82.33	17.45
UNet+ATM (GSA+EA)	83.07	69.89	82.28	17.30
UNet+ATM (GSA+Dcn+EA)	82.87	69.88	82.27	17.18
UNet+ATM (GSA+WDatt+EA)	83.15	70.41	82.64	17.01
UNet+ATM (GSA+WDatt+Dcn)	82.97	70.12	82.43	17.10
UNet+ATM (GSA+WDatt+Dcn+EA)(ours)*	**83.10**	**70.68**	**82.82**	**16.55**

[1] ATM with the modules listed in the mentioned in parentheses. * The best results are highlighted in bold.

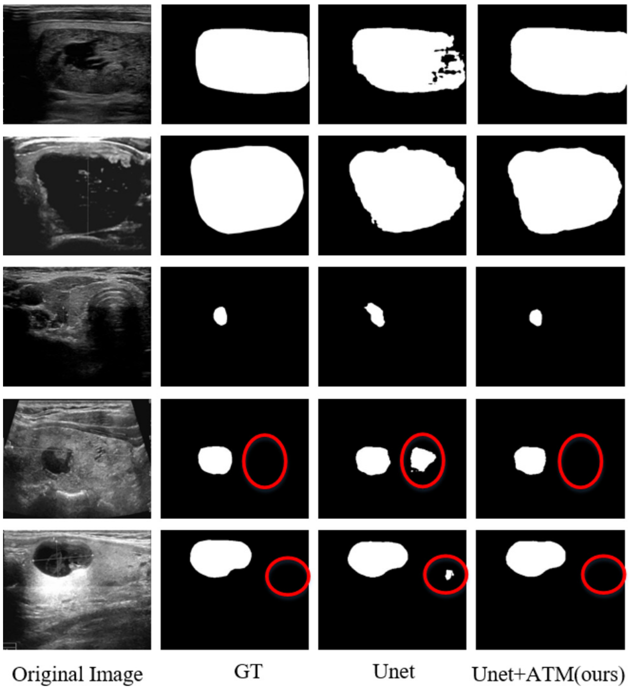


Fig. 14. Comparison between the results of our Adapted Multi-scale Feature Fusion Module and the original mode.

Among them, UNet+ATM(WDatt+Dcn+EA) represents the only local branch and UNet+AMT(GSA+EA) means the only global branch. The improved DSC and HD95 reveal the ability of deformable computing and global attention. It can also be seen that after parallel fusion of global and local branches, all indicators have significantly improved. Thus, a complementary feature relationship exists between the two branches. That is to say, the feature information that global branches pay attention to is missing from local branches, and vice versa. The visual segmentation results are displayed in Fig. 14. The examples in the first three lines mark the role of locally deformable computation, which helps the model identify thyroid objects with variable shapes. In addition, the examples in the following two lines imply the ability of global attention to model nodules using information around them. In addition, We further dismember the local branch. Although using Dcn as local information alone may lose some performance compared with WDatt, the fusion of the two helps the model obtain higher performance, validating the necessity and effectiveness of introducing them. Posterior most, we conduct ablation experiments on EA, losing consideration of the relationship between samples inevitably results in a certain performance loss. This result indicates that attention to the relationship between different feature maps is crucial in downstream tasks of ultrasound thyroid imaging.

4. Discussion

This work builds a novel boundary-preserving assembly Transformer UNet (BPAT-UNet) for thyroid ultrasound image segmentation. BPAT-UNet focuses on handling the segmentation of the nodules with irregular shapes and with small-size.

For irregular shape issues, the proposed BPAT-UNet using BPSM module improves thyroid nodules boundary segmentation. In addition, the DP algorithm is used to select edge points that adversely affect segmentation and can handle segmenting thyroid nodules with large shape variance.

For small nodules, the proposed BPAT-UNet pays more attention to small features, using AMFFM with channel attention to enhance small-scale thyroid nodules features and using ATM to integrate high-frequency local and low-frequency global feature information. Both AMFFM and ATM help improve the detection of objects with large shape differences as they can characterize the correlation between features and deformable feature calculation.

Furthermore, the proposed BPAT-UNet adaptively fuses downsampled multi-scale features and combines deformable local features for small-sized and irregular nodules.

5. Conclusion

BPAT-UNet is proposed for thyroid ultrasound segmentation that introduces BPSM, AMFFM, and ASTM to extract more richer features that obtain accurate thyroid nodules. The experimental results show that the proposed method can achieve better segment thyroid nodules compare with the SOTA methods. The limitation of the proposed work is that we only conduct experiments on ultrasound thyroid images and have not yet verified the effectiveness of the model on other organ ultrasound images (such as breast and prostate) or other kind of medical images (CT, PETs or MRI images). In the future, we will use more medical images to validate BPAT-UNet and compare it with the latest segmentation algorithms like the Diffusion model on more challenging medical images.

Declaration of Competing Interest

All authors of this manuscript have directly participated in planning, execution, and/or analysis of this study. The contents of this manuscript are not now under consideration for publication elsewhere. The contents of this manuscript will not be copyrighted, submitted, or published elsewhere while acceptance by Computer Methods and Programs in Biomedicine is under consideration. There are no directly related manuscripts or abstracts, published or unpublished, by any authors of this manuscript. I am one author signing on behalf of all co-authors of this manuscript, and attesting to the above.

Hui Bi

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 62171125 and 62141401, in part by the 67th National Postdoctoral Program of China under Grants 2020M671277, in part by the General Program of Jiangsu Provincial Health Commission under Grants M2020006, in part by the Jiangsu Provincial Key Research and Development Program Social Development Project under Grants BE2022720, in part by the Key Laboratory of Computer Network and Information Integration (Southeast University) of the Ministry of Education under Grants K93-9-2021-08, in part by the Science and Technology Project of Changzhou City under Grants under CE20215045 and CJ20220136.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2023.107614.

References

[1] J. Chen, H. You, K. Li, A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images, Comput. Methods Programs Biomed. 185 (2020) 105329, doi:10.1016/j.cmpb.2020.105329.

[2] E. Kollorz, E. Angelopoulou, M. Beck, D. Schmidt, T. Kuwert, Using power watersheds to segment benign thyroid nodules in ultrasound image data, Bildverarbeitung für die Medizin (2011) 124–128, doi:10.1007/978-3-642-19335-4_27.

[3] W.H. Hesamian, W. Jia, X. He, P.J. Kennedy, Deep learning techniques for medical image segmentation: achievements and challenges, J. Digit. Imaging 32 (4) (2019) 582–596, doi:10.1007/s10278-019-00227-x.

[4] A. Taghanaki, K. Abhishek, J.P. Cohen, J. Cohen-Adad, G. Hamarneh, Deep semantic segmentation of natural and medical images: a review, Artif. Intell. Rev. (1) (2020) 137–178, doi:10.1007/s10462-020-09854-1.

[5] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (7) (2022) 3523–3542, doi:10.1109/TPAMI.2021.3059968.

[6] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 640–651, doi:10.1109/TPAMI.2016.2572683.

[7] O. Ronneberger, P. Fischer, T. Brox, UNEt: convolutional networks for biomedical image segmentation, Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. 9351 (2015) 234–241, doi:10.1007/978-3-319-24574-4_28.

[8] H. Ding, X. Jiang, A.Q. Liu, N.M. Thalmann, G. Wang, Boundary-aware feature propagation for scene segmentation, Proc. IEEE Int. Conf. Comput. Vis. (2019) 6818–6828, doi:10.1109/ICCV.2019.00692.

[9] K. Kamnitsas, Efficient multi-scale 3d CNN with fully connected CRF for accurate brain lesion segmentation, Med. Image Anal. (2016) 36–61, doi:10.1016/j.media.2016.10.004.

[10] T. Takikawa, D. Acuna, V. Jampani, S. Fidler, Gated-SCNN: gated shape CNNs for semantic segmentation, Proc. IEEE Int. Conf. Comput. Vis. (2019) 5228–5237, doi:10.1109/ICCV.2019.00533.

[11] H.J. Lee, J.U. Kim, S. Lee, H.G. Kim, Y.M. Ro, Structure boundary preserving segmentation for medical image with ambiguous boundary, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2020) 4816–4825, doi:10.1109/CVPR42600.2020.00487.

[12] J. Chen, TransUNet: transformers make strong encoders for medical image segmentation, CoRR abs/2102.04306 (2021) 1–13.

[13] R. Dong, X. Pan, F. Li, DenseUNet-based semantic segmentation of small objects in urban remote sensing images, IEEE Access 7 (2019) 65347–65356, doi:10.1109/ACCESS.2019.2917952.

[14] G. Chen, A survey of the four pillars for small object detection: multiscale representation, contextual information, super-resolution, and region proposal, IEEE Trans. Syst. Man Cybern.: Syst. 52 (2) (2022) 936–953, doi:10.1109/TSMC.2020.3005231.

[15] H. Li, P. Xiong, H. Fan, J. Sun, DFANEt: deep feature aggregation for real-time semantic segmentation, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2019) 9514–9523, doi:10.1109/CVPR.2019.00975.

[16] A. Kirillov, R. Girshick, K. He, P. Dollár, Panoptic feature pyramid networks, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2019) 6392–6401, doi:10.1109/CVPR.2019.00656.

[17] M. Yang, K. Yu, C. Zhang, Z. Li, K. Yang, DenseASPP for semantic segmentation in street scenes, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2018) 3684–3692, doi:10.1109/CVPR.2018.00388.

[18] J. Schlemper, Attention gated networks: learning to leverage salient regions in medical images, Med. Image Anal. 53 (2019) 197–207, doi:10.1016/j.media.2019.01.012.

[19] K. Deng, Transbridge: a lightweight transformer for left ventricle segmentation in echocardiography, ASMUS 2021 : Simplifying Medical Ultrasound 12967 (2021) 63–72, doi:10.1007/978-3-030-87583-1_7.

[20] H. Gong, Multi-task learning for thyroid nodule segmentation with thyroid region prior, Proc. IEEE Int. Symp. Biomed. Imag. (2021), doi:10.1109/ISBI48211.2021.9434087.

[21] A. Tsai, A shape-based approach to the segmentation of medical imagery using level sets, IEEE Trans. Med. Imaging 22 (2) (2003) 137–154, doi:10.1109/TMI.2002.808355.

[22] X. Xiao, S. Lian, Z. Luo, S. Li, Weighted res-UNet for high-quality retina vessel segmentation, Proc. Int. Conf. Infor. Tech. Mech. Eng. (2018) 327–331, doi:10.1109/ITME.2018.00080.

[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2016) 770–778, doi:10.1109/CVPR.2016.90.

[24] Z. Zhou, M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNEt++: a nested UNet architecture for medical image segmentation, Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. Workshop 11045 (2018), doi:10.1007/978-3-030-00889-5_1.

[25] X. Li, H. Chen, X. Qi, Q. Dou, C.W. Fu, P.A. Heng, H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes, IEEE Trans. Med. Imaging 37 (12) (2018) 2663–2674, doi:10.1109/TMI.2018.2845918.

[26] A. Ma, J. Wang, Y. Zhong, Z. Zheng, Factseg: foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–16, doi:10.1109/TGRS.2021.3097148.

[27] J. Sun, F. Darbeha, M. Zaidi, B. Wang, SAUNEt: shape attentive UNet for interpretable medical image segmentation, Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. 12264 (2020) 797–806, doi:10.1007/978-3-030-59719-1_77.

[28] T. Gonçalves, I. Rio-Torto, L.F. Teixeira, J.S. Cardoso, A survey on attention mechanisms for medical applications: are we moving toward better algorithms? IEEE Access 10 (2022) 98909–98935, doi:10.1109/ACCESS.2022.3206449.

[29] O. Petit, N. Thome, R. Clément, L. Soler, UNEt transformer: self and cross attention for medical image segmentation, Int. Conf. Mach. Learn. Mach. Intell. Workshops 12966 (2021) 267–276, doi:10.1007/978-3-030-87589-3_28.

[30] B. Chen, Y. Liu, Z. Zhang, G. Lu, D. Zhang, TransattUNet: multi-level attention-guided UNet with transformer for medical image segmentation, arXiv (2021), doi:10.48550/arXiv.2107.05274.

[31] Y. Gao, M. Zhou, D.N. Metaxas, UTNEt: a hybrid transformer architecture for medical image segmentation, Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. 12903 (2021) 61–71, doi:10.1007/978-3-030-87199-4_6.

[32] A. Vaswani, Attention is all you need, Proc. Adv. Neural Inf. Process. Syst., Long Beach, CA, USA (2017) 5998–6008.

[33] A. Dosovitskiy, An image is worth 1616 words: transformers for image recognition at scale, Proc. Int. Conf. Learn. Represent. (2021) 1–5.

[34] C. Matsoukas, J.F. Haslum, M. Soderberg, K. Smith, Is it time to replace CNNs with transformers for medical images? arXiv (2021), doi:10.48550/arXiv.2108.09038.

[35] Y. Li, GT UNEt: a UNet like group transformer network for tooth root segmentation, Int. Conf. Mach. Learn. Mach. Intell. Workshops 12966 (2021) 386–395, doi:10.1007/978-3-030-87589-3_40.

[36] W. Wang, PVT V2: improved baselines with pyramid vision transformer, Comp. Visual Media (8) (2022) 415–422, doi:10.1007/s41095-022-0274-8.

[37] B. Chen, GLIt: neural architecture search for global and local image transformer, Proc. IEEE Int. Conf. Comput. Vis. (2021) 12–21, doi:10.1109/ICCV48922.2021.00008.

[38] Z. Pan, J. Cai, B. Zhuang, Fast vision transformers with hilo attention, arXiv (2022), doi:10.48550/arXiv.2205.13213.

[39] C. Si, Inception transformer, arXiv (2022), doi:10.48550/arXiv.2205.12956.

[40] J.M.J. Valanarasu, P. Oza, I. Hacihaliloglu, V.M. Patel, Medical transformer: gated axial-attention for medical image segmentation, Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. 12901 (2021), doi:10.1007/978-3-030-87193-2_4.

[41] Z. Liu, Swin transformer: hierarchical vision transformer using shifted windows, Proc. IEEE Int. Conf. Comput. Vis. (2021) 9992–10002, doi:10.1109/ICCV48922.2021.00986.

[42] Q. Hou, L. Zhang, M.M. Cheng, J. Feng, Strip pooling: rethinking spatial pooling for scene parsing, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2020) 4002–4011, doi:10.1109/CVPR42600.2020.00406.

[43] S. Ren, D. Zhou, S. He, J. Feng, X. Wang, Shunted self-attention via multi- scale token aggregation, arXiv (2021), doi:10.48550/arXiv.2111.15193.

[44] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, IEEE Trans. Pattern Anal. Mach. Intell. 42 (8) (2020) 2011–2023, doi:10.1109/TPAMI.2019.2913372.

[45] J. Dai, Deformable convolutional networks, Proc. IEEE Int. Conf. Comput. Vis. (2017) 764–773, doi:10.1109/ICCV.2017.89.

[46] Z. Xia, X. Pan, S. Song, L.E. Li, G. Huang, Vision transformer with deformable attention, arXiv (2022), doi:10.48550/arXiv.2201.00520.

[47] A. Stergiou, R. Poppe, G. Kalliatakis, Refining activation downsampling with softpool, Proc. IEEE Int. Conf. Comput. Vis. (2021) 10337–10346, doi:10.1109/ICCV48922.2021.01019.

[48] Y. Sha, Y. Zhang, X. Ji, L. Hu, Transformer-UNet: raw image processing with UNet, arXiv (2021), doi:10.48550/arXiv.2109.08417.

[49] L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, arXiv (2018), doi:10.48550/arXiv.1802.02611.

[50] K. Trebing, T. Sta, S. Mehrkanoon, Smaat-UNet: precipitation nowcasting using a small attention-UNet architecture, Pattern Recogn. Lett. 145 (2021) 178–186, doi:10.1016/j.patrec.2021.01.036.