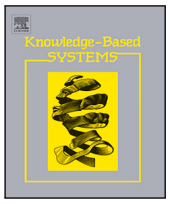
Knowledge-Based Systems 324 (2025) 113781

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



DRTNet: Dual-route transformer network for thyroid ultrasound segmentation based on Bbox-supervised learning

Hui Bi [a,b,c], Chengjie Cai [a], Jiawei Sun [b,d,e], Shihao Ge [a], Huazhong Shu [f,g], Xinye Ni [b,d,e,*]

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

ARTICLE INFO

Keywords:
Ultrasound image segmentation
Thyroid nodule segmentation
Transformer-based network
Bbox-supervised segmentation
DRTNet

ABSTRACT

***Background and Objective*:** Accurate nodule delineation plays a significant role in the intelligent diagnosis of thyroid disease. However, the labels accessing is difficult since it is time-consuming and laborious. To mitigate the over-dependence of the segmentation accuracy on the labels, we proposed a dual-route transformer network (DRTNet) based on Bbox-supervised learning that only requires the rough bounding rectangle as labels instead of a precise boundary for thyroid ultrasound segmentation.

***Methods*:** DRTNet incorporates double-branch foreground class activation mappings (CAMs) into Transformers to combine key areas. Meanwhile, double-branch architecture dynamically adjusts the feature distribution of nodules of different sizes in frequency channels and spatial dimensions, effectively addressing the localization of nodules of different sizes. Moreover, ultrasound prior background-aware pooling (UPBAP) is proposed in both branches to deal with the ambiguous boundary of thyroid nodules. Finally, adaptive uncertainty estimate multi-scale consistency (AUEMC) is proposed to help mitigate the risk of excessive over-fitting because of pseudo annotations, which further guarantees consistency among nodules with diverse resolutions.

***Results*:** Substantial improvement of segmentation accuracy is shown on the public thyroid dataset of TN3k and DDTI dataset with Dice similarity coefficient (DSC) of 84.94% and 83.98%, with 95% of the asymmetric Hausdorff distance (HD95) of 27.69 and 29.18, respectively. And our private dataset has a DSC of 84.39% and HD95 of 14.53.

***Conclusions*:** The proposed DRTNet used rectangular box labeled for thyroid ultrasound images based on Bbox-supervised learning. The experimental results show that the DRTNet is comparable to these fully supervised methods. Code is available at https://github.com/ccjcv/DRTNet.

1. Introduction

The incidence rate of thyroid nodules, one of the most common types of nodular lesions, has seen a steady increase in recent years [1]. Furthermore, malignant nodules are inextricably linked with cancer, which can be treated through early diagnosis and treatment [2]. Therefore, accurate segmentation of nodules plays a key role in thyroid diagnosis. Ultrasound imaging has become the preferred technology for examining and diagnosing thyroid nodules because of its outstanding advantages of no radiation, low cost, and real-time performance [3,4]. However, ultrasound images often have low contrast, speckles, shadows, and a substantial amount of noise that result in blurred edges and unexpectedly varied boundaries of thyroid nodules, making it considerably challenging to segment ultrasound thyroid nodules [5].

Recently, several weakly supervised semantic segmentation(WSSS) approaches works have been proposed to alleviate subjective bias, reduce time consumption, and alleviate labor-intensive processes for segmentation of thyroid nodule segmentation of ultrasound images [6–9]. Convolutional neural networks (CNNs) are known to have wide-ranging applications in the field of image segmentation because of their powerful processing scale invariance and ability to model the inductive deviation of images on the pixel-level label [10]. In addition,

* Corresponding author at: The Affiliated Changzhou NO.2 People's Hospital of Nanjing Medical University, Changzhou, 213003, China.
E-mail addresses: bihui@cczu.edu.cn (H. Bi), caichengjie666@163.com (C. Cai), 921173049@qq.com (J. Sun), 1098942584@qq.com (S. Ge), shu.list@seu.edu.cn (H. Shu), nxy@njmu.edu.cn (X. Ni).

https://doi.org/10.1016/j.knosys.2025.113781
Received 14 August 2023; Received in revised form 18 April 2025; Accepted 15 May 2025
Available online 1 June 2025

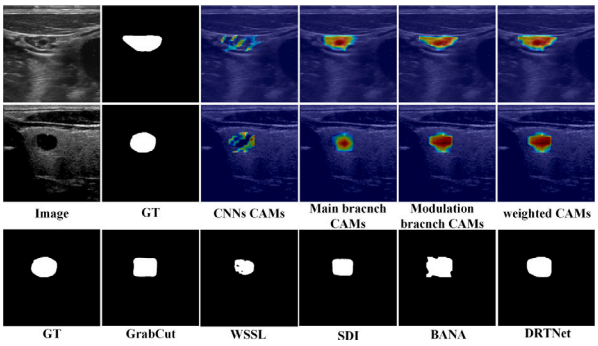


Fig. 1. Visual comparison of pseudo masks and our dual-branch weighted CAMs visualization. Our approach generates better pseudo masks than other WSSS methods using object Bbox. CAMs of the modulation branch make up for the lack of CAMs of main branch.

to highlight discriminative areas, class activation mappings (CAMs) derived from CNNs for image classification using global average pooling (GAP) [11] are a conventional technique. However, when applied to thyroid ultrasound image segmentation, WSSS based on CNN and GAP still has limitations, including Bbox supervised segmentation [12]. As shown in Fig. 1, the existing approaches is cannot accurately enough for locating thyroid nodules.

(1) Hard to activate integral object regions. CNN CAMs in Fig. 1 demonstrate a typical example in which CNNs combined with CAMs based on weak supervision cannot accurately achieve integral object regions, considerably affecting the quality of pseudo masks. This can be attributed to the fact that CNNs fail to properly explore the global feature relations properly, and classification-oriented CAMs often lack customized optimization for the entire object area [11]. To address this issue, some schemes have been attempted to expand the corresponding differentiated areas. For instance, Wei [13] focuses on other areas by erasing CAM areas and retraining the network. Seam [14] operates by regularizing different transformed images to expand the CAM area.

(2) Hard to deal with the disseminated boundary. As demonstrated in the third row of Fig. 1, existing weakly supervised ultrasound image pseudo masks fail to maintain accurate boundaries. The reason is that the gray value of the pixels on the edge of the thyroid nodules is usually very close to the surrounding pixels, which can easily lead to errors in distinguishing the foreground and background areas. Neglecting consideration for fuzzy boundaries in methods increases the likelihood of missing the border area. Most existing solutions are based on ready-made solutions such as GrabCut and MCG [15,16], which make achieving the expected results difficult when applied to ultrasound images. BANA leverages background-aware pooling (BAP) [17] to distinguish objects within the Bbox from the background during the classification stage, pursuing more precise object boundaries. However, the direct application of BAP on thyroid ultrasound images with ultrasound characteristics has failed to yield satisfactory results, as demonstrated by the BANA shown in Fig. 1.

(3) Easy to overfit with pseudo labels. The pseudo label generated by DRTNet in Fig. 1 shows discrepancies as compared with the ground truth(GT) labels. The labels obtained from the WSSS of the target Bbox belong to noisy labels, and directly applying them to model supervision will inevitably cause certain performance losses. To increase the precision of the final segmentation, some methods are polished via regularization [18,19].

To simultaneously resolve the aforementioned three issues in thyroid ultrasound segmentation, we propose a dual-route transformer network (DRTNet) for thyroid ultrasound segmentation based on Bbox-supervised learning. This design mainly focuses on the effective feature representation of the thyroid from the incomplete nodule internal region, blurring nodule edges, and the inferior pseudo nodule mask.

The proposed DRTNet incorporates innovations in three aspects. First, for the dual-branch modulation strategy, we dynamically compensate for the features missed by the main branch from both the frequency channel and spatial dimensions. This modulation scheme is novel and exhibits cross-task applicability, including the mining of complex features for most downstream tasks, not being limited to image segmentation scenarios. Additionally, the regularization strategy for box-supervised segmentation is equipped with adaptive weighting of channel information, which further enhances the consistency of segmentation results across different resolutions. This approach is also applicable to weakly supervised segmentation schemes in general scenarios.

The main contributions of this study are as follows:

(1) The proposed DRTNet for thyroid ultrasound segmentation introduces a double-branch foreground CAMs modulation with a transformer backbone to generate more accurate thyroid masks based on Bbox learning.

(2) Ultrasound prior BAP (UPBAP) is adopted to define the finer boundary of the thyroid nodules and focus on a relatively small difference between the nodules and the partial background.

(3) The adaptive uncertainty estimate multi-scale consistency (AUEMSC) is also introduced to avoid over-fitting of the segmentation results due to pseudo annotations.

(4) As per the quantitative results of the thyroid public dataset TN3k [20], DDTI [21] and our private dataset, the proposed framework demonstrates significant accuracy advantages compared with the other state-of-the-art techniques.

The rest of this study is organized as follows. Section 2 briefly reviews related studies published in recent years. Section 3 provides a detailed description of the architecture of the proposed DRTNet and its modules for thyroid ultrasound nodule segmentation. Section 4 presents our experimental results and compares the proposed method with other advanced segmentation methods. Furthermore, the effects of the key elements involved in our method are analyzed through a series of ablation studies. Lastly, the discussion and conclusions are presented in Sections 5 and 6, respectively.

2. Related studies

2.1. Weakly supervised semantic segmentation

Currently, in the field of ultrasound image segmentation, pixel-level labels are dominant [22–24]. However, pixel-level labels based on fully supervised learning are prone to subjective bias, time-consuming, and labor-intensive, and prone to subjective bias. Thus, to reduce the burden of segmentation and labeling, WSSS methods have emerged that use different levels of labels such as image-level and Bbox labels [25,26], which help save labeling time. The segmentation performance of the WSSS is comparable to that of full supervision [27,28].

WSSS methods typically leverage CAMs obtained from CNNs for image classification using GAP to generate pseudo masks for pixel-level semantic segmentation [11]. However, the pseudo masks generated from this method can only highlight critical regions. Thus, efforts to improve pseudo masks focus on two perspectives: expanding CAMs' receptive fields and refining target boundaries.

There are numerous ways to expand the receptive field of CAM in the pixel-level supervised segmentation, and of course, this improvement approach can also be implemented in the Bbox supervised mode. To broaden the receptive fields of CAMs in the corresponding area, Wei extends the target area by paralleling the expansion convolution with different expansion rates [29]. Moreover, Fan combines similarities and differences across image semantics [30]. Activation modulation and re-calibration (AMR) outlines two parallel branches for modulation compensation to obtain weighted CAMs, wherein the compensation branches rearrange feature importance from the channel space sequence [31]. The effectiveness of AMR has been demonstrated

using natural images. However, disparities in terms of the distribution between ultrasound and natural images hinder the expected performance of AMR in ultrasound images. The same limitation applies to the aforementioned method.

In terms of Bbox supervised segmentation, to refine target boundaries, most solutions use existing segmentation methods to generate pseudo labels, such as GrabCut. However, these segmentation methods are find it difficult to meet expectations in ultrasound thyroid images, and there are also methods that exploit classification networks to obtain pseudo labels. BAP distinguishes the foreground and background areas within the object boundary box from the background outside the box. Introducing segmentation inevitably requires substantial time and computational resources. The pool performance of BAP can be attributed to the distinctive characteristics of ultrasound images, necessitating further improvements for enhanced results.

2.2. Transformer and attention mechanism

The appearance of the vision transformer (ViT) [32] marks the first pure transformer architecture used in visual tasks. Subsequently, numerous image downstream tasks took it as the first choice, achieving astonishing performance [33,34]. The purpose of the attention mechanism aims to strengthen the global context, including self-attention [35]. Furthermore, the CAMs generated by the self-attention mechanism in transformers can model the global feature relationship that could claim a more complete target area to avoid the inherent defects of CNN [36]. Gao [37] proposed the token semantic coupled attention map, which is the first weakly supervised target location method based on a transformer. From this aforementioned study, it can be observed that transformers demonstrate advantageous capabilities in generating CAMs, thereby providing promising prospects for weakly supervised learning.

To capture the spatial and channel information within global features, channel and spatial attention mechanisms are proposed [38]. Both are equally crucial and cannot be disregarded as compared with the transformer's self-attention. SE [39] first noticed the importance of different channel features. CBAM [38] simultaneously solicits the characteristics of channels and spaces. Unlike existing attention, FcaNet [40] promotes channel attention in the frequency domain. Considering the varying sizes of thyroid nodules, channel and spatial attention mechanisms can dynamically adjust the feature distribution of objects. Therefore, we believe that these two attention mechanisms will also be beneficial for capturing thyroid nodule objects in our study.

2.3. Learning from noisy labels

WSSS methods achieve pixel-level semantic segmentation using pseudo masks obtained from weak supervision. A substantial discrepancy exists between the pseudo masks obtained from weak supervision and the manual annotations, resulting in model over-fitting and posing a significant challenge to the task.

SDI [15] extracts two segmentation schemes to obtain the same area of the pseudo masks. Box2Seg [41] generates an attention map to adjust the cross-entropy loss to eliminate incorrect labels.

TSMAN [42] uses two mutually interested network information to mine noisy gradients to mitigate the impact of noisy labels. SeCoST [43] uses teacher–student distillation to correct labels and employs a joint training mechanism between parallel models. OCR [44] designed an out-of-candidate correction mechanism to achieve a more accurate and complete recall of the position of each class.

In contrast to the abovementioned scheme, this study adopts the regularization commonly used in noise image classification [18,45]. URPC [46] corrects the uncertainty of the prediction results of different scales in the same model to achieve consistency. ADELE [47] inputs three groups of images with different scales into three segmentation networks and acquires multi-scale consistency through uncertainty estimation.

The abovementioned studies significantly improve the adaptability of segmentation with noisy labels on natural images. However, these advancements remain insufficient for thyroid ultrasound images, primarily due to the disparities in the distribution between ultrasound and natural images.

3. Methods

In this section, we present the overall architecture of the proposed DRTNet framework. We provide a detailed description of the DRTNet in the subsequent subsections. Our framework primarily comprises three stages, as shown in Fig. 2. In stage 1, we construct a dual-route Bbox-supervised transformer network for the classification of ultrasound thyroid nodules, with an auxiliary branch that integrates the dynamic modulation module (DMM). In addition, the UPBAP is adopted to achieve better separation of the foreground and background. Furthermore, we perform a weighted summation of the foreground and background CAMs from both pathways to obtain a more comprehensive representation of the object region. In stage 2, pseudo masks are generated using DenseCRF [48], a dense conditional random field method. In stage 3, AUEMSC is introduced into the transformer network for thyroid nodule segmentation of pseudo masks.

3.1. Dual-route classification using UPBAP

Our dual-route weak supervised classification network consists of a main branch and an auxiliary branch, as shown in Fig. 2, inspired based on AMR. The main branch primarily serves the purpose of emphasizing the distinctive region of the target object. Unlike AMR, the DMM of the auxiliary branch captures the missed nodule areas overlooked by the main branch (Fig. 3). To improve the foreground and background classification, we introduce BAP to build a more precise and accurate boundary area and exclude invalid areas in the background of thyroid ultrasound images [11] widely used in the field of weak supervision.

3.1.1. DMM

As shown in Fig. 3, by employing dynamic kernel convolution with channel and spatial attention, the DMM not only improves secondary features in the main branch of our framework but also captures large-range variations in thyroid nodules [49]. The secondary features of the main branch represent the features of other regions besides the significant area of nodules that the main branch focuses on.

First, cross supervision is performed on the CAMs generated by the dual branches to avoid concentration of CAMs being concentrated in critical or incorrect positions, achieving optimization of the output CAMs. Second, the DMM is equipped with dynamic weights between the channel and spatial levels to rearrange the distribution of activation features, promoting consistent attention to thyroid nodules throughout the model.

The DMM in the auxiliary branch comprises three essential components: the frequency channel dynamic kernel convolution branch (FCDCB), the spatial dynamic kernel convolution branch (SDCB), and residual connections. The DMM is located between two transformer layers, and the feature F^{in} derived from the former transformer layer is fed into the DMM. F^{c} originates from F^{in} through the 1×1 convolutional layer and it is subsequently passed into the frequency channel dynamic kernel convolution branch (FCDCB) to achieve F^{FCDCB} and the spatial dynamic kernel convolution branch (SDCB) to achieve F^{SDCB}, respectively. Then, we will get F^{DMM} is generated for the latter next transformer layer.

(a) FCDCB

The channel branch, as shown in Fig. 3, is proposed based on the key concept of dynamic routing, which is straightforward and concise. Building upon the utilization of multi-spectral channel attention [40], we further investigate the channel information in the frequency domain.

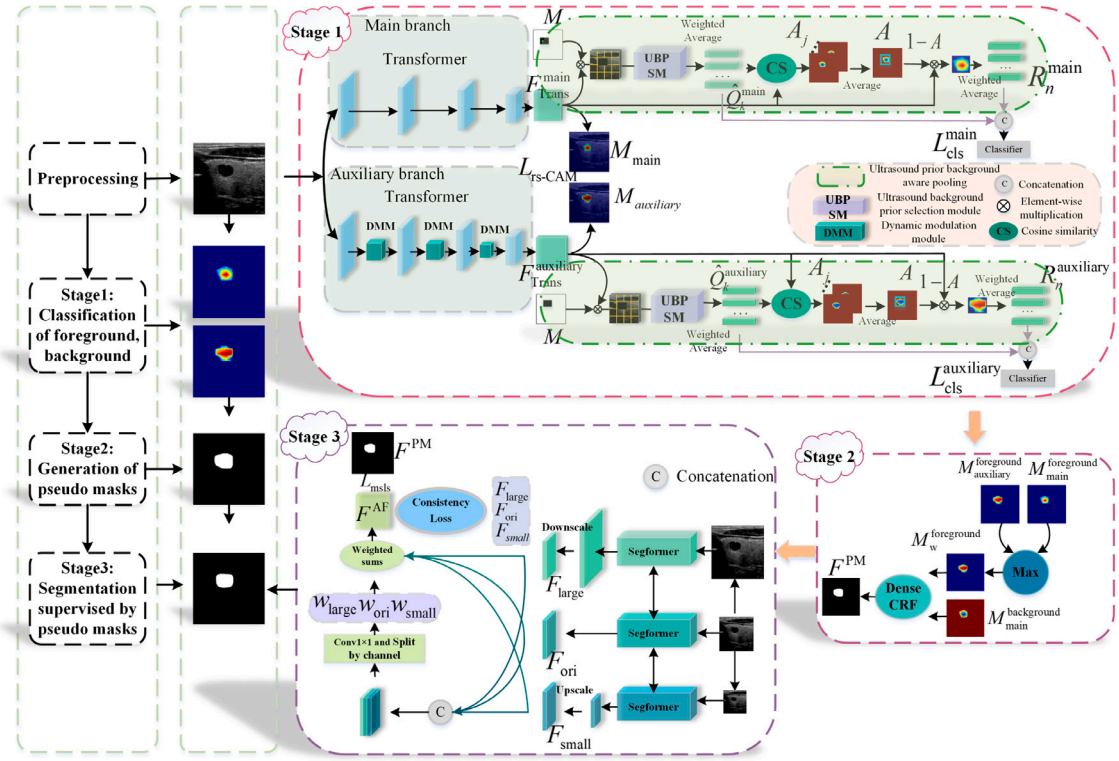


Fig. 2. Overall architecture of our proposed DRTNet.

In this study, discrete cosine transform (DCT) is commonly used as a component of the frequency channel attention weights, which are used to dynamically weight and sum the channel branches in DMM. This compensates for the possibility of missing other frequency feature components in separate GAP ultrasound information.

To ease the computational cost, we decomposed the convolutional feature $F^{c} \in R^{C\times H\times W}$ for subsequent operations, which we borrowed from MALUNet [50]. More specifically, it is evenly partitioned into four groups based on the feature channels, resulting in the formation of feature $F_i^{c} \in R^{(C/4)\times H\times W}(i = 0, 1, 2, 3)$. For each group, we assign a specific 2D DCT frequency component, which is expressed as follows:

$$\begin{aligned}\text{Freq}_{\,i} &= \text{DCT}_{2D}^{u,v}\left(F_i^{c}\right) \\ &= 2\sum_{h=0}^{H-1}\sum_{v=0}^{W-1} F_{i(:,h,w)}^{c} D_{h,w}^{u,v}, (i = 0, 1, 2, 3)\end{aligned} \tag{1}$$

where $[u, v]$ is the 2D index of the frequency component, $Freq_i$ is a vector of $\frac{C}{4}$ dimension, and $D_{h,w}^{u,v}$ represents the basic functions of 2D DCT, which is mathematically expressed as follows:

$$B_{h,w}^{u,v} = \cos(\frac{\pi h}{H}(u+\frac{1}{2}))\cos(\frac{\pi w}{W}(v+\frac{1}{2})) \tag{2}$$

Subsequently, all the vectors are concatenated in the channel dimension to obtain the multi-spectral vector.

$$Freq = Concat([Freq_0, Freq_1, Freq_2, Freq_3]) \tag{3}$$

where $Concat$ denotes the concatenation operation in the channel dimension. Subsequently, the multi-spectral vector is fed into the fully connected layer, followed by a sigmoid function, to generate the frequency attention feature map:

$$Z^{\text{MSATT}} = Sig(FC(Freq)) \tag{4}$$

where $FC(\bullet)$ denotes fully connected layer and $Sig(\bullet)$ denotes sigmoid function. Then, Z^{MSATT} is partitioned by the channel average to achieve the frequency attention weight $w_i^{\text{FA}} \in R^{(C/4)\times H\times W}(i = 0, 1, 2, 3)$, similar to ASFF [51].

Furthermore, for the feature $F_i^{c} \in R^{(C/4)\times H\times W}$, $(i = 0, 1, 2, 3)$ divided based on channel averaging, we perform four kernel convolution operations of different sizes. In the presence of thyroid nodules with significant size variations, we configure four different-sized convolutional kernels in a pyramidal arrangement; meanwhile, ESKNet only has two sizes [52]. Pyramid feature maps F_i^{PC}, $(i = 0, 1, 2, 3)$ captured through convolutional operations with four different kernel sizes are denoted as follows:

$$\begin{aligned}F_0^{\text{PC}} &= W_{3\times3} * F_0^{\text{cin}} \\ F_1^{\text{PC}} &= W_{5\times5} * F_1^{\text{cin}} \\ F_2^{\text{PC}} &= W_{7\times7} * F_2^{\text{cin}} \\ F_3^{\text{PC}} &= W_{9\times9} * F_3^{\text{cin}}\end{aligned} \tag{5}$$

where $*$ means the convolution operation, and $W_{k\times k}$ represents the convolution matrix whose kernel size is equal to k, $(k = 0, 1, 2, 3)$. PC denotes pyramid convolution.

Multiplication of the multi-scale convolutional features with the corresponding frequency channel attention weights $w_i^{\text{FA}}(i = 0, 1, 2, 3)$ allows for dynamic adjustment of the feature distribution of objects in the channel dimension. The output F^{FCDCB} of the FCDCB is expressed as follows:

$$\begin{aligned}F^{\text{FCDCB}} = Concat(&w_0^{\text{FA}} \times F_0^{\text{PC}}, w_1^{\text{FA}} \times F_1^{\text{PC}}, \\ &w_2^{\text{FA}} \times F_2^{\text{PC}}, w_3^{\text{FA}} \times F_3^{\text{PC}})\end{aligned} \tag{6}$$

(b) SDCB

Given the crucial role of spatial relationship modeling, the SDCB is also incorporated into our framework. Specifically, we begin by computing the spatial average pooling function of F^{c}, followed by convolutional operations to obtain the spatial attention map w^{SA}. The implementation process can be formulated as follows:

$$w^{\text{SA}} = Conv(SAP(F^{c})) \tag{7}$$

where $Conv$ denotes convolutional operation and SAP denotes spatial average pooling. Like FCDCB, the feature $F_i^{\text{PC}}(i = 0, 1, 2, 3)$ is also achieved after multi-scale convolutional processing.

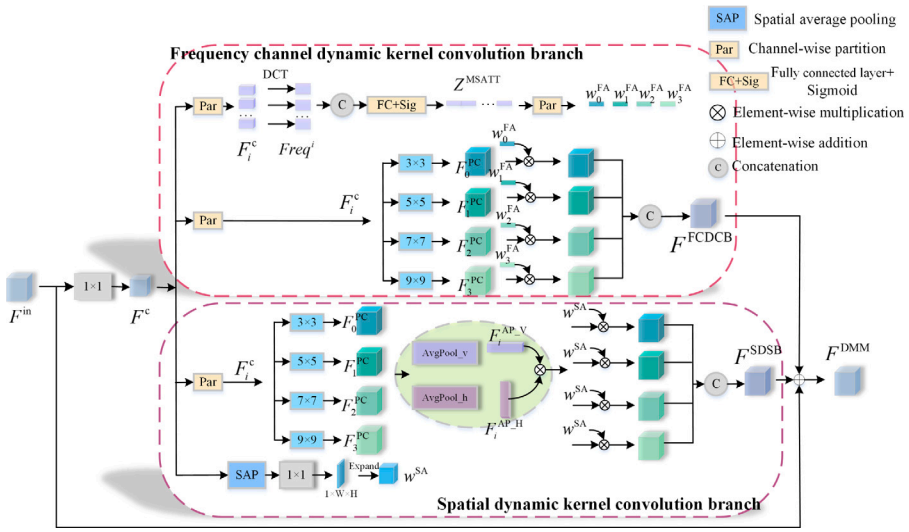


Fig. 3. Dynamic modulation module of auxiliary branch of our DRTNet in detail.

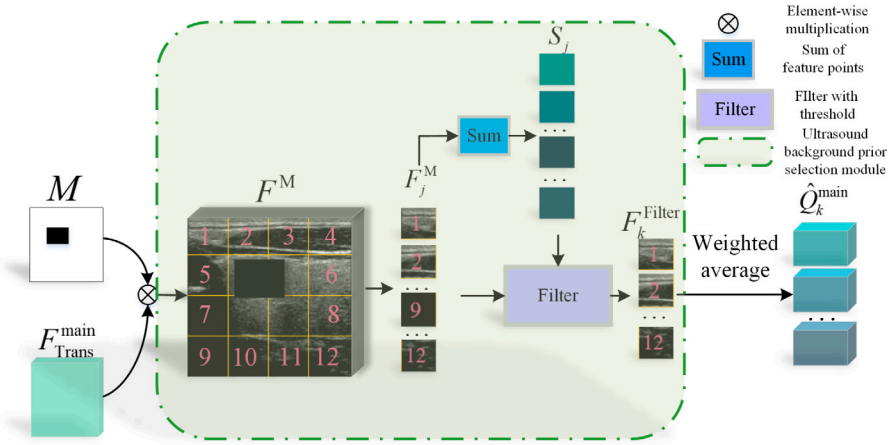


Fig. 4. Details of the key points in our UPBAP.

Furthermore, to enhance the attention in both spatial dimensions for the features, we perform average pooling operations in the vertical and horizontal directions to obtain $F_i^{\text{AP\_V}} \in R^{(C/4)\times w\times 1}$ and $F_i^{\text{AP\_H}} \in R^{(C/4)\times 1\times h}, (i = 0,1,2,3)$ [53]. The feature $F_i^{\text{AP\_VH}} \in R^{(C/4)\times w\times h}, (i = 0,1,2,3)$ of the strip spatial information enhancement is derived as follows:

$$F_i^{\text{AP\_VH}} = \left(AvgP\_v\left(F_i^{\text{PC}}\right)\right)\times \left(AvgP\_h\left(F_i^{\text{PC}}\right)\right), (i = 0,1,2,3) \tag{8}$$

where $AvgP\_v$ denotes vertical pooling and $AvgP\_h$ denotes horizontal pooling. Next, we perform a similar operation on the FCDCB to concatenate the features obtained by multiplying the attention weight $\boldsymbol{w}^{\text{SA}}$ with feature $F_i^{\text{AP\_VH}}$. Consequently, the feature distribution of objects in the spatial dimension can be dynamically adjusted. The output F^{SDCB} of the SDCB is represented as follows:

$$F^{\text{SDCB}} = Concat(w^{\text{SA}}\times F_0^{\text{AP\_VH}}, w^{\text{SA}}\times F_1^{\text{AP\_VH}}, w^{\text{SA}}\times F_2^{\text{AP\_VH}}, w^{\text{SA}}\times F_3^{\text{AP\_VH}}) \tag{9}$$

In summary, the output of the DMM is formulated as follows:

$$F^{\text{DMM}} = F^{\text{in}} + F^{\text{FCDCB}} + F^{\text{SDCB}} \tag{10}$$

3.1.2. UPBAP

To establish more desirable boundaries for the targets in the classification stage, the UPBAP, is developed to incorporate the background prior to ultrasound images into BAP. The additional filtering operations are added to specific region features based on the characteristics of thyroid ultrasound images. UPBAP not only inherits the excellent ability of BAP, but also filters the regions affecting the correlation calculation for the poor performance of the thyroid in ultrasound images. The inference of boundary regions is applied for the division of foreground and background regions. This integration is shown in Fig. 2.

(a) BAP

BAP considers the distinction between the foreground and background within the target Bbox as a retrieval task. In this context, the main branch serves as the BAP data flow introduction. Specifically, the transformer feature $F_{\text{Trans}}^{\text{main}}$ is divided into $N\times N$ grids; each grid feature is represented by $G(j), \left(1 \le j \le N^2\right)$. Then, we aggregate the features in each grid as queries for retrieval as follows:

$$Q_j^{\text{main}} = \frac{\sum_{p\in G(j)} M(p) F_{\text{Trans}}^{\text{main}}(p)}{\sum_{p\in G(j)} M(p)} \tag{11}$$

where $M(\bullet)$ represents the mask of the background outside the Bbox and $M(p) = 1$ is the position $p \in (1,2,\ldots,N^2)$ that falls outside any object boundary box B; otherwise, it is 0. Given the queries, we obtain the attention map for the background area within the Bbox as follows:

$$A(p) = \frac{1}{J}\sum_j A_j(p), (1 \le j \le N^2) \tag{12}$$

where J refers to the number of cells obtained by removing the cells outside the Bbox B that do not overlap with the background. The cosine similarity between the features within the Bbox and the query Q_j^{main} is then computed, and the resulting value is thereafter normalized to the range [0, 1] using the rectified linear unit (ReLU) function, yielding $A_j(p)$.

$$A_j(p) = \begin{cases} \text{ReLU}\left(\frac{F_{\text{Trans}}^{\text{main}}(p)}{\left\|F_{\text{Trans}}^{\text{man}}(p)\right\|} \cdot \frac{Q_j^{\text{main}}}{\left\|Q_j^{\text{main}}\right\|}\right), p \in B \\ 1, p \notin B \end{cases} \tag{13}$$

We then utilize an attention map A to obtain the foreground features R_n^{main} within the Bbox:

$$R_n^{\text{main}} = \frac{\sum_{p \in B_n} (1 - A(p)) F_{\text{Trans}}^{\text{main}}(p)}{\sum_{p \in B_n} (1 - A(p))} \tag{14}$$

Here, B_n indicates the nth target box in each feature map.

(b) UPBAP

By examining a series of thyroid ultrasound images, we have arrived at an intriguing observation: numerous black areas exist in the background region outside the boundary box of the thyroid nodules. Remarkably, these black areas exhibit close proximity to the pixels of the thyroid nodule objects, which can be attributed to the distinctive characteristics inherent in ultrasound imaging. Directly applying BAP incurs performance degradation because the black background outside the target Bbox increases the possibility of a black foreground.

Considering the stochastic nature of black regions outside the object Bbox, a simplistic region of interest (ROI) operation is inherently inadequate for effectively removing all background areas with similar pixel characteristics. Here, we propose the ultrasound background prior selection module as a solution to eliminate the need for extensive and laborious manual efforts. Details of the key points of our UPBAP are presented in Fig. 4.

In this context, the main branch also serves as the UPBAP data flow introduction. To further aggregate the background features outside of the Bbox, we represent the query for the background feature blocks as $\hat{Q}_k^{\text{main}}(1 \le k \le T)$ and the final number of selected grid cells as T. Specifically, the mask feature F^{M} is subjected to a grid division similar to BAP, resulting in $N \times N$ grids. The mask feature F^{M} is then obtained by multiplying the deep transformer feature $F_{\text{Trans}}^{\text{main}}$ with the mask M. The grid features $F_j^{\text{M}}(1 \le j \le N^2)$ outside the Bbox are thereafter aggregated by summing the pixel-wise features, resulting in the feature sum S_j of the corresponding region. Additionally, we set a threshold for the background grid regions outside the object Bbox and filtering grids with a significant amount of black regions.

Individual features after filtering are represented by $F_k^{\text{Filter}}(1 \le k \le T)$. The aforementioned threshold is determined by conducting multiple experimental tests, and the final selected threshold successfully excludes the black background areas. Subsequently, we perform a weighted average operation on $F_k^{\text{Filter}}(1 \le k \le T)$ to obtain individual features $\hat{Q}_k^{\text{main}}(1 \le k \le T)$ for the background regions. Compared with the original BAP approach, we achieved significant performance improvements by employing this simple region-based threshold selection operation. The abovementioned process can be formulated by the following equation:

$$F_k^{\text{Filter}} = Filter(F_j^{\text{M}})(1 \le k \le T, 1 \le j \le N^2) \tag{15}$$

$$\hat{Q}_k^{\text{main}} = WA(F_k^{\text{Filter}})(1 \le k \le T) \tag{16}$$

where $Filter(\bullet)$ means filter with threshold and $WA(\bullet)$ stands for weighted average.

3.1.3. Loss function of stage 1

For the foreground and background classification of stage 1, we employ conventional cross-entropy loss to optimize the prediction of the final softmax classification. During the training of the two-branch model, taking the main branch as an example, a two-class softmax classifier is applied to the features R_n^{main} and $\hat{Q}_k^{\text{main}}$ extracted from the foreground and background regions. $L_{\text{cls}}^{\text{main}}$ and $L_{\text{cls}}^{\text{auxiliary}}$ represent the supervision loss of the classification process in the main branch and auxiliary modulation branches, respectively. In addition, in order to maximize the CAMs of these two branches, we adopt a CAMs regularization supervision strategy. Supervised loss $L_{\text{rs-CAM}}$ is then applied to evaluate the consistency of the CAMs generated by both branches.

$$L_{\text{rs-CAM}} = \| M_{\text{main}} - M_{\text{auxiliary}} \|_1 \tag{17}$$

where M_{main} and $M_{\text{auxiliary}}$ refer to the CAMs generated by the main and auxiliary modulation branches, respectively. These mappings are obtained by multiplying the deep features of the corresponding transformer of the branch with the foreground class weights, followed by ReLU activation.

In summary, the total classification loss L_{ALL} in stage 1 is as follows:

$$L_{\text{ALL}} = L_{\text{cls}}^{\text{main}} + L_{\text{cls}}^{\text{auxiliary}} + L_{\text{rs-CAM}} \tag{18}$$

3.2. Pseudo mask generation

In stage 2 of pseudo mask generation, we leverage CAMs for each nodule Bbox obtained from the dual-route transformer classification network using UPBAP (Fig. 2). The compensation foreground CAMs $M_{\text{auxiliary}}^{\text{foreground}}$ obtained from the auxiliary branch contribute to the calibration of the main foreground CAMs $M_{\text{main}}^{\text{foreground}}$, enabling the generation of the final weighted CAMs $M_{\text{w}}^{\text{foreground}}$ demonstrated as follows:

$$M_{\text{w}}^{\text{foreground}} = Max(M_{\text{auxiliary}}^{\text{foreground}}, M_{\text{main}}^{\text{foreground}}) \tag{19}$$

where $Max(\bullet)$ represents the selection of the maximum value of pixels from two CAMs. Similar to the CAMs in stage 1, $M_{\text{auxiliary}}^{\text{foreground}}$ and $M_{\text{main}}^{\text{foreground}}$ are also generated by multiplying the deep features of the corresponding transformer of the branch with the foreground class weights, followed by a ReLU activation. Note that the deep features of the transformer in this stage must be restored to the original input size, which differs from that in stage 1.

The background CAMs $M_{\text{main}}^{\text{background}}$ are then acquired by multiplying the deep feature of the transformer of the main branch with the background weights, followed by a ReLU operation. Subsequently, the weighted CAMs $M_{\text{w}}^{\text{foreground}}$ and background CAMs $M_{\text{main}}^{\text{background}}$ are concatenated along the channel dimension to define a unary term for DenseCRF. Ultimately, the optimization of DenseCRF culminates in the generation of the final pseudo mask F^{PM}. The calculation of the abovementioned process is performed as follows:

$$F^{\text{PM}} = DenseCRF(Concat(M_{\text{w}}^{\text{foreground}}, M_{\text{main}}^{\text{background}})) \tag{20}$$

3.3. AUEMSC

In stage 3, considering that the pseudo masks are noisy, it is deemed inappropriate to directly apply a segmentation model [54]. Based on the semi-supervised literature [46,55], we average the model outputs corresponding to multiple-scale parallel samples to achieve label correction. Moreover, we incorporate consistency loss as regularization to provide additional supervisory signals to the network, thereby preventing over-fitting to the noisy regions of the labels. Furthermore, we introduce adaptive feature fusion [56] to further enhance the parallel multi-scale model outputs and achieve adaptability.

Specifically, as shown in Fig. 2, considering that Liu [56]'s selection of these three scaling parameters resulted in excellent experimental

results, we set the sample scales of the three transformer segmentation models to 0.7, 1, and 1.5. These three parameters represents the scaling parameters of the parallel three models, as shown by the "Downscale" and "Upscale" in Fig. 2. Additionally, ablation experiments are conducted on these three parameters to confirm the reasons for their selection. The outputs of the three branches are thereafter restored to the same size and concatenated in the channel dimension. After reducing the convolutional dimension, the features are divided into weighted weights corresponding to the scale of the feature along the channel dimension. The weighted coefficients (w_{large}, w_{ori}, w_{small}) for the feature maps at different scales are defined as follows:

$$w_{\text{large}} = \text{SP}_{c1}(Conv_{1\times1}(Concat(F_{\text{large}}, F_{\text{ori}}, F_{\text{small}}))) \tag{21}$$

$$w_{\text{ori}} = \text{SP}_{c2}(Conv_{1\times1}(Concat(F_{\text{large}}, F_{\text{ori}}, F_{\text{small}}))) \tag{22}$$

$$w_{\text{small}} = \text{SP}_{c3}(Conv_{1\times1}(Concat(F_{\text{large}}, F_{\text{ori}}, F_{\text{small}}))) \tag{23}$$

The model's outputs of the corresponding proportional samples are bilinearly interpolated to maintain the original size, and then F_{large}, F_{ori}, F_{small} are obtained. Moreover, SP_{c1}, SP_{c2}, SP_{c3} stand for the operation of partitioning by channel, and $Conv_{1\times1}$ denotes 1×1 convolution. The output of the adaptive fusion operation F^{AF} is expressed as follows:

$$F^{AF} = w_{\text{large}} \times F_{\text{large}} + w_{\text{ori}} \times F_{\text{ori}} + w_{\text{small}} \times F_{\text{small}} \tag{24}$$

Ultimately, we construct the regularization term loss L_{consis} to ensure the consistency between the outputs of the different scale models and the adaptive fused output F_m. The corresponding loss is formulated as follows:

$$L_{\text{consis}} = -\frac{1}{3}\sum_{k=1}^{3} KL(F_m \parallel F^{\text{AF}}) \tag{25}$$

where $KL(\bullet)$ refers to the Kullback–Leibler divergence; F_m corresponds to $F_{\text{large}}, F_{\text{ori}}, F_{\text{small}}$; and k represents the corresponding index of different scale models.

Based on the dynamic channel attention weights, the multi-scale label supervision loss L_{msls} is composed of the standard cross-entropy loss and denoted as follows:

$$\begin{aligned} L_{\text{msls}} = 1/3[&\varphi_{ce}(w_{\text{large}} \times F_{\text{large}}, F^{\text{PM}}) + \\ &\varphi_{ce}(w_{\text{ori}} \times F_{\text{ori}}, F^{\text{PM}}) + \\ &\varphi_{ce}(w_{\text{small}} \times F_{\text{small}}, F^{\text{PM}})] \end{aligned} \tag{26}$$

where φ_{ce} indicates cross-entropy loss and F^{PM} means pseudo mask. The consistency correction loss and the cross-entropy loss are weighted through λ_c to obtain the final loss L_{SUM}, which is defined as follows:

$$L_{\text{SUM}} = L_{\text{msls}} + \lambda_c * L_{\text{consis}} \tag{27}$$

4. Experiments

4.1. Datasets

To evaluate our model, we use two public datasets, the TN3K and DDTI datasets, and our own private thyroid dataset.

(1) **TN3k dataset**: The TN3k dataset comprises 2879 grayscale ultrasound images collected from 2421 patients. In this dataset, 2303 images were selected as the training set and the remaining 576 images as the testing set. The images and labels are all resized to 512×512.

(2) **DDTI dataset**: The DDTI dataset includes 637 ultrasound thyroid images with pixel-wise labels from a single device. Similarly,in this dataset, 509 images were selected as the training set and the remaining 128 images as the testing set. To remove patient privacy and other irrelevant information, we crop the images and labels to 512×512.

(3) **Our own thyroid dataset**: Our own thyroid dataset consists of 1426 ultrasound images collected from the Affiliated Changzhou No. 2 People's Hospital of Nanjing Medical University (No.2020_KY146-01). Images were acquired using three commercial scanners from Philips Healthcare (Best, Netherlands); Siemens Healthineers (Erlangen, Germany); GE Healthcare (Chicago, USA).

We adopt 1141 samples as training sets and 285 as testing sets. Here, we divide the datasets in a ratio of 8:2 for training and testing. Delineation of the nodules was performed by three physicians with extensive clinical experience in the ultrasound department. Here, we convert contours to contours and binary masks as ground truths. Similarly, the images and labels are cropped to 512×512.

Regarding the details of the private dataset, we classify the features based on morphology, margin, aspect ratio (A/T), echogenicity, calcification, and cystic component. For morphology, 87 cases exhibited regular shapes, 451 cases displayed irregular shapes, and 888 cases had undefined shapes. Regarding margin clarity, 127 cases demonstrated clear margins, 742 cases showed unclear margins, and 557 cases presented indistinct or blurry margins. Regarding the A/T ratio, 589 cases featured a ratio less than or equal to 1, whereas 837 cases displayed a ratio greater than 1. Echogenicity analysis revealed no echogenic cases, 14 cases with isoechoic or hyperechoic features, 1272 cases with hypoechoic features, 113 cases with markedly hypoechoic features, and 27 cases with mixed echogenicity. Calcification assessment showed 57 cases with coarse calcifications, 8 cases with eggshell calcifications, and 948 cases with microcalcifications. Regarding the cystic component, no purely cystic cases were identified; meanwhile, 14 cases exhibited solid components, and 1412 cases displayed mixed cystic and solid components. This classification scheme was applied to the dataset for comprehensive scientific analysis.

4.2. Evaluation metrics

We employ the following metrics to evaluate the segmentation performance of the framework: intersection over union (IoU), dice coefficient score (DSC), Hausdorff distance (HD), F1-score (F1), accuracy, and area under the receiver operating characteristic curve (AUC). The metrics are calculated as follows:

$$IoU = \frac{Y \cap P}{Y \cup P} \tag{28}$$

$$DSC(A, B) = \frac{2|Y \cap P|}{|Y| + |P|} \tag{29}$$

$$HD95(Y, P) = Max\left\{\begin{matrix} \underset{x\subset A}{sup}\ \underset{y\subset B}{inf}\|x - y\|, \\ \underset{y\subset B}{sup}\ \underset{x\subset A}{inf}\|x - y\| \end{matrix}\right\} \tag{30}$$

where Y and P are the segmentation results and GT, respectively, and x and y are the voxels in Y and P, respectively. IoU and DSC are sensitive to thyroid nodule areas, while HD95 is sensitive to the shape.

$$Recall = \frac{TP}{TP + FN}, Precision = \frac{TP}{TP + FP} \tag{31}$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{32}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{33}$$

where TP, FP, FN, TN are true positive, false positive, false negative, and true negative, respectively.

4.3. Implementation details

We train the proposed framework using PyTorch 1.10.0 and a server equipped with an NVIDIA TITAN 3080Ti GPU. All experiments of the TN3k dataset are conducted under five-fold cross-validation to provide a more extensive evaluation. Given the limited size of the DDTI dataset and our own dataset, conducting five-fold cross-validation could result in a low number of samples per fold, potentially compromising the accuracy and reliability of the model. Therefore, we validate the model on a randomly partitioned test set.

Table 1
Quantitative results of improvements in our framework on the TN3k dataset.

Model	IoU (%)	DSC (%)	HD95	F1 (%)	ACC (%)	AUC (%)
Supervision:Bbox labels						
WSSL [57]	66.17 ± 0.71	79.64 ± 0.52	42.00 ± 1.89	78.76 ± 0.56	94.23 ± 0.07	89.63 ± 0.41
BoxSup [58]	66.49 ± 0.78	79.87 ± 0.57	41.01 ± 2.33	78.95 ± 0.46	94.22 ± 0.14	90.14 ± 0.49
SDI [15]	67.14 ± 1.19	80.33 ± 0.85	38.95 ± 2.45	79.77 ± 0.74	94.67 ± 0.16	87.68 ± 0.79
BCM [59]	67.25 ± 0.64	80.42 ± 0.46	36.61 ± 1.83	78.19 ± 1.19	94.46 ± 0.22	88.15 ± 1.78
BANA(VGG) [17]	67.85 ± 0.40	80.85 ± 0.29	48.03 ± 3.42	77.88 ± 0.59	94.29 ± 0.07	88.93 ± 0.24
BANA(SegFormer) [17]	68.48 ± 0.96	81.29 ± 0.67	38.05 ± 1.77	80.53 ± 0.57	94.64 ± 0.12	89.43 ± 0.42
BCM(FRS) [26]	69.73 ± 0.57	82.16 ± 0.40	33.67 ± 0.56	79.48 ± 0.24	94.97 ± 0.16	88.61 ± 0.78
DRTNet[a]	**73.82 ± 0.47**	**84.94 ± 0.32**	**27.69 ± 0.84**	**83.31 ± 0.14**	**95.74 ± 0.20**	**91.32 ± 0.37**
Supervision: Pixel-level labels						
SegFormer [60]	73.04 ± 0.26	84.42 ± 0.17	35.70 ± 2.93	82.90 ± 0.46	95.72 ± 0.09	89.53 ± 0.37

[a] The best results are highlighted in bold.

Fig. 5. Five-fold cross-validation result chart on the TN3k dataset.

(1) Classification procedure for stage 1

We adopt the backbone network from SegFormer (MiT) [60] with ImageNet-1K pretraining weight as the classification backbone of stage 1. The classification network is trained for 24,000 iterations with a batch size of 2 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 then decayed using a warm-up cosine annealing algorithm. Here, we augment the training set with horizontal flipping, random cropping, random scaling, and color jittering.

(2) Segmentation procedure for stage 3

We exploit SegFormer for semantic segmentation with ImageNet-1K pretraining weight as the segmentation network of stage 3. Following a learning rate strategy similar to that of the classification network, we train the segmentation network for 23,805 iterations using a batch size of 4. The image augmentation strategy also comprises horizontal flipping, random cropping, random scaling, and color jittering. The training image size is then randomly cropped to 256, and the test size remains as 512.

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

We compare our proposed DRTNet with several existing box-level supervised semantic segmentation methods, including WSSL [57], BoxSup [58], SDI [15], BCM [59], BANA [17] and BCM(FRS) [26]. Full-supervised SegFormer [60] is also included to compare the differences between box labels and pixel-level labels. Here, we also compare the proposed DRTNet with the aforementioned box-level supervised semantic segmentation methods. For a fair comparison, except for BCM, all

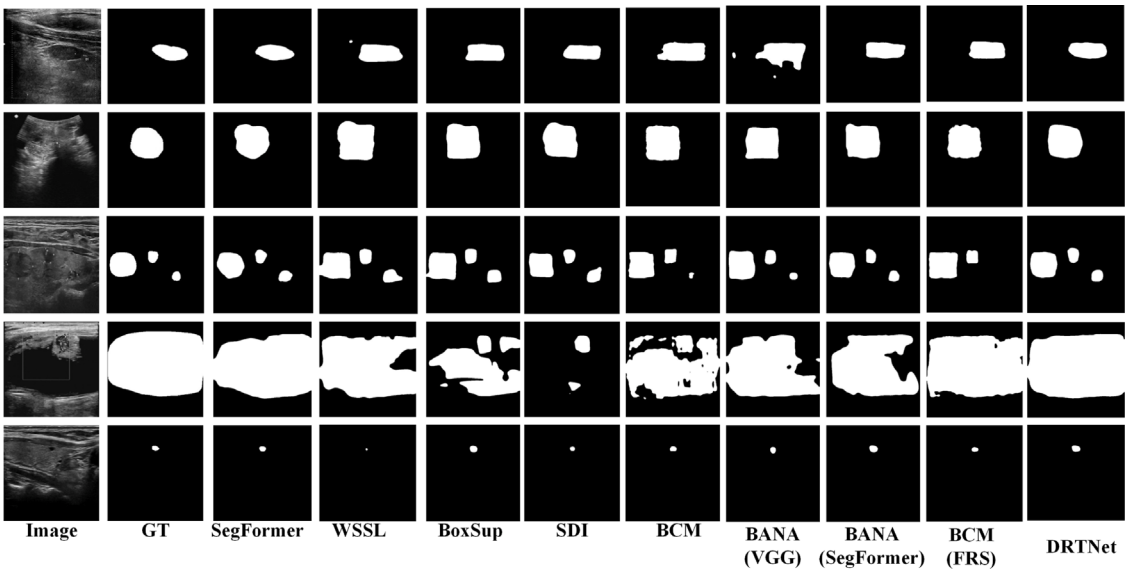


Fig. 6. Segmentation results on the TN3k dataset.

Table 2
Quantitative results of improvements in our framework on the DDTI dataset.

Model	IoU (%)	DSC (%)	HD95	F1 (%)	ACC (%)	AUC (%)
Supervision:Bbox labels						
WSSL [57]	62.39	76.84	45.32	76.41	92.73	84.98
BoxSup [58]	62.44	76.88	46.08	73.35	93.00	82.08
SDI [15]	65.66	79.27	41.19	75.73	92.72	87.59
BCM [59]	63.14	77.41	37.10	73.19	93.25	81.72
BANA(SegFormer) [17]	67.50	80.59	40.51	77.66	93.41	86.97
BCM(FRS) [26]	66.88	80.15	34.98	79.42	93.00	90.19
DRTNet[a]	**72.38**	**83.98**	**29.18**	**80.77**	**94.25**	**91.13**
Supervision: Pixel-level labels						
SegFormer [60]	69.92	82.30	37.55	79.43	94.41	86.20

[a] The best results are highlighted in bold.

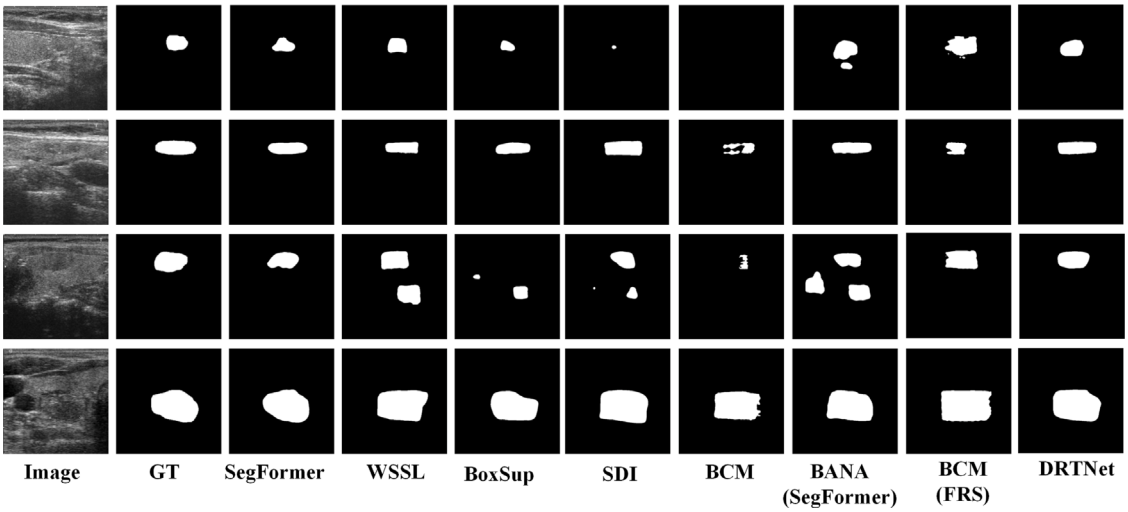


Fig. 7. Segmentation result on the DDTI dataset.

box-level supervised segmentation schemes all apply SegFormer without regularization. Moreover, BCM and BCM(FRS) employ DeepLab V2 (ASPP) with pretrained ResNet-101.

4.4.1. Results on the TN3k dataset

Because of the numerous indicators, we mainly compare multiple methods for representative indicators such as DSC, HD95, F1, ACC, and AUC scores. The quantitative results of the comparison experiments conducted on the TN3k dataset are presented in Table 1.

Remarkably, the proposed DRTNet with the improved regularization scheme reaches 73.82% on IoU, 84.94% on DSC, 27.69 on HD95, 83.31% on F1, 95.74% on ACC, and 91.32% on AUC, showing that our proposed DRTNet achieves superior segmentation performance on evaluation metrics compared with other aforementioned schemes.

The performance of the WSSL segmentation method is the lowest among all the comparison methods, as shown in Table 1. This phenomenon highlights that relying solely on DenseCRF to address all regions within the Bbox is insufficient, and the generated pseudo masks are inherently prone to inaccuracies.

Several pseudo label generation schemes have been implemented to improve the quality of pseudo masks. We conduct experiments to compare our approach with these representative methods. SDI introduces pseudo masks for recursive training with a slight improvement. BoxSup uses the GrabCut pseudo mask generation method, which is 0.48%, 0.28%, 0.19%, and 0.56% higher than WSSL on IoU, DSC, F1, and AUC, respectively. The BCM employs box-driven class-wise masking, which can implement spatial masking on the features of each class. The BCM segmentation results, as summarized in Table 1, also show the feasibility of corresponding operations on the ultrasound thyroid dataset. BCM (FRS) is based on BCM and introduces an anchor-based filling rate movement strategy to refine the filling rate of thyroid nodule objects. By shifting the fill rate, the negative impact of mislabeled pixels can be effectively reduced. Compared with BCM, both DSC and HD95 indicators have shown significant improvement. Based on these results, we can observe that pseudo label generation schemes can improve the quality of pseudo masks and segmentation accuracy.

Other methods employ pseudo masks generated through foreground–background classification, as shown in the last four rows of Table 1. BANA (VGG) is the method used in the original study using pretrained VGG16 DeepLab-V1 (LargeFOV) forto segmentation. Our comparison with BANA (SegFormer) suggests that the global modeling ability of a transformer is effective in thyroid ultrasound image segmentation.

Furthermore, we compared DRTNet with the SegFormer of pixel-level label and found that our DRTNet achieved 0.78% IoU, 0.52% DSC, 8.01 HD95, and 0.41% F1 improvements. From these results, we can see that the proposed DRTNet achieves the most accurate thyroid nodule segmentation results.

The median values of our models are in the leading position, as shown in the box diagram in Fig. 5. Concurrently, the range between the upper and lower quartiles is also relatively small, indicating the adaptability of our model in handling discrete data distributions and substantial data disparities.

We also evaluate the proposed DRTNet from a visual perspective. Fig. 6 shows the evidence that the predictions of our model are more consistent with the GT as compared to previous methods and can rival the performance of a fully supervised SegFormer. As shown in Fig. 6, we can see that the segmentation result is relatively ideal because of the applicability of the BAP of BANA in ultrasound thyroid images. By visualizing the segmentation results in Fig. 6, we can observe that BAP yields more consistent boundaries as compared to previous classical methods. This improvement can be attributed to the utilization of background features outside the thyroid nodule Bbox by BAP, effectively enhancing the discrimination between the foreground and background. Notably, the application of the BAP in ultrasound images interferes with abnormal pixel values. Our model has conducted in-depth analysis on this aspect and improves IoU by 2.03%, DSC by 1.41%, HD95 by 2.27, and F1 by 0.72% under the same segmentation strategy. We argue that this is because of the utility of our UPBAP, as it eliminates interfering characteristics outside the Bbox of the ultrasound thyroid nodule objects.

Moreover, the exceptional perceptual capability of the proposed DRTNet can detect accurate nodules of various sizes and thus provide compelling evidence that gives credit to our DMM.

4.4.2. Results on the DDTI dataset

To demonstrate the robustness of the proposed framework on different datasets, we conduct another comparison experiment on the DDTI dataset. Noted that the irrelevant regions of the thyroid images in this dataset are manually cropped, which is different than that done in the TN3K dataset. Unrelated regions represent regions with black background and various ethical information, rather than global information on the surrounding thyroid tissue area. The quantitative obtained on the DDTI dataset are shown in Table 2, the proposed framework performs consistently on different datasets and excels in numerous evaluated metrics.

The quantitative results indicate that DRTNet can reach 83.98% on DSC and 29.18 on HD95 scores. The experimental results are slightly lower than those obtained on TN3K, which can be attributed to the smaller dataset size of DDTI and the increased presence of ultrasound noise interference.

Notably, the performance of our non-regularized framework is comparable with that of fully supervised SegFormer. This observation sufficiently demonstrates the effectiveness of our dual-branch transformer modulation mechanism and UPBAP on the DDTI dataset.

Among the weakly supervised segmentation methods, only BCM and BCM(FRS) obtain relatively ideal HD95 results. This phenomenon can be attributed to box driven regularization operations and fill rate shifting strategies, which compensate for the limitations of pseudo masks. This finding highlights the positive impact of regularization on segmentation with noisy label supervision, moreover, our improved regularization strategy demonstrates superior performance.

The performance of our framework on different-sized thyroid nodule objects in the DDTI dataset is demonstrated through the visualized segmentation results in Fig. 7. It can be observed that our framework exhibits predictions that are more consistent with the GT as compared to existing weakly supervised segmentation methods. This highlights its outstanding segmentation capabilities on ultrasound thyroid datasets with more complex target noise.

Moreover, the misidentification of irrelevant regions is demonstrated by the error prediction examples of the existing models in the third row of Fig. 7. In contrast, neither of our frameworks is affected by similar regions but rather accurately segments the nodule area.

4.5. Ablation study and analysis

In this section, we demonstrate the effectiveness of the principal components of our DRTNet, i.e., transformer backbone, modulation compensation of two branches, UPBAP, and adaptive uncertainty estimation multi-scale consistency. The experiments of this ablation study are mainly conducted on the public ultrasound thyroid dataset TN3k.

4.5.1. Effect of transformer backbone

To verify the superior capability of the transformer over CNN, we compare the quality of CAMs generated by single-branch networks and the segmentation performance of pseudo masks in stage 3. The CNN and transformer backbone are pretrained VGG16 [61] and MiT with ImageNet1k pretraining weights, respectively. The visualization results shown in Fig. 8 demonstrate that CAMs generated with transformers activate more integral regions than CNNs owing to the effective improvement brought about by the self-attention mechanism of transformers in modeling and exploring global information, which addresses the issue of incomplete target regions captured by CNNs. Further, we train the third stage using the generated pseudo masks, and the quantitative results of the five-fold cross-validation are summarized in Table 3. The introduction of the transformer backbone assists the network in achieving 1.09%, 0.74%, 0.99, and 0.7% improvement in IoU, DSC, HD95, and F1, respectively. Despite the higher coverage of regions activated by a single transformer network compared with CNN, the CAMs visualizations in rows 1, 2, 3, and 5 of Fig. 8 still suffer from region neglect. This phenomenon proves that implementing a dual-branch modulation compensation strategy is effective.

Table 3
Quantitative results of two different backbones in a single-path method on the TN3k dataset.

Model	IoU (%)	DSC (%)	HD95	F1 (%)
CNN backbone	71.82 ± 1.51	83.59 ± 1.03	29.32 ± 2.37	82.06 ± 1.09
Transformer backbone[a]	**72.91 ± 0.54**	**84.33 ± 0.36**	**28.33 ± 0.76**	**82.76 ± 0.19**

[a] The best results are highlighted in bold.

Table 4
Quantitative results of different modules on the auxiliary branch in our method on the TN3k dataset.

Model	IoU (%)	DSC (%)	HD95	F1 (%)
Single branch	72.91 ± 0.54	84.33 ± 0.36	28.33 ± 0.76	82.76 ± 0.19
Auxiliary branch+AMR	73.30 ± 0.45	84.59 ± 0.30	28.23 ± 0.50	82.92 ± 0.32
Auxiliary branch+Our DMM[a]	**73.82 ± 0.47**	**84.94 ± 0.32**	**27.69 ± 0.84**	**83.31 ± 0.14**

[a] The best results are highlighted in bold.

Table 5
Quantitative results of different channel attention weights in our method on the TN3k dataset.

Model	IoU (%)	DSC (%)	HD95	F1 (%)
Standard channel attention weight	73.19 ± 0.33	84.52 ± 0.22	28.58 ± 0.75	83.08 ± 0.12
DWT channel attention weight	73.40 ± 0.40	84.66 ± 0.26	27.98 ± 0.70	82.87 ± 0.35
DFT channel attention weight	73.43 ± 0.24	84.68 ± 0.16	28.14 ± 1.10	83.09 ± 0.26
DCT channel attention weight[a]	**73.82 ± 0.47**	**84.94 ± 0.32**	**27.69 ± 0.84**	**83.31 ± 0.14**

[a] The best results are highlighted in bold.

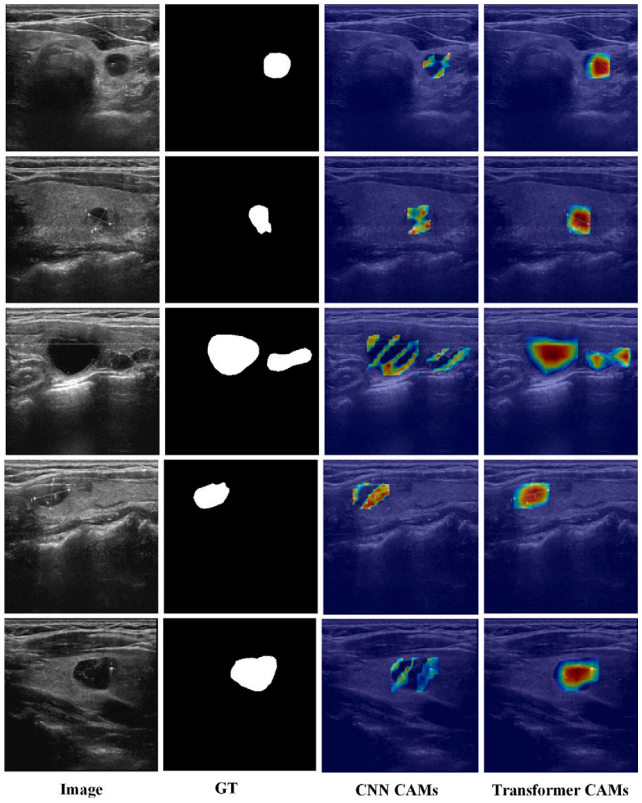


Fig. 8. Visual CAMs results of two different backbones in a single-path method on the TN3k dataset.

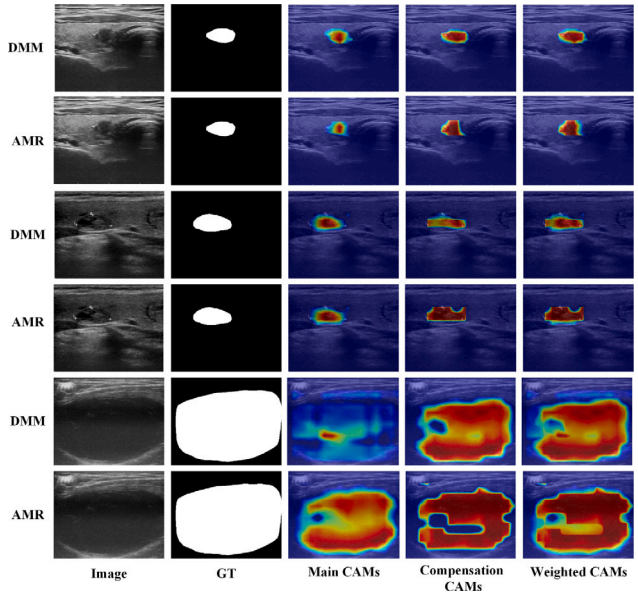


Fig. 9. Visual results of different modules on the auxiliary branch of the TN3k dataset.

4.5.2. *Effect of modulation compensation of two branches*

To verify that our design of dual-branch modulation compensation improves the experimental results, we present the segmentation indicators in Table 4 and the visual CAMs in Fig. 9. The advantages of the dual-branch architecture are displayed in Table 4. The introduction of the auxiliary branch, whether DMM or AMR, leads to significant improvements in all metrics as compared to the single-branch counterparts. In particular, the proposed DRTNet can clinch superior performance in all segmentation evaluation metrics, scoring 73.82% in IoU, 84.94% in DSC, 27.69 in HD95, and 83.31% in F1. This result may stem from the fact that DMM can assist the main branch in dynamically adjusting the feature distribution of objects of different sizes, whether large or small. Alternatively, AMR adopts the Gaussian function to redistribute the activation value of feature mapping to compensate for the defect of single-branch CAMs. Although the effectiveness of AMR modulation has been confirmed by evaluation indicators, visual CAMs demonstrate that AMR is at the end of its rope when faced with thyroid nodule objects with large size differences. Conversely, our DMM can handle this challenge with skill and ease. Lines 1–4 in Fig. 9 are CAM cases of smaller objects, while lines 5 and 6 are CAM cases of larger objects. Among them, the weighted CAMs of smaller objects based on DMM are closer to the GT visually. CAMs comparison diagram of the bottom two lines verifies that our DMM helps the main branch in activating more responses while the auxiliary

Table 6
Quantitative results of different pooling methods in our method on the TN3k dataset.

Model	IoU (%)	DSC (%)	HD95	F1 (%)
GAP	71.52 ± 0.63	83.39 ± 0.42	29.79 ± 1.21	81.52 ± 0.64
BAP	73.14 ± 0.17	84.49 ± 0.11	27.96 ± 0.95	82.59 ± 0.42
UPBAP(−300)	73.04 ± 0.54	84.42 ± 0.37	28.35 ± 1.07	82.65 ± 0.29
UPBAP(−320)	72.95 ± 0.25	84.38 ± 0.19	28.01 ± 0.87	82.78 ± 0.12
UPBAP(−336)	73.58 ± 0.20	84.78 ± 0.16	27.97 ± 0.32	82.98 ± 0.47
UPBAP(−342)	73.30 ± 0.30	84.59 ± 0.20	28.27 ± 1.23	82.54 ± 0.44
UPBAP(−330)[a]	**73.82 ± 0.47**	**84.94 ± 0.32**	**27.69 ± 0.84**	**83.31 ± 0.14**

[a] The best results are highlighted in bold.

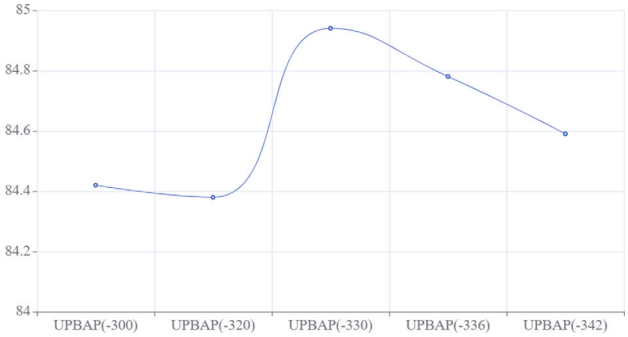


Fig. 10. Sensitivity analysis of UPBAP threshold parameters.

branch adds attention to the areas ignored by the main branch. The activation of the middle region is missing from the AMR results in the last row of Fig. 9, and our DMM aids in promptly rectifying this issue.

The reason for using frequency channel attention instead of standard channel attention in this study is mainly attributed to the characteristics of ultrasound images. The weight of the standard channel attention adopts GAP, which is a mean operation, similar to using only the simplest spectrum. Using separate GAP information on ultrasound images may miss other frequency components, resulting in the loss of important ultrasound features. Therefore, we adopt multi-spectral frequency channel attention weights to capture these features. The relevant ablation experiments based on DCT are shown in Table 5. We can observe that the scheme using standard channel attention weights performs poorly in evaluating metrics as compared to the scheme using DCT, which precisely confirms our hypothesis.

Both DFT [62] and DCT can examine images in the frequency domain, therefore, we replaced DCT with DFT in the DRTNet architecture. Despite the DFT channel attention weights providing more comprehensive frequency information (amplitude and phase), results shown in Table 5 indicate that an excessive focus on low-frequency global and high-frequency local features does not lead to performance improvements. This phenomenon is attributed to the overlapping effects of using them as weighted coefficients for multi-scale convolutional features, resulting in a performance decline. In contrast, the DCT channel attention weights exhibit a favorable energy concentration property, effectively enabling the attention mechanism to focus on ultrasound features. Applying them to multi-scale convolution operations yielded results consistent with our expectations.

Moreover, DWT [63] decomposes the original input into approximation (LL), horizontal detail (LH), vertical detail (HL), and diagonal detail (HH) subbands. Aggregating these subband components allows the consideration of high-frequency and low-frequency information. Experimental results have similarly confirm that the inferior performance of DFT features as weights compared with DCT can be attributed to this phenomenon.

4.5.3. Effect of the UPBAP module

The introduction of the UPBAP involves harvesting more detailed boundary areas in the foreground–background classification. Thus, to provide insight into the utility of the module we designed, UPBAP is compared with GAP and BAP on evaluation metrics, visual CAMs, and pseudo masks. Table 6 shows that after replacing GAP with BAP, the model achieves 1.62%, 1.10%, and 1.07% improvements in IoU, DSC, and F1, respectively. The HD95 index, which evaluates boundary delineation, shows a reduction of 1.83 when comparing BAP with GAP. This finding validates the effectiveness of utilizing background information outside the Bbox for foreground–background separation, resulting in improved delineation of object boundaries.

Fig. 12 presents some cases of thyroid pseudo masks of the nodule with blurred edges, wherein the performance of GAP on thyroid ultrasound images is less satisfactory. Its extensive coverage range results in the loss of shape description. Conversely, the pseudo masks generated by BAP exhibit a closer resemblance to GT, indicating its capability to achieve more precise boundaries compared to GAP.

Moreover, Table 6 illustrates that the segmentation results increase by 0.68% on IoU, 0.45% on DSC, and 0.72% on F1 when UPBAP is considered in our framework instead of BAP. Simultaneously, HD95 achieves a performance improvement of 0.27. Given that certain regions of the background outside the Bbox in some thyroid ultrasound images have pixel values like those of the nodules, we implemented this improvement to address this concern. Three examples of background regions in proximity to nodule pixels are displayed in Fig. 11. Neglecting areas of pixel proximity inevitably bring about performance degradation, and simple ROI operations to delete related areas inevitably lead to an increase in workload. Conversely, our background prior selection operation is relatively easy to implement. Apart from evaluating indicators, we also demonstrate the advantages of pseudo masks and CAMs for visual analysis. From the corresponding CAMs and pseudo masks in columns 4 and 7, the ultrasound background prior added based on BAP removes the influence of similar pixels and obtains more outstanding pseudo masks.

The threshold ablation study of the ultrasound background selection module is shown in Table 6 and Fig. 10 . Finding an optimal threshold is crucial for achieving desirable performance because excessively high and low thresholds can induce sub-optimal results or even a declined performance. In the sensitivity analysis of UPBAP thresholds, we first conducted model inference without feature block filtering and computed the feature values in the black areas outside the bounding boxes, obtaining a value of −330. Then, to improve the robustness, the feature values are scaled proportionally to −300 to −342, and a range of sample values within this scaled interval were selected for comparative efficiency testing. Subsequently, multiple values within this range were chosen for efficiency comparison experiments, and the maximum value was estimated by fitting a curve. Based on the relevant experiments, the threshold was ultimately determined to be -330.

4.5.4. Effect of the AUEMSC module

The quantitative results of the ablation experiments regarding improvements in stage 3 in our DRTNet are shown in Table 7 and Fig. 12. Various modifications use the SegFormer segmentation scheme without regularization in the second row of the table as the baseline. For the weighted weights of the three SegFormer branch outputs, we investigate three options (i.e., average weight, trainable weight,

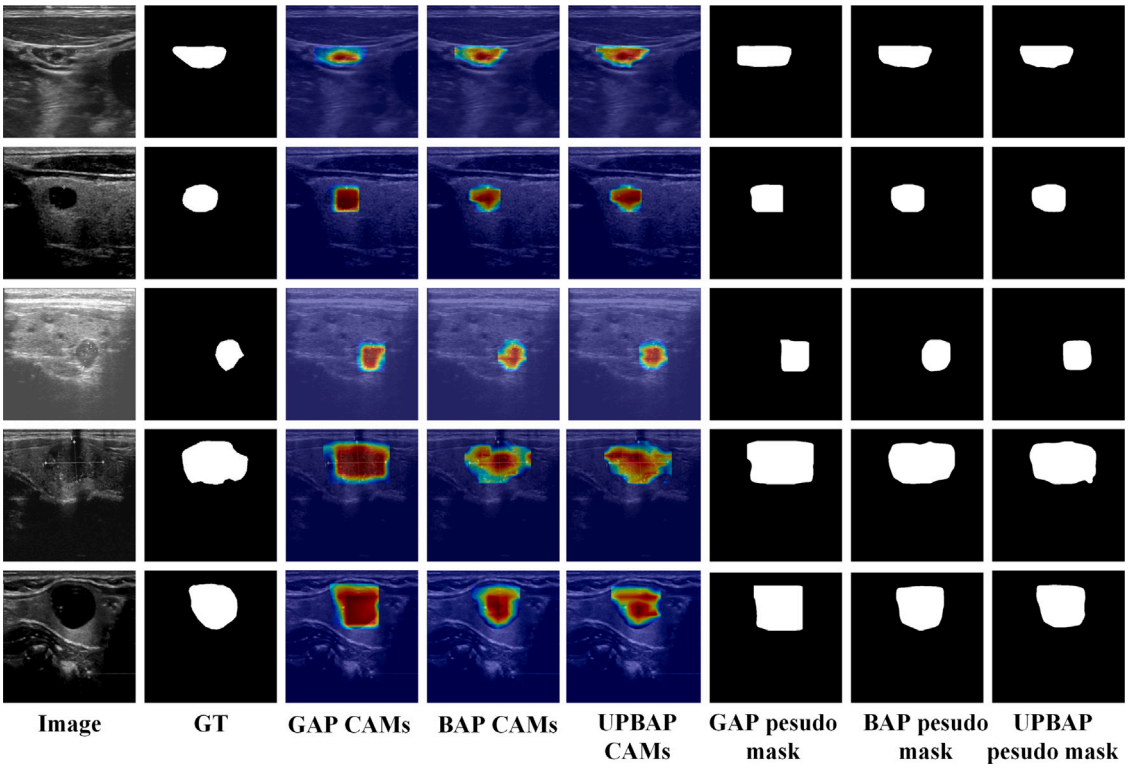


Fig. 11. Visual results of different pooling methods on the TN3k dataset.

Table 7
Quantitative results of different regularization methods on the TN3k dataset.

Model	IoU (%)	DSC (%)	HD95	F1 (%)
No regularization	70.51 ± 0.52	82.70 ± 0.36	35.78 ± 1.75	81.25 ± 0.50
Average weighted addition	72.87 ± 0.35	84.31 ± 0.24	29.59 ± 1.58	82.84 ± 0.29
Weighted addition of trainable weights	73.44 ± 0.42	84.68 ± 0.27	28.78 ± 1.20	82.97 ± 0.42
Our adaptive weighted addition[a]	**73.82 ± 0.47**	**84.94 ± 0.32**	**27.69 ± 0.84**	**83.31 ± 0.14**

[a] The best results are highlighted in bold.

Table 8
Quantitative results of different regularization parameters in our method on the TN3k dataset.

Scaling Parameters	IoU (%)	DSC (%)	HD95	F1 (%)
0.8,1,1.2	71.67 ± 0.39	83.50 ± 0.26	32.95 ± 1.02	82.31 ± 0.50
0.8,1,1.4	72.56 ± 0.66	84.09 ± 0.44	29.13 ± 1.09	82.33 ± 0.37
0.6,1,1.6	73.39 ± 0.48	84.65 ± 0.32	27.60 ± 1.09	82.93 ± 0.29
0.7,1,1.5[a]	**73.82 ± 0.47**	**84.94 ± 0.32**	**27.69 ± 0.84**	**83.31 ± 0.14**

[a] The best results are highlighted in bold.

and adaptive weight). No matter which weight regularization strategy is added, the improvement compared with the baseline method is tremendous. Therefore, introducing regularization consistency from the semi-supervised aspect for segmentation with noisy masks is essential. Our adaptive weighting method further promotes multi-scale consistency and significantly improves each evaluation metric as compared to the other two weighting methods (lines 3 and 4) in Table 7. Particularly, for the average weighted method, the adaptive scheme yields 0.95% IoU, 0.63% DSC, 1.90 HD95, and 0.47% F1 metric improvements, which is attributed to the excellent ability to maintain consistent output of multiple-scale models through channel dimension characteristics as weights.

Table 8 shows that using smaller-scale transformation parameter values undoubtedly has poor performance, which is attributed to the fact that scale-based regularization relies on the differences in scales between multiple images. In contrast, excessive scale differences not only lead to performance degradation but also increase computational complexity.

5. Discussion

The proposed DRTNet is a weakly supervised method based on Bbox for thyroid ultrasound image segmentation. DRTNet aims at achieving precise nodules and releasing over-dependent on pixel-level label for medical ultrasound fields. Transformer double-branch modulation, UPBAP and AUEMSC of DRTNet are applied to Better extraction of thyroid nodule features. It is validated on two public dataset and one private dataset. The experimental results show that DRTNet outperform other weakly-supervised methods and perform equivalent to pixel-level labeled. Finally, we will elaborate on the shortcomings of this work and discuss recent and future related work.

Segmentation based on weak annotations, such as Bbox, can significantly reduce the annotation burden on medical professionals, thereby enhancing diagnostic efficiency. Although numerous box-supervised segmentation methods have achieved significant success on natural images, designing an approach that performs well for weakly supervised segmentation of ultrasound images remains a challenging task. Thus, we propose a dual-route transformer network for thyroid ultrasound segmentation based on Bbox-supervised learning: DRTNet. Furthermore, the effectiveness of DRTNet was evaluated using three ultrasound thyroid datasets to validate its performance.

Intensive ablation studies and comparisons with other state-of-the-art approaches provided substantial support for the superiority of our

Table 9
Quantitative results of improvements in our framework on our own dataset.

Model	IoU (%)	DSC (%)	HD95	F1 (%)	ACC (%)	AUC (%)
Supervision:Boxes labels						
WSSL [57]	62.59	76.99	25.96	76.39	97.11	92.33
BoxSup [58]	63.79	77.89	22.98	81.08	97.63	91.61
SDI [15]	67.52	80.61	21.32	81.00	97.66	91.75
BCM [59]	67.57	80.65	18.33	79.38	97.55	91.92
BANA(Segformer) [17]	67.63	80.69	21.71	81.19	97.73	92.13
BCM(FRS) [26]	67.78	80.81	16.68	81.67	97.25	92.74
DRTNet[a]	**73.00**	**84.39**	**14.53**	**83.56**	**97.98**	**95.01**
Supervision: Pixel-level labels						
Segformer [60]	71.50	83.38	18.56	84.58	98.14	90.88

[a] The best results are highlighted in bold.

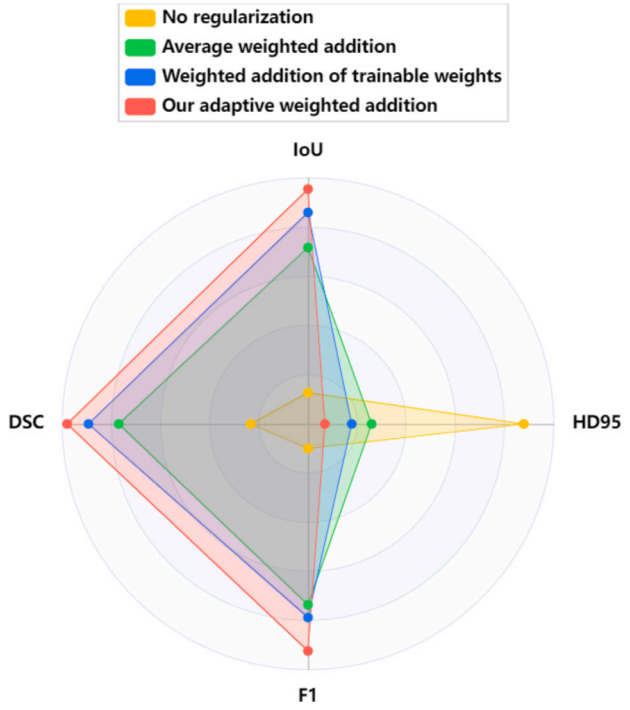


Fig. 12. Radar chart of adaptive uncertainty estimation multi-scale consistency.

Table 10
Quantitative results of different training weights.

Pretrained weights	IoU (%)	DSC(%)	HD95	F1 (%)
TN3k weights	67.30	80.45	23.39	78.98
DDTI weights	45.28	62.34	40.88	67.62

DRTNet, which mainly emphasizes effective features representation of the thyroid from the incomplete nodule internal region, blurring nodule edges, and the inferior pseudo thyroid mask. For the incomplete nodule internal region, the proposed DRTNet design is a double-branch foreground modulation of CAMs based on the transformers backbone to generate more accurate thyroid masks based on Bbox learning. Directed against the edges of the blurring nodule, the proposed DRTNet adopts the UPBAP to define the finer boundary of thyroid nodules and focus on the relatively small difference between the nodules and the partial background. With the inferior pseudo thyroid mask, the proposed DRTNet introduces adaptive AUEMSC to avoid over-fitting of segmentation results due to pseudo annotations.

In this chapter, we discuss in depth the experimental results obtained on private datasets. Given its exceptional performance in terms of pixel-level supervised segmentation, SegFormer stands as a robust benchmark model for comparing various supervision approaches. The segmentation results are shown in Table 9 and Fig. 13. Compared with SegFormer, existing segmentation schemes based on target box supervision only require rough annotation of nodule positions using rectangular boxes. Furthermore, by reducing the pressure on doctors to label, there will be no significant loss of accuracy. We can observe that the WSSL performs poorly on private datasets due to its overly simplistic pseudo mask generation scheme. Although BoxSup achieved further performance improvement with its iterative update strategy, it is still not ideal. BANA benefits from the BAP mechanism that utilizes the background outside the target box to distinguish foreground and background. BCM and BCM(FRS) both have a significant improvement in the HD95 metric due to the use of regularization strategies for nodule objects, which confirms that regularization strategies can also refine boundaries. At the same time, BCM with added FRS technology has a significant improvement in other indicators compared to ordinary regularization schemes. The performance of DRTNet meets expectations, which is attributed to the dual-branch modulation strategy and UPBAP. Finally, DRTNet optimized by AUEMSC outperforms the pixel-level segmentation method SegFormer in all indicators and visualizations.

To evaluate the clinical utility of the trained model on a publicly available thyroid ultrasound dataset, we directly used models trained with TN3K and DDTI to infer private datasets, and the corresponding results are summarized in Table 10. However, the results of the direct testing fail to meet clinical needs, which may be due to significant differences in terms of feature distribution between datasets.

The proposed framework was initially developed for thyroid ultrasound images but demonstrates remarkable cross-domain applicability.

Considering the large costs of medical image collection and dense annotation is a common problem with other medical images, the box labels used by DRTNet can greatly reduce annotation costs.

For homogeneous modalities, breast, liver, and other ultrasound images share similar imaging principles and partial structural characteristics with thyroid ultrasound images, yet exhibit significant differences in anatomical details and lesion features. For instance, the liver is surrounded by other organs such as the pancreas, kidneys, and gallbladder, with similar grayscale values and low contrast, while the inherent diversity and inconsistency of imaging information further increase the challenge of liver and tumor segmentation. This characteristic is analogous to the blurred boundaries of thyroid ultrasound nodules. The proposed algorithm, leveraging the feature extraction capability of Transformers and its superior mechanism for modulating feature distributions in both frequency-channel and spatial dimensions, can effectively distinguish targets from the background, thus making it equally competent for segmentation tasks in other anatomies of the same modality. Additionally, liver segmentation tasks may involve multi-target interactive segmentation [64], where the prompts used are similar to the bounding-box labels employed in this study.

For heterogeneous modalities, such as brain tumor segmentation in magnetic resonance imaging (MRI) and liver/pancreatic tumor segmentation in computed tomography (CT). Although the imaging principles differ significantly from ultrasound and present fewer noise-related

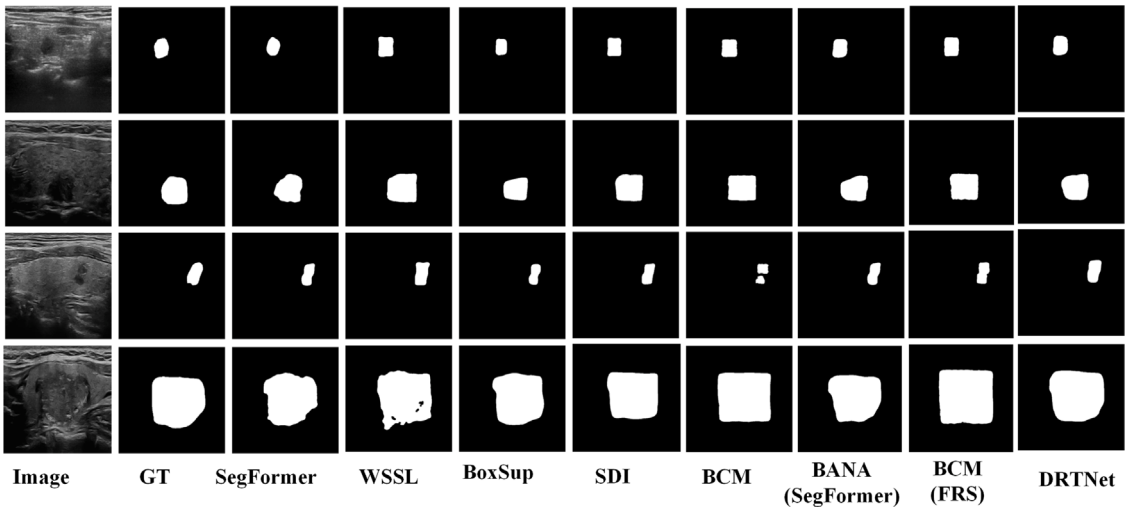


Fig. 13. Segmentation result on our own dataset.

challenges, they involve complex anatomical structures, mixed pixels in small structures, and artifact-related issues [65]. Nevertheless, the Transformer's feature extraction capability and dual-channel parallel modulation remain applicable in these scenarios. Furthermore, the adaptive multi-scale regularization mechanism effectively enhances performance without requiring explicit consideration of these challenges.

The limitation of this proposed study is that we only conducted experiments on ultrasound thyroid images and have not yet verified the effectiveness of this model on other organ ultrasound images (such as breast and prostate). Furthermore, models trained by TN3K and DDTI may not meet expectations when tested directly on private datasets. This is attributable to the fact that the quantity of the training set does not significantly exceed that of the test set, coupled with the presence of distribution disparities among the datasets. In the future, we will aim to gather more ultrasound images to validate DRTNet and examine the distribution differences in features between different datasets.

Finally, we will also discuss some recent work. Considering the successful applications of large-scale models (SAM and Med-SA) in medical image segmentation [66,67], we conducted a study focused on this area. Interestingly, the prompts required for this interactive segmentation bear a resemblance to the target Bbox utilized by DRTNet, albeit applied to different objects. SAM, a segmentation model based on the ViT encoder with prompt-based capabilities, progresses through manual annotation, semi-automatic, and fully automatic stages to achieve zero-shot transfer to other tasks or datasets. However, its performance in medical image segmentation is sub-optimal. Med-SA, building upon SAM extensively pretrained on a vast dataset, incorporates prompt-conditioned adaptation for fine-tuning to address this shortfall.

This type of prompt-based segmentation model includes Bbox and points, and compared to point-based prompts, Bbox prompts exhibit superior performance. Furthermore, Bbox prompts only require providing the top-left and bottom-right points of the nodule, which aligns with the labels used for training our DRTNet. When manual annotation of images is scarce, large models serve as reliable pseudo-label generators, thereby leading to generation of labels. Despite the superior performance of these large models and the promising prospects of interactive segmentation in medical contexts, their substantial memory consumption raises concerns with regard to their practical deployment in clinical settings. Thus, further investigation is required to assess their viability in real-world clinical applications. In contrast, our DRTNet requires only a single 8-GB graphics card.

Furthermore, we present recent research on the currently popular medical diffusion segmentation models [68], which typically generate segmentation results by simply passing randomly sampled noise through the learned denoising process. The corresponding results are shown in Fig. 14, the segmentation results listed are consistent with the GT. The impressive performance of diffusion segmentation models on ultrasound thyroid dataset deserves further research in the field of weakly supervised segmentation. In the future, we will conduct in-depth research on the application of diffusion segmentation models in ultrasound thyroid image segmentation.

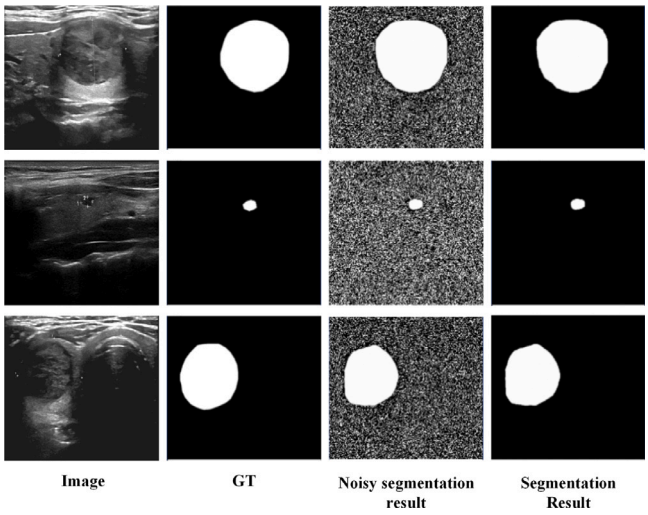


Fig. 14. Segmentation result of diffusion segmentation model on the TN3k dataset.

6. Conclusion

In this work, we are committed to investigate the segmentation of target box supervision and develop a dual-route transformer network for thyroid ultrasound segmentation. We introduce double-branch foreground CAMs' modulation, Transformer backbone and UPBAP to shape more accurate pseudo masks. AUEMSC ensures that even inferior pseudo masks possess satisfactory supervisory performance. Extensive experiments have been conducted to validate the proposed DRTNet on three different ultrasound thyroid datasets, Moreover, based on our findings, our method can outperform other state-of-art segmentation algorithms.

CRediT authorship contribution statement

Hui Bi: Writing – review & editing, Methodology, Conceptualization. **Chengjie Cai:** Writing – original draft, Software. **Jiawei Sun:** Visualization, Validation. **Shihao Ge:** Visualization, Validation. **Huazhong Shu:** Supervision, Funding acquisition, Formal analysis. **Xinye Ni:** Supervision, Project administration, Funding acquisition.

Declaration of competing interest

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

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62171125, 62371243 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 CE20215045 and CJ20220136.

Data availability

The authors do not have permission to share data.

References

[1] W.M. Alrubaidi, B. Peng, Y. Yang, Q. Chen, An interactive segmentation algorithm for thyroid nodules in ultrasound images, in: Intelligent Computing Methodologies: 12th International Conference, 2016, pp. 107–115, http://dx.doi.org/10.1007/978-3-319-42297-8_11.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: Cancer J. Clin. 68 (6) (2018) 394–424, http://dx.doi.org/10.3322/caac.21492.

[3] 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, http://dx.doi.org/10.1016/j.cmpb.2020.105329.

[4] E. Kollorz, E. Angelopoulou, M. Beck, D. Schmidt, T. Kuwert, Deep semantic segmentation of natural and medical images: A review, in: Bildverarbeitung Für Die Medizin 2011, 2011, pp. 124–128, http://dx.doi.org/10.1007/978-3-642-19335-4_27.

[5] G. Russ, S.J. Bonnema, M.F. Erdogan, C. Durante, R. Ngu, L. Leenhardt, European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS, Eur. Thyroid. J. 6 (5) (2017) 225–237, http://dx.doi.org/10.1159/000478927.

[6] M. Pandey, S. Lazebnik, Scene recognition and weakly supervised object localization with deformable part-based models, in: 2011 International Conference on Computer Vision, 2011, pp. 1307–1314, http://dx.doi.org/10.1109/ICCV.2011.6126383.

[7] A. Vezhnevets, V. Ferrari, J.M. Buhmann, Weakly supervised semantic segmentation with a multi-image model, in: 2011 International Conference on Computer Vision, 2011, pp. 643–650, http://dx.doi.org/10.1109/ICCV.2011.6126299.

[8] R. Huang, M. Lin, H. Dou, Z. Lin, Q. Ying, X. Jia, D. Ni, Boundary-rendering network for breast lesion segmentation in ultrasound images, Med. Image Anal. 80 (2022) 102478, http://dx.doi.org/10.1016/j.media.2022.102478.

[9] W. Shen, Z. Peng, X. Wang, H. Wang, J. Cen, D. Jiang, Q. Tian, A survey on label-efficient deep segmentation: Bridging the gap between weak supervision and dense prediction, 2022, arXiv preprint arXiv:2207.01223.

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

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929, http://dx.doi.org/10.1109/CVPR.2016.319.

[12] L. Yang, Y. Zhang, Z. Zhao, et al., Boxnet: Deep learning based biomedical image segmentation using boxes only annotation, 2018, arXiv preprint arXiv:1806.00593.

[13] Y. Wei, J. Feng, X. Liang, M.M. Cheng, Y. Zhao, S. Yan, Object region mining with adversarial erasing: A simple classification to semantic segmentation approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1568–1576, http://dx.doi.org/10.1109/CVPR.2017.687.

[14] Y. Wang, J. Zhang, M. Kan, S. Shan, X. Chen, Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12275–12284, http://dx.doi.org/10.1109/CVPR42600.2020.01229.

[15] A. Khoreva, R. Benenson, J. Hosang, M. Hein, B. Schiele, Simple does it: Weakly supervised instance and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 876–885, http://dx.doi.org/10.1109/CVPR.2017.181.

[16] H. Bi, C. Cai, J. Sun, Y. Jiang, G. Lu, H. Shu, X. Ni, BPAT-UNet: Boundary preserving assembled transformer UNet for ultrasound thyroid nodule segmentation, Comput. Methods Programs Biomed. (2023) 107614.

[17] Y. Oh, B. Kim, B. Ham, Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6913–6922, http://dx.doi.org/10.1109/CVPR46437.2021.00684.

[18] D. Tanaka, D. Ikami, T. Yamasaki, K. Aizawa, Joint optimization framework for learning with noisy labels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5552–5560, http://dx.doi.org/10.1109/CVPR.2018.00582.

[19] K. Yi, J. Wu, Probabilistic end-to-end noise correction for learning with noisy labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7017–7025, http://dx.doi.org/10.1109/CVPR.2019.00718.

[20] H. Gong, G. Chen, R. Wang, X. Xie, M. Mao, Y. Yu, G. Li, Multi-task learning for thyroid nodule segmentation with thyroid region prior, in: 2021 IEEE 18th International Symposium on Biomedical Imaging, ISBI, 2021, pp. 257–261, http://dx.doi.org/10.1109/ISBI48211.2021.9434087.

[21] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, E. Romero, An open access thyroid ultrasound image database, in: 10th International Symposium on Medical Information Processing and Analysis, Vol. 9287, 2015, pp. 188–193, http://dx.doi.org/10.1117/12.2073532.

[22] O. Ronneberger, T. Brox, P. Fischer, Gt u-net: A u-net like group transformer network for tooth root segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, 2015, pp. 234–241, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[23] Y. Li, S. Wang, J. Wang, G. Zeng, W. Liu, Q. Zhang, Y. Wang, Deep residual learning for image recognition, in: Machine Learning in Medical Imaging: 12th International Workshop, 2021, pp. 386–395, http://dx.doi.org/10.1007/978-3-030-87589-3_40.

[24] M.Y. Ansari, I.A.C. Mangalote, P.K. Meher, et al., Advancements in deep learning for B-mode ultrasound segmentation: A comprehensive review, IEEE Trans. Emerg. Top. Comput. Intell. (2024).

[25] S. Asgari Taghanaki, K. Abhishek, J.P. Cohen, J. Cohen-Adad, G. Hamarneh, Deep semantic segmentation of natural and medical images: a review, Artif. Intell. Rev. 54 (2021) 137–178, http://dx.doi.org/10.1007/s10462-020-09854-1.

[26] C. Song, W. Ouyang, Z. Zhang, Weakly supervised semantic segmentation via box-driven masking and filling rate shifting, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[27] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, 2018, pp. 3–11, http://dx.doi.org/10.1007/978-3-030-00889-5_1.

[28] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, L. Soler, U-net transformer: Self and cross attention for medical image segmentation, in: Machine Learning in Medical Imaging: 12th International Workshop, pp. 267–276, http://dx.doi.org/10.1007/978-3-030-87589-3_28.

[29] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, T.S. Huang, Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7268–7277, http://dx.doi.org/10.1109/CVPR.2018.00759.

[30] J. Fan, Z. Zhang, C. Song, T. Tan, Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4283–4292, http://dx.doi.org/10.1109/CVPR42600.2020.00434.

[31] J. Qin, J. Wu, X. Xiao, L. Li, X. Wang, Activation modulation and recalibration scheme for weakly supervised semantic segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 2117–2125, http://dx.doi.org/10.1609/aaai.v36i2.20108, (2).

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021, pp. 1–5.
[33] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, L. Zhang, Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890, http://dx.doi.org/10.1109/CVPR46437.2021.00681.
[34] R. Ranftl, A. Bochkovskiy, V. Koltun, Vision transformers for dense prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12179–12188, http://dx.doi.org/10.1109/ICCV48922.2021.01196.
[35] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803, http://dx.doi.org/10.1109/CVPR.2018.00813.
[36] L. Ru, Y. Zhan, B. Yu, B. Du, Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16846–16855, http://dx.doi.org/10.1109/CVPR52688.2022.01634.
[37] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, Q. Ye, Ts-cam: Token semantic coupled attention map for weakly supervised object localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2886–2895, http://dx.doi.org/10.1109/ICCV48922.2021.00288.
[38] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19, http://dx.doi.org/10.1007/978-3-030-01234-2_1.
[39] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141, http://dx.doi.org/10.1109/TPAMI.2019.2913372.
[40] Z. Qin, P. Zhang, F. Wu, X. Li, Fcanet: Frequency channel attention networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 783–792, http://dx.doi.org/10.1109/ICCV48922.2021.00082.
[41] V. Kulharia, S. Chandra, A. Agrawal, P. Torr, A. Tyagi, Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation, in: Computer Vision–ECCV 2020, 16th European Conference, 2020, pp. 290–308, http://dx.doi.org/10.1007/978-3-030-58583-9_18.
[42] M. Shaobo, C. Xuejin, Z. Zheng-Jun, W. Feng, Z. Yongdong, A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 4578–4585.
[43] A. Kumar, V.K. Ithapu, SeCoST: Sequential CoSupervision for weakly labeled audio event detection, 2019, arXiv preprint arXiv:1910.11789.
[44] Z. Cheng, P. Qiao, K. Li, et al., Out-of-candidate rectification for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23673–23684.
[45] X. Xia, T. Liu, B. Han, C. Gong, N. Wang, Z. Ge, Y. Chang, Robust early-learning: Hindering the memorization of noisy labels, in: International Conference on Learning Representations, 2021.
[46] X. Luo, G. Wang, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency, Med. Image Anal. 80 (2022) 102517, http://dx.doi.org/10.1016/j.media.2022.102517.
[47] S. Liu, K. Liu, W. Zhu, Y. Shen, C. Fernandez-Granda, Adaptive early-learning correction for segmentation from noisy annotations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2606–2616, http://dx.doi.org/10.1109/CVPR52688.2022.00263.
[48] P. Krähenbühl, V. Koltun, Efficient inference in fully connected crfs with gaussian edge potentials, Adv. Neural Inf. Process. Syst. 24 (2011).
[49] Z. Wu, S. Li, C. Chen, H. Qin, A. Hao, Salient object detection via dynamic scale routing, IEEE Trans. Image Process. 31 (2022) 6649–6663, http://dx.doi.org/10.1109/TIP.2022.3214332.
[50] J. Ruan, S. Xiang, M. Xie, et al., MALUNet: A multi-attention and light-weight unet for skin lesion segmentation, in: IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2022, pp. 1150–1156.
[51] D. Jiang, B. Sun, S. Su, Z. Zuo, P. Wu, X. Tan, FASSD: A feature fusion and spatial attention-based single shot detector for small object detection, Electron. 9 (9) (2020) 1536, http://dx.doi.org/10.3390/electronics9091536.
[52] G. Chen, J. Zhang, Y. Liu, J. Yin, X. Yin, L. Cui, Y. Dai, ESKNet-An enhanced adaptive selection kernel convolution for breast tumors segmentation, 2022, arXiv preprint arXiv:2211.02915.
[53] X. He, Y. Zhou, J. Zhao, et al., Swin transformer embedding unet for remote sensing image semantic segmentation, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15.
[54] J. Xu, Z. Chen, T.Q. Quek, K.F.E. Chong, Fedcorr: Multi-stage federated learning for label noise correction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10184–10193, http://dx.doi.org/10.1109/CVPR52688.2022.00994.
[55] T. Miyato, S.I. Maeda, M. Koyama, S. Ishii, Virtual adversarial training: a regularization method for supervised and semi-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 41 (8) (2018) 1979–1993, http://dx.doi.org/10.1109/TPAMI.2018.2858821.
[56] S. Liu, D. Huang, Y. Wang, Learning spatial fusion for single-shot object detection, 2019, arXiv preprint arXiv:1911.09516.
[57] G. Papandreou, L.C. Chen, K.P. Murphy, A.L. Yuille, Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2016, pp. 1742–1750.
[58] J. Dai, He. K., J. Sun, Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1635–1643, http://dx.doi.org/10.1109/ICCV.2015.191.
[59] C. Song, Y. Huang, W. Ouyang, L. Wang, Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3136–3145.
[60] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.
[61] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[62] H.H. Yang, C.H.H. Yang, Y.C.F. Wang, Wavelet channel attention module with a fusion network for single image deraining, in: 2020 IEEE International Conference on Image Processing, ICIP, Abu Dhabi, United Arab Emirates, 2020, pp. 883–887, http://dx.doi.org/10.1109/ICIP40778.2020.9190720.
[63] J. Lee-Thorp, J. Ainslie, I. Eckstein, et al., Fnet: Mixing tokens with fourier transforms, 2021, arXiv preprint arXiv:2105.03824, 2021.
[64] Y. Ding, L. Li, W. Wang, et al., Clustering propagation for universal medical image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3357–3369.
[65] J. Chen, J. Mei, X. Li, et al., TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers, Med. Image Anal. 97 (2024) 103280.
[66] A. Kirillov, E. Mintun, N. Ravi, et al., Segment anything, in: IEEE/CVF International Conference on Computer Vision, ICCV, 2023, http://dx.doi.org/10.1109/ICCV51070.2023.0037.
[67] J. Wu, R. Fu, H. Fang, et al., Medical sam adapter: Adapting segment anything model for medical image segmentation, 2023, arxiv preprint arxiv:2304.12620, 2023.
[68] J. Wu, R. Fu, H. Fang, Medsegdiff: Medical image segmentation with diffusion probabilistic model, in: Medical Imaging with Deep Learning, PMLR, 2024, pp. 1623–1639.