Medical Image Analysis 58 (2019) 101555

Contents lists available at ScienceDirect

Medical Image Analysis

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

# Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks

Tianjiao Liu[a,b], Qianqian Guo[c], Chunfeng Lian[b], Xuhua Ren[d], Shujun Liang[e], Jing Yu[g], Lijuan Niu[c,*], Weidong Sun[a,*], Dinggang Shen[b,f,*]

[a] *State Key Laboratory of Intelligent Technology and Systems, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*
[b] *Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA*
[c] *National Cancer Center/Cancer Hospital of Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China*
[d] *The Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China*
[e] *Department of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China*
[f] *Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea*
[g] *College of Computer Science and Technology, Beijing University of Technology, Beijing 100124, China*

ARTICLE INFO

*Article history:*
Received 17 January 2019
Revised 10 June 2019
Accepted 4 September 2019
Available online 5 September 2019

*Keywords:*
Ultrasound image
Thyroid nodule
Convolutional neural networks
Clinical knowledge

ABSTRACT

Accurate diagnosis of thyroid nodules using ultrasonography is a valuable but tough task even for experienced radiologists, considering both benign and malignant nodules have heterogeneous appearances. Computer-aided diagnosis (CAD) methods could potentially provide objective suggestions to assist radiologists. However, the performance of existing learning-based approaches is still limited, for direct application of general learning models often ignores critical domain knowledge related to the specific nodule diagnosis. In this study, we propose a novel deep-learning-based CAD system, guided by task-specific prior knowledge, for automated nodule detection and classification in ultrasound images. Our proposed CAD system consists of two stages. First, a multi-scale region-based detection network is designed to learn pyramidal features for detecting nodules at different feature scales. The region proposals are constrained by the prior knowledge about size and shape distributions of real nodules. Then, a multi-branch classification network is proposed to integrate multi-view diagnosis-oriented features, in which each network branch captures and enhances one specific group of characteristics that were generally used by radiologists. We evaluated and compared our method with the state-of-the-art CAD methods and experienced radiologists on two datasets, i.e. Dataset I and Dataset II. The detection and diagnostic accuracy on Dataset I were 97.5% and 97.1%, respectively. Besides, our CAD system also achieved better performance than experienced radiologists on Dataset II, with improvements of accuracy for 8%. The experimental results demonstrate that our proposed method is effective in the discrimination of thyroid nodules.

## 1. Introduction

Thyroid nodules are very common in the general population, which can be detected with high-resolution ultrasound for about 19%-68% randomly selected individuals (HaugenBryan et al., 2016). As one of the most common types of endocrine carcinoma, the prevalence of thyroid cancer has increased worldwide in the past 30 years, making it the fifth most common cancer among women (Gabriella et al., 2013). Ultrasonography is a very important tool for thyroid nodule evaluation. Due to its low cost and high sensitivity, ultrasonography has been widely used in various tasks, including fine-needle aspiration (FNA) biopsy guidance, postoperative evaluation, nodule detection, and diagnosis (Danese et al., 1998; Kouvaraki et al., 2003; Gul et al., 2010). However, radiologists usually perform diagnosis based on sonographic characteristics of nodules in ultrasound images, which is relatively subjective and highly depends on the clinical experience of radiologists. To tackle this challenge, many computer-aided diagnosis (CAD) systems have been proposed for automated and objective classification of thyroid nodules. Typically, a CAD system consists of three fundamental components, i.e., (1) nodule detection, (2) imaging feature extraction, and (3) classifier construction. Traditional methods usually predefine

* Corresponding authors.
*E-mail addresses:* ltj14@mails.tsinghua.edu.cn (T. Liu), niulijuan8197@126.com (L. Niu), wdsun@tsinghua.edu.cn (W. Sun), dgshen@med.unc.edu (D. Shen).

https://doi.org/10.1016/j.media.2019.101555

nodule locations by radiologists, and then extract hand-crafted features (e.g., textural and morphology features) and construct classifier (e.g., using AdaBoost and decision tree) (Koundal et al., 2013; Agarwal et al., 2013; Chang et al., 2010; Dorin et al., 2013; Ye et al., 2009; Ali Abbasian et al., 2015; Ding et al., 2014; Si et al., 2011). For instance, Singh and Jindal (2012) used gray level co-occurrence matrix features to construct a k-nearest neighbor (KNN) model for thyroid nodule classification. Keramidas et al. (2008) extracted fuzzy local binary pattern as noise-resistant textural features and adopted support vector machine (SVM) as a classifier. Acharya et al. (2014) extracted grayscale features based on stationary wavelet transform and compared several common classifiers. In Tsantis et al. (2009), a set of morphological features (such as Mean radius, Radius entropy, and Radius standard deviation) were extracted from segmented nodules to describe the shape and the boundary regularity of each nodule. However, since thyroid nodules vary in shape, size and internal characteristics, the low-level hand-crafted features used in these traditional CAD methods can only have limited discriminative capacity due to their inherent simplicity and locality.

On the other hand, deep learning models, especially convolutional neural networks (CNNs), have shown dominant performance than conventional learning methods in various visual recognition tasks, such as object detection and image classification (Lian et al., 2019; Jean et al., 2016; Xu et al., 2015; Lian et al., 2018; Gao et al., 2016). By learning hierarchical feature representations in a task-oriented manner, CNNs can well capture semantic characteristics from input images. Due to this critical advantage, multiple CNN-based CAD methods have been recently proposed for thyroid nodule diagnosis. For example, Chi et al. (2017) fine-tuned a pre-trained GoogLeNet to extract high-level features from ultrasound images, based on which they further trained a cost-sensitive Random Forest classifier to differentiate between malignant and benign cases. Ma et al. (2016) trained two complementary patch-based CNNs of different depths to extract both low-level and high-level features, and then fused their feature maps to classify thyroid nodules. Gao et al. (2017) proposed a CAD system based on a multi-scale CNN model, achieving competitive diagnostic sensitivity compared with an experienced radiologist using the Thyroid Imaging Reporting and Data Systems (TI-RADS) (Tessler et al., 2017) and the American Thyroid Association (ATA) guidelines (HaugenBryan et al., 2016). Song et al. (2018) proposed a cascaded network for thyroid nodule detection and recognition based on multi-scale SSD network and spatial pyramid architecture. Although existing deep learning methods are generally more powerful than traditional methods, their diagnostic performance is still limited. Similar to most traditional methods using hand-crafted features, existing deep learning methods usually treat natural images and medical images equally, and directly apply general CNN models (previously designed in computer vision community) for thyroid nodule classification. This kind of operation improperly ignores important domain and expert knowledge related to the specific task of medical diagnosis.

In this study, to make the CAD process be more consistent with radiologists' diagnostic consideration, we propose a new deep learning framework guided by specific domain knowledge and radiologist experience for automated detection and classification of thyroid nodules in ultrasound images. Our proposed CAD framework consists of two main steps: (1) detecting the nodule locations by a multi-scale pyramidal network; (2) classifying detected nodules with an expert-knowledge-guided multi-branch network. In the first step, a pyramidal network is constructed to detect nodules at different scales, in which a top-down module is designed to integrate high-level semantic information with lower-level feature maps for small nodule detection. Also, to further improve detection performance, we refer to an important clinical constraint, i.e., the real distributions of nodule size and shape, to design reliable anchors of each feature scale in our proposal generating for better box anchoring. The clinical prior knowledge helps optimizing the initialization and filtering of proposals, which leads to more precise detection. In the second step, we captured the sonographic characteristics that radiologists focus on, to complementarily construct our multi-branch classification network. Specifically, in addition to a basic branch to extract semantic features from input patches, our multi-branch network also integrates a context branch and a margin branch to extract enhanced contextual and marginal features that are strongly correlated with the malignancy of nodules. Also, we introduce an aspect-ratio-preserving strategy to process the input image patches, which can preserve nodule's aspect ratio from resizing for better classification performance. The experimental results demonstrate that our approach is effective in diagnosing thyroid nodules, with superior diagnostic performance compared with both existing CAD methods and experienced radiologists.

## 2. Materials and methods

### 2.1. Image acquisition

The images used in our research work were from two databases, both supplied by the Cancer Hospital of the Chinese Academy of Medical Sciences, acquired during January 2016 to October 2017. Dataset I consists of 7690 thyroid nodule images acquired from 4279 patients, of which 5139 cases are malignant and 2551 cases are benign. Dataset II consists of 450 thyroid nodule images acquired from 376 patients, of which 322 cases are malignant and 128 cases are benign. All images were acquired using the ultrasound machine (GE Logiq E9, S7) with the probe frequency set as 5-12mhz or 8-15mhz. Typical thyroid nodules are shown in Fig. 1. For each image from both Datasets I and II, experienced radiologists draw the ground-truth region of interest (ROI) for nodule detection, such as the red box in Fig. 1(a), and also annotated the respective ground-truth label (i.e., benign/malignant) for diagnosis. The malignant or benign of a nodule was defined according to FNA biopsy and pathological results. To further compare automated diagnosis performance with radiologists under different situations, nodules in Dataset II were divided into multiple groups according to 6 sonographic characteristics: nodule size, margin, shape, aspect ratio, composition, and calcification. Specially, from Fig. 1, we can observe that benign and malignant nodules may have very similar sonographic characteristics. That is, it is actually very challenging

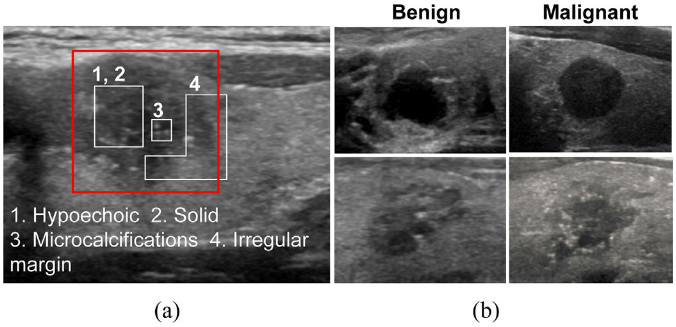

**Fig. 1.** Illustration of typical thyroid nodules: (a) Typical characteristics used by radiologists for thyroid nodule diagnosis: internal echo, component, calcification, and margin. The example nodule in (a) is a solid hypoechoic nodule with microcalcifications and irregular margin, and the red rectangle identifies the nodule position. (b) Subjects from different categories may have similar appearances in morphology, which causes confusion to diagnosis and increases intra-/inter-observer variabilities. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

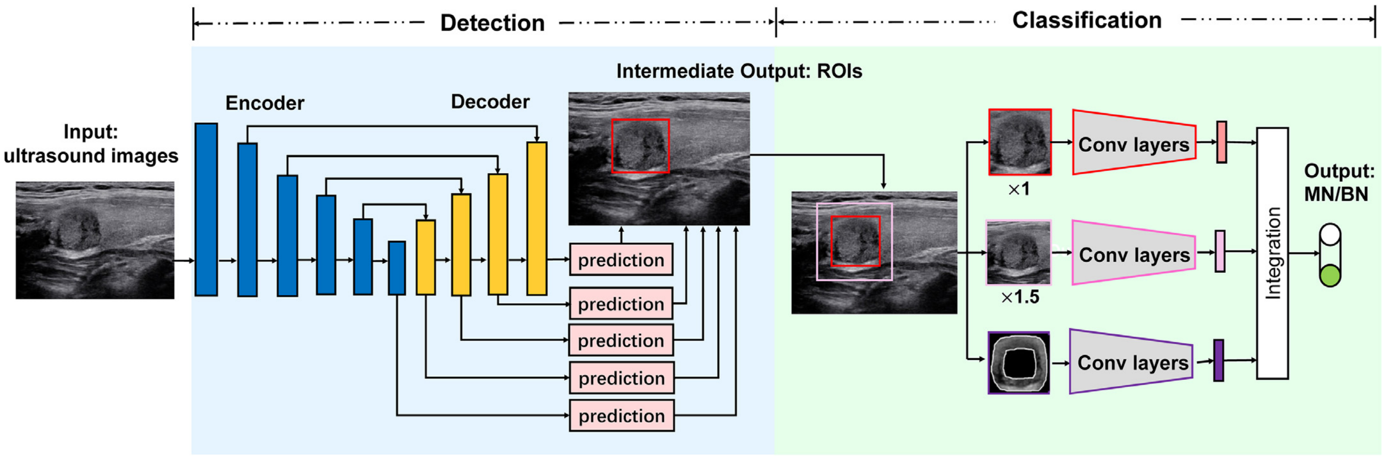

**Fig. 2.** Schematic diagram of our proposed deep learning framework for automated thyroid nodule detection and classification.

for radiologists to perform diagnosis using solely those sonographic characteristics.

### 2.2. Clinical-knowledge-guided detection and classification networks

The proposed method integrates key sonographic characteristics (that radiologists focus on) with deep convolutional networks to accurately detect and diagnose thyroid nodules. Fig. 2 shows the pipeline of our proposed method, where a multi-scale pyramidal network is first designed to automatically locate nodules in ultrasound images, and then a multi-branch CNN is built to extract multi-view task-oriented features for nodule classification.

#### 2.2.1. Multi-scale detection network

Region-free object detection methods (e.g., SSD Liu et al., 2016) and region-proposal-based detection architectures (e.g., Faster R-CNN Ren et al., 2015) have shown promising performance on detection accuracy in various object detection tasks. Compared with SSD, region-based methods generally have higher accuracy and lower omission rate, which is a practically meaningful advantage in the nodule detection and diagnosis (e.g., microcarcinoma). Region-based methods usually require two stages: (1) generate candidate bounding boxes (i.e., region proposals) that potentially contain target objects, and (2) run classifier and regressor on the proposed boxes to predict a precise ROI for each target object. As one of the state-of-the-art detection methods, Faster R-CNN unifies the above two stages into an efficient end-to-end framework, where a region proposal network (RPN) is first applied to generating proposals, based on which an R-CNN detector that shares backbone with the RPN is designed to predict the final ROI. However, since Faster R-CNN only considers mono-scale feature representations, it may fail to detect objects with varying appearance, a very practical challenge that should be tackled in our task of thyroid nodule detection. Specifically, clinical prior knowledge shows that nodules (both benign and malignant) have a wide variety of sizes. Fig. 3 summarizes the size distribution of the thyroid nodules included in our study, from which we can observe that some nodules could be very small (e.g., Fig. 3(b)) while some others could be large (e.g., Fig. 3 (c)). In order to robustly detect those nodules in different sizes, an intuitive way is to make full use of the multi-scale feature hierarchy inherently learned by successive convolutional layers. Therefore, we propose to combine the idea of feature pyramid network (FPN) (Lin et al., 2017) with Faster R-CNN to learn multi-scale region proposals for nodule detection. The architecture of our proposed multi-scale detection network is presented in Fig. 4. Briefly, relatively large instances will be detected in our network using feature maps produced by the top layers; while, leveraging an efficient top-down strategy, high-level semantics will also be back-propagated and combined with low-semantic feature maps at the bottom layers to detect relatively small instances. In the rest of this subsection, we will introduce in detail the architecture of our multi-scale detection network, a clinical-knowledge-based strategy for defining detection anchors, and a cost-sensitive hybrid loss for training our detection network.

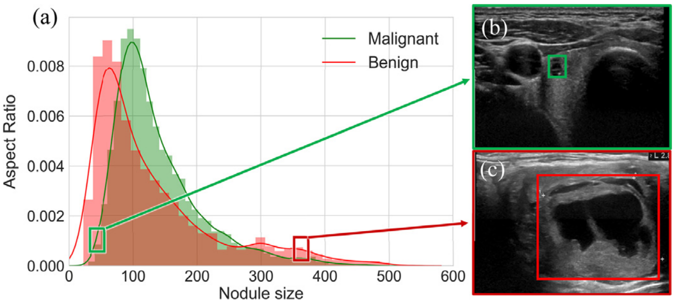

**Fig. 3.** (a) The scale histogram (in pixels) of the thyroid nodules. (b) A malignant nodule with relatively small size. (c) A benign nodule with relatively large size. The green and red boxes in (a) indicate the nodule size locations in the area distribution for (b) and (c). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

*1) Network architecture*: As shown in Fig. 4, we used ResNet-50 (He et al., 2016) as the backbone of our detection network. It consists of one convolutional layer (i.e., Conv1 using 7 × 7 kernels), and four Residual block (i.e., Conv2_x to Conv5_x). Each of them is followed by a max pooling layer to downsample the feature maps by a scale factor of 2. So, similar to that in FPN, the backbone here forms a *bottom-up pathway*, which computes a pyramidal feature hierarchy with progressively increased abstraction but decreased spatial resolution. The specific architecture of those residual blocks is kept the same as that in the original paper: Conv2_x to Conv5_x have 3, 4, 6, and 3 building blocks, respectively. Each building block contains 3 convolutional layers using 1 × 1, 3 × 3, and 1 × 1 kernels, respectively.

To reliably detect objects with flexible sizes, we design a *top-down pathway* to hierarchically fuse high-level backbone features with low-level backbone features, and then detect objects with different sizes using the features merged at different scales. Specifically, from Conv2_x to Conv5_x, a top-down module (TDM) is injected between any two adjacent blocks. In each TDM, the higher-level feature maps are upsampled by a factor of 2 using a deconvolutional layer, which are then concatenated with the transferred

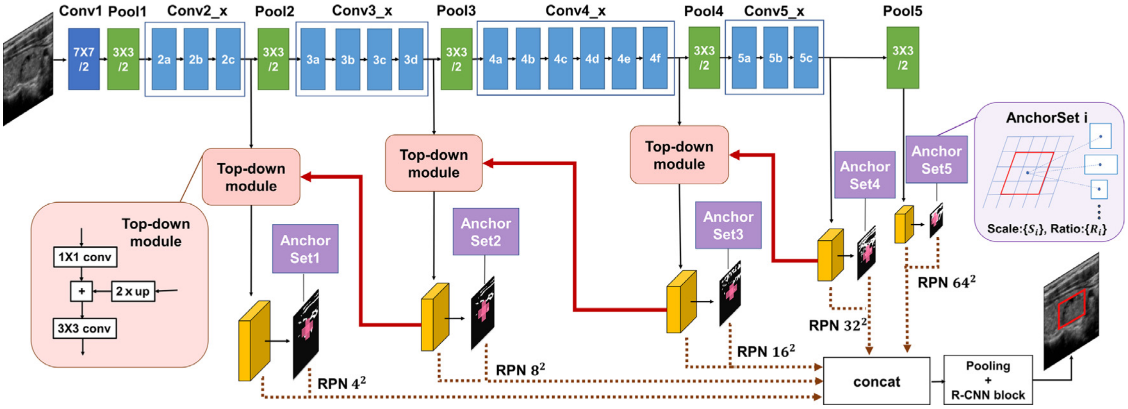

**Fig. 4.** Multi-scale detection network based on FPN. We choose the anchor boxes based on the scales and aspect ratio prior of thyroid nodules. 'Conv' block represents convolutional layer; 'Pool' block represents max-pooling layer; 'Conv2' block represents residual module; Specially, '3 × 3/2' denotes a filter kernel size of 3 and a stride of 2; Symbol '+' in the top-down module represents feature map concatenation; At different scale (named as $\mathbf{4^2}$,$\mathbf{8^2}$), RPN uses different anchor set as defined in Table 1.

(by 1 × 1 convolutions) lower-level feature maps. The concatenated feature maps are finally processed by a 3 × 3 convolutional layer to generate the fused feature map with reduced aliasing effect for detecting objects at this specific scale. In this way, our proposed network yields region proposals at five different scales (i.e., Conv2_x, Conv3_x, Conv4_x, Conv5_x, and Pool5), for which the feature map sizes are 1/4, 1/8, 1/16, 1/32 and 1/64 of the original input size, respectively, both in width and height. Together with the corresponding feature maps, those region proposals are finally fed into the R-CNN block to detect nodules with different sizes.

*2) Detection anchors restricted by prior knowledge*: Similar to the original Faster R-CNN, our multi-scale detection network also predicts region proposals for object detection using anchors with designed sizes and aspect ratios (i.e., height divided by width). Generally, anchors are the size-fixed proposal candidates defined on feature maps. After shifting and resizing, these anchors are mapped back to the original image to get proposals. Intuitively, how to define these anchors is a key issue that directly influences the quality of the predicted region proposals. To ensure that the anchors used in our method can cover the variabilities of real nodules, we define anchors differently at different scales (or resolution levels) based on clinical prior knowledge.

Specifically, according to expert knowledge and clinical studies, thyroid nodules usually expand or infiltrate isotopically and rarely grow in only one direction. Thus, their aspect ratios actually lie in a certain range. Inspired by this fact, we pre-computed aspect ratio distribution in terms of the nodule size on our training dataset with the result shown in Fig. 5. From Fig. 5, we have two observations: (1) the nodule aspect ratio roughly ranges from 0.2 to 2.5, which indicates that the aspect ratio for anchors and proposed regions should also be selected within this range. (2) The aspect ratio decreases along the increase of nodule size, which implies that anchors defined at bottom layers (for small object detection) should have a relatively wider range of aspect ratios than those defined at top layers. Therefore, for ensuring the size and shape of region proposals to be consistent with the real nodule distribution, we constrain anchor boxes at lower feature maps to have smaller sizes and a larger range of aspect ratios than those at higher feature maps. In total, for each scale, we generate anchor boxes with 3 optional sizes and 3 optional aspect ratios at each feature map location. The sizes (i.e., height in voxels) and aspect ratios for anchors defined at different scales, which can finely cover the different nodule sizes in our current images, are shown in Table 1. The scale is represented by the number of voxels in the original image that one voxel in a specific feature map corresponds to. Therefore, based on the prior knowledge on the aspect ratios and sizes of the thyroid nodules, we can effectively improve the reliability of region proposals for automated nodule detection.

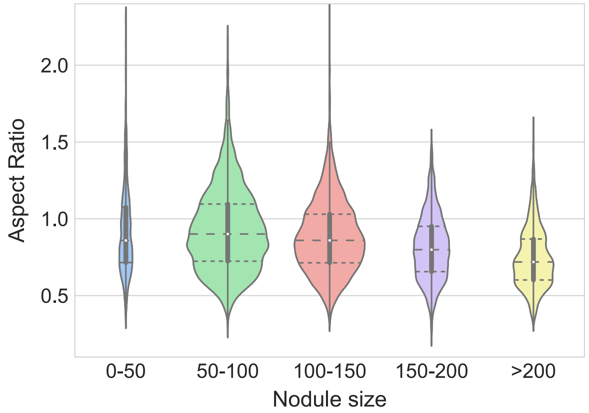

**Fig. 5.** The distribution of aspect ratios with respect to the sizes of the thyroid nodules. Here, we use the square of area (in voxels) to measure nodule size. Each violin shows the distribution of aspect ratios for a specific range of nodule sizes, where the dotted lines mark the median and quartiles positions.

**Table 1**
Anchors' scales and aspect ratios at each feature map location.

| Stride i | 4 | 8 | 16 | 32 | 64 |
|---|---|---|---|---|---|
| $S_i$ | (4,8,16) | (4,8,16) | (8,16,32) | (8,16,24) | (8,12,16) |
| $p_i$ | (2/5,1,5/2) | (2/5,1,5/2) | (1/2,1,2) | (2/3,1,3/2) | (2/3,1,3/2) |

*3) Loss function*: To train our multi-scale detection network, a multi-task loss function for end-to-end region proposal and nodule detection is defined as:

$$\begin{aligned} L = \alpha\left(\frac{1}{N_{cls_{rpn}}}\sum_i L_{cls_{rpn}}(p_i, p_i^*) + \lambda\frac{1}{N_{reg_{rpn}}}\sum_i p_i^* L_{reg_{rpn}}(t_i, t_i^*)\right) \\ +(1-\alpha)\left(\frac{1}{N_{cls_{ob}}}\sum_j L_{cls_{ob}}\left(p_j^{obj}, p_j^{obj*}\right) + \lambda\frac{1}{N_{reg_{ob}}}\sum_j L_{reg_{ob}}(t_j, t_j^*)\right) \end{aligned} \tag{1}$$

$$L_{cls_{ob}}\left(p_j^{obj}, p_j^{obj*}\right) = -\sum_{m=1}^{K} p_{j,m}^{obj*}\log(p_{j,m}^{obj})c_{y,m} \tag{2}$$

$$\{c_{y,m}\} = \begin{bmatrix} 0 & 1 & 1 \\ 2 & 0 & 1 \\ 2 & 2 & 0 \end{bmatrix} \tag{3}$$

The overall loss includes 4 parts. Specifically, $L_{cls_{rpn}}(p_i, p_i^*)$ represents classification loss for proposal box $i$, where $p_i$ is the predicted probability of proposal $i$ being a nodule, and $p_i^*$ is the ground-truth label (which equals 1 if $i$ is a positive region proposal, and 0 otherwise). $L_{reg_{rpn}}(t_i, t_i^*)$ is the regression loss between the proposal box offsets $t_i = (t_{i,x}, t_{i,y}, t_{i,w}, t_{i,h})$ and the ground truth offsets $t_i{}^* = (t_{i,x}^*, t_{i,y}^*, t_{i,w}^*, t_{i,h}^*)$ with respect to the ground-truth nodule bounding box. It targets at minimizing the smoothed $L_1$ loss defined in Girshick (2015), and is only activated for positive region proposals (i.e., weighted by $p_i^*$). The two terms are normalized by $N_{cls_{rpn}}$ and $N_{reg_{rpn}}$, and weighted by a balancing parameter $\lambda$. Similarly, $L_{cls_{ob}}(p_j^{obj}, p_j^{obj*})$ represents the final classification loss for each box over K+1 categories, where $p_j^{obj} = \{p_{j,0}^{obj}, p_{j,1}^{obj}, \ldots, p_{j,K}^{obj}\}$ and $p_j^{obj*} = \{p_{j,0}^{obj*}, p_{j,1}^{obj*}, \ldots, p_{j,K}^{obj*}\}$ are the predicted probability and ground-truth label for the proposal $j$, respectively. In our task, K = 2 (i.e., malignant nodule *vs.* benign nodule). $L_{reg_{ob}}$ is the corresponding regression loss between the predicted and ground-truth offsets for region proposals. It is worth noting that, to improve the detection precision and reduce the false negatives, the final classification loss is cost-sensitive in our implementation, as shown in Eq. (2). The elements in the first row represent the cost of classifying background proposal into background, benign nodules, and malignant nodules, respectively. The second row and third row represent the cost for classifying benign nodules and malignant nodules, respectively. According to the cost-sensitive matrix $C = \{c_{y,m}\}$ defined in Eq. (3), we assume that misclassifying malignancies into other categories (i.e., benign/context) will lead to the largest cost while misclassifying the context as benign/malignant nodule will yield the least cost. In this way, we can effectively reduce the false negative rate of malignant nodules and the misdetection cost. In addition, since we focus on detecting nodules in this stage, we empirically set $\alpha = 0.6, \lambda = 6$.

### *2.2.2. Thyroid nodule classification*

It is noteworthy that the classification results yielded by our multi-scale detection network (i.e., by minimizing Eq. (2)) is relatively coarse, considering that the proposal classification is only included as a supplementary task to assist the primary task of nodule detection in the first step. But, those classification results still provide valuable diagnosis advice. The samples incorrectly classified by our multi-scale detection network could be challenging nodules that cannot be well discriminated with those relatively ‘diagnosis-weak’ features learned primarily for nodule detection. This implies that, to further improve the diagnostic performance, a more sophisticated classification network is desired, in which the perspectives and attention of radiologists (especially in diagnosing challenging nodules) should be included to guide the learning of more discriminative features for automated nodule recognition.

Important sonographic characteristics that radiologists use to analyze the nodules include component, internal echo, aspect ratio, context texture, margin smoothness, margin shape, and so on Girshick (2015). By extension, component, internal echo and aspect ratio reflect the basic features of a nodule. For example, most cystic nodules are benign while malignant nodules tend to break through the capsule, making their height more likely to be larger than their width (i.e., aspect ratio larger than 1) (Hee Jung et al., 2011). Besides, context texture which reflects the relationship between nodules and surrounding issues is also taken into account. For instance, the difference between internal echo and external echo can reflect the ways of cell proliferation, which different between malignant and benign nodules. Moreover, margin region aggregates category-sensitive features which are essential complements to basic inner characteristics, since that malignant nodules tend to have blur and irregular lobulated margin due to their rapid infiltrating growth (Anil et al., 2011). In contrast to previous clinical studies that performed diagnosis using low-level sonographic characteristics (e.g., aspect ratio used in Hee Jung et al. (2011)), we propose a triple-branch CNN to learn multi-view classification-oriented features to deal with above three specific groups (i.e. basic, context and margin) accordingly.

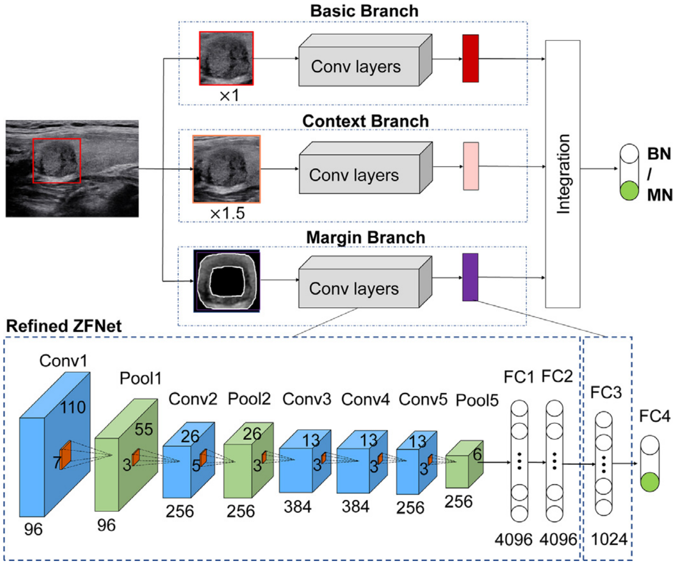

**Fig. 6.** The structure of our multi-branch network for thyroid nodule classification network. ‘BN’ represents benign nodule and ‘MN’ malignant nodule.

Specifically, our multi-branch CNN adopts three complementary branches to learn multi-view features from different regions according to the attention of radiologists. The structure of our multi-branch CNN is shown in Fig. 6, where sub-networks in all branches have the same structure but with different parameters. We refined ZFnet (Zeiler and Fergus, 2014) architecture as the backbone of our classification network. It inherits 5 convolutional layers (i.e., Conv1 using 7 × 7 kernels, Conv2 using 5 × 5 kernels, Conv3-Conv5 using 3 × 3 kernels), two fully-connected layers with 4096 units, and two additional fully-connected layers with 1024 and 2 outputs, respectively. The max-pooling with the window size of 3 × 3 is applied after Conv1, Conv2 and Conv5.

The first branch inputs the original nodules. Notably, CNNs usually need standard-sized inputs, requiring the original ROIs to be first resized. To preserve the original aspect ratio information for nodules with different shapes, we use an aspect-ratio-preserving cropping strategy to resize the original ROIs. For each nodule with size $h \times w$, we crop it and fill it into a square patch (size:max$(h, w)$) with zero padding. In this way, nodules will not be stretched in only one direction, and the aspect ratio information will be preserved for CNN-based nodule classification.

In addition to the information within the nodule, characteristics of the surrounding tissue with respect to the nodule, such as diffuse sclerosis and internal and external echogenicity differences mentioned above, also have an impact on nodule classification. To consider such kind of contextual information, larger image patches (i.e., 1.5 times the size of the original nodule ROI) located at the center of the original ROI are cropped as the inputs for the second branch, named as Context branch.

As mentioned above, margin patterns (i.e., smoothness, and shape) are also important characteristics that should be referred to in thyroid nodule diagnosis. Thereby, in the third branch of our network, we aim at enhancing the margin region in the ROIs to learn more salient features for the classification task. To this end, we employ the GGVF-snake (Xu and Prince, 1998) to process the nodule region and obtain a coarse nodule contour. As most nodules

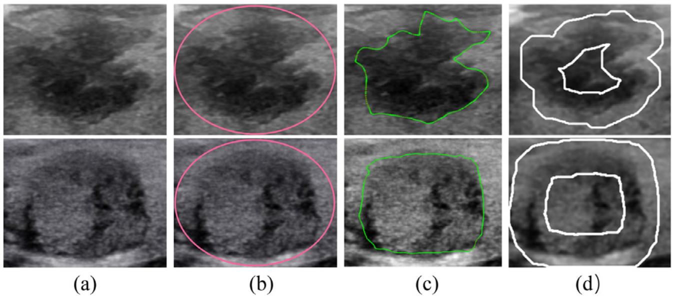

**Fig. 7.** Illustration of the extraction of nodule margin region. (a) Nodule patches; (b) Original ellipse snakes; (c) Nodule contours produced by the GVF-snake method; (d) The corresponding margin region.

are oval shaped, the circumscribed ellipse of the bounding box determines the original snake position. Then, we dilate the contour to obtain a ribbon mask with a certain width along the contour line to cover the margin region of the nodule, as shown in Fig. 7. The ribbon mask includes the region both within and beyond the nodule in order to preserve the characteristics difference across the margin. Finally, the ribbon-masked margin region is used to train the third branch, which provides coarse attention for the nodule margin area.

The multi-view features from the three parallel branches (i.e., the output of FC3 in each refined ZFNet) are finally concatenated and fused by a FC layer (with 256 units) for automated nodule classification using the softmax classifier.

### *2.3. Implementation details*

As discussed above, our proposed deep learning framework for automated thyroid nodule detection and classification contains two stages: (1) Automatically detect the nodule location; (2) Automatically diagnose the detected nodules using our multi-branch classification network. In our implementation, the backbones in multi-scale detection network and multi-branch classification network were initialized by the ResNet-50 and ZFNet pre-trained from the 1000 class general object classification (ImageNet) (Krizhevsky et al., 2012), respectively. Training dataset and testing dataset were divided by patients, i.e., one patient will not be included in both datasets.

For our multi-scale detection network, we adopted a specific data augmentation strategy to improve the generalization ability during network training. That is, considering the directional nature of nodule growth, we performed horizontal flip and rotation (with small angles of 4, 8, 12, 16, 20 degrees) on each training image. The detection network was first trained with 0.00003 learning rate for 40k iterations, and then with 0.0003 for another 40k iterations, with the batch size kept as 2. For our classification network, we first trained each complementary branch independently, and then jointly refined all components of the network. We trained the model with 0.001 learning rate for 30k iterations, and then with 0.0001 for another 50k iterations. Both the weight decay and bias decay were set as 0.0005, and the batch size was set as 32.

## 3. Results

### *3.1. Performance measurement and experimental settings*

The average precise (AP) was used to quantitatively evaluate the detection performance. Specifically, we looked at two types of AP where the first metric (denoted as detAP) measures the detection accuracy of all nodules, regardless of their categories (i.e., malignant/benign nodules), while the second metric measures the detection accuracy per category (denoted as AP-BN and AP-MN) and the mean accuracy across two categories (denoted as mAP). To quantitatively evaluate the classification performance, four metrics were used: (1) Accuracy= (TP+TN)/(TP+TN+FP+FN); (2) Sensitivity=TP/(TP+FN); (3) Specificity =TN/(TN+FP); (4) area under the ROC curve (AUC). Here, TP (true positive) and TN (true negative) represent the numbers of positive and negative samples that are correctly classified, respectively, while FP (false positive) and FN (false negative) are the numbers of negative and positive samples that are incorrectly classified, respectively. In the classification of the thyroid nodules, positive samples are the malignant nodules, and vice versa.

In the rest of this section, we first evaluate the automated detection performance of our proposed multi-scale detection network on Dataset I, and compare it with the state-of-the-art detection methods, including Faster R-CNN, SSD+FPN and Faster R-CNN+FPN without clinical guidance. Then, we evaluate the nodule recognition performance of our proposed multi-branch classification network on Dataset I, and compare it with a traditional hand-crafted-feature-based method, a fuse-feature method (Liu et al., 2017), and also two variants of our triple-branch network, i.e., a single-branch network and a double-branch network. After that, we apply our networks trained on Dataset I to classify samples from Dataset II, and compare our classification performance with the diagnosis results by senior radiologists.

### *3.2. Detection results*

In this group of experiments, we compare our clinical-knowledge-guided multi-scale detection network with Faster R-CNN+FPN method and SSD+FPN method. The corresponding results are summarized in Table 2, where the nodules were divided into four groups (i.e., $<$ 50, 50 - 100, 100 - 200, and $>$ 200) according to the nodule size represented by the square root of the areas. From Table 2, two main conclusions can be drawn. *First*, both the FPN-based methods and our proposed method led to better detection performance than the original Faster R-CNN on all groups of nodules. For example, compared with the Faster R-CNN, our proposed method improved the overall mAP by 3.1% (0.916 *vs.* 0.947). In particular, the mAP for small nodules (i.e., $<$ 50) was significantly increased by 5.2% (0.887. vs. 0.939). The reason why Faster R-CNN had better performance on large nodules than smaller ones is that all nodules were detected on the highest feature map while larger nodules benefited from high-level features with enough res-

**Table 2**
Performance of the multi-scale detection network with different nodule sizes.

| Model | Size | detAP | AP-BN | AP-MN | mAP |
|---|---|---|---|---|---|
| Faster R-CNN | all | 0.925 | 0.911 | 0.921 | 0.916 |
| | <50 | 0.905 | 0.899 | 0.874 | 0.887 |
| | 50–100 | 0.914 | 0.906 | 0.913 | 0.910 |
| | 100–200 | 0.939 | 0.912 | 0.931 | 0.922 |
| | >200 | 0.956 | 0.935 | 0.942 | 0.939 |
| Faster R-CNN + FPN | all | 0.965 | 0.931 | 0.943 | 0.938 |
| | <50 | 0.944 | 0.937 | 0.918 | 0.928 |
| | 50–100 | 0.957 | 0.932 | 0.937 | 0.935 |
| | 100–200 | 0.970 | 0.929 | 0.950 | 0.940 |
| | >200 | 0.971 | 0.938 | 0.944 | 0.941 |
| SSD + FPN | all | 0.954 | 0.926 | 0.942 | 0.934 |
| | <50 | 0.915 | 0.903 | 0.881 | 0.892 |
| | 50–100 | 0.942 | 0.922 | 0.930 | 0.926 |
| | 100–200 | 0.969 | 0.932 | 0.957 | 0.945 |
| | >200 | **0.980** | **0.944** | **0.955** | **0.950** |
| Proposed method | all | **0.977** | **0.942** | **0.951** | **0.947** |
| | <50 | **0.959** | **0.946** | **0.932** | **0.939** |
| | 50–100 | **0.974** | **0.942** | **0.948** | **0.945** |
| | 100–200 | **0.978** | **0.939** | **0.957** | **0.948** |
| | >200 | 0.979 | **0.944** | 0.952 | 0.948 |

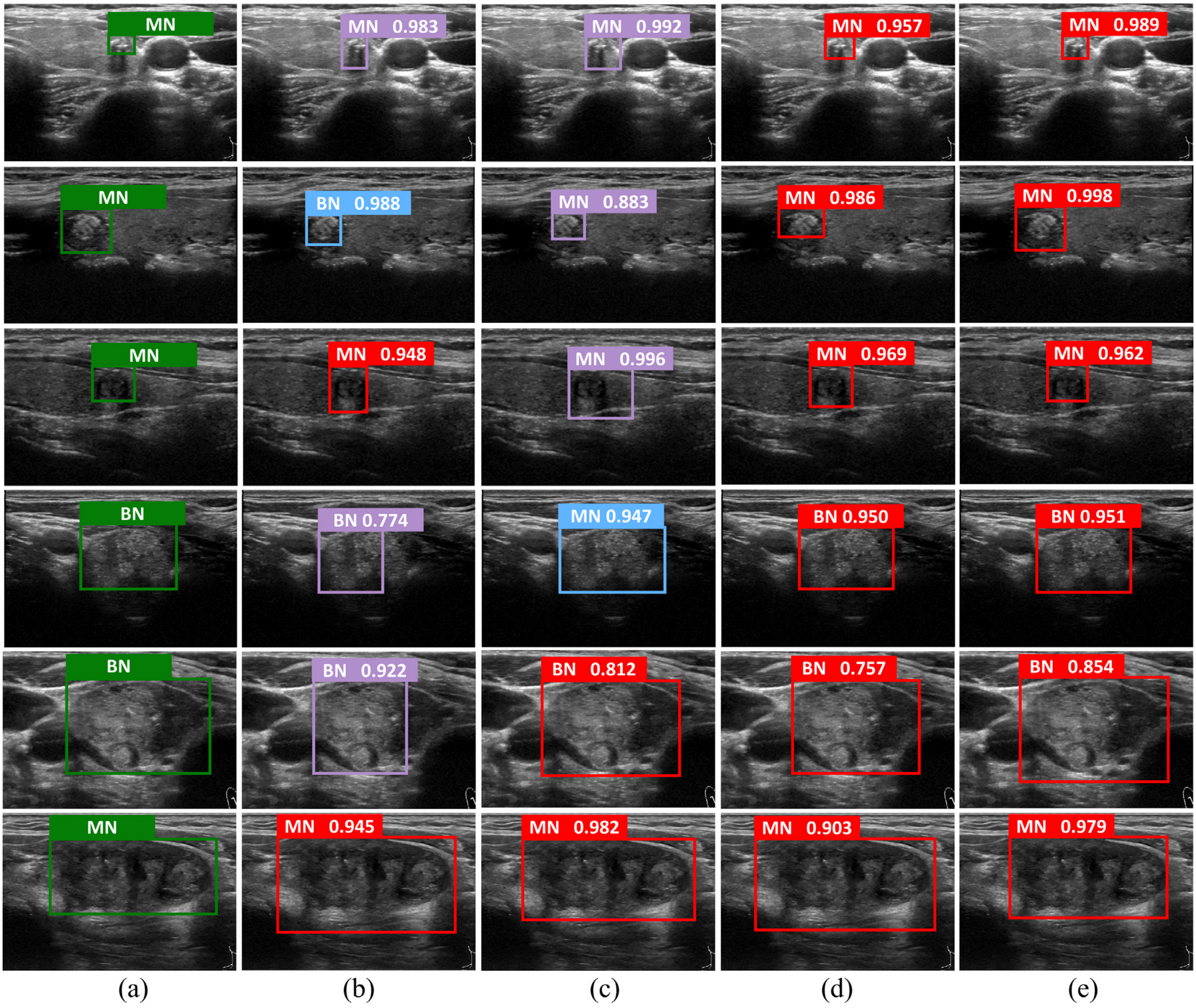

**Fig. 8.** Illustration of the improvement benefiting from our clinical-knowledge-guided multi-scale detection network. The ground-truth ROIs and class labels for typical thyroid nodule images are shown in (a). The automated detections and their corresponding classification scores obtained by Faster R-CNN, Faster R-CNN + FPN, SSD + FPN, and our proposed method are shown in (b), (c), (d) and (e), respectively. Green boxes illustrate the ground-truth ROIs and class labels, red boxes illustrate successfully detected and classified nodules, blue boxes illustrate successfully detected but misclassified nodules, and purple boxes illustrate wrongly detected nodules. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

olution. The improvement suggests that, being superior to mono-scale features used in the Faster R-CNN, leveraging multi-scale pyramidal features does effectively improve the detection accuracy, especially for very small nodules. *Second*, our multi-scale detection network optimized with clinical prior knowledge yielded better performance compared with two general FPN-based methods, e.g., the overall mAP was improved by 0.9% (0.938 *vs.* 0.947) and 1.3% (0.934 *vs.* 0.947) for Faster R-CNN+FPN and SSD+FPN, respectively. It indicates that including clinical prior knowledge and radiologist's experience could effectively refine the feature learning to improve the nodule detection accuracy. Notably, in addition to the overall performance, the promotion was mainly reflected in the detection of small nodules (i.e., $<$ 50), e.g., the detAP was increased by 1.5% (0.944 *vs.* 0.959) compared to Faster R-CNN+FPN, and 4.4% (0.915 *vs.* 0.959) compared to SSD+FPN. Apart from the quantitative comparison presented in Table 2, we also qualitatively compared the detection results, with typical examples shown in Fig. 8. Consistent with quantative results, the detection examples presented in Fig. 8 also demonstrate that our proposed method can detect nodule locations more accurately, especially for small cases.

### 3.3. Classification results

In this group of experiments, we compare our multi-branch classification network with the other four CAD methods, including (1) **HLS-SVM** that trained an SVM classifier with histogram of oriented gradient (HOG), local binary patterns (LBP), and scale-invariant feature transform (SIFT) (Koundal et al., 2013) (2) **Fuse-feature** method (Liu et al., 2017) that combine HOG, LBP and SIFT features with deep features extracted from fine-tuned VGG-F net (Parkhi et al., 2015); (3) **Basic**, a variant of our method, that adopted the first branch of our proposed classification network; and (4) **Context**, a variant of our method, that included only the first two branches of our complete multi-branch network.

**Table 3**
The thyroid nodule classification results obtained by different CAD methods.

| Model | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| HLS-SVM | 0.875 | 0.889 | 0.847 |
| Fuse-feature | 0.948 | 0.968 | 0.908 |
| Basic | 0.934 | 0.96 | 0.883 |
| Context | 0.961 | 0.974 | 0.933 |
| Proposed | **0.971** | **0.982** | **0.951** |

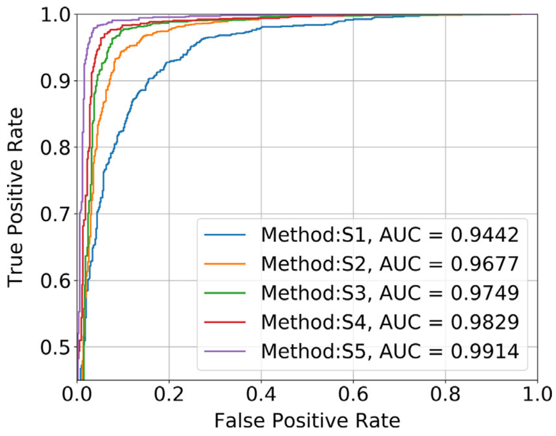

**Fig. 9.** ROC of thyroid nodule classification using different CAD methods. The methods under comparison include the HLS-SVM method (S1), basic CNN (S2), fuse-feature method (S3), CNN with context information(S4), and our full method (S5).

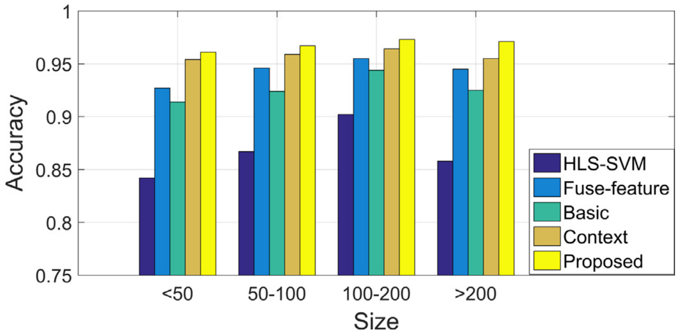

**Fig. 10.** Comparison of classification accuracy of different methods for nodules of different sizes.

**Table 4**
Impact of ratio-preserve strategy.

| Model | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| Non-ratio-preserving | 0.966 | 0.978 | 0.939 |
| Ratio-preserving | **0.971** | **0.982** | **0.951** |

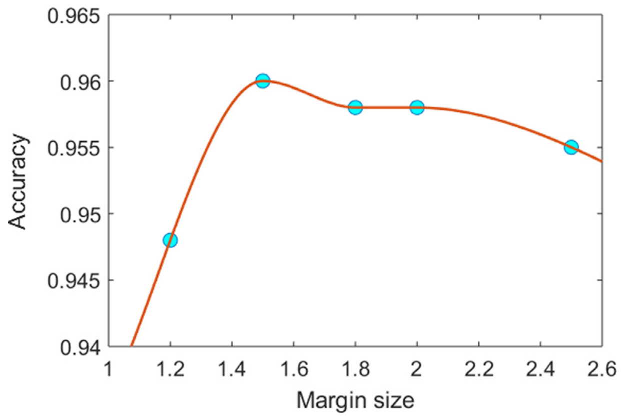

**Fig. 11.** Size selection of the image patch in the context branch.

The classification results obtained by different methods are summarized in Table 3 and Fig. 9, from which we have the following observations. *First*, compared with **HLS-SVM**, the four deep learning methods (i.e., **Fuse-feature, Basic, Context**, and our multi-branch network) obtained better performance, indicating that the learning task-oriented features is beneficial for nodule classification. *Second*, our multi-branch network as well as its variant (i.e., **Context**) consistently outperformed the other two deep learning methods (i.e., **Basic** and **Fuse-feature**), which implies that integrating expert knowledge into deep neural networks could effectively boost the discriminative capacity of automatically-learned features for nodule classification. *Third*, our proposed multi-branch network yielded the best performance in terms of all metrics. For example, it brought 2.2% and 0.8% improvements in identifying malignant nodules compared with Basic (i.e., 0.960 *vs.* 0.982) and Context (i.e., 0.974 *vs.* 0.982), respectively. This verifies the benefit of integrating the contextual and margin information (i.e., the typical characteristics that radiologists pay attention to) in optimizing our CNN-based CAD framework.

To analyze the classification performance in more detail, we further compare those methods on nodules with different sizes, with the corresponding results summarized in Fig. 10. By comparing **Context** and our proposed method with other three methods, we can see that the contextual information captured by the second branch of our network mainly improved the classification of small nodules, e.g., the accuracy on the $<$ 50 group was increased from 0.914 to 0.954 for **Basic** *vs.* **Context**. This is mainly because the neighborhood information provided by surrounding tissues is critical for those small nodules with very limited internal features. On the other hand, by comparing our proposed method with other methods, we can observe that the margin information captured by the third branch of our network effectively improved the recognition of larger nodules, e.g., the accuracy on the $>$ 200 group was increased from 0.955 to 0.971 for Context vs. Proposed. This is due to the fact that the classification of larger nodules relies more on margin knowledge, typically with clearer margin and detailed appearance.

As discussed in Section 2.2.2, an aspect-ratio-preserving strategy is designed to crop ROIs from ultrasound images for training our multi-branch classification network. The motivation is to preserve the original aspect ratio information for nodules with different shapes, considering its value for the classification task. To evaluate its effectiveness, Table 4 compares the classification results obtained by our proposed method with and without this aspect-ratio-preserving strategy. It is found that this strategy effectively improved the performance in terms of all metrics, verifying that the retained aspect ratio and shape information are beneficial for nodule recognition.

To select the proper size of image patches in the context branch, we have compared 1.2, 1.5, 1.8, 2.0 and 2.5 margins of detected nodule and the corresponding diagnosis accuracies obtained by the Context method was shown in the Fig. 11. It can be seen that the accuracy turned out to be the highest around 1.5 margin and kept at a relatively high level in the range from 1.5 to 2.0. Therefore 1.5 margin was used for diagnosis in our method.

### 3.4. Automated classification versus radiologists diagnosis

Apart from the comparison with CAD methods, we further compare our proposed method with senior radiologists on Dataset II to evaluate its clinical values. The same as that in Dataset I, all

**Table 5**
The diagnostic results obtained by experienced radiologists and our proposed CAD method on Dataset II.

| Method | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| Radiologists | 0.867 | 0.931 | 0.711 |
| Proposed CAD | **0.949** | **0.972** | **0.891** |

nodules in Dataset II were pathologically verified by FNA biopsy and histology evaluation to define the binary label (i.e., malignant/benign). To evaluate the diagnosis performance under different situations in detail, nodules in Dataset II were further separated as multiple groups by 6 sonographic characteristics: nodule size in terms of maximum diameter (i.e., < 0.5 cm, 0.5-1.0 cm, or > 1.0 cm), margin smoothness(smooth, partly smooth, or indistinct), shape (i.e., irregular lobulation or round), aspect ratio (i.e., < 1 or > 1), composition (i.e., solid or cystic), and calcification (i.e., microcalcification or non-microcalcification). Notably, these characteristics are radiologist's attention when making a diagnostic decision. There were totally 5 radiologists attending nodule analysis, who have 6 to 10 years working experience on thyroid diagnosis, with an average annual diagnosis of more than 1000 patients. Each nodule was analyzed by at least two radiologists on the above sonographic characteristics and final diagnosed decision came from radiologists' discussion.

To more clearly analyze what kind of thyroid nodules our proposed CAD system can provide reliable suggestions, we first compared accuracy, sensitivity, and specificity the on the whole Dataset II to get overall performance comparison with radiologists. Then, we further analyze the diagnosis performance in detail on different feature groups of each sonographic characteristic (e.g. solid-group nodules in terms of characteristic 'composition').

The overall performance is summarized in Table 5. The result indicates that our proposed CAD method yielded an overall better performance than the attending radiologists, with improvements of accuracy, sensitivity, and specificity for 8%, 4%, and 18%, respectively.

The detailed results obtained by our proposed method and radiologists on each nodule group are summarized in Table 6. By comparing radiologist's performance across different nodule groups, we can observe that radiologists had relative high accuracy in diagnosing (1) nodules with size larger than 1 cm, (2) nodules with smooth margin, (3) cystic or partially cystic nodules, (4) nodules with irregular lobulation shape, and (5) nodules with aspect ratio larger than 1. Diagnosing those nodules are relatively easier in practice, considering that they usually have typical benign or malignant patterns that easy to be recognized. On these groups of nodules, by strengthening the learning of radiologists' attention, our proposed method achieved competitive high diagnosis accuracy. For example, by adding margin branch, our model successfully captured category-sensitive margin features and achieved high sensitivity on smooth nodules (i.e. 1.000) and irregular lobulation nodules (i.e. 0.968). On the other hand, radiologist's performance decreased when nodules do not have obvious benign/malignant patterns, especially when benign nodules have the sonographic characteristics very close to malignant ones, e.g., nodules with indistinct margin (radiologist's specificities were 0.425). In contrast, we can observe that our method consistently led to much better performance on these nodule groups. For example, our method promoted diagnostic accuracy for the nodules with indistinct boundaries, partly smooth boundaries, aspect ratio less than 1, and solid composition by 8.5%, 7.2%, 8.7%, and 6.6%, respectively. In addition, it is worth mentioning that radiologists were prone to give malignance suggestion when nodules have malignant characteristics. Our method significantly improved the specificity (i.e., reduced the false positive diagnosis), such as for nodules with irregular lobulation shape (0.214 vs. 0.821) and nodules with solid composition (0.548 vs. 0.790). The improvement came from the extraction and combination of high-level features that are hard to be described directly by radiologists, but discriminative for the differentiation between benign and malignant nodules. In conclusion, the experimental results presented in Table 6 supports our assumption that using clinical prior knowledge and radiologist's knowledge to guide the learning of high-level convolutional features could effectively enhance the influence of critical sonographic characteristics and finally improve the diagnostic performance.

### 3.5. Hard subjects analysis

We also analyzed the diagnostic performance of different CAD methods and experienced radiologists on challenging subjects. Specifically, from Dataset II, radiologists identified 207 suspicious cases (166 malignant and 41 benign in pathology) that are hard to be diagnosed (i.e., the evidence of malignant/benign is not obvious). The diagnostic results achieved by different methods on hard subjects are summarized in Table 7. We can observe that the misdiagnosis rate of benign nodules by radiologists was very high, considering that the specificity was only 0.366. The main reason could be that those benign nodules misdiagnosed by radiologists are follicular cells and thyroid adenomas which have one or more indications of malignancy, such as irregular lobes, blurred borders, or heterogeneous hypoechoic. In contrast, our proposed CAD

**Table 6**
Comparison of diagnosis results in different morphological features groups.

| Characteristics | | Radiologists | | | Proposed method | | |
|---|---|---|---|---|---|---|---|
| | | Accuracy | Sensitivity | Specificity | Accuracy | Sensitivity | Specificity |
| Size | <0.5cm | 0.766 | 0.729 | 0.849 | **0.925** | **0.932** | **0.909** |
| | 0.5-1.0cm | 0.845 | **0.964** | 0.520 | **0.925** | 0.956 | **0.860** |
| | >1.0cm | 0.909 | **0.991** | 0.711 | **0.958** | 0.982 | **0.898** |
| Margin | Smooth | 0.943 | 0.900 | **0.961** | **0.951** | **1.000** | 0.942 |
| | Indistinct | 0.866 | 0.940 | 0.425 | **0.952** | **0.972** | **0.749** |
| | Partly smooth | 0.829 | 0.902 | 0.474 | **0.901** | **0.935** | **0.737** |
| Shape | Irregular lobulation | 0.876 | 0.964 | 0.214 | **0.918** | **0.968** | **0.821** |
| | Round | 0.836 | 0.833 | 0.839 | **0.910** | **0.917** | **0.903** |
| Aspect ratio | >1 | 0.910 | 0.952 | 0.286 | **0.963** | **0.970** | **0.714** |
| | <1 | 0.860 | 0.896 | 0.694 | **0.947** | **0.953** | **0.917** |
| Composition | Solid | 0.859 | 0.918 | 0.548 | **0.925** | **0.965** | **0.790** |
| | Cystic/partially cystic | 0.931 | **0.938** | 0.929 | **0.944** | 0.938 | **0.946** |
| Calcification | Microcalcification | 0.889 | **0.968** | 0.231 | **0.940** | 0.968 | **0.762** |
| | Non-Microcalcification | 0.856 | 0.892 | 0.784 | **0.951** | **0.961** | **0.931** |

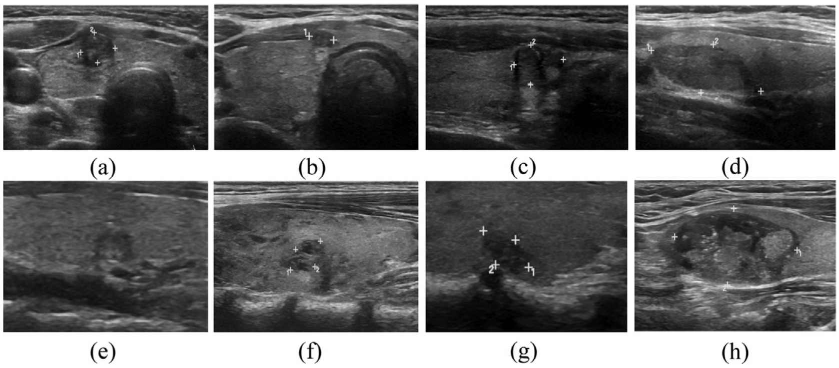

**Fig. 12.** Typical hard subjects that were correctly diagnosed by our proposed method while misdiagnosed by radiologists. (a)–(c) are malignant nodules, (d)-(h) are benign nodules. (a) papillary carcinoma with round shape; (b) papillary carcinoma with round shape and homogeneous echo; (c) hypoechoic papillary carcinoma; (d) follicle cell with smooth margin; (e) thyroid adenoma with blurred margin; (f) thyroid adenoma with irregular lobulation; (g) follicle cell with aspect ratio>1; (h) follicle cell with microcalcification.

**Table 7**
Diagnosis results obtained by radiologists and CAD methods for hard subjects from Dataset II.

| Method | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| Radiologists | 0.816 | 0.928 | 0.366 |
| HLS-SVM | 0.835 | 0.921 | 0.488 |
| Fuse-feature | 0.889 | 0.940 | 0.682 |
| Proposed | **0.928** | **0.964** | **0.780** |

method led to significantly better specificity (i.e., 0.780) and sensitivity (i.e., 0.964) than radiologists as well as the other two CAD methods, indicating its high reliability in diagnosing hard subjects from both the benign and malignant categories. Typical hard subjects misdiagnosed by radiologists but correctly-classified by our proposed method are shown in Fig. 12, from which we can have a consistent observation, i.e., our proposed method can correctly classify hard nodules without obvious malignant or benign sonographic characteristics. The above analyses further confirm that the proposed method can not only absorb radiologist's diagnostic experience to guide automated classification, but also extract more discriminative features that are difficult to be described by radiologists for robust nodule classification.

## 4. Discussion and conclusion

In this study, we proposed a deep-learning-based CAD system for automated detection and classification of thyroid nodules. Our system includes a nodule detection stage and a nodule categories classification stage (i.e., benign or malignant). We demonstrated that introducing clinical prior knowledge into multi-scale detection network did help improve detection accuracy. We also transfered radiologists' expert knowledge into automated nodule classification. The diagnostic accuracy achieved by our proposed system was higher (e.g., 0.971 on Dataset I) than existing CAD approaches for thyroid nodule classification. Furthermore, the accuracy, sensitivity and specificity of our CAD system were superior to that of experienced radiologists.

1) Advantages. Our proposed CAD system has multiple potential advantages in detecting and classifying thyroid nodules. First, it is robust to the nodule size in the task of nodule detection, due to the multi-scale detection network architecture. Experimental results shown in Section 3.2 demonstrated that our multi-scale detection network achieved better accuracy compared with the mono-scale detection method, especially for very small nodules. These results suggested that low-level features were complemented by semantic features to improve the ability of nodule description while maintaining the high-resolution property. Furthermore, we defined detection anchors according to the morphological distribution of nodules. Our experiments evidenced that clinical restriction in designing anchors optimized the initialization and filtering of proposals, which made it easier to regress positive proposals to the ground truth and led to more precise nodule locations. This result implied that our system could be competent to assist radiologists to screen nodules and avoid missing small nodules, saving their time.

Second, our method leads to superior classification accuracy, especially on the very challenging cases, mainly due to the complementary multi-view feature learning in our multi-branch classification network guided by radiologist's attention and experience. The contextual features captured by the second branch provided critical neighborhood information of those small nodules with very limited internal features, while the margin branch effectively improves the recognition of larger nodules since that their local aggressive behavior or other proliferative features which related to margin is usually obvious enough. Compared to other CNN-based and traditional hand-crafted-feature-based approaches for thyroid nodule classification (Chi et al., 2017; Ma et al., 2016; Tsantis et al., 2009), our method introduces clinical knowledge into deep learning model in an easy-to-use manner. Rather than directly modeling some clinical knowledge (i.e. irregular lobulation) into several specific indicators such as radius entropy (Tsantis et al., 2009), our method extractes sensitive-feature intensive regions (according to experts' attention) as CNN inputs to preserve the CNN's advantage of acquiring high-level semantic features. Experimental results shown in Section 3.3 demonstrates that our proposed multi-branch classification network consistently outperformed the state-of-the-art automated methods in diagnosing various types of nodules. Compared with radiologists, as shown in Sections 3.4 and 3.5 our method reached the high accuracy of radiologists in groups of nodules with obvious malignant features and significantly improved the accuracy of nodules without obvious features. It implied that our method effectively absorbs radiologist's diagnostic experience to guide automated classification while generating more distinctive deep features.

Noticing that the additional branches introduce more parameters which may bring performance improvement, we thus constructed an ensemble model using the same three-branch structure as our method to verify that the performance boost does benefit from the integration of the contextual and margin information. Each branch was trained with same data and different random initializer, and the final ensemble accuracy was 0.952, 1.9% lower than our proposed method. It indicates that although introducing more parameters do increase the performance to some extent, our

**Table 8**
Improvement of diagnosis accuracy with different backbones. ‘Original’ represents directly using backbone network.

| Backbone | Method | Accuracy | Improvement |
|---|---|---|---|
| Refined ZFnet | Original | 0.934 | 3.70% |
| | Our method | 0.971 | |
| DenseNet-121 | Original | 0.954 | 2.40% |
| | Our method | 0.978 | |

**Table 9**
Improvement of detection mAP with different backbones. ‘FRCNN+FPN’ represents Faster R-CNN+FPN method.

| Backbone | Method | mAP | Improvement |
|---|---|---|---|
| GoogLeNet | FRCNN+FPN | 0.929 | 1.30% |
| | Our method | 0.942 | |
| ResNet-50 | FRCNN+FPN | 0.938 | 0.90% |
| | Our method | 0.947 | |

method further boosts the performance improvements beyond the benefits from the use of more parameters.

2) Clinical benefit. As our method is able to effectively distinguish malignant nodules from benign nodules, practical clinical benefits in clinic are expected from three possible aspects. *First*, the increasing prevalence of thyroid nodules takes radiologists plenty of time to screen malignant nodules. For example, Cancer Hospital of the Chinese Academy of Medical Sciences accepts more than 100 first-visit-patient every day and each nodule would take radiologist at least three to five minutes to diagnose. High-quality nodule classification is conducive to liberating radiologists from the heavy reading work. *Second*, the diagnosis accuracy varies greatly across individuals. Our method learning from expert experience will potentially be a meaningful assistant for radiologists to reduce misdiagnosis rate, especially for inexperienced radiologists. *Third*, for those nodules which are difficult to diagnosis, doctors usually suggest analysis through FNA. However, FNA is an invasive examination which may bring extra harm including pain, haematoma and complications to patients. In addition, at least half of all biopsied nodules will be finally categorized as benign (Grani et al., 2018). As our method has high classification accuracy (especially in benign nodules), it will potentially reduce the unnecessary FNA biopsies. To further interpret this point statistically, we compared the reduction of false positive samples of different methods discussed in Section 3.5. Based on the ATA management guidelines (HaugenBryan et al., 2016), 88 nodules from the 207 suspicious samples need FNA biopsy according to the radiologist. Our proposed method reduced this number to 76 comparing to 83 of HLS-SVM and 81 of Fuse-feature method, which verified the feasibility of our method in reducing the amount of FNA biopsies.

3) Impact of backbone networks. Different backbone networks may influence the performance of our method. In this part, we mainly discuss about performance improvement of our proposed method over different backbone networks. In the classification task, we compared our method using refined-ZFnet with a more complicated backbone, i.e. DenseNet-121 (Iandola et al., 2014). DenseNet model increases the network depth while keeping the number of parameters small. The comparison experiment results in Table 8 showed that our proposed method increased accuracy of 2.4% compared with original DenseNet baseline. The results proved that the clinical prior knowledge we used can improve the diagnosis results under different networks. Similarly, we compared the performance of our detection method using GoogLeNet and ResNet as backbone in Table 9. It is worth mentioning that, our method had overall comparable performance on these two backbones and when using ResNet our method still got performance improvement.

4) Limitations. Although our proposed CAD system effectively improved thyroid nodule classification performance, it still has several limitations. *First*, the datasets (both Dataset I and Dataset II) used for building and evaluating our proposed CAD system are category imbalanced, which contain more malignant nodules than benign ones. This imbalance limits the improvement of specificity. The practical challenge is that high-quality large-scale medical image data is hard to obtain. Thus, we should further collect data and explore the methodological solution to deal with small sample and imbalance problem. *Second*, the annotation is coarse, as no specific margin is known. A more precise annotation (i.e. pixel annotation) of nodules margin may improve the diagnostic accuracy. In this direction, some of the recent works (Ma et al., 2017; Ying et al., 2018) made efforts on nodule segmentation. *Third*, it is noticed that the speed of the region-proposal-based detection method is lower than that of the SSD-like methods, which might be one potential shortcoming to be overcome in the future research.

In conclusion, a deep-learning-based CAD framework guided by clinical knowledge and expert experience, consisting of a multi-scale detection network and a multi-branch classification network, has been proposed for thyroid nodule detection and classification in ultrasound images. The proposed method has shown superior performance than the state-of-the-art automated methods as well as experienced radiologists in terms of both the detection and classification accuracy. Our proposed CAD method can be used as a reliable second opinion for radiologists to help them avoid misdiagnosis due to overloaded work. Furthermore, it could give helpful suggestions for junior radiologists without enough clinical experience.

## Declaration of Competing Interest

None.

## References

Acharya, U.R., Sree, S.V., Krishnan, M.M.R., Molinari, F., ZieleŸnik, W., Bardales, R.H., Witkowska, A., Suri, J.S., 2014. Computer-aided diagnostic system for detection of hashimoto thyroiditis on ultrasound images from a polish population. J. Ultrasound Med. 33 (2), 245–253.

Agarwal, D., Shriram, K.S., Subramanian, N., 2013. Automatic view classification of echocardiograms using histogram of oriented gradients. In: IEEE International Symposium on Biomedical Imaging, pp. 1368–1371.

Ali Abbasian, A., Akbar, G., Afshin, M., 2015. Classification of benign and malignant thyroid nodules using wavelet texture analysis of sonograms. J. Ultrasound Med. Off. J. Am. Inst. Ultrasound Med. 34 (11). 1983–9

Anil, G., Hegde, A., Chong, F.H.V., 2011. Thyroid nodules: risk stratification for malignancy with ultrasound and guided biopsy. Cancer Imaging 11 (1), 209–223.

Chang, C.Y., Chen, S.J., Tsai, M.F., 2010. Application of support-vector-machine-based method for feature selection and classification of thyroid nodules in ultrasound images. Pattern Recognit. 43 (10), 3494–3506.

Chi, J., Walia, E., Babyn, P., Wang, J., Groot, G., Eramian, M., 2017. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. J. Digital Imaging 30 (4), 477–486.

Danese, D., Sciacchitano, S., Farsetti, A., Andreoli, M., Pontecorvi, A., 1998. Diagnostic accuracy of conventional versus sonography-guided fine-needle aspiration biopsy of thyroid nodules. Thyroid 8 (1), 15–21.

Ding, J., Cheng, H.-D., Huang, J., Zhang, Y., 2014. Multiple-instance learning with global and local features for thyroid ultrasound image classification. In: 2014 7th International Conference on Biomedical Engineering and Informatics. IEEE, pp. 66–70.

Dorin, B., Luminita, M., Anjan, B., 2013. Thyroid nodule recognition based on feature selection and pixel classification methods. J. Digital Imaging 26 (1), 119–128.

Gabriella, P., Francesco, F., Concetto, R., Sebastiano, S., Riccardo, V., 2013. Worldwide increasing incidence of thyroid cancer: update on epidemiology and risk factors. J. Cancer Epidemiol. 2013 (1), 965212.

Gao, L., Liu, R., Jiang, Y., Song, W., Wang, Y., Liu, J., Wang, J., Wu, D., Li, S., Hao, A., 2017. Computer-aided system for diagnosing thyroid nodules on ultrasound: a comparison with radiologist-based clinical assessments. Head Neck 40 (4).

Gao, Y., Maraci, M.A., Noble, J.A., 2016. Describing ultrasound video content using deep convolutional neural networks. In: IEEE International Symposium on Biomedical Imaging, pp. 787–790.

Girshick, R., 2015. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448.

Grani, G., Lamartina, L., Ascoli, V., Bosco, D., Biffoni, M., Giacomelli, L., Maranghi, M., Falcone, R., Ramundo, V., Cantisani, V., et al., 2018. Reducing the number of unnecessary thyroid biopsies while improving diagnostic accuracy: toward the “right” tirads. J. Clin. Endocrinol.Metab. 104 (1), 95–102.

Gul, K., Ozdemir, D., Korukluoglu, B., Ersoy, P.E., Aydin, R., Ugras, S., Ersoy, R., Cakir, B., 2010. Preoperative and postoperative evaluation of thyroid disease in patients undergoing surgical treatment of primary hyperparathyroidism. Endocr. Pract. Official J. Am. Coll. Endocrinol. Am. Assoc. Clin. Endocrinologists 16, 7–13.

HaugenBryan, R., AlexanderErik, K., BibleKeith, C., DohertyGerard, M., MandelSusan, J., NikiforovYuri, E., RandolphGregory, W., SawkaAnna, M., SchuffKathryn, G., ShermanSteven, I., et al., 2016. 2015 American thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the american thyroid association guidelines task force on thyroid nodules and differentiated thyroid cancer. Thyroid.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hee Jung, M., Young, K.J., Eun-Kyung, K., Jung, K.M., 2011. A taller-than-wide shape in thyroid nodules in transverse and longitudinal ultrasonographic planes and the prediction of malignancy. Thyroid Off. J. Am. Thyroid Assoc. 21 (11), 1249.

Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., Keutzer, K., 2014. Densenet: implementing efficient convnet descriptor pyramids. arXiv:1404.1869v1.

Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. Science 353 (6301), 790.

Keramidas, E.G., Iakovidis, D.K., Maroulis, D., Dimitropoulos, N., 2008. Thyroid texture representation via noise resistant image features. In: Proceedings of the IEEE Symposium on Computer-Based Medical Systems, pp. 560–565.

Koundal, D., Gupta, S., Singh, S., 2013. Survey of computer-aided diagnosis of thyroid nodules in medical ultrasound images. Adv. Intell. Syst. Comput. 177, 459–467.

Kouvaraki, M.A., Shapiro, S.E., Fornage, B.D., Edeiken-Monro, B.S., Sherman, S.I., Vassilopoulou-Sellin, R., Lee, J.E., Evans, D.B., 2003. Role of preoperative ultrasonography in the surgical management of patients with thyroid cancer. Surgery 134 (6), 946–954.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.

Lian, C., Liu, M., Zhang, J., Shen, D., 2019. Hierarchical fully convolutional network for joint atrophy localization and alzheimer’s disease diagnosis using structural mri. IEEE Trans. Pattern Anal. Mach. Intell.

Lian, C., Zhang, J., Liu, M., Zong, X., Hung, S.-C., Lin, W., Shen, D., 2018. Multi-channel multi-scale fully convolutional network for 3d perivascular spaces segmentation in 7t mr images. Med. Image Anal. 46, 106–117.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: CVPR, 1, p. 4.

Liu, T., Xie, S., Yu, J., Niu, L., Sun, W., 2017. Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features. In: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, pp. 919–923.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. Ssd: single shot multibox detector. In: European Conference on Computer Vision. Springer, pp. 21–37.

Ma, J., Wu, F., Zhao, Q., Kong, D., et al., 2017. Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. Int. J. Comput. Assisted Radiol. Surg. 12 (11), 1895–1910.

Ma, J., Wu, F., Zhu, J., Xu, D., Kong, D., 2016. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. Ultrasonics 73, 221–230.

Parkhi, O.M., Vedaldi, A., Zisserman, A., et al., 2015. Deep face recognition.. In: BMVC, 1, p. 6.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39.

Si, L., Kim, E.H., Dighe, M., Kim, Y., 2011. Thyroid nodule classification using ultrasound elastography via linear discriminant analysis. Ultrasonics 51 (4), 425–431.

Singh, N., Jindal, A., 2012. Ultra sonogram images for thyroid segmentation and texture classification in diagnosis of malignant (cancerous) or benign (non-cancerous) nodules. Int. J. Eng. Innovative Technol. (IJEIT) 1 (5), 202–206.

Song, W., Li, S., Liu, J., Qin, H., Zhang, B., Shuyang, Z., Hao, A., 2018. Multi-task cascade convolution neural networks for automatic thyroid nodule detection and recognition. IEEE J. Biomed. Health Inf.

Tessler, F.N., Middleton, W.D., Grant, E.G., Hoang, J.K., Berland, L.L., Teefey, S.A., Cronan, J.J., Beland, M.D., Desser, T.S., Frates, M.C., et al., 2017. Acr thyroid imaging, reporting and data system (ti-rads): white paper of the acr ti-rads committee. J. Am. Coll. Radiol. 14 (5), 587–595.

Tsantis, S., Dimitropoulos, N., Cavouras, D., Nikiforidis, G., 2009. Morphological and wavelet features towards sonographic thyroid nodules evaluation. Comput. Med. Imaging Graphics 33 (2), 91–99.

Xu, C., Prince, J.L., 1998. Generalized gradient vector flow external forces for active contours1. Signal Process. 71 (2), 131–139.

Xu, Y., Jia, Z., Ai, Y., Zhang, F., Lai, M., Chang, E., 2015. Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation, pp. 947–951.

Ye, J., Chow, J.H., Jiang, C., Zheng, Z., 2009. Stochastic gradient boosted distributed decision trees. In: Acm Conference on Information & Knowledge Management, pp. 2061–2064.

Ying, X., Yu, Z., Yu, R., Li, X., Yu, M., Zhao, M., Liu, K., 2018. Thyroid nodule segmentation in ultrasound images based on cascaded convolutional neural network. In: International Conference on Neural Information Processing. Springer, pp. 373–384.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. Springer, pp. 818–833.