Multi-task Cascade Convolution Neural Networks for Automatic Thyroid Nodule Detection and Recognition

Wenfeng Song, Shuai Li, Ji Liu, Hong Qin, Bo Zhang, Shuyang Zhang, and Aimin Hao

***Abstract*—Thyroid ultrasonography is a widely-used clinical technique for nodule diagnosis in thyroid regions. However, it remains difficult to detect and recognize the nodules due to low contrast, high noise, and diverse appearance of nodules. In today's clinical practice, senior doctors could pinpoint nodules by analyzing global context features, local geometry structure, and intensity changes, which would require rich clinical experience accumulated from hundreds and thousands of nodule case studies. To alleviate doctors' tremendous labor in the diagnosis procedure, we advocate a machine learning approach to the detection and recognition tasks in this paper. In particular, we develop a multi-task cascade convolution neural network framework (MC-CNN) to exploit the context information of thyroid nodules. It may be noted that, our framework is built upon a large number of clinically-confirmed thyroid ultrasound images with accurate and detailed ground truth labels. Other key advantages of our framework result from a multi-task cascade architecture, two stages of carefully-designed deep convolution networks in order to detect and recognize thyroid nodules in a pyramidal fashion, and capturing various intrinsic features in a global-to-local way. Within our framework, the potential regions of interest after initial detection are further fed to the spatial pyramid augmented CNNs to embed multi-scale discriminative information for fine-grained thyroid recognition. Experimental results on 4309 clinical ultrasound images have indicated that, our MC-CNN is accurate and effective for both thyroid nodules detection and recognition. For the correct diagnosis rate of malignant and benign thyroid nodules, its mAP performance can achieve up to 98.2% accuracy, which outperforms the common CNNs by 5% on average. In addition, we conduct rigorous user studies to confirm that our MC-CNN outperforms experienced doctors, yet only consuming roughly 2% $(1/48)$ of doctors' examination time on average. Therefore, the accuracy and efficiency of our new method exhibit its great potential in clinical applications.**

***Index Terms*—Thyroid Nodules, Detection, Recognition, Pyramid Convolution Neural Networks.**

I. Introduction and Motivation

THYROID nodule is one of the most commonly-observed nodular lesions, with the prevalence of 19% to 68% in general population. Now, we have been witnessing about 240% increase in thyroid cancer during the past thirty years [1], which is one of the worst among all types of cancers [2], [3]. At the imaging front, ultrasonography has been a dominant and preferred screening modality towards the clinical diagnosis of thyroid nodules, which is also used as guidance for fine-needle aspiration biopsy (FNAB) and subsequent treatments [4], [5], [6]. Recently, many guidelines have been established for radiologists to evaluate thyroid nodules based on ultrasound characteristics [7], [8], [9]. However, since ultrasonography is susceptible to echo disturbances and speckle noises, ultrasonography based thyroid nodule diagnosis still heavily relies on rich experiences and delicate skills of senior radiologists. Less experienced practitioners may potentially have high misdiagnosis rate due to their inability of accurately comprehending ultrasonography characteristics. Mis-diagnosis might consequently call for unnecessary biopsy and surgery, that would make patients have much more pressure and anxiety, and at the same time unavoidably increase medical expense. To effectively leverage the high-quality diagnosis experiences gained by senior radiologists, smart thyroid diagnosis CADx system is urgently needed. Yet, the key success of the smart thyroid diagnosis CADx system build-up may be hindered by the fact that, the ultrasound thyroid's appearances are frequently influenced by internal content, shape, echogenicity, and many other factors, as shown in Fig. 1.

lishuaiouc@126.com

The benign nodules and the malignant nodules both have a wide variety of styles and layouts. Fig. 1(a) shows the benign nodules, and most of them have irregular shapes, smooth regions, and boundaries. Fig. 1(b) shows the malignant nodules, and most of them have irregular shapes, coarse regions, and boundaries. Therefore, the thyroid nodules are hard to be directly recognized based on color and shape features.

In recent years, there exist many studies that employ sonographic features for thyroid malignancy diagnosis, which can be roughly classified into two main categories: hand-crafted feature based classifiers [10], and the data-driven methods.

Hand-craft Feature Methods. The pipeline of these methods frequently involves feature extraction and classification. Typical methods in this category may include, GLCM, LBP, Discrete Wavelet Transform (DWT), K-Nearest Neighbor (K-NN), Probabilistic Neural Network (PNN), Decision Tree (DT), Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Adaboost classifier, [11], [12], [13], [14], [15], [16], [17] Bayesian classifier, GBDT [16] and random forest [17]. Despite their rapid development in recent years, handcrafted features in some sense can only exploit the low-level information, such as image texture [12], [13], geometry morphology [14], and statistical distributions [11]. Such methods usually need to further employ classifiers to conduct classification. Hence, only if given highly-discriminative features, such methods could well solve the recognition problem.

Recently, some works focused on the characteristics of malignant thyroid in high resolution ultrasound images and ultrasound elastography. For example, Acharya et al. [18]

proposed a novel Gabor transform based automated system for the classification of benign and malignant thyroid nodules using high resolution ultrasound (HRUS). Sun et al. [19] demonstrated that ultrasound elastography had high sensitivity and specificity for the identification of thyroid nodules. Meanwhile, Raghavendra et al. [20] fused the spatial gray level dependence features (SGLDF) with the fractal textures to decipher the intrinsic structure of benign and malignant thyroid lesions. Nevertheless, for thyroid nodules, the high variability of the ultrasound image makes it hard to effectively distinguish benign nodules from malignant ones. So it is even more critical to design and select the most significant features, let alone comprehensively fusing different scales of local and global features as experienced radiologists would do. Besides, available classifiers mostly tend to over-fit the training dataset,since the features locally designed at single scale and in single region are insufficient to encode critical information in order to determine different types of nodules.

Data-driven Learning Methods. As for the data-driven methods, recently the convolution neural networks (CNNs) can greatly improve the classification and detection performance on natural images without the need of hand-crafted feature description, such as, Alexnet [21], GoogLenet [22], Residual net [23], Faster RCNN [24], Single Shot Detection (SSD) [25]. etc. One salient advantage of CNNs is that, they could overcome the aforementioned difficulties by extracting multi-level features automatically. Now, even though it is possible to use hybrid CNNs to classify the thyroid nodules [2], it is still much more complex and redundant to extract features with multiple scales in CNNs. For example, existing methods oftentimes fail to recognize nodules of smaller scales or lower contrast, and this is especially true for thyroid nodules that would rely on proper recognition of their neighboring tissues towards correct diagnosis. The features translated by CNN are aggregated in multiple-level layers. The lower levels represent the shallow features like shapes, gradient, and color appearance, while the high level features represent the semantic discriminative features.

Compared with the traditional feature extraction methods, it is demonstrated in [25] that CNN has two advantages. (1) The detection based on CNN features is robust to distortions, including changes caused by camera lens, different lighting conditions, different poses, partial occlusions, horizontal and vertical shifts, etc.; (2) The computational cost of CNN-based feature extraction is relatively low, because the same coefficients in the convolutional layer are used across the input images. Motivated by the success in natural image recognition, some recent works were proposed to apply the CNNs to thyroid recognition.

However, the instinct limitation of existing CNNs is that, they only consider single-region features, wherein feature kernels locally focus on the single scale perception while ignoring the corresponding context information. Moreover, almost all of the existing methods tend to separately conduct detection and classification tasks, which may easily make the information isolated between the two tasks. However, the features should be shared and complemented with each other in both scales and context. In this circumstance, Liu et al. [25] propose a single shot detection (SSD) network to detect and recognize objects from high-quality natural images. Inspired by such method, we employ a coarse-to-fine pyramid framework E_c for the thyroid nodule classification on 2D ultrasound images. One advantage of the coarse-to-fine network is that, a multi-task joint network would enable to learn the detection and classification tasks at the same time in a mutually-supplementary way, just in the same way as what experienced radiologists would do. Moreover, the spatial pyramid network could extract features that would continuously change from global to local, which in fact imitates the process of simultaneously considering the neighboring regions and the high-level semantic features.

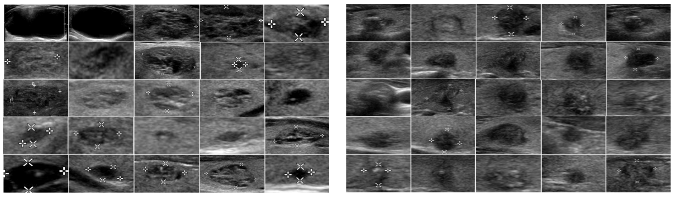

(a) Benign (b) Malignant

Fig. 1: Illustration of thyroid nodules: (a) Benign nodules; (b) Malignant nodules. Calipers labels are shown for better understanding only, and are not used for training and testing.

To achieve the goal of incorporating effective global features into our smart diagnosis system, we propose a multi-task cascade pyramid CNN framework (MC-CNN) to jointly learn multi-level features, as shown in Fig. 2 and detailed in Algorithm 1. In contrast to the existing CNN methods [21], [23], we extend the single-scale network to the pyramid based coarse-to-fine spatial convolution network, where the integrated local and global clues in concert could make the final prediction much more reliable. The critical clinical testing has shown that, our approach could achieve the state-of-the-art performance in a variety of real patients' datasets, including the datasets with different scales and from patients with different-age groups. Specifically, the salient contributions of this paper can be summarized as follows:

- We propose a multi-task cascade CNN framework, to jointly perform thyroid detection and recognition in a coarse-to-fine manner, which supports coarsely locating and classifying nodules on the entire ultrasound image to produce potential nodule proposals first, and then pinpointing nodules in a much-finer scale based on spatial pyramid CNNs in real time.
- We propose a multi-scale single shot detection (SS-D) network guided by the nodule prior, which could greatly simplify the diagnosis process and could well accommodate ultrasound datasets with different scales for patients of different-age groups compared with the currently available methods.
- We propose to embed a spatial pyramid module into traditional CNNs in order to refine thyroid nodule recognition, which could leverage both the global context and locally detailed information, thus giving rise to much better discrimination capability on nodule types.

Algorithm 1 Pipeline of MC-CNN framework

```
Input:
1: $x$: The set of samples with labels $y$ (class labels) and $l$ (bounding boxes position);
2: $E_0$: Pre-trained VGG-16 Network model on Imagenet dataset;
Output: Cascade models $E_c$ (Detection and Coarse Recognition model), $E_f$ (Fine Recognition Model)
3: repeat
4:     Training detection and coarse recognition model weight $\mathbf{w}$ with multi-scale SSD network $E_c$, and finely tuning on the VGG-16 network $E_0$;
5:     Using the trained model $E_c$ to predict the bounding boxes and types of the nodules;
6:     The results are further refined with a fine recognition network $E_f$;
7: until $(L_x < \delta(0.01))$
```

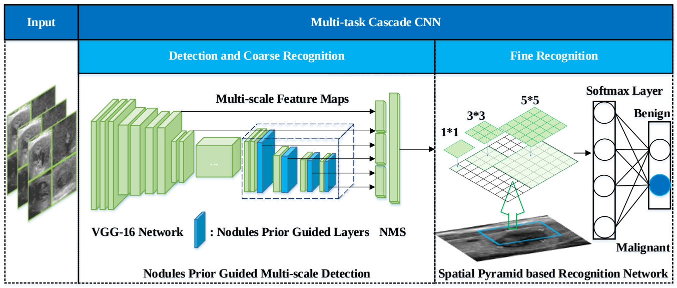

Input
Multi-task Cascade CNN
Detection and Coarse Recognition
Fine Recognition
Multi-scale Feature Maps
...
VGG-16 Network
: Nodules Prior Guided Layers
NMS
Nodules Prior Guided Multi-scale Detection
1*1
3*3
5*5
Softmax Layer
Benign
Malignant
Spatial Pyramid based Recognition Network

Fig. 2: Architecture of our MC-CNN framework. The framework contains two stages. (1) Detection and coarse recognition: to detect the nodule locations and coarsely recognize the nodules; (2) Fine recognition: to recognition the nodules finely.

II. Detection and Recognition based on Multi-Scale SSD Network

In construction of E_c in Algorithm 1, our training strategy is inspired by the SSD and multiple box framework [25]. Nonetheless, as shown in Fig. 3 (the pipeline of multiple scale detection network) and Fig. 5, we extend them to handle highly-varying thyroid nodules. We further re-construct the SSD detection network by adding multiple full convolution layers followed by nodules prior-guided anchors (extends from faster-rcnn [24]) generated layer to extract different scales of features from global to local. The detection approach is based on a feed-forward convolution network, which produces a fixed-size collection of bounding boxes and the corresponding class-assigning scores for the object instances in those boxes, followed by a non-maximum suppression (NMS) step to produce the final detection candidates(detailed in our supplement material). The shallow network layers are based on a standard architecture used for high quality image classification (truncated before any classification layers), wherein we leverage the base network used by SSD to extract feature maps. Different from the original SSD, we add multi-scale layers to fit for the thyroid nodules and arrive at coarse recognition.

Multi-scale Detection Network. Based on the two distributions of the thyroids' ratios and scales, which are shown in Fig. 4, the multi-scale SSD layers involve two-fold improvements.

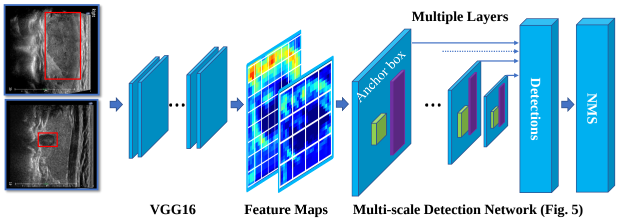
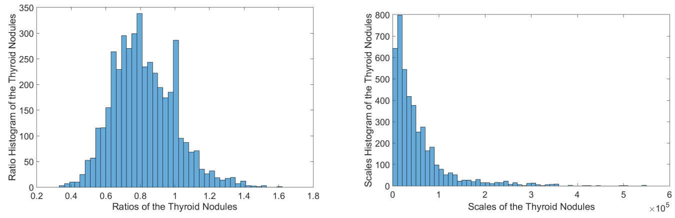


Fig. 3: Detection and recognition based on multi-scale SSD network. Based on the output of the feature maps from the base net of VGG16, we add the anchor boxes embedded with the thyroid scales and ratio prior (green and purple boxes). Furthermore, to cover the large and small scales of nodules, we concatenate all the newly added convolutional layers to the final detection layer.

The first one is to detect all the ratios of the nodules by using a set of anchor boxes ranging from 0.75 to 1.5, which are pre-computed in the training datasets. Specifically, the anchor boxes slide over the feature map in a convolutional manner, so that the nodule position relevant to the boxes will have a high response. In this way, we predict the offsets related to the default box shapes and the scores, which indicate the presence of a malignant nodule in each of those boxes. The newly-proposed anchor boxes can largely decrease the number of the false positive candidates.

The second one is suitable for the thyroid scales by incorporating multi-size perception fields of the anchors into multiple fully convolutional layers. Since it is hard to cover all the nodules with single-layer perception based conventional detection methods. We add multiple convolutional feature layers at the end of the truncated base network as [25] does. These layers gradually decrease in size, and allow multi-scale detections and predictions continually. Each of the newly-added convolutional layers predicts a result at different scales of perception fields. The size ranges from 5% (50*35) to 100%(1024*768) with respects to the input image (1024*768), and covers most of the nodules, whose sizes range from 35*40 pixels to nearly 774*573 pixels. All the feature maps are followed by a set of anchor boxes, which are determined by the distribution of the thyroid nodules. The anchors slide over the feature maps in the same way as the ratios guided anchor boxes do.

For a convolutional feature map with a size of $W \times H$, there are WHk anchors in a single nodules-guided layer. To add global high-level semantic features and local low-level detail features, this specially-defined layer is added after each full convolution layer to extract different-resolution features. Specially in Fig. 5, the left nodule has an aspect ratio close to 1:1. Our model adds several nodule prior guided feature maps at the end of the base network, which predicts the offsets to the thyroid boxes with different scales and aspect ratios, and predicts their associated confidences. The top 4 high confidence boxes are with the aspect ratios of 1:1, 1:2, 1:3, and 1:4, respectively. The top-ranked candidate boxes of the right nodule have similar aspect ratio with the nodules. We show two scales of the feature maps, which are 4*4 and 8*8.



(a) Ratio

(b) Scales

Fig. 4: Illustration of the thyroid nodules: (a) The ratio histogram of the thyroid nodules; (b) The scale histogram of the thyroid nodules (pixels). ‘Ratio’ means the width divided by the height. ‘Scales’ means the area of the nodules.

The newly-added layers are concatenated before the final loss layer. We compute c class scores and 4 offsets relevant to the original default box shape. This results in a total of $(c+4)k$ filters that are applied around at each location in the feature map, yielding $(c+4)kmn$ outputs for a $m \times n$ feature map.

Multi-task Loss. Each training image is annotated with a ground-truth class label y and a ground-truth bounding-box regression target l, which denotes a 4-dimension vector, (x position, y position, width, height). We use multi-task loss L^D and L^C on each labeled bounding box to jointly train the bounding-box regression and recognition: 1) Bounding box regression: for each candidate proposal, we predict its offset to the nearest ground truth. The learning objective function is formulated as a regression problem, and we employ the smooth L_1 norm proposed in [26] to make the predicted boxes of each sample x be close to its ground truth bounding box l, the L_i^D is the loss for coarse classification error:

$$L^D(l, l^*) = \sum_{r=1}^{R} smooth_{L1}(l_r - l_r^*), \tag{1}$$

where,

$$smooth_{L_1}(l_r - l_r^*) = \begin{cases} 0.5(l_r - l_r^*)^2 & \text{if } |l_r - l_r^*| < 1 \\ |l_r - l_r^*| - 0.5 & \text{otherwise} \end{cases}. \tag{2}$$

Here, l_r is the regression box offset and position, l_r represents the ground-truth label, while l_r^* denotes the predicted label. The smooth loss refines the regression in a continuous way, which is suitable for various shapes of nodules; 2) Coarse recognition of the thyroid nodules: the objective function learning is formulated as a binary classification problem. For each bounding box, we use the cross-entropy loss:

$$L^C(p(x|\mathbf{w}), y) = \sum_{r=1}^{R} -(y_r log(p_r)) + (1 - y_r)(1 - log(p_r)). \tag{3}$$

Here, p_r is the probability produced by the network, $\mathbf{w}$ is the trained weight of E_c, and it indicates that sample x_i is a thyroid nodule. The notation $y_r \in \{0, 1\}$ denotes the ground-truth label.

Specially, to accelerate the speed of convergence, we employ the VGG16 network E_0 as a feature extractor for its high performance in Imagenet classification tasks. Then, we add the multi-scale full convolution layers, similar to SSD. Following each feature map, we add the nodules prior-guided layer to generate anchors for each class. At first, we construct the detection model to find the proposals corresponding to the probable thyroid nodules, which outputs the location (x_{pos}, y_{pos}, w, h) and the class-assignment score s of the nodules. Then, we construct the fine recognition network to refine the recognition accuracy.

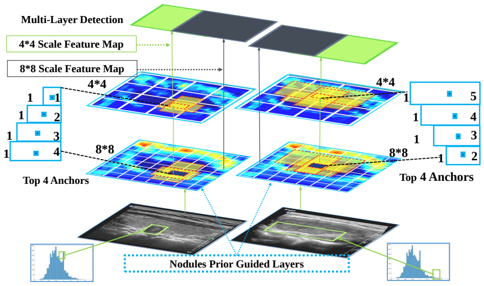


Fig. 5: SSD-based multi-scale network. The colored layers in the middle rows represent the added multi-scale layers. They are designed by considering the prior distributions of the thyroid nodules. Based on the multiple levels of feature maps, the top 4 anchor boxes with different ratios are represented with blue boxes. The distributions at different positions are illustrated in two examples.

III. Recognition Refinement based Spatial Pyramid Architecture

In construction of E_f in Algorithm 1, we further introduce our spatial pyramid architecture, which in fact represents an effective global contextual prior. In a deep neural network, the size of receptive field can roughly indicate to what extent we use contextual information. Although the theoretical receptive field of CNNs is already larger than the input image [23], and the existing work [27] proves that the empirical receptive field of CNN is much smaller than the theoretical one especially on high-level layers. This requirement makes single convolution neural networks insufficient to incorporate the momentous global context prior. We address this issue by embedding an effective global context aggregation network structure into original CNNs. Specially, the most direct intuition is to enlarge the perception. However, for thyroid ultrasound images with complex background and structure, this strategy is still not enough to cover necessary information. The annotated nodules in these thyroid images relate to many substrate and tissue locations. Directly fusing them to form a single vector may lose the spatial relation and cause ambiguity. However, the global context information along with the neighboring region’s context is helpful in this regard to distinguish among various categories. Thus, a more powerful representation should fuse information from different neighboring regions with these receptive fields. Similar conclusions are drawn in the classical works [28], [29] of natural RGB image classification. In [29], different-level feature maps generated by pyramid pooling are finally flattened and concatenated to be fed into a fully connected layer for classification. This global prior is designed to remove the fixed-size constraints of CNN for image classification. To further reduce context information loss among different regions, we propose multiple levels of global priors, containing information with different scales and varying among different regions. We call it spatial pyramid module, as illustrated in Fig. 6. The spatial pyramid module fuses features under three different pyramid scales, and is highlighted in blue for convolution operators and purple for pooling operators. The following pyramid levels separate the feature map into different neighboring regions and form the pooling representation for different locations. The different spatial pyramid level can produce the feature map with varied size. To maintain the weight of global feature, we use 1×1 convolution layer for each pyramid level to reduce the dimension of context representation to $1/N$ of the original one, if the level size of pyramid is N. Finally, different levels of features are concatenated as the final global feature. Specially, the feature maps of f_{c7} are fed to the spatial pyramid levels with multiple scales of convolution and pooling kernels.

$$f_p = \sum_i (f_{c7}(x * \mathbf{w}_i + b))(i = 1, 3, 5). \tag{4}$$

Here, $\mathbf{w}_i$ represents the kernel of the i_{th} operator, including the convolution and pooling kernels, and $\mathbf{b}$ is a bias parameter. Based on the learned weights over spatial pyramid, we can approximately obtain the semantic feature map of thyroid nodules as follows:

$$M(x) = ReLU(f_p(\mathbf{w}, x)). \tag{5}$$

We use the $ReLU$ active function to make the trained model none-linear. Here, $\mathbf{w}$ is the learned parameter for detection and recognition models E_c and E_f. Based on this activated feature map, we can further obtain the probability of the nodules' classes as follows:

$$p((y = j|x); \mathbf{w}) = \frac{exp(M_j)(x)}{\sum_{j=1}^{C} exp(M_j(x))}. \tag{6}$$

Here, j represents the j_{th} class. We further use the detected bounding box to refine the regions to be precisely classified. The probability vector for each class is p_j, and the loss of the crop region is defined as:

$$L^F(y, l) = \sum_j^C y_j(1 - log(p_j * p_c)). \tag{7}$$

Here, p_c is the coarsely predicted softmax probability of the binary classification under the loss of Eq. 3 and p_j is the softmax probability of the binary classification.

It may be noted that, the number of the pyramid levels and the size of each level are variables. They are related to the size of feature map that is fed into the pyramid pooling layer. The structure abstracts different nodule regions by adopting varying-size pooling kernels in a few strides. Thus, the multi-stage kernels should maintain a reasonable gap in representation. Our pyramid pooling module is a three-level one with kernel sizes of 1×1, 3×3, 5×5 respectively. For the type of spatial pyramid operation in different layers, we will conduct extensive experiments to show the differences in Section VI.

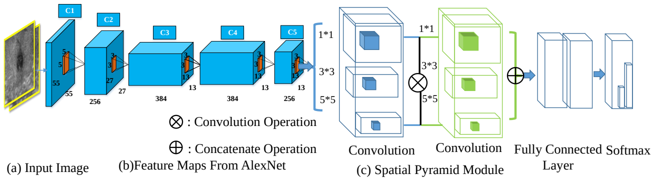


Fig. 6: Spatial pyramid network E_f. The spatial pyramid module consists of two convolution and pooling layers. The three layers are all with multiple scales of filters. Detailed network is described in our supplemental document.

IV. Coarse-to-fine Cascade Network

Specifically, the detector and classifier (including coarse and fine classification) involve bounding box regression, coarse recognition of the thyroid nodules, and local fine recognition of the detected nodules. The details of our cascade multiple tasks network are shown in Fig. 3. Such joint-task model shares the features for both detection and recognition, so that the detected nodule candidates can be refined in a coarse-to-fine manner from both low and high level of feature maps. The loss corresponding to the two tasks is formulated as:

$$L(x) = \sum_{i=1}^{N} [L_i^D(y_i, y_i^*) + L^C(l_i, l_i^*)] + L^F(y, y^*). \tag{8}$$

Here, the L_i^D is the loss for coarse classification error, y_i (y) is the ground truth label of the thyroid nodules, and y_i^* is the predicted label (labels) of the thyroid nodules. L^C is the loss for detected box location, where N is the number of the matched boxes. If $N = 0$, we set the loss to 0. The localization loss is a smooth L_1 loss between the predicted box and the ground truth box. Meanwhile, L^F is the fine classification error. The train process of MC-CNN is detailed in Algorithm 2.

Algorithm 2 MC-CNN Training Algorithm

```
while $L^D(x, l, l^*) < \delta$ and $L^C(p_r, y_r) < \delta$ do
    Extracting the basic features from VGGNet;
    Feeding the features into the multi-scales layers;
    Generating anchors $(x_{pos}, y_{pos}, w, h)$ via nodules prior guided layers;
    Optimizing loss $L^D(x, l, l^*)$ and $L_r^C(p_r, y_r)$ ;
end while;
Cropping detected images based on the predicted boxes $(x_{pos}, y_{pos}, w, h)$ with model $E_c$;
while $L^F(y, y^*) < \delta$ do
    Extracting feature maps from original CNNs;
    Refining feature maps $f_{c7}$ in spatial pyramid $f_p$;
end while;
```

V. Datasets and Implementation

A. Dataset Labeling for Network Training

One of the most important factors in any deep learning model is the training dataset labeling. Our training dataset is labeled by the senior doctors of Peking Union Medical College Hospital according to the pathology verification, which lasts two years in total. The thyroid nodules are marked with bounding boxes and the assigned benign/malignant class labels.

Besides, the training dataset covers most of the ultrasound images corresponding to different-aged patients and different-sized nodules. For all patients with malignant nodules and some patients with benign nodules who underwent surgery, the most accurate diagnosis is based on the histopathological examination results over the excised thyroid nodules. The ground truth is taken inside the same patient. The pixel depth ranges from 0 to 255. The images are collected in the clinical settings. Thus, the examination size is determined by the doctors' ROI of the thyroids.

To this end, all the thyroid instances involved in our training ultrasound image set have been examined via pathological examination. The involved ultrasound images are captured with Phillips HDI 5000, IU 22, GE Logiq 9 or Logiq 7 devices equipped with either a 5-12 MHz or an 8-15 MHz linear-array transducer. Among this initial cohort, only the patients who meet the following criteria are included: (1) older than 18 years of age, (2) with total or nearly total thyroidectomy or lobectomy, (3) with complete preoperative thyroid ultrasound images, and (4) with surgical pathology examination. =Moreover, the lesions and nodules that fail to meet the criteria for any pattern of ATA guidelines are excluded. Finally, a total of 1580 patients with 4309 images are obtained after surgery or FNA. Among these examples, 3128 thyroid nodules are benign and 3100 ones are malignant, and the mean nodule size is 2.4cm. Each thyroid nodules has several longitudinal cutting maps or cross cutting maps. In total, 6228 thyroid nodules images are obtained. Meanwhile, the boundary of the thyroid nodules in each image is manually delineated by the doctors. Thus, we can get 6228 labeling boxes, and these thyroid nodules images for training our MC-CNN.

B. Implementation

The framework contains two stages: (1) Detection and coarse recognition: to detect the nodule locations, we first employ a VGG-16 network (E_0) as the backbone to extract the high-level semantic feature maps from the ultrasound image. Then, we add convolutional layers at the top of the truncated base network. All the convolutional layers are followed by a nodule prior guided layer. Each of the multiple convolutional layers produces the detection and coarse recognition results at different scales. All the bounding boxes are processed by NMS. This step is to build up the detection and coarse recognition model (E_c); (2) Fine ecognition: to further improve the accuracy of classification, we add a spatial pyramid based recognition network to predict the category of the nodules at different scales. The fine recognition model is represented as E_f.

The stage (1) and (2) are the multi-task cascade CNN.

For our multi-scale detection network in the MC-CNN, we adapt a data-augmented strategy to improve the performance in the training procedure. The strategy leverages the entire original input image, sampled patches and randomly cropped patches, of which, the minimum Jaccard overlapping with the objects could be 0.1, 0.3, 0.5, 0.7, or 0.9, and the photo-metric distortions are similar to those described in [25]. We first train the models with 10^{-3} learning rate for 60k iterations, and then with 10^{-4} for another 60k iterations.

As for our spatial pyramid based recognition refinement network in the MC-CNN, the inputs of the base CNN are randomly sampled patches with size of $512 * 512$ from the entire thyroid nodules image, and the corresponding outputs are based on the average results of all the inputs. Our spatial pyramid CNN uses the multi-scale feature maps to train the softmax for thyroid nodules recognition. To demonstrate the efficiency of our MC-CNN, we also implement several commonly-used classification methods, including Nearest Neighbors, decision tree, random forest, adaboost, KNN, Bernouli Naive Bayesian, and GBDT. The classifiers are optimized with grid search in our experiments.

To evaluate the reliability and stability of the classifiers, we further train 7 classifiers with the feature vector extracted from fc7 layer in AlexNet [21] as input. In the AlexNet training process, both the weight decay and bias decay are 0.0005. The learning rate is set 0.02. The momentum is 0.9, linearly over 10 epochs. Besides, in our experiments, the thyroid dataset is split by $8:2$ for training and testing respectively. We use the Caffe lib to train the datasets on Tesla K80 GPUs.

VI. Result Analysis

In this section, we first conduct three experiments to separately demonstrate the significance of the multiple scale, the spatial pyramid, and the coarse-to-fine structure. Then, we compare the MC-CNN with the state-of-the-art methods. In addition, we conduct user studies in practice to demonstrate the applicability of the MC-CNN. Finally, to test the generalization ability of our model, we conduct unsupervised transfer learning experiments on publically-available datasets.

A. Experimental Design

Evaluations of our Multi-Scale SSD Network. Our multi-scale detection network is designed to detect and coarsely recognize the nodules. The detection results have a significant improvement. In the results, all the small-scale and large-scale nodules, which were lost in the original SSD network, are correctly detected. The results are shown in Fig. 7. The quantitative performances are shown in Table I and Table II. All the experiments are conducted with a 5-fold cross validation. The quantitative results show that, our adaptive thyroid detection is effective for all the scales of ultrasound images captured from different-aged patients. From the view point of thyroid nodule size, the nodules larger than 3cm can benefit most, which are highlighted in bold.

Furthermore, as shown in Table I, the mAPs of the three age groups are improved ranging from 0.7% to 4.9%, compared with the original SSD network. In particular, in different-aged groups, the young group has the most discriminate features for detection. The mAPs of the three-scale (small, middle, large) thyroid nodules are improved ranging from 0.1% to 4%. Especially, the large and small scales of thyroid nodules, which are mis-detected in the original SSD, could be detected accurately by our M-SSD. The mAPs of malignant ones are

TABLE I: Performance of the M-SSD network in 5-fold cross validation (mean ± standard variance). IOU=0.5 (Intersection over Union), "M-SSD" represents the multi-scale SSD.

Model	mAP	Benign(0)	Malignant(1)	Dataset
SSD	0.945 ± 0.04	0.965 ± 0.02	0.934 ± 0.03	Size:Large
M-SSD	0.985 ± 0.01	0.982 ± 0.01	0.987 ± 0.01	Size:Large
SSD	0.938 ± 0.05	0.916 ± 0.05	0.954 ± 0.02	Size:Middle
M-SSD	0.944 ± 0.05	0.972 ± 0.01	0.982 ± 0.01	Size:Middle
SSD	0.942 ± 0.04	0.967 ± 0.01	0.924 ± 0.05	Size:Small
M-SSD	0.988 ± 0.01	0.989 ± 0.01	0.987 ± 0.01	Size:Small
SSD	0.952 ± 0.01	0.955 ± 0.04	0.952 ± 0.04	Age:Old
M-SSD	0.980 ± 0.01	0.979 ± 0.02	0.983 ± 0.01	Age:Old
SSD	0.969 ± 0.02	0.975 ± 0.02	0.939 ± 0.06	Age:Middle
M-SSD	0.972 ± 0.02	0.987 ± 0.01	0.955 ± 0.02	Age:Middle
SSD	0.949 ± 0.05	0.962 ± 0.03	0.929 ± 0.06	Age:Small
M-SSD	0.971 ± 0.02	0.972 ± 0.02	0.968 ± 0.03	Age:Small

TABLE II: Dataset analysis.

Large	**Middle**	**Small**	**Old**	**Middle**	**Young**
(>3cm)	(1cm-3cm)	(0cm-1cm)	(49-78)	(38-49)	(18-38)

improved ranging from 0.3% to 4.9%, which are larger than that of the benign ones (ranging from 0%-1.2%).

Evaluations of our Spatial Pyramid Architecture. To evaluate our spatial pyramid based recognition refinement network, we further conduct three experiments. The first one is to compare with the CNNs without the spatial pyramid layers. We have compared it with AlexNet (conv5-small), GoogLenet. Here we compare the spatial layers in different CNN layers: after the 5th pooling (pool5) layer, after the 5th convolution layer (conv5). The result is shown in Fig. 8, and it indicates that, the network with the spatial pyramid layer added after the 5th convolution layer, performs the best among the 4 networks. The network 'pool5' ranks the second among all 4 networks, which adds the spatial pyramid layer after the 5th pooling layer. The original AlexNet ranks the third. Meanwhile, the GoogLenet ranks the forth. This phenomenon states that, the spatial pyramid layer can help improve the performance of the thyroid nodules recognition, but the GoogLenet has the most parameters, which leads to an over-fitting result. To further study what happened in the spatial pyramid networks, we visualize the AlexNet and the AlexNet with spatial pyramid network after the 5th convolution layer. The result is shown in Fig. 10. The results demonstrate that, the spatial pyramid structure can activate larger perceptions than the original AlexNet for nodules in the feature map layers.

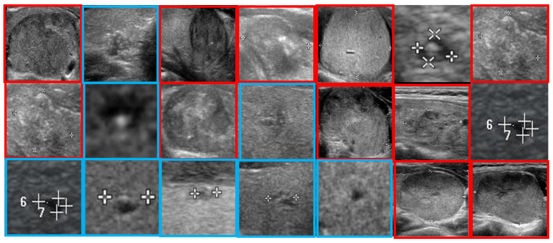


Fig. 7: Illustration of the improvement benefiting from our multi-scale SSD, which can detect the missing small-scale and large-scale nodules. The large and small sizes of thyroids are beyond the detection ability of the original SSD network. The blue boxes represent the large thyroids, while the red boxes represent the small ones. Some more cases are in Fig. 13.

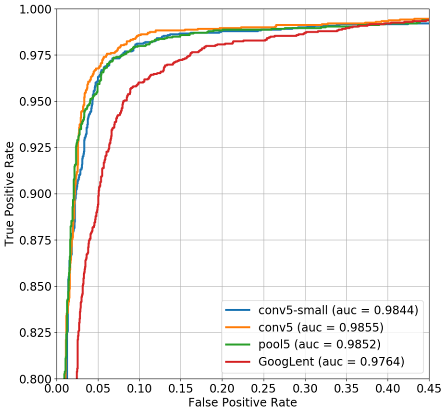


Fig. 8: ROC of CNNs without classifiers.

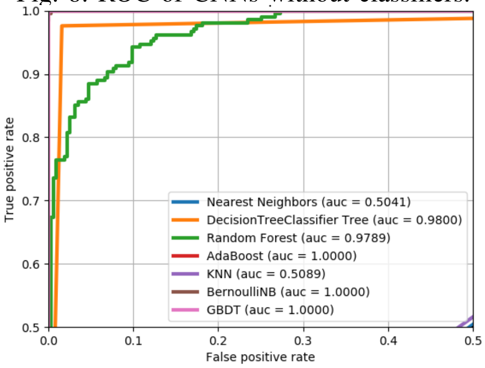


Fig. 9: MC-CNN features with classifiers.

The second experiment is to compare it with 7 existing classifiers, including nearest neighbors, decision tree, random forest, adaboost, KNN, Bernouli Naive Bayesian, and GBDT. All the classifiers are trained with grid search to find the proper hyper-parameters. The results are shown in Fig. 11. For the AUC, 6 of 7 classifiers are lower than spatial pyramid networks. The random forest is close to spatial pyramid networks, however, the ROC curve of spatial pyramid is more smooth than that of random forest. Thus, the spatial pyramid furnished CNN achieves the best performance in most classifiers.

In our third experiment, we combine our MC-CNN framework with the 7 classifiers, which gives rise to significant performance improvement. We fed the classifiers with the fc7 layer's output of the spatial pyramid network. The result is shown in Fig. 9. Three of the 7 classifiers improve dramatically to 100% over the testing dataset, including Adaboost, Bernouli Naive Bayesian, and GBDT. Moreover, other classifiers are improved more or less from 0.6% to 2.68%.

Comparisons with State-of-the-art Methods. Our coarse-to-fine framework can improve the performance of thyroid nodules recognition, which is demonstrated through three experiments: coarse classification network from detection network, the single task classification network from the doctor-cropped results, and the MC-CNN framework. The quantity performance is documented in Table III. The first row is the coarse recognition result. The 2-4 rows are the single-task networks derived from AlexNets and GoogLenet, with different layers embed into the spatial pyramid module. The result shows that, our MC-CNN achieves great performance improvement compared with the single coarse and single fine

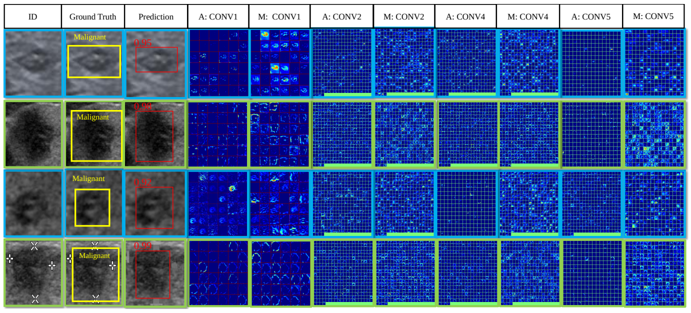


Fig. 10: Comparisons between the original AlexNet and the MC-CNN equipped AlexNet. The first column shows the input images. The boxes with the same color are the feature maps extracted from the original and spatial pyramid equipped AlexNets. 'A' means AlexNet feature, while 'M' means MC-CNN feature. We visualize the feature maps, including shallow features such as 'CONV1', 'CONV2' and high level features such as 'CONV4' and 'CONV5'. The MC-CNN network can activate a larger perception scope for high-level recognition tasks, so that the high level features could focus more on the thyroid nodules related regions.

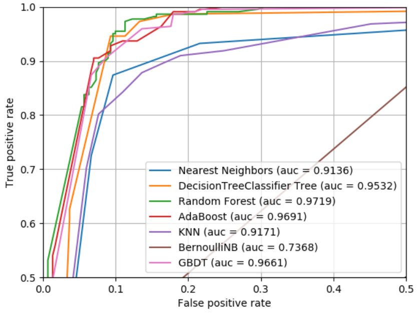


Fig. 11: ROC of the 7 classifiers with the AlexNet features extracted from fc7.

classification networks. The average accuracy is improved ranging from 6% to 10.1%, the sensitivity and the Specificity is improved from 7.5% to 12.5%. It demonstrates our coarse-to-fine framework can extract consistent features from the detection results for the classification task. For the single classification based on the input patches cropped by doctors, its performance is slightly lower than that based on the detection boxes, which is 6% lower in accuracy, and 7.5% lower in sensitivity. Moreover, the results show that the multi-task framework performs better than the original SSD. Besides, for the cases failed in the original SSD, we further analyze their improvement benefiting from our method by visualizing their concrete nodules in Fig. 12 and Fig. 13. We can conclude from the two sets with the scales larger than 5cm and smaller than 0.5cm improve most.

B. User Studies of Our MC-CNN Framework

To evaluate our MC-CNN, we compare our results with those from the senior doctors. We make a dataset that contains 360 images covering three-age stages and three scales, which are separately collected and labeled, which is described in Section V-A.

The doctors referred to the features of malignant based on ATA, and drew conclusions based on experience. The results are documented in Table IV, and it states that our MC-CNN performs better than human doctors in terms of both time efficiency and accuracy. Without our CADx system, the doctors become tired and impatient when facing a dataset with more than 360 (187 benign nodules and 180 malignant nodules) nodules. Furthermore, in case of complex ultrasound images, the result becomes even more distinct since our MC-CNN achieves higher accuracy than the doctors, of which, the improvement in accuracy is 12%, the improvement in sensitivity is 13% and the improvement in sensitivity in 8%, with only 2.1% time consumption on average compared to

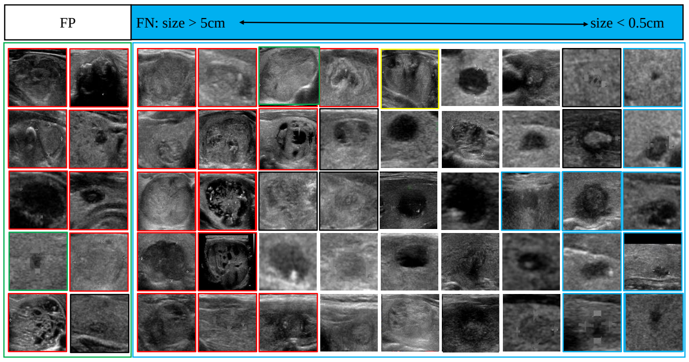


Fig. 12: Illustration of the failed recognition cases produced by the original SSD [25]. It mis-classifies the small thyroid nodules ($< 0.5cm$) (blue boxes) and large ($> 5cm$) thyroid nodules (red boxes).

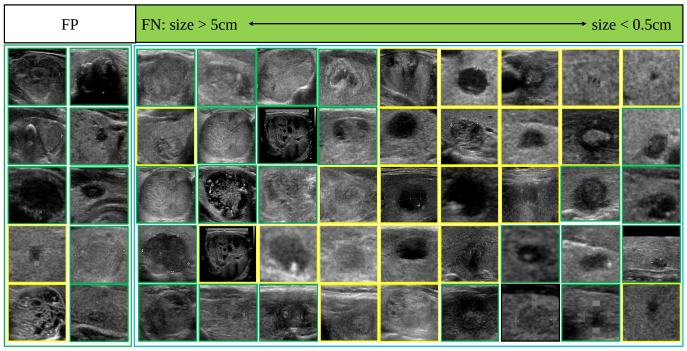


Fig. 13: Illustration of the improvement benefiting from our method. It shows some of the successful cases which are failed in the coarse recognition network E_c. The nodule proposals are fed into our spatial pyramid network. Meanwhile, the classification accuracy improves greatly. The green boxes are the improved cases in the coarse network, while the yellow are the hard cases which are also mis-classified by our coarse-to-fine framework.

TABLE III: Comparisons among coarse and fine classification networks. 'Pool5' and 'Conv5' indicate to add spatial pyramid module after this layer. 'Coarse' means the classification results are directly obtained from the detection stage. 'Fine' means the doctors' hand-cropped candidates are fed into our spatial pyramid module furnished network (mean ± standard variance).

Method	Accuracy	Sensitivity	Specificity
Coarse: Detection	0.903 ± 0.05	0.892 ± 0.04	0.920 ± 0.04
Fine: AlexNet	0.926 ± 0.04	0.908 ± 0.05	0.981 ± 0.01
Fine: Pool5	0.919 ± 0.05	0.894 ± 0.05	0.983 ± 0.01
Fine: GoogLenet	0.889 ± 0.08	0.845 ± 0.06	0.976 ± 0.02
Fine: Conv5	0.930 ± 0.04	0.925 ± 0.02	0.981 ± 0.01
MC-CNN	0.982 ± 0.01	0.983 ± 0.01	0.980 ± 0.01

TABLE IV: User study based performance evaluations. (mean)

Method	Accuracy	Sensitivity	Specificity	Time(s)	AUC
Human	0.87	0.86	0.91	12.00	0.88
MC-CNN	0.98	0.98	0.98	0.25	0.98

TABLE V: User study based performance evaluations over different datasets. (mean over 5-fold cross validation)

Method	Accuracy	Sensitivity	Specificity	Dataset
Doctor	0.897	0.955	0.87	Age: Old
MC-CNN	0.985	0.971	0.982	Age: Old
Doctor	0.96	0.941	0.977	Age: Middle
MC-CNN	0.963	0.992	0.930	Age: Middle
Doctor	0.816	0.799	0.852	Age: Young
MC-CNN	0.962	0.942	0.973	Age: Young
Doctor	0.841	0.85	0.831	Size: Large
MC-CNN	0.946	0.942	0.955	Size: Large
Doctor	0.913	0.918	0.908	Size: Middle
MC-CNN	0.986	0.992	0.979	Size: Middle
Doctor	0.891	0.806	0.982	Size: Small
MC-CNN	0.962	0.972	0.938	Size: Small

the doctors. Table V documents the performances in different stages and can handle thyroid nodules with different sizes. The nodules from old patients and the large-sized nodules are improved most, ranging from 9.1% to 12.5%.

C. Evaluations using Public Dataset

In order to demonstrate the generalization ability of our MC-CNN, we verify it by conducting transferred learning experiments on the open dataset [30], without the need of additional training on this dataset. This dataset has a total number of 299 patients, including 270 women and 29 men, whose ages vary as 57.35 ± 16.2 years. We treat the labels in triads (following [30]) over "(4c) Three or four suspicious ultrasound features" and "(5) Five suspicious features" as the malignant nodules while treating "Normal" thyroid and "Benign" level as benign ones. Finally, we obtain 111 malignant ultrasound images and 41 benign ultrasound images. Our MC-CNN outperforms most of the previous works with the significant advantage, as shown in Table VI. The number of the benign dataset is half of that of the malignant dataset. The classifiers, such as Naive Bayesian, GBDT, and MLP, are sensitive to such data distribution, thus the sensitivity is lower than the Specificity. At the same time, our MC-CNN is stable in both classes even if the distribution is heavily unbalanced.

D. Discussion

In this section, we mainly describe the advantage, the comparison with the state of the art, in what it could be improved, which could be further applications.

Advantages. MC-CNN has three main advantages as follows. (1) Our methods includes two tasks, and the nodules can be detected and recognized at the same time. In our framework, the thyroid US images based detection and recognition tasks are jointly learned for sharing the common features while distinguishing the malignant, the benign, and the background. However, this multi-task network cannot perform classification with high performance, and it only provides precise bounding boxes of the nodules. To address this issue, we propose to learn the task in a coarse-to-fine manner, that uses the spatial pyramid module to improve the recognition accuracy. With the coarse-to-fine framework, the thyroid with different scales, especially the extremely large and small ones, are detected and correctly classified. Meanwhile, it suggests that the size of kernels of CNN filters indeed affects the performance of the recognition task. We only extend the simple CNN structure to improve the efficiency; (2) MC-CNN can detect and recognize a wider range of the nodules in multiple scales. The scale prior of the nodules is the important information both for detection and recognition. Although MC-CNN could be divided into serval scales to learn different scale features, the concatenated features ignore the correlation among different scales. To overcome these problems, we propose the spatial pyramid module to learn the multiple scales of features in a single module. The spatial pyramid can effectively fuse different sizes to generate a complementary effect, while the AlexNet is easy to be fine-tuned and has a simple structure to be extended conveniently. Moreover, the Alexnet is not easily overfitted on the limited scales of datasets. Therefore, MC-CNN can effectively represent the thyroid nodule features for classification, which has been proved in the clinically setting involving doctors' studies; (3) MC-CNN could be generalized to handle more datasets without training. Thyroid nodule detection and recognition for US images are solved by a novel MC-CNN, which has the advantages over a simple network architecture with good performance. In this study, we first evaluate the feasibility of MC-CNN for US images, and the results on all the ages and sizes of nodules show that MC-CNN achieves competitive results compared with AlexNet, GoogleNet, and conventional classifiers such as, GDBT, SVM, etc.

TABLE VI: Performance evaluation on public dataset.

Method	Accuracy	Sensitivity	Specificity
MC-CNN	**0.921**	**0.941**	**0.962**
Naive Bayesian	0.737	0.631	0.746
GBDT [16]	0.717	0.478	0.863
MLP	0.737	0.529	0.818
AlexNet [21]	0.784	0.625	0.854
GoogLenet [22]	0.750	0.586	0.847

Limitations. Our MC-CNN still has some limitations. For example, in Fig. 12 and Fig. 13, for the extremely small $(< 0.01cm)$ and extremely large $(> 10cm)$ nodules, our MC-CNN tends to produce failure cases, which are marked with yellow boxes, because the scales of the nodules are critical for accurate detection and recognition. Despite our prior and on-going efforts, currently high-quality training datasets are still insufficient compared with natural images that have been widely employed in various deep learning applications. The essential reason is that, the rapid acquisition of high-quality, high-volume medical datasets remains a bottleneck, it needs a time-consuming and rigorous verification process to label ground truth, which must consume huge labor of experienced radiologists. Therefore, we would further introduce transfer learning into the pre-trained models to make our model easily be transferred to other different scales of unmanifested nodules.

VII. Conclusion and Future Works

In this paper, we have advocated a novel MC-CNN framework that can learn thyroid nodule detection and classification on ultrasound images. The new learning architecture affords the detection and classification tasks to share commonly-needed features, with an objective of better distinguishing benign nodules from malignant nodules, as well as the complex background. In order for such goal to be easily accomplished, in our new learning architecture we must add a multi-scale layer to improve the detection performance for thyroid nodules that could be varying significantly in scales. Consequently, the detected nodule candidates are fed back into the spatial pyramid augmented AlexNet to further improve the classification performance. As a result, our MC-CNN has shown superior advantages over the original single shot detection and other single task classification methods based on our comprehensive experiments.

In the near future, we will continue to conduct extensive user studies and evaluations towards possible clinical trials. It is our expectation that, our MC-CNN architecture could achieve more promising performance for smart thyroid nodule diagnosis on ultrasound images, which leads to a better and greater potential in ultrasound-based clinical applications in the future. Meanwhile, with the scale increase of the collected datasets, we plan to study more compacted CNN structures to extract more discriminative features efficiently. Specifically, more comprehensive evaluations on how to adapt our MC-CNN to other popular networks also deserve our immediate efforts.

Acknowledgements. This research is supported in part by NSFC of China (NO. 61672077 and 61532002), Applied Basic Research Program of Qingdao (NO. 161013xx) , National Science Foundation of USA (NO. IIS-0949467, IIS-1047715, IIS-1715985, and IIS-1049448), and capital health research and development of special 2016-1-4011.

References

[1] C. Zhu, T. Zheng, B. A. Kilfoy, X. Han, S. Ma, Y. Ba, Y. Bai, R. Wang, Y. Zhu, and Y. Zhang, "A birth cohort analysis of the incidence of papillary thyroid cancer in the united states, 1973–2004," *Thyroid*, vol. 19, no. 10, pp. 1061–1066, 2009.

[2] J. Ma, F. Wu, T. Jiang, J. Zhu, and D. Kong, "Cascade convolutional neural networks for automatic detection of thyroid nodules in ultrasound images," *Medical Physics*, vol. 44, no. 5, pp. 1678–1691, 2017.

[3] J. A. Sipos, "Advances in ultrasound for the diagnosis and management of thyroid cancer," *Thyroid*, vol. 19, no. 12, pp. 1363–1372, 2009.

[4] F. Pacini, M. Castagna, L. Brilli, G. Pentheroudakis, and E. G. W. Group, "Thyroid cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up," *Annals of oncology*, vol. 21, no. suppl_5, pp. v214–v219, 2010.

[5] P. R. Larsen, "New guidelines for patients with thyroid nodules and differentiated thyroid cancer," *Nature Clinical Practice Endocrinology & Metabolism*, vol. 2, no. 6, pp. 297–298, 2006.

[6] L. Davies and H. G. Welch, "Increasing incidence of thyroid cancer in the united states, 1973-2002," *Jama*, vol. 295, no. 18, pp. 2164–2167, 2006.

[7] E. Horvath, S. Majlis, R. Rossi, C. Franco, J. P. Niedmann, A. Castro, and M. Dominguez, "An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management," *The Journal of Clinical Endocrinology & Metabolism*, vol. 94, no. 5, pp. 1748–1751, 2009.

[8] J.-Y. Park, H. J. Lee, H. W. Jang, H. K. Kim, J. H. Yi, W. Lee, and S. H. Kim, "A proposal for a thyroid imaging reporting and data system for ultrasound features of thyroid carcinoma," *Thyroid*, vol. 19, no. 11, pp. 1257–1264, 2009.

[9] J. Y. Kwak, K. H. Han, J. H. Yoon, H. J. Moon, E. J. Son, S. H. Park, H. K. Jung, J. S. Choi, B. M. Kim, and E.-K. Kim, "Thyroid imaging reporting and data system for us features of nodules: a step in establishing better stratification of cancer risk," *Radiology*, vol. 260, no. 3, pp. 892–899, 2011.

[10] U. R. Acharya, G. Swapna, S. V. Sree, F. Molinari, S. Gupta, R. H. Bardales, A. Witkowska, and J. S. Suri, "A review on ultrasound-based thyroid cancer tissue characterization and automated classification," *Technology in cancer research & treatment*, vol. 13, no. 4, pp. 289–301, 2014.

[11] C.-Y. Chang, S.-J. Chen, and M.-F. Tsai, "Application of support-vector-machine-based method for feature selection and classification of thyroid nodules in ultrasound images," *Pattern recognition*, vol. 43, no. 10, pp. 3494–3506, 2010.

[12] N. Singh and A. Jindal, "Ultra sonogram images for thyroid segmentation and texture classification in diagnosis of malignant (cancerous) or benign (non-cancerous) nodules," *Int. J. Eng. Innov. Technol.*, vol. 1, pp. 202–206, 2012.

[13] D. Bibicu, L. Moraru, and A. Biswas, "Thyroid nodule recognition based on feature selection and pixel classification methods," *Journal of digital imaging*, vol. 26, no. 1, pp. 119–128, 2013.

[14] S. Tsantis, N. Dimitropoulos, D. Cavouras, and G. Nikiforidis, "Morphological and wavelet features towards sonographic thyroid nodules evaluation," *Computerized Medical Imaging and Graphics*, vol. 33, no. 2, pp. 91–99, 2009.

[15] J. Ding, H. Cheng, C. Ning, J. Huang, and Y. Zhang, "Quantitative measurement for thyroid cancer characterization based on elastography," *Journal of Ultrasound in Medicine*, vol. 30, no. 9, pp. 1259–1266, 2011.

[16] J. Ye, J.-H. Chow, J. Chen, and Z. Zheng, "Stochastic gradient boosted distributed decision trees," in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 2061–2064.

[17] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[18] U. R. Acharya, P. Chowriappa, H. Fujita, S. Bhat, S. Dua, J. E. W. Koh, L. W. J. Eugene, P. Kongmebhol, and K. H. Ng, "Thyroid lesion classification in 242 patient population using gabor transform features from high resolution ultrasound images," *Knowledge-Based Systems*, vol. 107, pp. 235–245, 2016.

[19] Y. Hong, X. Liu, Z. Li, X. Zhang, M. Chen, and Z. Luo, "Rea time ultrasound elastography in the differential diagnosis of benign and malignant thyroid nodules," *Journal of Ultrasound in Medicine*, vol. 28, no. 7, pp. 861–867, 2009.

[20] U. Raghavendra, U. R. Acharya, A. Gudigar, J. H. Tan, H. Fujita, Y. Hagiwara, F. Molinari, P. Kongmebhol, and K. H. Ng, "Fusion of spatial gray level dependency and fractal texture features for the characterization of thyroid lesions," *Ultrasonics*, vol. 77, pp. 110–120, 2017.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[26] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, ser. Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[27] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *arXiv preprint arXiv:1412.6856*, 2014.

[28] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2. IEEE, 2006, pp. 2169–2178.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*. Springer, 2014, pp. 346–361.

[30] L. Pedraza, C. Vargas, F. Narvaez, O. Duran, E. Munoz, and E. Romero, "An open access thyroid ultrasound image database," in *Tenth International Symposium on Medical Information Processing and Analysis*. International Society for Optics and Photonics, 2015, pp. 92 870W–92 870W.