Fully Automated Thyroid Nodule Detection, Segmentation, and Classification in Ultrasound Images Using Deep Learning and Image Processing

by

Atefeh Shahroudnejad

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Medical Sciences - Radiology and Diagnostic Imaging

University of Alberta

Abstract

Thyroid cancer has a high prevalence all over the world. Accurate thyroid nodule detection and diagnosis in early stages leads to effective treatment and decreases the mortality rate. However, thyroid nodule detection and assessment using ultrasound imaging is a very challenging task, even for experienced radiologists, due to the ultrasound image characteristics and variations in thyroid nodule sizes and appearances. Existing Computer-Aided Diagnosis (CAD) systems are not fully automated and also have limited performances. This thesis presents a fully automated thyroid CAD system to assist radiologists. The proposed CAD system consists of four components: nodule detection, nodule segmentation, thyroid segmentation, and nodule classification from thyroid ultrasound scans acquired through ultrasound examination of the thyroid. For nodule detection, a novel one-stage detection network, TUN-Det, is proposed, which introduces Residual U-blocks (RSU) to built the TUN-Det backbone, and presents a newly designed multi-head architecture comprised of three parallel RSU variants to replace the plain convolution layers of both the classification and regression heads. Residual blocks enable each stage of the backbone to extract both local and global features, and the multi-head design embeds the ensemble strategy into one end-to-end module to improve the accuracy and robustness by fusing multiple outputs generated by diversified sub-modules. TUN-Det achieves very competitive results against the state-of-the-art models on the overall Average Precision (AP) metric and outperforms them in terms of AP_{35} and AP_{50}. For nodule segmentation, a residual dilated U-Net,

resDUnet, is proposed, which has a residual structure, and also dilated convolution layers are embedded in the bottleneck part of the network. Residual connections lead to consistent training and dilated convolution layers generate richer multi-scale features. Our resDUnet achieves a high Dice score and much smooth visual results. For thyroid gland segmentation in ultrasound sweeps, LSTM-UNet is proposed, which uses time-distributed convolution blocks and bidirectional convolutional LSTM in the U-Net. The building blocks extract spatial-temporal information and consider the inter-frame correlation of consecutive frames. LSTM-Unet avoids the under-segmentation problem, which is a common issue in thyroid segmentation methods. For the nodule classification component, two rule-based classifiers are proposed for nodule composition and nodule margin, which use different image processing techniques and decide based on the pre-defined rules. The rules are defined based on the clinical definitions. All Experimental results indicate the promising performance of the proposed CAD system in clinical applications.

Preface

This study was approved by the health research ethics board of the University of Alberta.

Acknowledgements

First, I would like to thank the industrial collaborator, Medo.ai, and Dr. Dornoosh Zonoobi to support me and give me the chance to work on their valuable project. I learned a lot from their amazing AI team and I am so grateful that I could be part of that.

I would also like to thank my supervisors Dr. Kumar Punithakumar and Dr. Pierre Boulanger, and Servier Virtual Cardiac Centre director, Dr. Michelle Noga. They steered me in the right direction all the time and without their passionate help and support, this project could not have been successfully conducted.

Finally, I especially thank my lovely family who always encourages and supports me to accomplish my goals and follow my dreams.

Table of Contents

List of Tables

List of Figures

Acronyms

Computer Aided Diagnosis (CAD)

Convolutional Neural Network (CNN)

Feature Pyramid Network (FPN)

Fine Needle Aspiration (FNA)

Fully Convolutional Network (FCN)

Long Short-Term Memory (LSTM)

Non-Maximum Suppression (NMS)

Recurrent Neural Network (RNN)

Region Proposal Network (RPN)

Region Of Interest (ROI)

Residual Dilated UNet (resDUnet)

ReSidual U-blocks (RSU)

Sagital (SAG)

Support Vector Machine (SVM)

Thyroid Ultrasound Nodule Detection (TUN-Det)

Thyroid Imaging, Reporting, and Data System (TIRADS)

Transverse (TRX)

Ultrasound (US)

Weighted Boxes Fusion (WBF)

Chapter 1

Introduction

1.1 Overview

Over the last few decades, the incidence of thyroid cancer has rapidly increased all over the world [85], [92]. In 2019, United States reported almost 52,070 adults (37,810 women, 14,260 men) diagnosed with thyroid cancer, which resulted in a total of 2170 deaths [85]. Canadian Cancer Statistics has also reported 230 deaths among 8200 adults (6100 women and 2100 men) who were diagnosed with thyroid cancer [10]. According to these statistics, women are more likely to develop thyroid cancer than men, and its prevalence increases with age. A thyroid nodule is the main concept in thyroid cancer. When cells grow abnormally within the thyroid gland, they form thyroid nodules that can be benign or malignant (cancerous) [3]. As part of clinical workflow in thyroid sonography, thyroid nodules are measured, and their sizes are monitored over time as significant growth could be a sign of thyroid cancer. Thyroid nodules are most common in the general population and occur up to 71% of people [97]. Based on epidemiological studies, the prevalence of thyroid nodules detected by high-resolution ultrasound varies from 19% to 68% in a random population[33]. Almost 90% of these nodules are benign and are unlikely to grow in size and become cancerous, even if they grow [28]. However, if they become malignant, they lead to a high mortality rate [42]. The mortality rate could be reduced if these nodules are diagnosed and treated in the early stages.

1.2 Challenges and Goal

Thyroid nodules larger than 1 cm are considered suspicious based on their echogenic texture, and they are recommended for biopsy via Fine Needle Aspiration (FNA). However, FNA is highly invasive, costly, and sometimes inconclusive. When combined with the low specificity of physical examinations, this results in over-diagnosis and over-treatment of thyroid nodules, which is a frequent source of anxiety for the patient and clinician, and meanwhile, a major financial burden on healthcare systems.

Ultrasound (US) is the primary diagnostic modality for thyroid examination (for both the detection and characterization of thyroid nodules), which is performed in both transverse (TRX) and sagittal (SAG) orientations using an ultrasound probe (Figure. 1.1). In addition to being non-invasive, safe,

(a) TRX (b) SAG

Figure 1.1: Thyroid ultrasound TRX and SAG views.

portable, inexpensive, and fast (in terms of acquisition time) for thyroid examination, many thyroid nodules are detectable in ultrasonography, even if they are too small to be detected by palpation [101](high sensitivity). However, ultrasound images have variable spatial resolutions and heavy noises such as speckle noise, which make the detection, segmentation, and classification tasks complicated. Thyroid nodules have diverse sizes, shapes, and appearances and they sometimes look very similar to the thyroid tissues and hard to be defined by clear boundaries (*e.g.* ill-defined nodule). Some nodules have heterogeneous patterns due to diffuse thyroid disease, which makes these nodules hard to differentiate from each other and their backgrounds. Besides, the

occasional occurrence of multiple thyroid nodules within the same image, and large thyroid nodules with complex interior textures, which could be considered internal nodules, further increase the difficulty of the nodule detection task. Figure. 1.2 shows some challenging thyroid nodules. To reduce subjective errors by experts, avoid unnecessary biopsies and surgeries, and precise rapid diagnosis in thyroid ultrasound images, there is an urgent need for a thyroid Computer-Aided Diagnosis (CAD) system to assist radiologists and increase the survival rate [14].

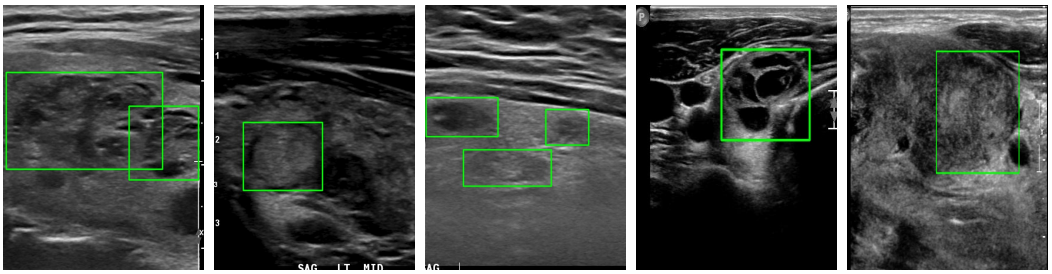


Figure 1.2: Challenging nodule examples.

1.3 Thesis Contributions

In this thesis, we propose a fully automated thyroid CAD system (Figure. 1.3). Our contribution consists of three folds:

1. Nodule detection: CAD systems require preliminary finding Region of Interest (ROI) of nodules for further processing. In traditional CAD systems, the ROIs are manually defined by experts, which is time-consuming and highly relies on the experiences of the radiologists and sonographers. To address this limitation, the first component of our proposed CAD system is automatic thyroid nodule detection, which predicts the bounding boxes of thyroid nodules from ultrasound images and plays a very important role in computer-aided thyroid cancer diagnosis.

2. Nodule and Thyroid segmentation: Thyroid nodule assessment requires a precise segmentation of the nodule's boundary and thyroid gland. Since nodules boundaries are often blurred due to noises and artifacts, manual segmentation of nodules is time-consuming, tedious, and it leads to

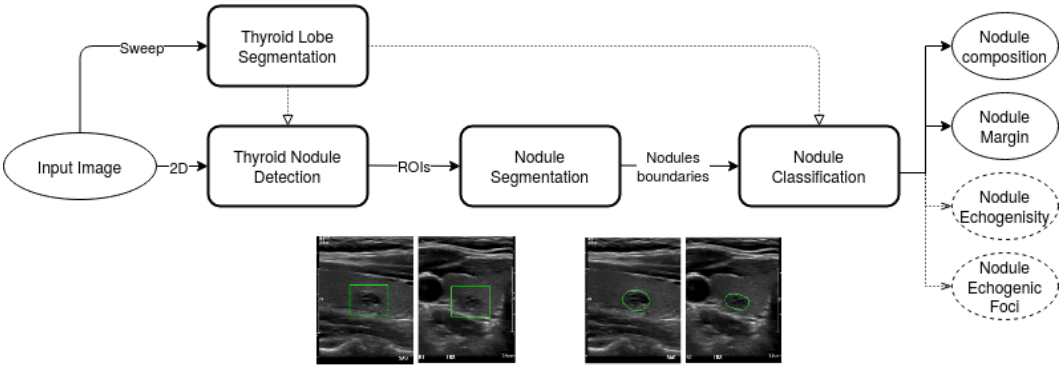


Figure 1.3: Block diagram of the proposed CAD system

intraobserver or interobserver variability. Moreover, the pixel intensity distribution inside the nodule varies considerably depending on the nodule composition, which makes the manual segmentation more difficult. Besides, to analyze the characteristics of the thyroid for nodule assessment and also filter the false detected nodules outside of the thyroid gland, it is important to segment the thyroid gland as well. Therefore, the next components of our CAD system are nodule and thyroid segmentation modules, which respectively segments nodules and thyroid, to eliminates all the mentioned shortcomings.

3. Nodule classification: Thyroid nodule diagnosis is conducted by nodule characteristics classification. Since ultrasound is subject to high variability in interpretation, in order to standardize the reporting and characterization of thyroid nodules, the American College of Radiology introduced the Thyroid Imaging Reporting and Data System (TIRADS) [31], which is based on five characteristics of the nodule including echogenicity, composition, shape, margin, and presence of calcification. Various modifications of the TIRADS, such as the Korean Society of Thyroid Radiology TIRADS [84], have also been proposed. Although these modified TIRADS reduce the variability in reporting nodules, the fundamental limitation is that individual characteristics of the nodule are determined manually, and this makes the assessment subjective. Based on the original TIRADS definition, composition and margin of nodules need to

be determined for fast and reliable risk stratification of them. Hence, the last component of the proposed CAD system is nodule classification which classifies the composition and margin of the segmented nodule.

1.4 Outline

This thesis includes 5 chapters:

- Chapter 2 reviews previous works in thyroid nodule detection, segmentation, and assessment. It also provides the literature on object detection and segmentation.
- Chapter 3 proposes our new fully automated thyroid CAD system including nodule detection, nodule and thyroid segmentation, and nodule classification.
- Chapter 4 presents the datasets that we used, the evaluation metrics, and our experimental results.
- Chapter 5 concludes the thesis with a summary and presents a direction for the future work.

Chapter 2

Literature Review

In this chapter, we look into the previous CAD systems for thyroid ultrasound nodule assessment, including nodule detection, segmentation, and classification. We also provide the literature on object detection and segmentation methods.

2.1 Feature-based Methods

These methods extract features from ultrasound images (or the nodule area) and then classify them using a classifier to evaluate the nodule.

2.1.1 Clinical Features

These methods use sonographic features (i.e., size, aspect ratio, shape, margin, echogenicity, internal composition, calcification, peripheral halo, capsule, cervical lymph node, and vascularity) to diagnose thyroid malignancy. Zhang et al. [110] feed a set of 11 features, collected from thyroid US examination, and one feature from Real-time Elastography (RTE) into the nine well-known machine learning classifiers to estimate malignancy of a nodule. Based on their experiments, the Random Forest classifier has the highest performance among the other classifiers. [68] compares the performance of three classifiers: Random Forest, Support Vector Machine (SVM) and Logistic Regression, on a feature set. In [103] the most discriminative sonographic features are selected using the Relief feature selection method and fed into Extreme Learning Machine (ELM) classifier to identify malignant nodules.

2.1.2 Hand-crafted Features

The pipeline of these traditional methods includes feature extraction from nodule ROI, an optional feature selection technique, and a classifier to discriminate the malignant nodules from the benign ones.

Here, hand-crafted features refer to spatial and frequency textural features, shape features, and statistical features. Textural features include Gray-Level Co-occurrence Matrix (GLCM) [12], [15], [26], [93], Gray-Level Run Length Matrix (GLRLM) [12], [15], [109], Spatial Gray-Level Dependence Features (SGLDF) [72], fractal textures [72], [109], Local Binary Patterns (LBP) [49], Gabor transform [3] Discrete Wavelet Transform (DWT) [4], Wavelet features [12], [95], Fourier Power Spectrum [109], and local Fourier coefficient [12]. Shape features include morphological features [95], [109]. Statistical features include Statistical feature matrix [12], [109], first-order statistics [6], [26], [109], Grey-Level Histograms (GLH) [68], and Higher Order Spectral (HOS) entropy [73].

Some feature selection techniques that have been used in this field are k-fold [12], Relief-F [3], and minimum redundancy–maximum relevance [26].

Classifiers include SVM [3], [12], [15], [26], [49], [72], [73], [93], [95], [109], Random Forest [70], K-Nearest Neighbor (K-NN) [3], Fuzzy KNN [56], Neural Networks [3], [95], Decision Tree [3], and Adaboost [4]. Due to the complex structure of thyroid nodules and the existing noise in ultrasound images, these methods can not achieve satisfactory results because they need very high discriminative local and global features from different scales to determine nodule malignancy. Moreover, overfitting usually happens to the classifiers due to insufficient features.

2.2 Conventional Methods

2.2.1 Thyroid Nodule Detection

There are a few conventional methods for automatic nodule detection. Keramidas et al. [40] define thyroid ROI and extract LBP features from the ROI

patches and classify these patches as normal thyroid or nodular tissue using a K-NN classifier. In [41], they extract Fuzzy LBP (FLBP) and Fuzzy GLH (FGLH) features and use SVM and K-NN classifiers.

2.2.2 Thyroid and Nodule Segmentation

There are several conventional methods for thyroid and nodule segmentation. Illanes et al. [38] extract texture features from US image patches using Continuous Wavelet Transformation(CWT) and parametrical modeling, then segment thyroid patches by K-means clustering. In [13], Radial Basis Function (RBF) neural network is used to segment thyroid by classifying US image patches based on the extracted textural and statistical features. Narayan et al. [65] perform speckle patch similarity estimation to segment thyroid. Tsantis et al. [94] create a hybrid multi-scale model (HMM) combined with wavelet edge detection and Hough transform to segment nodules. [36], [62], [66], [80] and [69] use active contours to segment nodules and thyroid, respectively. Variable Background Active contour (VBAC) based on Active contours without edges (ACWE) model [79], and Genetic-algorithm VBAC (GA VBAC) [37] have been proposed for nodule segmentation. In [47] Spatial neutrosophic clustering and level-set are used to segment nodules. Graph cut and Iterative random walks solver are used for thyroid segmentation in [69] and [22], respectively.

2.3 Deep Learning Methods

2.3.1 Thyroid Nodule Detection

Recently, few studies have proposed automatic thyroid nodule detection based on deep learning. Song et al. [88] propose a multi-scale SDD-based network embedded by nodule prior distribution guided layers. Liu et al. [53] propose multi-scale region-based detection by combining Feature Pyramid Network (FPN), Faster-RCNN, and a prior distribution for size and shape of nodules. [100], [111], [106], and [1] use YOLOv2, YOLOv3, SSD, and Mask-RCNN, respectively for nodule detection.

2.3.2 Thyroid and Nodule Segmentation

Several deep models have been proposed for thyroid and nodule segmentation [16]. Deep Convolutional Neural Network (CNN) models with multiple intermediate layers [59], [60], eight-layer Fully Convolutional Network (FCN) [50], deep VGG19 based model [108], and U-Net based model [113] have been applied for nodule segmentation. In [69], a 3D-UNet model is applied to segment the entire thyroid gland.

2.4 Object Detection Methods

In the past decades, large number of object detection approaches have been proposed [118]. Early approaches mainly use hand-crafted features to detect specific targets like faces [98], [99], humans [25]. Later, detection models have been extended to detect more general targets based on different features including image gradients (BING [21]), edges (EdgeBox [117]), image structures (Selective Search [96]), *etc*. However, due to the large variations of targets, traditional methods suffer from lack of accuracy and robustness.

In recent years, with the rapid development of machine learning and deep learning, object detection has achieved great improvements by introducing machine learning and deep learning techniques. These methods can be mainly categorized into the following groups: (1) Two-stage models; (2) One-stage models; (3) Anchor-free models.

2.4.1 Two-stage methods

In these detectors, regions of interest are generated by a region proposal method or a region proposal network, then these candidate regions are sent to the main pipeline for detection and classification. pioneer RCNN [30] uses an offline selective search as the first stage. In the second stage, it uses a CNN to extract features from the wrapped region proposals and classifies them by SVM. However, RCNN is very slow. Spatial pyramid pooling network (SPP-net) [34] was proposed to accelerate RCNN. SPP-net shares convolutional feature computation to avoid over-computation. It feeds an entire image to CNN

once and then extracts proposal features from feature map regions. Then, Fast-RCNN [29] was proposed to speed up training and testing even more and improve detection performance by using single-stage training. Similar to SPP-net, the computation sharing strategy is also employed. Although Fast-RCNN was significantly faster, the fixed region proposal algorithm was still a major bottleneck. Faster-RCNN [77] addresses this problem by proposing Region Proposal Network (RPN), in which proposals are predicted from features through a separate convolutional network. It utilizes multi-task loss to train RPN and detection networks together. In R-FCN [24], the computations after the ROI pooling layer are shared and moved after the last convolutional layer. Position-sensitive score maps were created in the following to increase detection accuracy. Since no region sub-network is applied on each ROI, R-FCN is faster than Faster-RCNN. Cascaded-RCNN [9] is another Faster-RCNN extension, which addresses overfitting problem during training and quality mismatch issue at inference time by increasing IoU thresholds to train a sequence of detectors.

2.4.2 One-stage methods

In these detectors, there is no region proposal preliminary stage, and the whole pipeline happens in a single stage, which makes these methods faster. OverFeat [83] is one of the pioneers in one-stage detectors, which uses image pyramid and multi-scale sliding window within a CNN and combines their predictions. OverFeat suffers from poor detection accuracy. YOLO (v1, v2, v3, v4, v5) [7], [74]–[76] and SSD [54] were proposed later to leverage the accuracy and meanwhile, improve the speed. They grid input image, consider a set of default boxes (anchor boxes) for each grid cell, and fed them all to a CNN once to score the presence of objects in these boxes. Here, there is no independent computation per region. extended versions of YOLO (v2, v3, v4, v5) [7], [75], [76] focus on achieving higher detection accuracy and faster speed. Unlike YOLO, SSD applies multi-scale feature maps to detect objects in different scales. Retinanet [51] is another representative model in the one-stage category, which achieved a high speed and detection accuracy by addressing the

extreme imbalance issue between foreground (contains object) and background (without object) classes during model training. A new Focal loss, introduced in Retinanet, puts more weight on a sparse set of hard misclassified examples and less weight on easy well-classified ones.

2.4.3 Anchor-free methods

Since both two-stage and one-stage methods are based on pre-defined anchor boxes, they are called anchor-based methods. The number of anchor boxes, their sizes, and aspect ratios highly affect detection performance in anchor-based methods. Therefore, they need to be determined carefully. Moreover, a large number of anchor boxes is required to guarantee a sufficient overlap with ground truth and as most of these boxes are labeled as background, it causes an imbalance issue between foreground and background during training. In order to overcome the aforementioned limitations in anchor-based detectors, anchor-free methods were proposed recently, which can be keypoint-based or center-based. In key-point based approaches, each object bounding box is represented by a set of keypoints, which are predicted through a CNN. CornerNet [48], CenterNet [27], ExtremeNet [114] and RepPoints [107] are from this category. Center-based approaches consider central points, such as FoveaBox [46], or region points, such as FCOS [91], inside the object bounding box as foreground and predict the distance to the box borders.

2.5 Object Segmentation Methods

Object segmentation problem has been studied for decades in medical and natural images [64]. Early approaches use various conventional techniques such as thresholding, edge detection, graph cut [8], active contours [11], level-set [61], and superpixels [2] to determine target boundaries [82]. However, these approaches perform well only when the target boundaries are clear and well-defined. Motivated by the success of deep learning in image segmentation tasks, recent approaches focus on deep learning techniques. FCN [58] is one of the pioneers which uses convolutional layers to generate the segmentation

mask for an input image. In FCN all fully-connected layers are replaced by fully-convolutional layers to covert classification scores to segmentation output. To produce more accurate segmentation, feature maps from shallow and deep layers are combined together through skip connections. Due to the large receptive field of FCN, it ignores global information and some useful semantic context. To address this issue, several methods use graphical models to feed semantic context into FCN [17], [57], [81], [112]. ParseNet [55] also adds global features to local feature maps by using global average pooling to make the segmentation much refined. Pyramid scene parsing network (PSPNet) learns global context information by extracting different patterns from an input image and feeding these feature maps into a pyramid pooling module to fuse features from different pyramid scales. DeepLab family [17]–[20] captures global context information from different scales by applying several dilated (atrous) convolutions with different dilation rates.

Another well-known deep learning segmentation framework is encoder-decoder architecture, such as SegNet [5], V-Net [63], W -Net [104], HRNet [89], U-Net [78] and its versions, including UNet++[116], Dense-UNet [32], Res-UNet [105], Attention-UNet [67], and U2-Net [71]. U-Net based models are the most popular methods in medical image segmentation. Their encoder (contracting) and decoder (expanding) paths are almost symmetric. The encoder is responsible for extracting features from inputs and capturing their contexts. The decoder makes localization precise and leads to a smooth segmentation. Skip connections append high-resolution feature maps from the encoder part to the corresponding up-sampled feature maps at the same level in the decoder part to keep pattern information.

Chapter 3

Proposed Methods

In this section, we will discuss each module in the proposed end-to-end pipeline of the fully automated thyroid CAD system. Modules include automatic thyroid nodule detection, automatic segmentation of nodule and thyroid gland boundaries, and nodule classification.

3.1 Thyroid Nodule Detection

Thyroid nodule detection in ultrasound images is a very challenging task in both medical image analysis and computer vision fields. As mentioned in Chapter 1, different characteristics of ultrasound images from natural images, and also thyroid nodules characteristics lead to high inter-observer variability among human readers, and analogous challenges for machine learning tools, which often lead to inaccurate or unreliable nodule detection. To address the above issues, the utilization of multi-scale features is very important. Therefore, we propose a novel one-stage thyroid ultrasound nodule detection model, called *TUN-Det*, whose backbone is built upon the ReSidual U-blocks (RSU) [71], which is capable of extracting richer multi-scale features from feature maps with different resolutions [71]. In addition, we design a multi-head architecture for both the nodule bounding boxes classification and regression in our TUN-Det to predict more reliable results. Each multi-head module is comprised of three different heads, which are variants of the RSU block and arranged in parallel. Each multi-head module outputs three separate outputs, which are supervised by losses computed independently in the training process.

In the inference step, these outputs of each multi-head module are fused by the Weighted Boxes Fusion (WBF) algorithm [87], which is similar to the ensemble strategy. This strategy is able to greatly improve the detection performance.

3.1.1 TUN-Det Architecture

Feature Pyramid Network (FPN) is one of the most popular architectures in object detection. The FPN architecture is able to efficiently extract high-level and low-level features from deeper and shallow layers, respectively. As we know, multi-scale features play very important roles in object detection. High-level features are responsible for predicting the classification scores while low-level features are used to guarantee the bounding boxes' regression accuracy. The FPN architectures usually take existing image classification networks, such as VGG [86], ResNet [35], and so on, as their backbones. However, each stage of these backbones is only able to capture single-scale features because image classification backbones are designed to perceive only high-level semantic meaning while paying less attention to the low-level or multi-scale features [71]. To capture more multi-scale features from different stages, we build the TUN-Det upon the RSU, which was first proposed in salient object detection U^2-Net [71]. Our proposed TUN-Det is also a one-stage FPN similar to RetinaNet [51].

Figure 3.1 illustrates the overall architecture of our newly proposed TUN-Det for ultrasound thyroid nodule detection. As we can see, the backbone of our TUN-Det consists of five stages. The first stage is a plain convolution layer with stride of two, which is used to reduce the feature map resolution. The second to the fifth stages are RSU-7, RSU-6, RSU-5 and RSU-4, respectively. There is a maxpooling operation between the neighboring stages. Compared with other plain convolutions, the RSUs are able to capture both local and global information from feature maps with arbitrary resolutions[71]. Therefore, richer multi-scale features $\{C_3, C_4, C_5\}$ can be extracted by the backbone built upon these blocks for supporting nodule detection. Then, an FPN [51] is applied on top of the backbone's features $\{C_3, C_4, C_5\}$ to create multi-scale pyramid features $\{P_3, P_4, P_5, P_6, P_7\}$, which will be used for bounding boxes

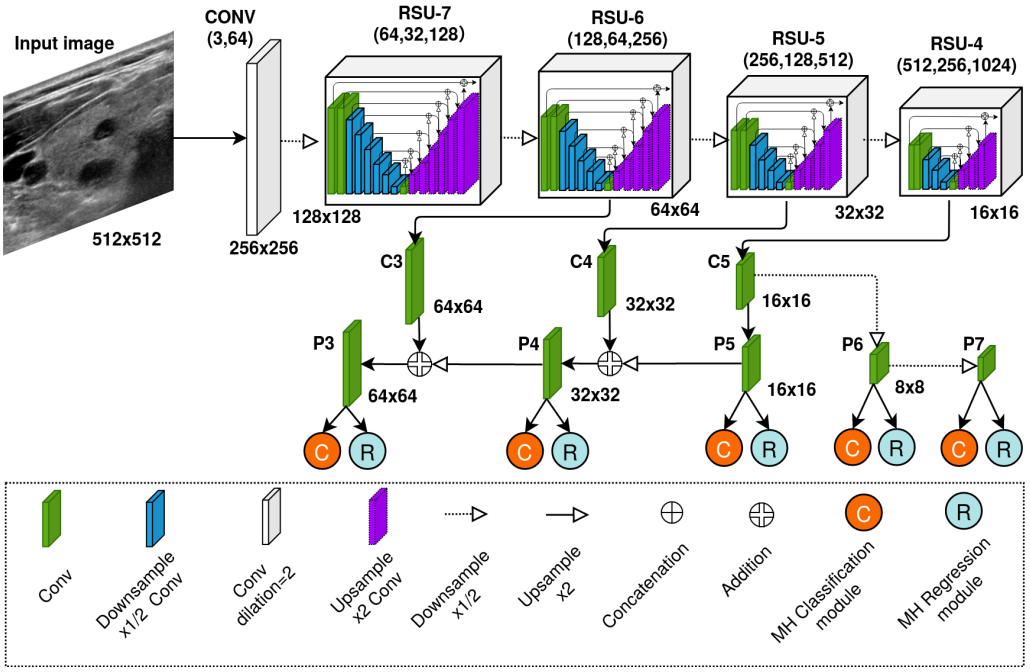


Figure 3.1: Architecture of the proposed TUN-Det.

regression and classification.

3.1.2 Multi-head Classification and Regression Module

After obtaining the multi-scale pyramid features $\{P_3, P_4, P_5, P_6, P_7\}$, the most important step is regressing the coordinates of bounding boxes and predicting their probabilities of being nodules. These two processes are usually implemented by a regression module $BBOX_i = R(P_i)$ and a classification module $CLAS_i = C(P_i)$, respectively. The regression outputs $\{BBOX_3, BBOX_4, ..., BBOX_7\}$ and the classification outputs $\{CLAS_3, CLAS_4, ..., CLAS_7\}$ from different features are then fused to achieve the final detection results by conducting non-maximum suppression (NMS).

To further reduce the False Positives (FP) and False Negatives (FN) in the detection results, a multi-model ensemble strategy [115] is usually considered. However, this approach is not preferable in real-world applications due to high computational and time costs. Hence, we design a multi-head (three-head) architecture for both classification and regression modules to address this issue. Particularly, each classification and regression module consists

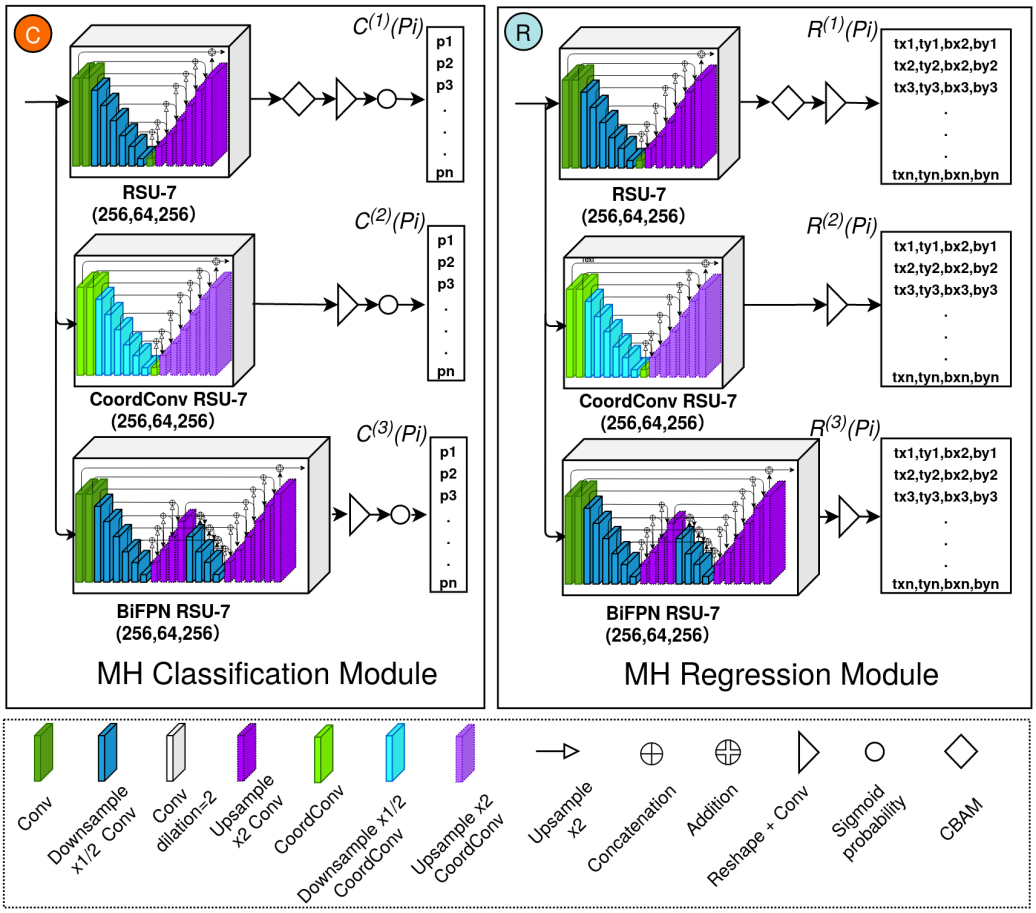


Figure 3.2: Multi-head classification and regression module.

of three parallel-configured heads, $\{C^{(1)}, C^{(2)}, C^{(3)}\}$, and $\{R^{(1)}, R^{(2)}, R^{(3)}\}$, respectively. Given a feature map P_i, three classification outputs, $\{C^{(1)}(P_i)$, $C^{(2)}(P_i)$, $C^{(3)}(P_i)\}$, and three regression outputs, $\{R^{(1)}(P_i)$, $R^{(2)}(P_i)$, $R^{(3)}(P_i)\}$, will be produced. In the training process, their losses will be computed separately and summed to supervise the model training. In the inference step, the Weighted Boxes Fusion (WBF) algorithm [87] are introduced to fuse the regression and classification outputs of generated different heads. This design embeds the ensemble strategy into both the classification and regression module so that it is able to improve the detection accuracy while avoiding training multiple models, which is a standard procedure in common ensemble methods.

Architectures of $R^{(i)}$ and $C^{(i)}$ are the same except for the last convolution layer (Figure. 3.2). To increase the diversity of the prediction results and hence

reducing the variance, three variants, CBAM RSU-7, CoordConv RSU-7, and BiFPN RSU-7, of RSU-7 are developed to construct the multi-head modules.

The first head is CBAM RSU-7, in which a Convolutional Block Attention Module (CBAM) [102] block is added after the standard RSU-7 block to refine features by channel (M_c) and spatial (M_s) attention maps. The formulation can be described as:

$$\begin{aligned} F_c &= M_c(F_{in}) \otimes F_{in}, \\ F_s &= M_s(F_c) \otimes F_c, \end{aligned} \tag{3.1}$$

where F_{in} is an input feature map, F_c and F_s are feature maps after refinement by the channel attention map and the following spatial attention map, respectively. Channel and spatial attention maps are computed as follows:

$$\begin{aligned} M_c(F_{in}) &= \sigma(MLP(AvgPool(F_{in})) + MLP(MaxPool(F_{in}))), \\ M_s(F_c) &= \sigma(f^{7\times7}([AvgPool(F_c); MaxPool(F_c)])), \end{aligned} \tag{3.2}$$

The second head is CoordConv RSU-7, which replaces the plain convolution layers in the original RSU-7 by Coordinate Convolution [52] layers to encode geometric information. CoordConv can be described as:

$$conv(concat(F_{in}, F_i, F_j)), \tag{3.3}$$

where $F_{in} \in \mathbb{R}^{(h\times w\times c)}$ is an input feature map, F_i and F_j are extra row and column coordinate channels, respectively.

The third head is BiFPN RSU-7, which expands RSU-7 by adding a bi-directional FPN (BiFPN) [90] layer between the encoding and decoding stages to improve multi-scale feature representation. BiFPN layer has a $\cap$-shape architecture consisted of bottom-up and top-down pathways, which helps to learn high-level features by fusing them in two directions. Here, we use a four-stage BiFPN layer to avoid complexity and reduce the number of trainable parameters.

3.1.3 Supervision

As shown in Figure 3.1, our newly proposed TUN-Det has five groups of classification and regression outputs. Therefore, the total loss is the summation

of these five groups of outputs:

$$\mathcal{L} = \sum_{i=1}^{5} \alpha_i \mathcal{L}_i, \tag{3.4}$$

where α_i and $\mathcal{L}_i$ are the weight and the loss of each group, respectively. all α_i values are set to 1.0 here. For every anchor, each group produces three classification outputs $\{C^{(1)}, C^{(2)}, C^{(3)}\}$ and three regression outputs $\{R^{(1)}, R^{(2)}, R^{(3)}\}$. Therefore, the loss of each group can be defined as

$$\mathcal{L}_i = \textstyle\sum_{j=1}^{3} \lambda_i^{C^{(j)}} \mathcal{L}_i^{C^{(j)}} + \sum_{j=1}^{3} \lambda_i^{R^{(j)}} \mathcal{L}_i^{R^{(j)}}, \tag{3.5}$$

where $\mathcal{L}_i^{C^{(j)}}$ and $\mathcal{L}_i^{R^{(j)}}$ are the corresponding losses for classification and regression outputs, respectively. $\lambda_i^{C^{(j)}}$ and $\lambda_i^{R^{(j)}}$ are their corresponding weights to determine the importance of each output. We set all the λ weights to 1.0 in our experiments. $\mathcal{L}_i^{C^{(j)}}$ is the focal loss [51] for classification. It can be defined as follows:

$$\begin{aligned} \mathcal{L}_i^{C^{(i)}} &= \text{Focal}(p_t) = \beta_t (1 - p_t)^{\gamma} \times BCE(p_c, y_c), \\ p_t &= \begin{cases} p_c & \text{if } y_c = 1 \\ 1 - p_c & \text{otherwise} \end{cases}, \quad \beta_t = \begin{cases} \beta & \text{if } y_c = 1 \\ 1 - \beta & \text{otherwise,} \end{cases} \end{aligned} \tag{3.6}$$

where p_c and y_c are predicted and target classes respectively. β and γ are focal weighting factor and focusing parameter, respectively that are set to 0.25 and 2.0, respectively.

$\mathcal{L}_i^{R^{(j)}}$ is the Smooth-L1 loss [29] for regression, which is less sensitive to outliers than L2 loss and it helps to avoid exploding gradients [29]. Smooth-L1 is defined as:

$$\mathcal{L}_i^{R^{(j)}} = \text{Smooth-L1}(p_r, y_r) = \begin{cases} 0.5(\sigma x)^2 & \text{if } |x| < \frac{1}{\sigma^2} \\ |x| - \frac{0.5}{\sigma^2} & \text{otherwise,} \end{cases}, \quad x = p_r - y_r \tag{3.7}$$

where p_r and y_r are predicted and ground-truth bounding boxes respectively. σ splits loss into L2 and L1 regions by defining the point where the regression loss changes from L2 to L1 loss. It is set to 3.0 in our experiments.

3.2 Thyroid Nodule Segmentation

After finding ROI by our proposed thyroid nodule detection model TUN-Det, the next key module in the CAD system is the precise segmentation of nodule

boundaries. As mentioned in Chapter 1, thyroid nodule segmentation in ultrasound images is also challenging due to the wide range of nodule textures and sizes, and also ultrasound images characteristics. Here, we propose residual dilated U-Net model, called *resDUnet*, to generate a segmentation mask for each nodule based on its ROI. Our resDUnet has a residual structure, which improves its learning ability, and accelerates the convergence. The additional dilated convolution layers are also applied to generate multi-scale features to map the encoding to decoding path of the U-Net.

3.2.1 resDUnet Architecture

The U-Net architecture uses a series of convolution and maxpooling operations in the encoding path to learn image features. The spatial and contextual information of these features is reconstructed in the decoding path through transposed convolutions and skip connections from the encoder. The newly proposed resDUnet improves the segmentation result by adding residual shortcut connections [35] in the building blocks and embedding dilated convolution layers in the bottleneck part of the network. Figure 3.3 illustrates the overall architecture of the proposed resDUnet for ultrasound thyroid nodule segmentation. Residual connections intend to eliminate vanishing and exploding gradient problems and lead to consistent training [35]. Dilated convolution layers apply 3×3 convolution with different dilation rates, which can be defined as:

$$(F *_l k)(p) = \sum_{s+lt=p} F(s)k(t), \tag{3.8}$$

where $*_l$ is the dilated convolution operator. F and k are discrete function and discrete 3×3 filter, respectively. Considering that dilated convolution increases receptive field while keeping the resolution, and also different dilation rates apply different receptive fields, more robust features are extracted in different scales.

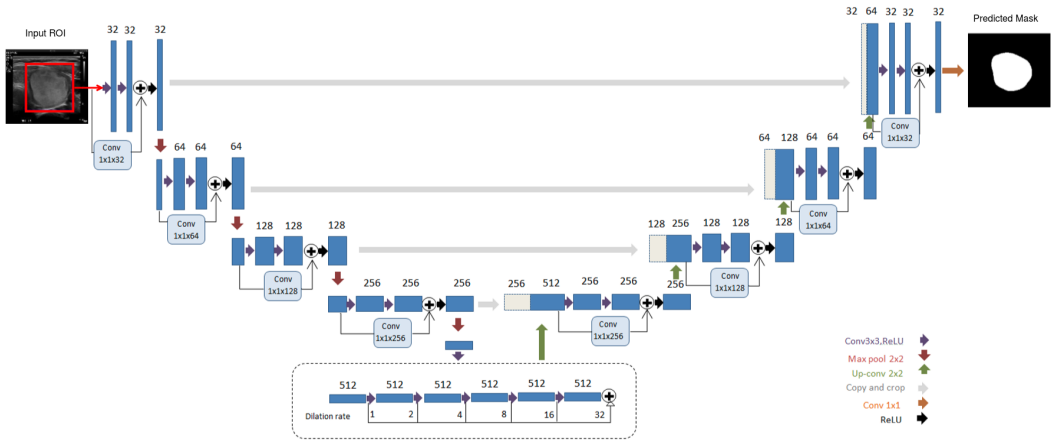


Figure 3.3: Schematic architecture of the proposed resDUnet.

3.2.2 Supervision

Our proposed resDUnet is trained with Dice coefficient loss function, which is defined as:

$$\text{Dice}(p, y) = -\frac{2 * (p \cap y) + \epsilon}{(p + y) + \epsilon}, \tag{3.9}$$

where p and y are the predicted segmentation mask and ground-truth mask, respectively. ϵ is added to ensure that the function is always defined.

3.3 Thyroid Gland Segmentation

Thyroid gland segmentation plays an important role in analyzing thyroid characteristics for nodule assessment and also filtering the false detected nodules outside of the thyroid gland. Moreover, it assists in detecting thyroid abnormalities. A thyroid ultrasound scan consists of a series of thyroid picture frames, which is called a sweep. Thyroid segmentation is also a challenging task because the thyroid gland is more sensitive to ultrasound image quality, noise and artifacts due to its size, which usually causes the under-segmentation problem. To address this issue, we propose LSTM-UNet which considers the inter-frame correlation information of consecutive frames by using time-distributed convolution blocks and embedding bidirectional convolutional LSTM (BiConvLSTM) in the U-Net.

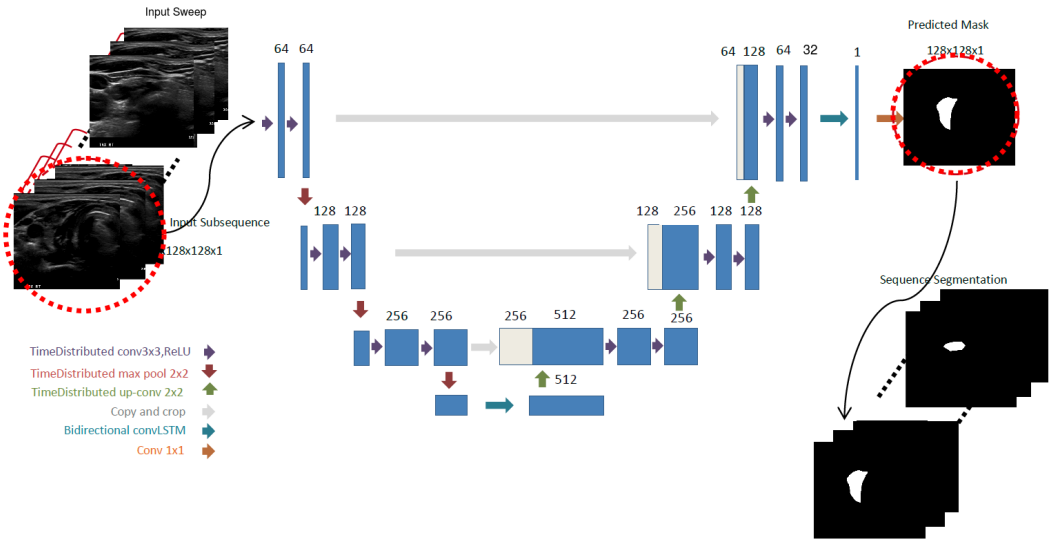


Figure 3.4: Schematic architecture of the proposed LSTM-UNet.

3.3.1 LSTM-UNet Architecture

There is a correlation between consecutive temporal frames in the sweep. Since sequential data can be processed using successful Recurrent Neural Networks (RNNs), we integrate time-distributed convolution blocks and BiConvLSTM into the U-Net to make its predictions more robust. The intuition behind using these building blocks is exploiting spatio-temporal correlation of consecutive frames due to the fact that the consecutive frames represent correlated features. Figure 3.4 shows the overall architecture of our proposed LSTM-UNet for ultrasound thyroid gland segmentation. As we can see, when input sequence length is 3 with sample rate 1, each frame in the sweep is fed to the network along with its previous and next frames as a 3D input sub-sequence to predict its segmentation mask. Time-distributed is a wrapper that applies a convolution layer to each temporal frame of the input sequence independently. BiConvLSTM is a forward and backward ConvLSTM, which its LSTM fully-connected layers have been replaced with convolutional layers to maintain spatial information and improve the prediction of image sequences. ConvLSTM

is defined as:

$$
\begin{aligned}
i_t &= \sigma(W_{hi} * H_{t-1} + W_{xi} * X_t) + b_i) \\
f_t &= \sigma(W_{hf} * H_{t-1} + W_{xf} * X_t) + b_f) \\
\tilde{C}_t &= tanh(W_{hc} * H_{t-1} + W_{xc} * X_t) + b_c) \\
C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\
o_t &= \sigma(W_{ho} * H_{t-1} + W_{xo} * X_t) + b_o) \\
H_t &= o_t \odot tanh(C_t),
\end{aligned}
\tag{3.10}
$$

where σ is the sigmoid function, $*$ is convolution operator, $\odot$ is element-wise multiplication. The input $\{X_1, X_2, ..., X_t\}$ updates hidden states $\{H_0, H_1, ..., H_{t-1}\}$ and cell states $\{C_1, C_2, ..., C_t\}$. $\{W_{hi}, W_{hf}, W_{hc}, W_{ho}\}$ and $\{W_{xi}, W_{xf}, W_{xc}, W_{xo}\}$ are convolution kernels corresponding to the hidden state and the input, respectively. $\{b_i, b_f, b_c, b_o\}$ are corresponding bias terms. Figure 3.5 shows BiConvLSTM, which includes two directions ConvLSTM.

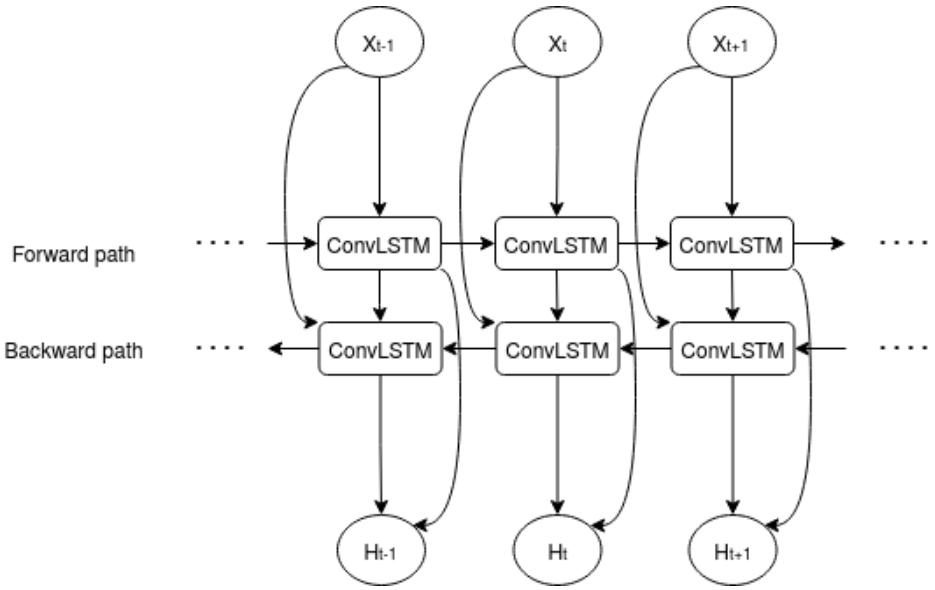


Figure 3.5: Block diagram of BiConvLSTM.

The LSTM-UNet is trained with Dice coefficient loss defined in Equation 3.9.

3.4 Thyroid Nodule Classification

The last component of the proposed CAD system is thyroid nodule classification. As mentioned in Chapter 1, TIRADS is the baseline for thyroid nodule classification. Figure 3.6[1] illustrates the five nodule characteristics based on TIRADS including Shape, Margin, Composition, Echigenicity, and Echogenic foci.

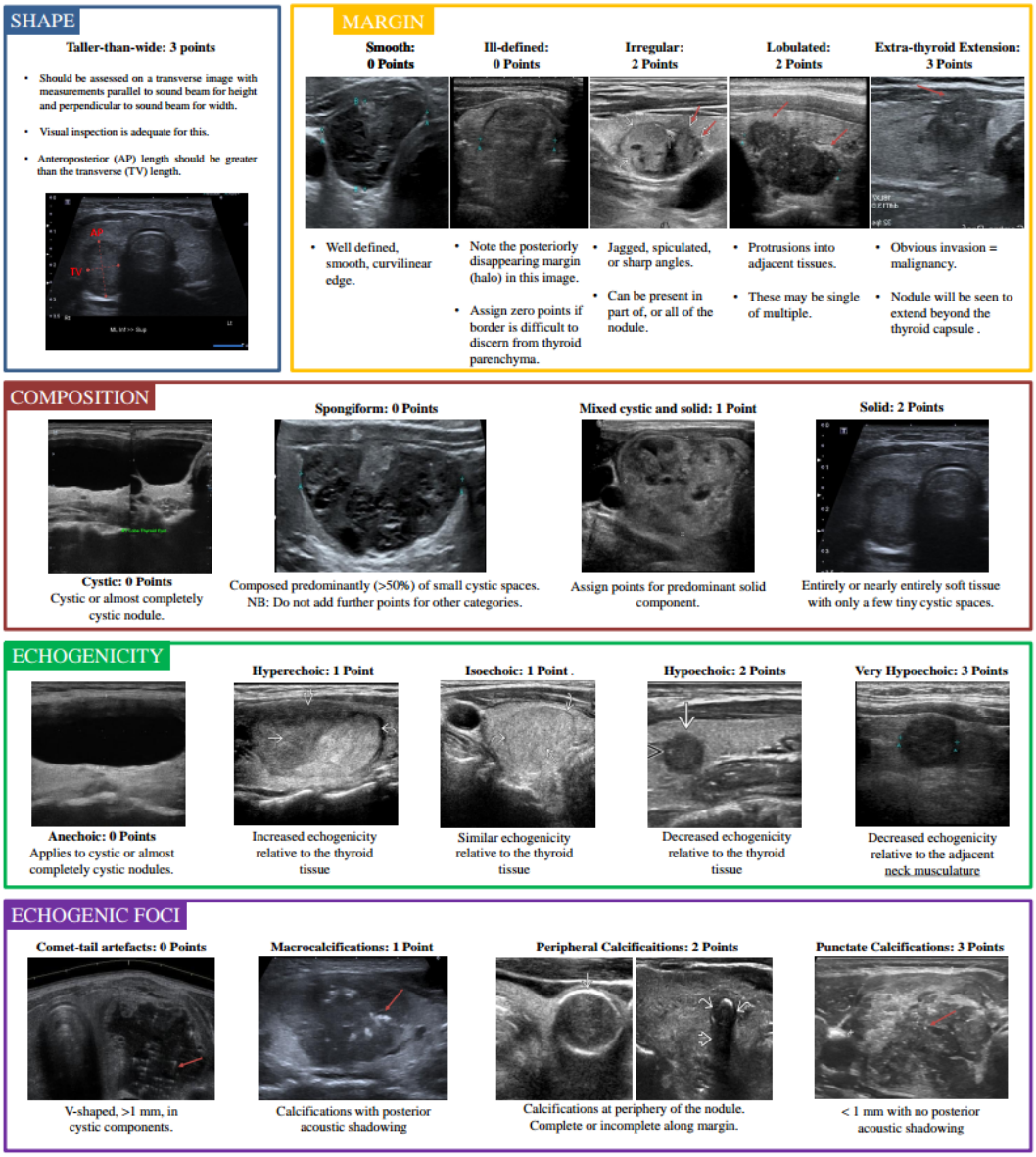


Figure 3.6: TIRADS five nodule characteristics and their definitions.

[1]Photo credit: ePosters - Pictorial Review of TI-RADS Scoring System for Thyroid Nodules

In this thesis, we only focus on composition and margin. Using the nodule predicted segmentation mask and pixel information in the original image, two rule-based classifiers are developed to categorize the composition and margin of the segmented nodule to estimate its malignancy risk.

3.4.1 Nodule Composition

Nodule composition is defined as the proportion of cystic (fluid) and solid (soft tissue) components. TIRADS classification defines composition categories as cystic, spongiform, mixed solid cystic, and solid:

- Cystic: When almost the entire nodule is filled by cystic components.
- Spongiform: When there are predominant small cystic spaces.
- Mixed solid cystic: When the nodule includes a fair ratio of cystic and solid components.
- Solid: When entirely or nearly entirely of the nodule is solid and only a few small cystic spaces might exist.

The proposed composition classifier has been designed based on the clinical rules. Figure 3.7 shows our rule-based classifier. It consists of five different image processing based stages including (1) morphological erosion to remove peripheral calcification, (2) hard thresholding to find solid percentage, (3) adaptive histogram equalization for image enhancement, (4) Isodata thresholding to find dark regions, and (5) morphological closing to find the number of liquid blobs. Results of these stages are compared to the pre-defined rules and the final decision about the composition category is made.

3.4.2 Nodule Margin

Nodule margin is described as the border of the nodule, which can be smooth, ill-defined, irregular or lobulated, and extra-thyroid extension based on the TIRADS definition:

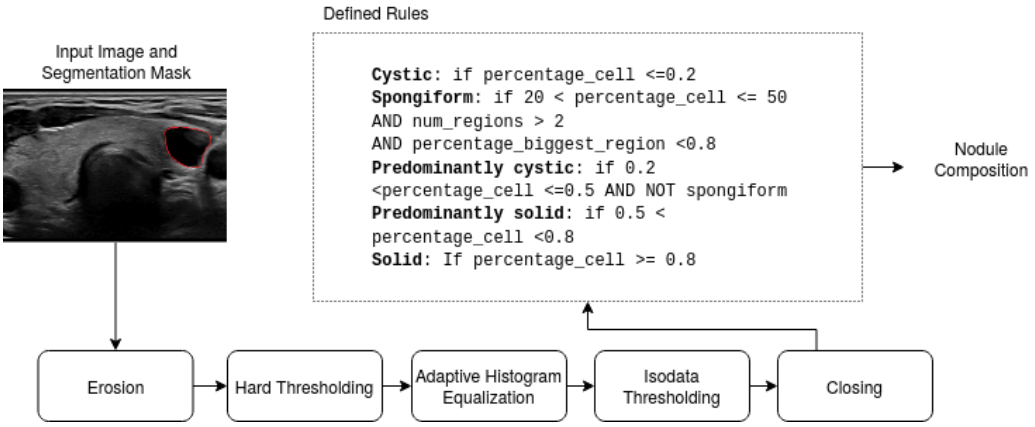


Figure 3.7: Rule-Based thyroid nodule composition classifier.

- Smooth: When the nodule border is well-defined, smooth, and curvilinear.
- Ill-defined: When there is no clear boundary for the nodule.
- Irregular: When a part of or the entire nodule border is jagged, speculated, or with sharp angles.
- Lobulated: When the nodule border includes single or multiple protrusions.
- Extra-thyroid extension: When the nodule invades beyond the thyroid gland.

The proposed margin classifier has been designed based on the clinical margin definitions. Figure 4.6 illustrates our rule-based classifier, which includes different image processing techniques. (1) morphological closing is applied on the segmentation mask of the nodule to remove probable peripheral calcification. (2) after finding the nodule contour, (3) Convex Hull and (4) best fit ellipsoid are computed for the contour. (5) Sobel edge detection is used to find image gradient. Results of these modules are compared to the pre-defined rules and the final decision about the margin category is made.

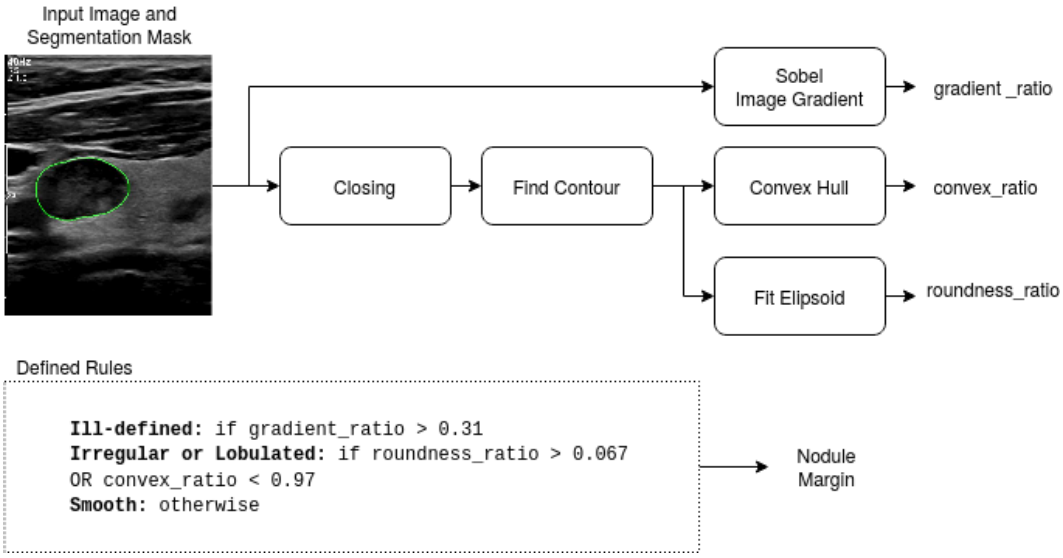


Figure 3.8: Rule-Based thyroid nodule margin classifier.

Chapter 4

Experimental Results

In this chapter, we evaluate the performance of the proposed modules in our automatic CAD system through the experimental results. We first introduce datasets that we have used for each module and present the evaluation metrics for each task. Due to the availability of ground truth, different datasets have been used for nodule detection, nodule segmentation, and thyroid segmentation tasks as described in the following section.

4.1 Datasets

4.1.1 Thyroid Nodule Detection

To validate the performance of our newly proposed TUN-Det on ultrasound thyroid nodule detection task, we built a new thyroid nodule detection dataset. The dataset was retrospectively collected from 700 patients aged between 18–82 years who presented at 12 different imaging centers for a thyroid ultrasound examination. Our retrospective study was approved by the health research ethics boards of the participating centers. There are a total of 3941 ultrasound images, which were extracted from 1924 transverse (TRX) and 2017 sagittal (SAG) scans. These images were split into three subsets for training (2534), validation (565), and testing (842) with 3554, 981, and 1268 labeled nodule bounding boxes, respectively. All nodule bounding boxes were manually labeled by five experienced sonographers (with ≥ 8 years of experience in thyroid sonography) and validated by three radiologists.

4.1.2 Thyroid Nodule Segmentation and Classification

The ultrasound image data used for training and testing our newly proposed resDUnet on ultrasound thyroid nodule segmentation task was obtained retrospectively from our data collection center after obtaining ethics approval. The data comprises SAG and TRX cine-sweeps acquired using Philips and GE ultrasound scanners. The training set includes a total of 4266 ultrasound image slices containing thyroid nodules which have been acquired from 63 SAG and TRX sweeps of 41 patients. The test set includes a total of 352 ultrasound images from 141 patients. The boundary of each nodule was manually delineated by experienced sonographers. Using the segmentation masks, we generated ROIs around the nodule which are used as the input to the network. In order to account for variability in the manually selected ROI, we randomly changed the centroid and dimensions of the ROI which generated an augmented dataset including 12798 images. Regarding the nodule classification task, nodules were categorized based on their composition and margin by medical experts.

4.1.3 Thyroid Gland Segmentation

To train and evaluate the performance of the proposed LSTM-UNet on the thyroid gland segmentation task, the dataset was retrospectively collected from 105 patients after obtaining ethics approval. There are a total of 1050 ultrasound SAG scans, which were split into training (824 sweeps of 86 patients) and testing (226 sweeps of 19 patients). All ground-truth masks were manually labeled by experienced sonographers.

4.2 Evaluation metrics

4.2.1 Average Precision

To evaluate the performance of our TUN-Det against other models, Average Precision (AP), which is one of the most frequently used metrics in object detection [51], has been used as the evaluation metric. To calculate AP, we

first need to determine whether the predicted bounding box is True Positive (TP), False Positive (FP), or False Negative (FN) using Intersection over Union (IoU), which is defined as:

$$\text{IoU} = \frac{P \cap G}{P \cup G}, \tag{4.1}$$

where P and G indicate the predicted bounding box and ground-truth bounding box, respectively. When IoU >specified threshold, it is considered as TP. If IoU <specified threshold or it is a duplicated bounding box, it is considered as FP. When a target is missing, FN happens. Each predicted bounding box has a confidence score, which is used in ranking it. The next step is calculating precision and recall over the test dataset, which are defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{4.2}$$

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{4.3}$$

Then, the area under the Precision-Recall curve is computed as AP.

4.2.2 Dice Score

To evaluate the performance of our proposed segmentation models (resDUnet for nodule segmentation and LSTM-UNet for thyroid gland segmentation) against other architectures, Dice Score has been used, which is one of the most popular evaluation metrics in object segmentation and it is defined as:

$$\text{Dice} = \frac{2 * |P' \cap G'|}{|P'| + |G'|}, \tag{4.4}$$

where P' and G' indicate the predicted segmentation mask and ground-truth mask, respectively.

4.2.3 Hausdorff Distance

Hausdorff Distance (HD) is another metric that has been used for evaluating the nodule segmentation model. HD measures the distance between the predicted segmentation mask contour points (P'_c) and ground-truth mask contour points (G'_c). It is defined as:

$$\mathrm{HD}(P'_c, G'_c) = \max\left\{\max_{p\in P'_c}\min_{g\in G'_c}\|p-g\|_2, \quad \max_{g\in G'_c}\min_{p\in P'_c}\|p-g\|_2\right\}. \tag{4.5}$$

4.2.4 Root Mean Square Error

Root Mean Square Error (RMSE) is also a distance metric for segmentation evaluation which is defined as:

$$\mathrm{RMSE}(P'_c, G'_c) = 0.5 * (\frac{1}{n}\sum_{i=1}^{n}\|p-\tilde{g}\|_2 + \frac{1}{m}\sum_{i=1}^{m}\|\tilde{p}-g\|_2), \tag{4.6}$$

where $p \in P'_c$ and $g \in G'_c$. $\tilde{P}'_c$ and $\tilde{G}'_c$ are sub-sample of P'_c and G'_c contours. $\tilde{p} \in \tilde{P}'_c$ and $\tilde{g} \in \tilde{G}'_c$.

4.2.5 Confusion Matrix and Kappa Score

To evaluate the performance of the nodule composition classifier, Confusion Matrix (error matrix), which is the popular metric for statistical classification, has been used. Rows and columns in the confusion matrix correspond to ground-truth and predicted class. Kappa score measures inter-rater reliability, which means how much ground-truth and classifier agree with each other. Kappa is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{4.7}$$

where P_o and P_e are probability of agreement (confusion matrix diagonal) and probability of random agreement, respectively.

4.2.6 Accuracy

Accuracy is another metric for classification, which is defined as:

$$\mathrm{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.8}$$

Accuracy has been used to evaluate the performance of the nodule margin classifier.

4.3 Thyroid Nodule Detection

4.3.1 Implementation Details

Our proposed TUN-Det was implemented in Tensorflow 1.14 and Keras. The input images are resized to 512×512 and the batch size is set to 1. The model parameters are initialized by Xavier and Adam optimizer [45] with default parameters is used to train the model. Both our training and testing processes were conducted on a 12-core, 24-thread PC with an AMD Ryzen Threadripper 2920x 4.3 GHz CPU (128 GB RAM) with an Nvidia GTX 1080 Ti GPU (11GB memory). The model converged after 200 epochs and took 20 hours in total. The average inference time per image (512×512) was 94 ms.

4.3.2 Ablation Study

To validate the effectiveness of our proposed architecture, ablation studies are conducted on different configurations and the results are summarized in Table 4.1. AP_{35}, AP_{50}, AP_{75} are average precision at the fixed 35%, 50%, 75% IoU thresholds, respectively. AP is the average of APs computed over ten different IoU thresholds from 50% to 95% [AP_{50}, AP_{55}, $\cdots$, AP_{95}]. The first two rows show the comparison between the original RetinaNet and the RetinaNet-like detection model with our newly developed backbones built upon the RSU-blocks. As we can see, our new adaptation greatly improves the performance against the original RetinaNet. The bottom part of the table illustrates the ablation studies on different configurations of classification and regression modules. It can be observed that our multi-head classification and regression modules, CoordConv-CBAM-BiFPN, shows better performance against other configurations in terms of the AP, AP_{35} and AP_{50}.

Table 4.1: Ablation on different backbones and heads configurations.

Model	AP	AP_{35}	AP_{50}	AP_{75}
RetinaNet w/ ResNet-50 backbone (baseline) [51]	39.50	74.03	69.07	41.39
w/ RSU backbone	40.73	79.56	74.81	41.62
w/ RSU + CBAM-RSU heads	42.63	80.92	75.49	**45.58**
w/ RSU + CoordConv-RSU heads	41.85	79.62	75.24	43.55
w/ RSU + BiFPN-RSU heads	41.70	80.11	74.20	43.54
w/ RSU + CoordConv-CBAM-BiFPN MH (Our TUN-Det)	**42.75**	**81.22**	**75.66**	45.53

4.3.3 Comparisons against State-of-the-arts

Quantitative Comparisons.

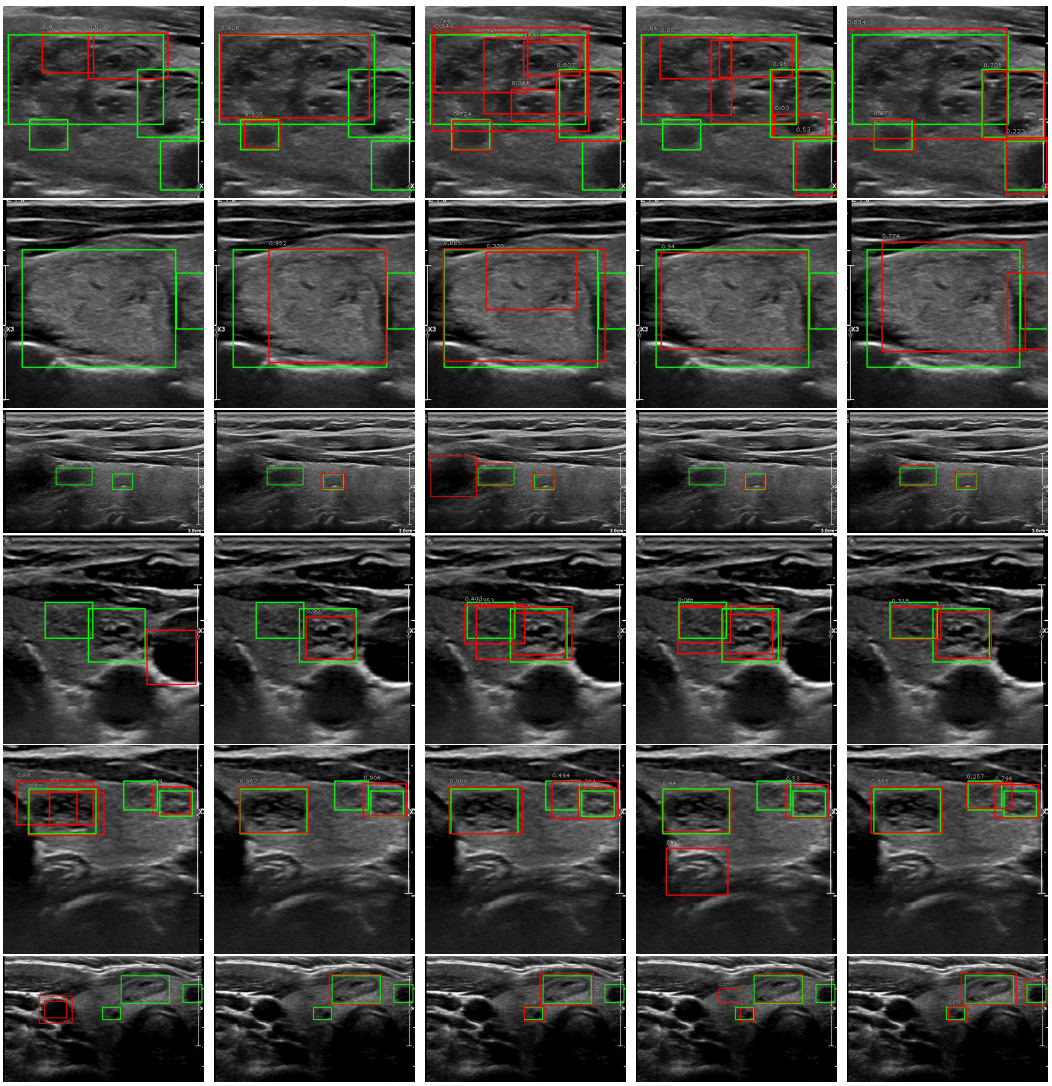
To evaluate the performance of our newly proposed TUN-Det, we compare our model against six typical state-of-the-art detection models including (i) Faster-RCNN [77] as a two-stage model; (ii) RetinaNet [51], SSD [54], YOLO-v4 [7] and YOLO-v5 [39] as one-stage models; and (iii) FCOS [91] as an anchor-free model. As shown in Table 4.2, our TUN-Det greatly improves the AP, AP_{35}, AP_{50} and AP_{75} against Faster-RCNN, RetinaNet, SSD, YOLOV4 and FCOS. Compared with YOLO-v5, our TUN-Det achieves better performance in terms of AP_{35} and AP_{50} but produces inferior results in terms of AP and AP_{75}. It is worth noting that 35% and 50% are usually selected as the threshold in our practical applications to achieve balanced results. Because higher or lower score usually generates very unbalanced precision and recall scores (one of them is close to 100% and the other one is close to 0%). Therefore, we believe higher scores on AP_{35} and AP_{50} are preferable, while AP ($AP_{50:95}$), AP_{75} and AP computed with higher IoU threshold are usually reported to show the average performance, which is not practically useful.

Table 4.2: Comparisons against the state-of-the-arts.

Model	Backbone	AP	AP_{35}	AP_{50}	AP_{75}
Faster-RCNN [77]	VGG16	0.91	42.13	29.65	2.58
SSD [54]	VGG16	19.05	40.10	36.55	18.10
FCOS [91]	ResNet-50	33.15	62.74	58.67	32.44
RetinaNet [51]	ResNet-50	39.50	74.03	69.07	41.39
YOLO-v4 [7]	CSPDarknet-53	40.43	78.21	72.48	42.04
YOLO-v5 [39]	CSPNet	**45.19**	78.71	74.74	**50.90**
TUN-Det	RSU	42.75	**81.22**	**75.66**	45.53

Qualitative Comparisons.

Figure 4.1 shows the qualitative comparison of our TUN-Det with other SOTA models on sampled SAG scans (first 2 rows) and TRX scans (last 2 rows). Each column shows the result of one method. The ground-truth is shown with green and detection result is shown in red. Figure 4.1 (1st row) shows that TUN-Det



(a) Faster RCNN (b) FCOS (c) RetinaNet (d) YOLOv5 (e) Our TUNDet

Figure 4.1: Qualitative comparison of ground-truth (green) and detection results (red) for different methods. Each column shows the result of one method.

can correctly detect the the challenging case of a non-homogeneous large hypoechoic nodule, while all other methods fail. The 2nd row demonstrates that TUN-Det can detect big, and partially visible nodules. The 3th row illustrates that TUN-Det performs well in detecting nodules with ill-defined boundaries, while others miss them. The 4rd, 5th and 6th rows highlight that our TUN-Det successfully excludes the false positive and false negative nodules. The last column of Fig.4.1 signifies that our TUN-Det produces the most accurate nodule detection results.

4.4 Thyroid Nodule Segmentation

4.4.1 Implementation Details

Our proposed resDUnet was implemented in Keras with Tensorflow backend. The input images are resized to 128×128 and the batch size is set to 1. The model has been trained with the Dice coefficient loss function and Adam optimizer with an initial learning rate of 10^{-5}. Keras EarlyStopping strategy is used to avoid over-fitting. Both our training and testing processes were conducted on the same system as the nodule detection task. The total number of trainable parameters of resDUnet is 1741833.

4.4.2 Comparisons against other segmentation models

Quantitative Comparisons.

To evaluate the performance of the proposed resDUnet, we compare our model with standard U-Net[78] and one of its variants, UNet++[116]. As we can see in Table. 4.3, the average Dice score of resDUNet is 82% on the entire test set, which is higher than U-Net and UNet++. To better realize the performance of resDUnet on different sizes of nodules, we categorize the nodules into three sub-groups based on their sizes (greater than 50k pixels, as large, 10k–50k pixels, as medium and less than 10k pixels, as small nodules) and compare the average Dice score in each sub-group. Our resDUnet improves the Dice score of small nodules. Boxplot diagram of the Dice scores has been presented in Figure 4.2.

Table 4.3: Comparisons against U-Net and UNet++.

Model	Large >50k pixels 51 images	Medium 10k-50k pixels 150 images	Small <10k pixels 151 images	Total 352 images		
	Dice	Dice	Dice	Dice	HD	RMSE
U-Net [78]	90%	87%	73%	81%	17.69	5.77
UNet++[116]	91%	87%	71%	80%	19.62	6.1
resDUnet	91%	87%	**74%**	**82%**	17.8	5.8

Qualitative Comparisons.

Figure 4.3 shows the qualitative comparison of resDUnet with U-Net and UNet++, as a variant of U-Net. The ground-truth and segmentation result

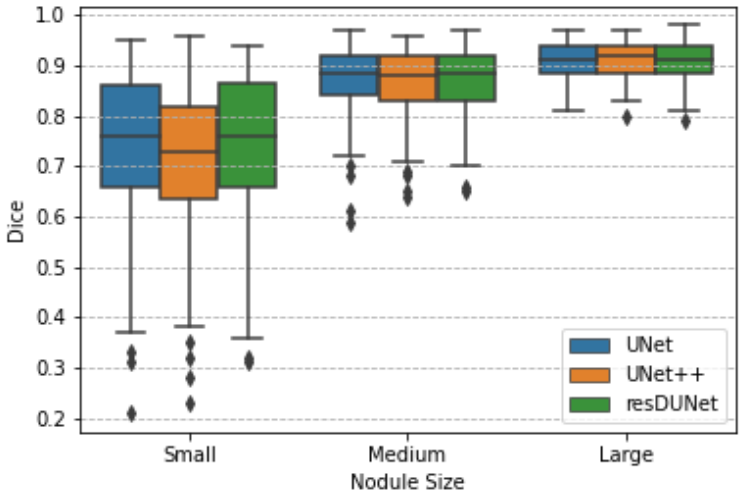
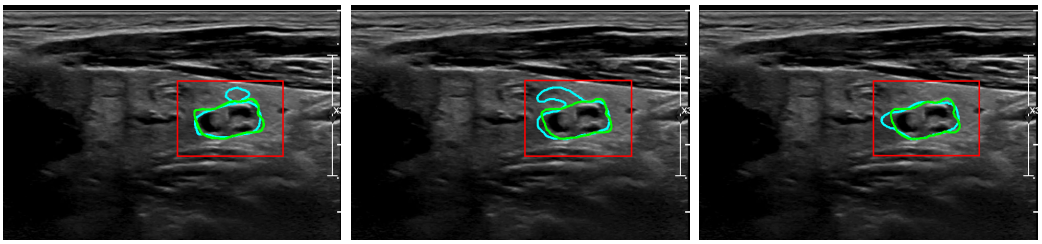


Figure 4.2: Boxplot of Dice scores of U-Net, UNet++, and resDUnet on three categories of nodules (small, medium, and large), and the entire test set.

are shown in green and blue, respectively. The last column of Figure 4.3 illustrates that resDUnet segmentation result is closer to the ground-truth rather than the two other networks. Moreover, it is able to delineate the nodule even with a relatively large ROI.



(a) U-Net (b) UNet++ (c) Our resDUNet

Figure 4.3: Qualitative comparison of ground-truth (green) and segmentation results (blue) for different methods. ROI has shown by a red rectangle.

4.5 Thyroid Gland Segmentation

4.5.1 Implementation Details

The proposed LSTM-UNet was implemented in Keras with Tensorflow backend. We set input sequence length to 3 with sample rate 1. Therefore, to

predict the mask of each frame, that frame along with its next and previous frames are fed to the network as a 3D input. Each frame is resized to 128×128 and the batch size is set to 1. The model uses Adam optimizer with an initial learning rate of 10^{-5}. Keras EarlyStopping strategy is used to avoid over-fitting. Both training and testing processes were conducted on the same system as the nodule detection task.

4.5.2 Comparisons against other segmentation models

Quantitative Comparisons.

To evaluate the performance of the proposed LSTM-UNet, We compare our model with standard U-Net[78], 3D-UNet[23] and Residual 3D-UNet. As shown in Table 4.4, LSTM-UNet achieves a better average Dice score (81%) compared to the other models.

Table 4.4: Comparisons against the state-of-the-arts.

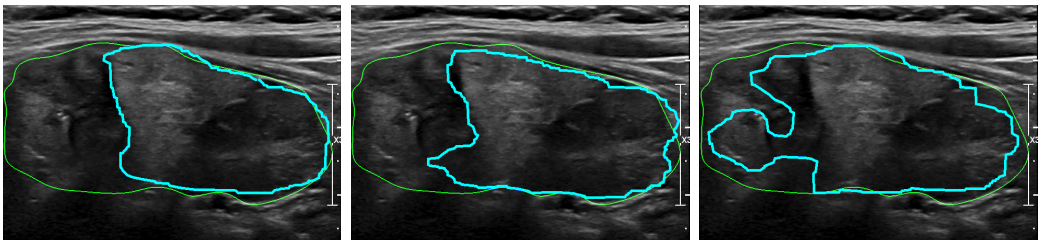
Model	Model size (MB)	Sequence size	Dice	Inference time (ms)
U-Net [78]	93.3	-	77%	2
Res 3D-UNet	67	32	77%	3
3D-UNet[23]	65.3	32	78%	1.5
LSTM-UNet	80	3	**81%**	17

Qualitative Comparisons.

Figure 4.4 shows the qualitative comparison of our LSTM-UNet with other models. The ground-truth and segmentation result are shown in green and blue, respectively. The last column of Figure 4.4 shows that LSTM-UNet tries to segment the whole area and avoid the inferior result.

4.6 Thyroid Nodule Classification

Nodule composition and margin rule-based classifiers were implemented using OpenCV 4.3.0 on the same system as the nodule detection task.



(a) U-Net (b) 3D-UNet (c) Our LSTM-UNet

Figure 4.4: Qualitative comparison of ground-truth (green) and segmentation results (blue) for different methods.

Table 4.5: Nodule composition confusion matrix.

Ground-truth*Prediction*	Solid	Cystic	Mixed	Spongiform
Solid	146	1	29	0
Cystic	0	5	9	0
Mixed solid cystic	19	2	51	0
Spongiform	2	0	1	0

4.6.1 Nodule Composition

Table 4.5 displays the performance of our rule-based composition classifier through the confusion matrix between the predicted class and ground-truth. The dataset includes a total of 265 validated nodules: 176 solid, 14 cystic, 72 mixed solid cystic, and 3 spongiform nodules. The Kappa score is 51%, which shows a relatively high agreement between the predictions and ground-truth assessments by radiologists. It is worth mentioning that the agreement between medical experts ranges between $(0.59 - 0.64)$ and the agreement between non-expert readers ranges between $(0.18 - 0.36)$ for thyroid nodule composition[43], [44]. Therefore, the agreement of our classifier with an expert radiologist is close to the medical experts' agreement. Figure 4.5 shows the limitation of the composition classifier, which is relatively subjective due to the trade off between solid and mixed solid cystic categories.

4.6.2 Nodule Margin

Margin classification is very challenging due to the lack of agreement between medical experts, which leads to low accuracy and inconsistencies in labeled data. Another issue is that the nodule segmentation mask needs to be com-

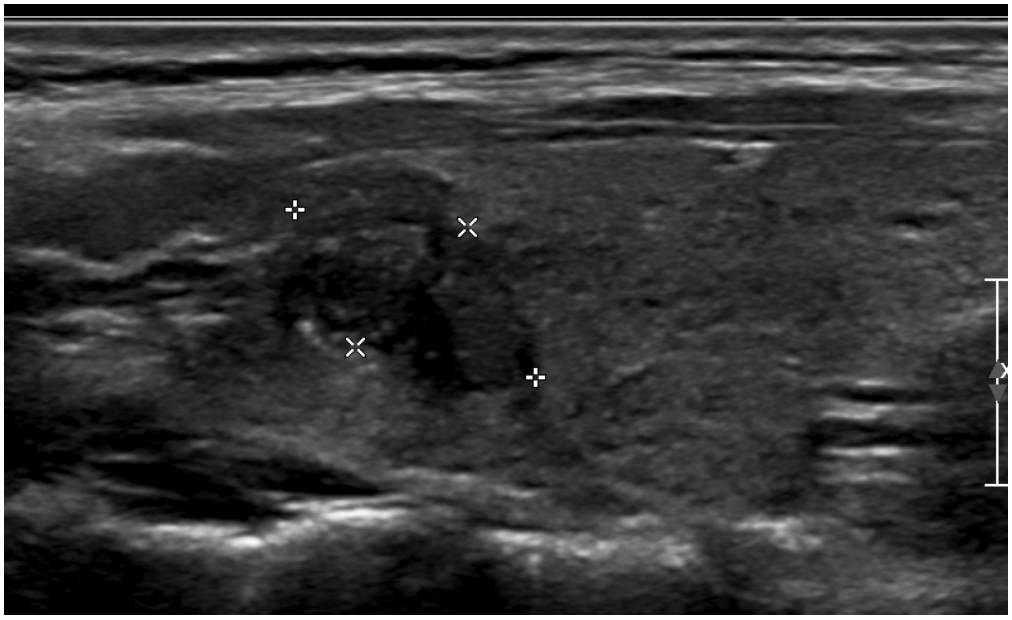
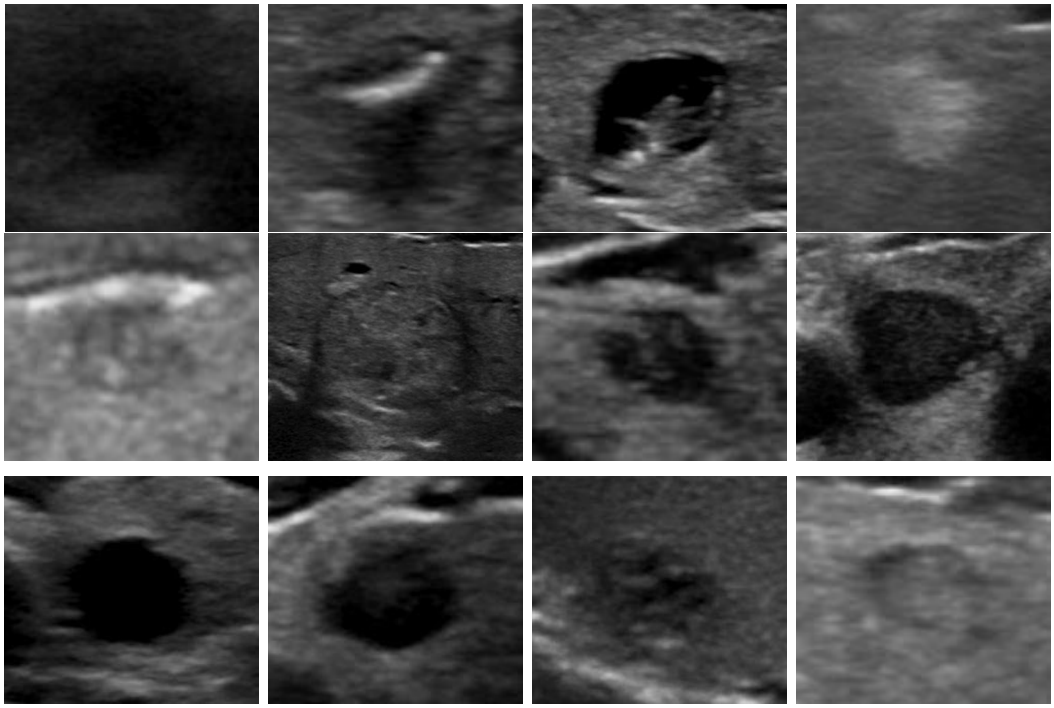


Figure 4.5: Trade off between solid and mixed solid cystic categories. The predicted class is mixed solid cystic, but the ground-truth label is solid.

pletely perfect and captures all small sharp angles, jagged, and protrusions, which barely happens, even when an expert annotates it manually. Figure 4.6 shows some difficult and challenging examples in each margin category. Therefore, we tested the margin classifier on a small dataset including a total of 139 validated nodules: 43 smooth, 40 ill-defined, and 56 irregular or lobulated nodules. The classifier correctly classified 27 smooth, 29 ill-defined, and 38 irregular or lobulated nodules. Hence, the accuracy is 68%.



(a) Smooth (b) Ill-defined (c) Irregular (d) Lobulated

Figure 4.6: Difficult thyroid nodule margin examples.

Chapter 5

Conclusion and Future Works

In this thesis, a fully automated thyroid CAD system has been presented to assist radiologists in thyroid nodule detection and diagnosis in thyroid ultrasound scans. The newly proposed CAD system consists of four modules: nodule detection, nodule segmentation, thyroid segmentation, and nodule classification. The proposed detection network, TUN-Det, includes a novel backbone built upon the RSU blocks to extract richer multi-scale features, and a newly proposed multi-head architecture for both classification and regression heads to fuse outputs from diversified sub-modules and a further improvement in the nodule detection performance. The proposed nodule segmentation network, resDUnet, includes residual shortcut connections in the U-Net architecture to make the training process consistent, and additional dilated convolution layers to extract more multi-scale features without losing the resolution. The proposed thyroid segmentation network, LSTM-UNet, replaces the plain convolution layers in the U-Net by time-distributed convolution blocks and bidirectional convolutional LSTM in U-Net to capture the spatial-temporal information by considering the correlation between consecutive temporal frames in the sweep. Two rule-based classifiers have been designed for nodule composition and nodule margin classification. They are based on several image processing techniques.

Experimental results demonstrate that TUN-Det achieves very competitive results against the state-of-the-art models on the overall Average Precision (AP) metric and outperforms them in terms of AP_{35} and AP_{50}. The resDUNet

achieves a high Dice score and much smooth visual results, compared to U-Net and UNet++. LSTM-UNet outperforms U-Net and 3D-UNet in terms of average Dice score.

5.1 Future Works

There is still a lot of room to improve the CAD system performance. In the near future, we will focus on:

- Improving the detection consistency between neighboring slices of 2D sweeps and exploring new representations for describing nodules merging and splitting in 3D space.
- Improving nodule segmentation for big ROIs, ill-defined nodules, and challenging textures.
- Registering nodules in 2D sweeps to improve the segmentation consistency.
- Improving thyroid gland segmentation for challenging thyroid textures in both SAG and TRX views.
- Decreasing inference time for all the models.
- Completing TIRADS classification.

References

[1] F. Abdolali, J. Kapur, J. L. Jaremko, M. Noga, A. R. Hareendranathan, and K. Punithakumar, "Automated thyroid nodule detection from ultrasound imaging using deep convolutional neural networks," *Computers in Biology and Medicine*, vol. 122, p. 103 871, 2020.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels," Tech. Rep., 2010.

[3] U. R. Acharya, P. Chowriappa, H. Fujita, S. Bhat, S. Dua, J. E. Koh, L. 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.

[4] U. R. Acharya, O. Faust, S. V. Sree, F. Molinari, and J. S. Suri, "Thyroscreen system: High resolution ultrasound thyroid image characterization into benign and malignant classes using novel combination of texture and discrete wavelet transform," *Computer methods and programs in biomedicine*, vol. 107, no. 2, pp. 233–241, 2012.

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[6] 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.

[7] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[8] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, IEEE, vol. 1, 2001, pp. 105–112.

[9] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[10] CanadianCancerStatisticsCommittee, *Canadian cancer statistics 2019*, URL `http://cancer.ca.login.ezproxy.library.ualberta.ca/Canadian-Cancer-Statistics-2019-EN`, (Accessed 12 December 2019), 2019.

[11] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *International journal of computer vision*, vol. 22, no. 1, pp. 61–79, 1997.

[12] 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.

[13] C.-Y. Chang, Y.-F. Lei, C.-H. Tseng, and S.-R. Shih, "Thyroid segmentation and volume estimation in ultrasound images," *IEEE transactions on biomedical engineering*, vol. 57, no. 6, pp. 1348–1357, 2010.

[14] T.-C. Chang, "The role of computer-aided detection and diagnosis system in the differential diagnosis of thyroid lesions in ultrasonography," *Journal of Medical Ultrasound*, vol. 23, no. 4, pp. 177–184, 2015.

[15] Y. Chang, A. K. Paul, N. Kim, J. H. Baek, Y. J. Choi, E. J. Ha, K. D. Lee, H. S. Lee, D. Shin, and N. Kim, "Computer-aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: A comparison with radiologist-based assessments," *Medical physics*, vol. 43, no. 1, pp. 554–567, 2016.

[16] J. Chen, H. You, and K. Li, "A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images," *Computer methods and programs in biomedicine*, vol. 185, p. 105 329, 2020.

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[19] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[21] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3286–3293.

[22] D. China, A. Illanes, P. Poudel, M. Friebe, P. Mitra, and D. Sheet, "Anatomical structure segmentation in ultrasound volumes using cross frame belief propagating iterative random walks," *IEEE journal of biomedical and health informatics*, vol. 23, no. 3, pp. 1110–1118, 2018.

[23] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, Springer, 2016, pp. 424–432.

[24] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, IEEE, vol. 1, 2005, pp. 886–893.

[26] 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.

[27] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6569–6578.

[28] C. Durante, G. Costante, G. Lucisano, R. Bruno, D. Meringolo, A. Paciaroni, E. Puxeddu, M. Torlontano, S. Tumino, M. Attard, *et al.*, "The natural history of benign thyroid nodules," *Jama*, vol. 313, no. 9, pp. 926–935, 2015.

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

[30] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[31] E. G. Grant, F. N. Tessler, J. K. Hoang, J. E. Langer, M. D. Beland, L. L. Berland, J. J. Cronan, T. S. Desser, M. C. Frates, U. M. Hamper, *et al.*, "Thyroid ultrasound reporting lexicon: White paper of the acr thyroid imaging, reporting and data system (tirads) committee," *Journal of the American college of radiology*, vol. 12, no. 12, pp. 1272–1279, 2015.

[32] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis, "Fully dense unet for 2-d sparse photoacoustic tomography artifact removal," *IEEE journal of biomedical and health informatics*, vol. 24, no. 2, pp. 568–576, 2019.

[33] B. R. Haugen, E. K. Alexander, K. C. Bible, G. M. Doherty, S. J. Mandel, Y. E. Nikiforov, F. Pacini, G. W. Randolph, A. M. Sawka, M. Schlumberger, *et al.*, "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*, vol. 26, no. 1, pp. 1–133, 2016.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, IEEE Computer Society, 2016, pp. 770–778.

[36] D. K. Iakovidis, M. A. Savelonas, S. Karkanis, and D. E. Maroulis, "Segmentation of medical images with regional inhomogeneities," in *18th International Conference on Pattern Recognition (ICPR'06)*, IEEE, vol. 3, 2006, pp. 976–979.

[37] D. K. Iakovidis, M. A. Savelonas, S. A. Karkanis, and D. E. Maroulis, "A genetically optimized level set approach to segmentation of thyroid ultrasound images," *Applied Intelligence*, vol. 27, no. 3, pp. 193–203, 2007.

[38] A. Illanes, N. Esmaeili, P. Poudel, S. Balakrishnan, and M. Friebe, "Parametrical modelling for texture characterization—a novel approach applied to ultrasound thyroid segmentation," *PloS one*, vol. 14, no. 1, e0211215, 2019.

[39] G. Jocher, A. Stoken, J. Borovec, NanoCode012, ChristopherSTAN, L. Changyu, Laughing, tkianai, A. Hogan, lorenzomammana, yxNONG, AlexWang1900, L. Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, F. Ingham, Frederik, Guilhen, Hatovix, J. Poznanski, J. Fang, L. Yu, changyu98, M. Wang, N. Gupta, O. Akhtar, PetrDvoracek, and P. Rai, *ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements*, version v3.1, Oct. 2020. DOI: 10.5281/zenodo.4154370. [Online]. Available: https://doi.org/10.5281/zenodo.4154370.

[40] E. G. Keramidas, D. K. Iakovidis, D. Maroulis, and S. Karkanis, "Efficient and effective ultrasound image analysis scheme for thyroid nodule detection," in *International Conference Image Analysis and Recognition*, Springer, 2007, pp. 1052–1060.

[41] E. G. Keramidas, D. Maroulis, and D. K. Iakovidis, "Tnd: A thyroid nodule detection system for analysis of ultrasound images and videos," *Journal of medical systems*, vol. 36, no. 3, pp. 1271–1281, 2012.

[42] V. P. Kharchenko, P. M. Kotlyarov, M. S. Mogutov, Y. K. Alexandrov, A. N. Sencha, Y. N. Patrunov, and D. V. Belyaev, *Ultrasound diagnostics of thyroid diseases*. Springer Science & Business Media, 2010.

[43] H. G. Kim, J. Y. Kwak, E.-K. Kim, S. H. Choi, and H. J. Moon, "Man to man training: Can it help improve the diagnostic performances and interobserver variabilities of thyroid ultrasonography in residents?" *European journal of radiology*, vol. 81, no. 3, e352–e356, 2012.

[44] S. H. Kim, C. S. Park, S. L. Jung, B. J. Kang, J. Y. Kim, J. J. Choi, Y. I. Kim, J. K. Oh, J. S. Oh, H. Kim, *et al.*, "Observer variability and the performance between faculties and residents: Us criteria for benign and malignant thyroid nodules," *Korean journal of radiology*, vol. 11, no. 2, pp. 149–155, 2010.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[46] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "Foveabox: Beyound anchor-based object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 7389–7398, 2020.

[47] D. Koundal, S. Gupta, and S. Singh, "Automated delineation of thyroid nodules in ultrasound images using spatial neutrosophic clustering and level set," *Applied Soft Computing*, vol. 40, pp. 86–97, 2016.

[48] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[49] I. Legakis, M. A. Savelonas, D. Maroulis, and D. K. Iakovidis, "Computer-based nodule malignancy risk assessment in thyroid ultrasound images," *International Journal of Computers and Applications*, vol. 33, no. 1, pp. 29–35, 2011.

[50] X. Li, S. Wang, X. Wei, J. Zhu, R. Yu, M. Zhao, M. Yu, Z. Liu, and S. Liu, "Fully convolutional networks for ultrasound image segmentation of thyroid nodules," in *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, IEEE, 2018, pp. 886–890.

[51] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[52] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the coordconv solution," *arXiv preprint arXiv:1807.03247*, 2018.

[53] T. Liu, Q. Guo, C. Lian, X. Ren, S. Liang, J. Yu, L. Niu, W. Sun, and D. Shen, "Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks," *Medical image analysis*, vol. 58, p. 101 555, 2019.

[54] 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.

[55] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[56] D.-Y. Liu, H.-L. Chen, B. Yang, X.-E. Lv, L.-N. Li, and J. Liu, "Design of an enhanced fuzzy k-nearest neighbor classifier based computer aided diagnostic system for thyroid disease," *Journal of medical systems*, vol. 36, no. 5, pp. 3243–3254, 2012.

[57] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1377–1385.

[58] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[59] 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.

[60] J. Ma, F. Wu, T. Jiang, Q. Zhao, and D. Kong, "Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks," *International journal of computer assisted radiology and surgery*, vol. 12, no. 11, pp. 1895–1910, 2017.

[61] R. Malladi, J. A. Sethian, and B. C. Vemuri, "Shape modeling with front propagation: A level set approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 17, no. 2, pp. 158–175, 1995.

[62] D. E. Maroulis, M. A. Savelonas, D. K. Iakovidis, S. A. Karkanis, and N. Dimitropoulos, "Variable background active contour model for computer-aided delineation of nodules in thyroid ultrasound images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 5, pp. 537–543, 2007.

[63] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, IEEE, 2016, pp. 565–571.

[64] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *arXiv preprint arXiv:2001.05566*, 2020.

[65] N. S. Narayan, P. Marziliano, J. Kanagalingam, and C. G. Hobbs, "Speckle patch similarity for echogenicity-based multiorgan segmentation in ultrasound images of the thyroid gland," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 172–183, 2015.

[66] H. A. Nugroho, A. Nugroho, and L. Choridah, "Thyroid nodule segmentation using active contour bilateral filtering on ultrasound images," in *2015 International Conference on Quality in Research (QiR)*, IEEE, 2015, pp. 43–46.

[67] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[68] M. Patrıcio, C. Oliveira, and F. Caseiro-Alves, "Differentiating malignant thyroid nodule with statistical classifiers based on demographic and ultrasound features," in *2017 IEEE 5th Portuguese Meeting on Bioengineering (ENBENG)*, IEEE, 2017, pp. 1–4.

[69] P. Poudel, A. Illanes, D. Sheet, and M. Friebe, "Evaluation of commonly used algorithms for thyroid ultrasound images segmentation and improvement using machine learning approaches," *Journal of healthcare engineering*, vol. 2018, 2018.

[70] A. Prochazka, S. Gulati, S. Holinka, and D. Smutek, "Patch-based classification of thyroid nodules in ultrasound images using direction independent features extracted by two-threshold binary decomposition," *Computerized Medical Imaging and Graphics*, vol. 71, pp. 9–18, 2019.

[71] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, p. 107 404, 2020.

[72] 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.

[73] U. Raghavendra, A. Gudigar, M. Maithri, A. Gertych, K. M. Meiburger, C. H. Yeong, C. Madla, P. Kongmebhol, F. Molinari, K. H. Ng, *et al.*, "Optimized multi-level elongated quinary patterns for the assessment of thyroid nodules in ultrasound images," *Computers in biology and medicine*, vol. 95, pp. 55–62, 2018.

[74] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[75] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[76] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[77] 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.

[78] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[79] M. Savelonas, D. Maroulis, D. Iakovidis, S. Karkanis, and N. Dimitropoulos, "A variable background active contour model for automatic detection of thyroid nodules in ultrasound images," in *IEEE International Conference on Image Processing 2005*, IEEE, vol. 1, 2005, pp. I–17.

[80] M. A. Savelonas, D. K. Iakovidis, I. Legakis, and D. Maroulis, "Active contours guided by echogenicity and texture for delineation of thyroid nodules in ultrasound images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 519–527, 2008.

[81] A. G. Schwing and R. Urtasun, "Fully connected deep structured networks," *arXiv preprint arXiv:1503.02351*, 2015.

[82] H. Seo, M. Badiei Khuzani, V. Vasudevan, C. Huang, H. Ren, R. Xiao, X. Jia, and L. Xing, "Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications," *Medical physics*, vol. 47, no. 5, e148–e167, 2020.

[83] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[84] J. H. Shin, J. H. Baek, J. Chung, E. J. Ha, J.-h. Kim, Y. H. Lee, H. K. Lim, W.-J. Moon, D. G. Na, J. S. Park, *et al.*, "Ultrasonography diagnosis and imaging-based management of thyroid nodules: Revised korean society of thyroid radiology consensus statement and recommendations," *Korean journal of radiology*, vol. 17, no. 3, pp. 370–395, 2016.

[85] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 1, pp. 7–34, 2019.

[86] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[87] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*, pp. 1–6, 2021.

[88] W. Song, S. Li, J. Liu, H. Qin, B. Zhang, S. Zhang, and A. Hao, "Multi-task cascade convolution neural networks for automatic thyroid nodule detection and recognition," *IEEE journal of biomedical and health informatics*, vol. 23, no. 3, pp. 1215–1224, 2018.

[89] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *arXiv preprint arXiv:1904.04514*, 2019.

[90] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *CVPR*, 2020, pp. 10 781–10 790.

[91] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 9627–9636.

[92] D. Topstad and J. A. Dickinson, "Thyroid cancer incidence in canada: A national cancer registry analysis," *CMAJ open*, vol. 5, no. 3, E612, 2017.

[93] S. Tsantis, D. Cavouras, I. Kalatzis, N. Piliouras, N. Dimitropoulos, and G. Nikiforidis, "Development of a support vector machine-based image analysis system for assessing the thyroid nodule malignancy risk on ultrasound," *Ultrasound in medicine & biology*, vol. 31, no. 11, pp. 1451–1459, 2005.

[94] S. Tsantis, N. Dimitropoulos, D. Cavouras, and G. Nikiforidis, "A hybrid multi-scale model for thyroid nodule boundary detection on ultrasound images," *Computer methods and programs in biomedicine*, vol. 84, no. 2-3, pp. 86–98, 2006.

[95] 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.

[96] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[97] S. Vaccarella, S. Franceschi, F. Bray, C. P. Wild, M. Plummer, L. Dal Maso, *et al.*, "Worldwide thyroid-cancer epidemic? the increasing impact of overdiagnosis," *N engl j med*, vol. 375, no. 7, pp. 614–617, 2016.

[98] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, IEEE, vol. 1, 2001, pp. I–I.

[99] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[100] L. Wang, S. Yang, S. Yang, C. Zhao, G. Tian, Y. Gao, Y. Chen, and Y. Lu, "Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the yolov2 neural network," *World journal of surgical oncology*, vol. 17, no. 1, pp. 1–9, 2019.

[101] M. J. Welker and D. Orlov, "Thyroid nodules," *American Family Physician*, vol. 67, no. 3, pp. 559–566, 2003.

[102] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.

[103] J. Xia, H. Chen, Q. Li, M. Zhou, L. Chen, Z. Cai, Y. Fang, and H. Zhou, "Ultrasound-based differentiation of malignant and benign thyroid nodules: An extreme learning machine approach," *Computer methods and programs in biomedicine*, vol. 147, pp. 37–49, 2017.

[104] X. Xia and B. Kulis, "W-net: A deep model for fully unsupervised image segmentation," *arXiv preprint arXiv:1711.08506*, 2017.

[105] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-unet for high-quality retina vessel segmentation," in *2018 9th international conference on information technology in medicine and education (ITME)*, IEEE, 2018, pp. 327–331.

[106] S. Xie, J. Yu, T. Liu, Q. Chang, L. Niu, and W. Sun, "Thyroid nodule detection in ultrasound images with convolutional neural networks," in *2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, IEEE, 2019, pp. 1442–1446.

[107] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9657–9666.

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

[109] Q. Yu, T. Jiang, A. Zhou, L. Zhang, C. Zhang, and P. Xu, "Computer-aided diagnosis of malignant or benign thyroid nodes based on ultrasound images," *European Archives of Oto-rhino-laryngology*, vol. 274, no. 7, pp. 2891–2897, 2017.

[110] B. Zhang, J. Tian, S. Pei, Y. Chen, X. He, Y. Dong, L. Zhang, X. Mo, W. Huang, S. Cong, *et al.*, "Machine learning–assisted system for thyroid nodule diagnosis," *Thyroid*, vol. 29, no. 6, pp. 858–867, 2019.

[111] L. Zhang, Y. Zhuang, Z. Hua, L. Han, C. Li, K. Chen, Y. Peng, and J. Lin, "Automated location of thyroid nodules in ultrasound images with improved yolov3 network," *Journal of X-Ray Science and Technology*, no. Preprint, pp. 1–16, 2020.

[112] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1529–1537.

[113] S. Zhou, H. Wu, J. Gong, T. Le, H. Wu, Q. Chen, and Z. Xu, "Mark-guided segmentation of ultrasonic thyroid nodules using deep learning," in *Proceedings of the 2nd International Symposium on Image Computing and Digital Medicine*, 2018, pp. 21–26.

[114] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 850–859.

[115] Z.-H. Zhou, "Ensemble learning.," *Encyclopedia of biometrics*, vol. 1, pp. 270–273, 2009.

[116] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, Springer, 2018, pp. 3–11.

[117] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*, Springer, 2014, pp. 391–405.

[118] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.