TUN-Det: A Novel Network for Thyroid Ultrasound Nodule Detection

Atefeh Shahroudnejad[1,2], Xuebin Qin[1,2(✉)], Sharanya Balachandran[1,2], Masood Dehghan[1,2], Dornoosh Zonoobi[2], Jacob Jaremko[1,2], Jeevesh Kapur[2,3], Martin Jagersand[1], Michelle Noga[1], and Kumaradevan Punithakumar[1]

[1] University of Alberta, Edmonton, Canada
xuebin@ualberta.ca
[2] Medo.ai, Alberta, Canada
[3] National University Hospital, Singapore, Singapore

Abstract. This paper presents a novel one-stage detection model, TUN-Det, for thyroid nodule detection from ultrasound scans. The main contributions are (i) introducing Residual U-blocks (RSU) to build the backbone of our TUN-Det, and (ii) 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, which plays an important role in detection of nodules with different sizes and appearances. 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. Experimental results conducted on 1268 thyroid nodules from 700 patients, show that our newly proposed RSU backbone and the multi-head architecture for classification and regression heads greatly improve the detection accuracy against the baseline model. Our TUN-Det also achieves very competitive results against the state-of-the-art models on overall Average Precision (AP) metric and outperforms them in terms of AP_{35} and AP_{50}, which indicates its promising performance in clinical applications. The code is available at: https://github.com/Medo-ai/TUN-Det.

Keywords: Thyroid nodule detection · Deep convolutional networks · Ultrasound image · Multi-scale features · Multi-head architecture

1 Introduction

Ultrasound (US) is the primary diagnostic tool for both the detection and characterization of thyroid nodules. 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. Hence, finding Region of Interest (ROI) of nodules for further processing becomes the preliminary step

A. Shahroudnejad and X. Qin—Equal contribution.

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12901, pp. 656–667, 2021.
https://doi.org/10.1007/978-3-030-87193-2_62

of the Computer-Aided Diagnosis (CAD) systems. In traditional CAD systems, the ROIs are manually defined by experts, which is time-consuming and highly relies on the experience of the radiologists and sonographers. Therefore, automatic thyroid nodule detection, which predicts the bounding boxes of thyroid nodules, from ultrasound images could play a very important role in computer aided thyroid cancer diagnosis [11,33].

Thyroid nodule detection in ultrasound images is an important yet challenging task in both medical image analysis and computer vision fields [4,18,26,29]. In the past decades, many traditional object detection approaches have been proposed [7,34,35,40], such as BING [5], EdgeBox [39] and Selective Search [32]. However, due to the large variations of the targets, there is still significant room for the improvements of traditional object detection approaches in terms of accuracy and robustness. In recent years, object detection has achieved great improvements by introducing machine learning and deep learning techniques. These methods can be mainly categorized into three groups: (**i**) two-stage models: such as RCNN [10], Fast-RCNN [9], Faster-RCNN [24], SPP-Net [12], R-FCN [6], Cascaded-RCNN [3] and so on; (**ii**) one-stage models: such as OverFeat [25], YOLO (v1, v2, v3, v4, v5) [1,2,21–23], SSD [19], RetinaNet [16] and so on; (**iii**) anchor-free models, such as CornerNet [15], CenterNet [8], ExtremeNet [38], RepPoints [37], FoveaBox [14] and FCOS [31]. As we know, the two-stage models are originally more accurate but less efficient than one-stage models. However, with the development of new losses, e.g. focal loss [16] and training strategies, one-stage models are now able to achieve comparable performance against two-stage models while requires less time costs. The anchor-free models relies on the object center or key points, which are relatively less accessible in ultrasound images.

Almost all of the above detection models are originally designed for object detection from natural images, which have different characteristics than ultrasound images. Particularly, ultrasound images have variable spatial resolution, heavy speckle noise, and multiple acoustic artifacts, which make the detection task challenging. In addition, thyroid nodules have diverse sizes, shapes and appearances. Sometimes, thyroid nodules are very similar to the thyroid tissue and are not defined by clear boundaries (e.g. ill-defined nodule). Some nodules are heterogeneous due to diffuse thyroid disease, which makes these nodules difficult to differentiate from each other and their backgrounds. In addition, 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. These 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, multi-scale features are very important. Therefore, we propose a novel one-stage thyroid nodule detection model, called *TUN-Det*, whose backbone is built upon the ReSidual U-blocks (RSU) [20], which is able to extract richer multi-scale features from feature maps with different resolutions. In addition, we design a multi-head architecture for both the nodule

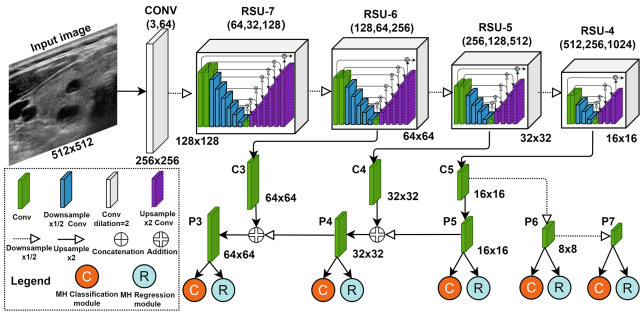


Fig. 1. Architecture of the proposed TUN-Det.

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, multi-head outputs are combined to achieve better detection performance. The Weighted Boxes Fusion (WBF) algorithm [28] is introduced to fuse the outputs of each multi-head module. In summary, our contributions are threefold: (i) a novel one-stage thyroid nodule detection network, TUN-Det, built upon the Residual U-blocks [20]; (ii) a novel multi-head architecture for both bounding boxes classification and regression heads, in which the ensemble strategy is embedded; (iii) Very competitive performance against the state-of-the-art models on our newly built thyroid nodule detection dataset.

2 Proposed Method

2.1 TUN-Det Architecture

Feature Pyramid Network (FPN) is one of the most popular architecture in object detection. Because 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 VGG [27], ResNet [13] 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[20]. To capture more multi-scale features from different stages, we build our TUN-Det upon the Residual U-blocks (RSU), which was first proposed in salient object detection U^2-Net [20]. Our proposed TUN-Det is also a one-stage FPN similar to RetinaNet [16].

Figure 1 illustrates the overall architecture of our newly proposed TUN-Det for 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 maps 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 convolution, the RSUs are able to capture both local and global information from feature maps with arbitrary resolutions[20]. Therefore, richer multi-scale features $\{C_3, C_4, C_5\}$ can be extracted by the backbone built upon these blocks for supporting the nodule detection. Then, an FPN [16] 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 regression and classification.

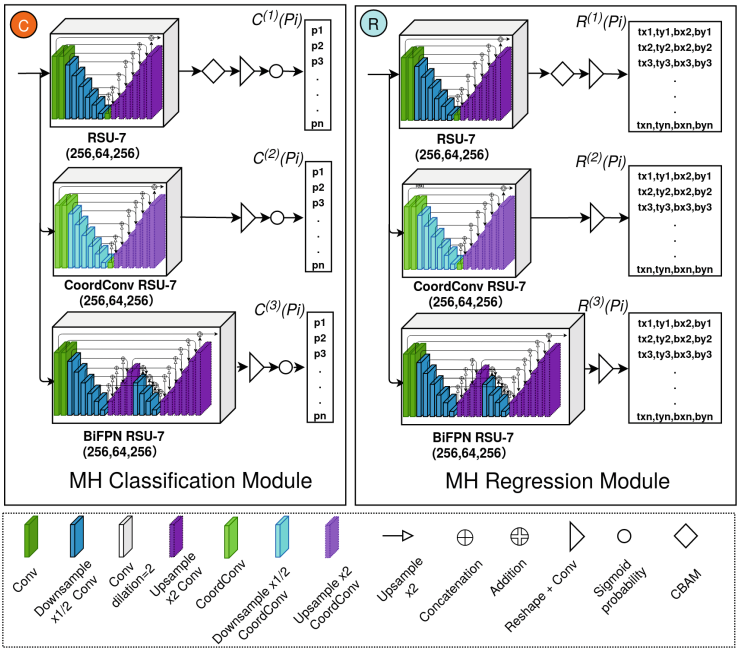


Fig. 2. Multi-head classification and regression module.

2.2 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 bounding boxes' coordinates 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 $\{BBOX_3, BBOX_4, \ldots, BBOX_7\}$ and the

classification outputs $\{CLAS_3, CLAS_4, \ldots, CLAS_7\}$ 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, multi-model ensemble strategy 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 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 [28] is used to fuse the regression and classification outputs of different heads. This design embeds the ensemble strategy into both the classification and regression module to improve the detection accuracy while avoiding training multiple models, which is a standard procedure in common ensemble methods.

In this paper, the architectures of $R^{(i)}$ and $C^{(i)}$ are the same except for the last convolution layer (see Fig. 2). To increase the diversity of the prediction results and hence reducing the variance, three variants of RSU-7 (**CBAM RSU-7**, **CoordConv RSU-7** and **BiFPN 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) [36] block is added after the standard RSU-7 block to refine features by channel (M_c) and spatial (M_s) attention. The formulation can be described as $F_c = M_c(F_{in}) \otimes F_{in}$ and $F_s = M_s(F_c) \otimes F_c$. The second head is **CoordConv RSU-7**, which replaces the plain convolution layers in the original RSU-7 by Coordinate Convolution [17] layers to encode geometric information. CoordConv can be described as $conv(concat(F_{in}, F_i, F_j))$, 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 bi-directional FPN (BiFPN) [30] 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 four-stage BiFPN layer to avoid complexity and reduce the number of trainable parameters.

2.3 Supervision

As shown in Fig. 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: $\mathcal{L} = \sum_{i=1}^{5} \alpha_i \mathcal{L}_i$, where α_i is the weight of each group (all α 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

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

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 [16] for classification. It can be defined as follows:

$$\begin{aligned} \mathcal{L}_i^{C^{(i)}} &= \text{Focal}(p_t) = \alpha_t (1 - p_t)^{\gamma} \times BCE(p_c, y_c), \\ p_t &= \begin{cases} p_c & \text{if } y_c = 1 \\ 1 - p_c & \text{otherwise} \end{cases}, \quad \alpha_t = \begin{cases} \alpha & \text{if } y_c = 1 \\ 1 - \alpha & \text{otherwise,} \end{cases} \end{aligned} \tag{2}$$

where p_c and y_c are predicted and target classes respectively. α and γ are focal weighting factor and focusing parameters that are set to 0.25 and 2.0, respectively. $\mathcal{L}_i^{R^{(j)}}$ is the Smooth-L1 loss [9] for regression, which is defined as:

$$\mathcal{L}_i^{R^{(j)}} = \text{Smooth-L1}(p_r, y_r) = \begin{cases} 0.5(\sigma x)^2 & \text{if } |x| < \frac{1}{\sigma^2} \\ |x| - \frac{0.5}{\sigma^2} & \text{otherwise,} \end{cases}, \quad x = p_r - y_r \tag{3}$$

where p_r and y_r are predicted and ground truth bounding boxes respectively. σ defines where the regression loss changes from L2 to L1 loss. It is set to 3.0 in our experiments.

3 Experimental Results

3.1 Datasets and Evaluation Metrics

To validate the performance of our newly proposed TUN-Det on ultrasound thyroid nodule detection task, we build 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 are extracted from 1924 transverse (TRX) and 2017 sagittal (SAG) scans. These images are split into three subsets for training (2534), validation (565) and testing (842) with 3554, 981, and 1268 labeled nodule bounding boxes, respectively. There is no common patient in the training, validation and testing sets. All nodule bounding boxes are manually labeled by 5 experienced sonographers (with $\geq$8 years of experience in thyroid sonography) and validated by 3 radiologists. To evaluate the performance of our TUN-Det against other models, Average Precision (AP) [16] is used as the evaluation metric. The validation set is only used to select the model weights in the training process. All the performance evaluation conducted in this paper is based on the testing set.

3.2 Implementation Details

Our proposed TUN-Det is 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 with default parameters is used to train the model. Both our training and testing process are 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 1080Ti GPU (11GB memory). The model converges after 200 epochs and takes 20 h in total. The average inference time per image (512×512) is 94 ms.

Table 1. Ablation on different backbones and heads configurations. AP_{35}, AP_{50}, AP_{75} are average precision at the fixed 35%, 50%, 75% IoU thresholds, respectively. AP is the average of AP computed over ten different IoU thresholds from 50% to 95% [AP_{50}, AP_{55}, $\cdots$, AP_{95}].

Model	AP	AP_{35}	AP_{50}	AP_{75}
RetinaNet w/ ResNet-50 backbone (baseline) [16]	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

3.3 Ablation Study

To validate the effectiveness of our proposed architecture, ablation studies are conducted on different configurations and the results are summarized in Table 1. 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}.

3.4 Comparisons Against State-of-the-Arts

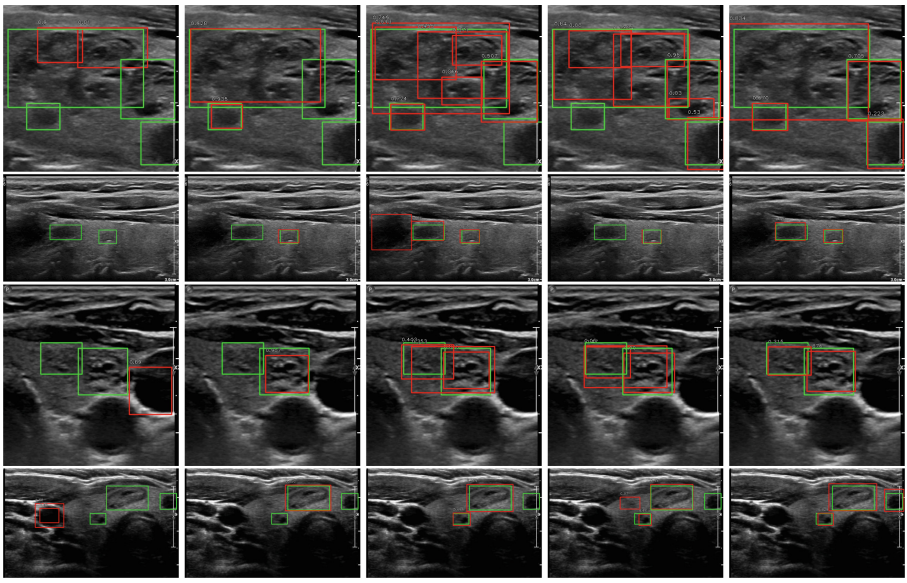
Quantitative Comparisons. To evaluate the performance of our newly proposed TUN-Det, we compare our model against six typical state-of-the-art detection models including (i) Faster-RCNN [24] as a two-stage model; (ii) RetinaNet

[16], SSD [19], YOLO-v4 [2] and YOLO-v5 [1] as one stage models; and (iii) FCOS [31] as an anchor-free model. As shown in Table 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 Although our model is inferior in terms of AP_{75}, it is doing a better job in terms of FN (i.e. our Average Recall at 75%, AR_{75}, is 45.5 vs. 40.3 in YOLO-v5), which is a priority in the context of thyroid nodule detection to not missing any nodules. Having low Recall with high Precision is unacceptable as it would miss many cancers. Regarding AP, it is usually reported to show the average performance. However, in practice we seek a threshold for achieving final detection results in real-world clinical applications. According to the experiments, our model achieves the best performance under different IoU thresholds (e.g. 35%, 50%), which means our model is more applicable to clinical workflow.

Table 2. Comparisons against the state-of-the-arts.

Model	Backbone	AP	AP_{35}	AP_{50}	AP_{75}
Faster-RCNN [24]	VGG16	0.91	42.13	29.65	2.58
SSD [19]	VGG16	19.05	40.10	36.55	18.10
FCOS [31]	ResNet-50	33.15	62.74	58.67	32.44
RetinaNet [16]	ResNet-50	39.50	74.03	69.07	41.39
YOLO-v4 [2]	CSPDarknet-53	40.43	78.21	72.48	42.04
YOLO-v5 [1]	CSPNet	**45.19**	78.71	74.74	**50.90**
TUN-Det (ours)	RSU	42.75	**81.22**	**75.66**	45.53

Qualitative Comparisons. Figure 3 shows the qualitative comparison of our TUN-Det with other SOTA models on sampled sagittal scans (first two rows) and transverse scans (last two rows). Each column shows the result of one method. The ground truth is shown with green and detection result is shown in red. Figure 3 (1st row) shows that TUN-Det can correctly detect the challenging case of a non-homogeneous large hypo-echoic nodule, while all other methods fail. The 2nd row illustrate that TUN-Det performs well in detecting nodules with ill-defined boundaries, while others miss them. The 3rd and 4th rows highlight that our TUN-Det successfully excludes the false positive and false negative nodules. The last column of Fig. 3 signifies that our TUN-Det produces the most accurate nodule detection results.



(a) Faster-RCNN (b) FCOS (c) RetinaNet (d) YOLOv5 (e) Our TUN-Det

Fig. 3. Qualitative comparison of ground truth (green) and detection results (red) for different methods. Each column shows the result of one method. (Color figure online)

4 Conclusion and Discussion

This paper proposes a novel detection network, TUN-Det. The novel backbone, built upon the RSU blocks, of our TUN-Det greatly improves the detection accuracy by extracting richer multi-scale features from feature maps with different resolutions. The newly proposed multi-head architecture for both classification and regression heads further improves the nodule detection performance by fusing outputs from diversified sub-modules. Experimental results show that our TUN-Det achieves very competitive performance against existing detection models on overall AP and outperforms other models in terms of AP_{35} and AP_{50}, which indicates its promising performance in practical applications. We believe that this architecture is also promising for other detection tasks on ultrasound images. 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.

References

1. Ultralytics/yolov5. https://github.com/ultralytics/yolov5. Accessed Oct 2020
2. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOv4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)

3. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–62 (2018)
4. Chen, J., You, H., Li, K.: A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images. Comput. Methods Programs Biomed. **185**, 105329 (2020)
5. Cheng, M.M., Zhang, Z., Lin, W.Y., Torr, P.: Bing: Binarized normed gradients for objectness estimation at 300fps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3286–3293 (2014)
6. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: Object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems, pp. 379–387 (2016)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
8. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: CenterNet: keypoint triplets for object detection. In: Proceedings of the IEEE international Conference on Computer Vision, pp. 6569–6578 (2019)
9. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
10. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
11. Haugen, B.R., et al.: 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer. Thyroid **26**(1), 1–133 (2016)
12. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE Computer Society (2016)
14. Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., Shi, J.: FoveaBox: beyond anchor-based object detection. IEEE Trans. Image Process. **29**, 7389–7398 (2020)
15. Law, H., Deng, J.: CornerNet: detecting objects as paired keypoints. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 765–781. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_45
16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
17. Liu, R., et al.: An intriguing failing of convolutional neural networks and the CoordConv solution. arXiv preprint arXiv:1807.03247 (2018)
18. Liu, T., et al.: Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks. Med. Image Anal. **58**, 101555 (2019)
19. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2

20. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O., Jagersand, M.: U2-net: going deeper with nested U-structure for salient object detection, vol. 106, p. 107404 (2020)
21. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
22. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
23. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
25. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: OverFeat: integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
26. Sharifi, Y., Bakhshali, M.A., Dehghani, T., DanaiAshgzari, M., Sargolzaei, M., Eslami, S.: Deep learning on ultrasound images of thyroid nodules. Biocybern. Biomed. Eng. (2021)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: ensembling boxes from different object detection models. Image Vis. Comput. **107**, 1–6 (2021)
29. Song, W., et al.: Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition. IEEE J. Biomed. Health Inform. **23**(3), 1215–1224 (2018)
30. Tan, M., Pang, R., Le, Q.V.: EfficientDet: scalable and efficient object detection. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition, pp. 10781–10790 (2020)
31. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9627–9636 (2019)
32. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. Int. J. Comput. Vis. **104**(2), 154–171 (2013)
33. Vaccarella, S., Franceschi, S., Bray, F., Wild, C.P., Plummer, M., Dal Maso, L., et al.: Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis. N. Engl. J. Med. **375**(7), 614–617 (2016)
34. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, p. I. IEEE (2001)
35. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
36. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
37. Yang, Z., Liu, S., Hu, H., Wang, L., Lin, S.: RepPoints: point set representation for object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9657–9666 (2019)

38. Zhou, X., Zhuo, J., Krahenbuhl, P.: Bottom-up object detection by grouping extreme and center points. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 850–859 (2019)
39. Zitnick, C.L., Dollár, P.: Edge boxes: locating object proposals from edges. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 391–405. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_26
40. Zou, Z., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: a survey. arXiv preprint arXiv:1905.05055 (2019)