IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 January 2024, accepted 23 January 2024, date of publication 9 February 2024, date of current version 15 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3359367

RESEARCH ARTICLE

Automatic Thyroid Nodule Detection in Ultrasound Imaging With Improved YOLOv5 Neural Network

DAQING YANG[1], JIANFU XIA[1], RIZENG LI[1], WENCAI LI[1], JISHENG LIU[1], RONGJIAN WANG[1], DONG QU[2], AND JIE YOU[3]

[1]Department of General Surgery, The Second Affiliated Hospital of Shanghai University (Wenzhou Central Hospital), Wenzhou 325000, China
[2]Department of Radiology, The Second Affiliated Hospital of Shanghai University (Wenzhou Central Hospital), Wenzhou 325000, China
[3]Department of Thyroid Surgery, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou 325000, China

Corresponding authors: Dong Qu (qu-dong@sohu.com) and Jie You (youjie@wmu.edu.cn)

This work was supported by the Basic Research Project of Wenzhou City under Grant Y20220893.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Medical Ethics Committee of Wenzhou Central Hospital under Application No. 20231224182900029061.

ABSTRACT With the incidence of thyroid cancer increasing dramatically, the burden of sonographic diagnosis lies heavy for radiologists. An automatic computer-aided diagnosis system with both precision and efficiency is in demand. This retrospective study included 191 ultrasound images of 171 patients (85 benign and 86 malignant) in Wenzhou Central Hospital. An improved You Only Look Once version 5 neural network (improved YOLOv5) is proposed in this work. It comprises the coordinate attention (CA) module and the label smoothing regularization (LSR) module, in which the CA module enables the network ability of positional information extraction. The improved neural network correctly recognizes the lesion area and nodule type with a mean average precision (mAP) of 95.3% in 8.4 ms on the test set. The ablation experiment demonstrates that the integration of the CA and the LSR module cost 1.3 ms extra inference time per image in exchange for raising the mAP by 4.4%. Afterward, 10 ultrasound images with wrong nodule types are added to the dataset for training, the result shows that the LSR module can significantly prevent the network from being misled by the bug data. Compared with other state-of-the-art networks, the improved network has superiority in both precision and robustness for diagnosing benign/malignant thyroid nodules with only a small dataset. The proposed network also has the potential to transfer to other sonographic diagnosis tasks.

INDEX TERMS Computer-aided diagnosis, thyroid nodules, ultrasound, YOLOv5 network, attention module, label smoothing.

I. INTRODUCTION

Thyroid nodules are common tumors in adults, among which women are about 3 times as likely as men to be diagnosed with thyroid cancer [1]. In the past three decades, the incidence of thyroid cancer has tripled or more in several high-income countries [2], [3]. Thyroid nodules are the early manifestations of thyroid cancer. Accurate screening of benign and malignant nodules is of great significance in improving the survival rate of thyroid cancer patients.

Ultrasonography is a primary, non-invasive, non-radioactive and inexpensive technique for the screening of thyroid nodules. The existing difficulties in ultrasound thyroid diagnosis are listed as follows: 1) the usual size of thyroid nodules is small and has vague margins; 2) the diagnosis is relatively subjective and mainly depends on the knowledge and experience of radiologists. It's difficult to accurately judge complicated thyroid nodules. Furthermore, the ever-growing number of patients will greatly increase the

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

workload of radiologists, thereby leading to more misdiagnosis [4]; 3) A fine needle aspiration biopsy (FNAB) would be applied in clinical practice, for secondary identification of suspected malignant nodules in ultrasound nodule images [5]. But FNAB is expensive and has the risk of causing nerve and blood vessel damage. Recent studies also indicate that about 10∼30% of nodules remain indeterminate after FNAB [6]. Therefore, an accurate and efficient automated thyroid ultrasound image diagnosis tool is needed urgently, with which the unnecessary FNAB and misdiagnosis would be efficiently reduced.

Deep learning technology is considered a promising alternative to manual diagnosis [7], [8], [9], [10], [11], [12]. In recent years, great advances are made in computer vision with convolutional neural networks (CNN) [13], [14], [15]. The texture extraction ability of CNN has exceeded human eyes. Ultrasound image diagnosis is a typical object detection task in the computer vision field. Traditionally, object detection task would be performed with two different models, which is the texture extraction model and the classification model. In 2017, Liu et al. [16] trained a VGG-F [17] net as a feature extraction model and the feature was then fed to a support vector machine (SVM) for nodule classification. Many researchers applied this kind of two-model-combined solution [18], [19]. But the gradient of the feature extraction model and the classification model can not be descended at the same time in the architecture. The whole training routine is quite complicated and time-consuming. The situation is improved with the proposal of the region-based CNN (R-CNN) series [20], which comprises two stages in one model. In the first stage, thousands of regions are proposed, where there might be an object. And the second stage predicts the class of the object and refines the bounding box. Li et al. [21] proposed a deep-learning approach for thyroid papillary cancer detection. By using layer con-catenation in Faster R-CNN [22], more detailed features of low-resolution images were extracted. Abdolali et al. [23] utilized a deep-learning framework based on the multi-task model Mask R-CNN. Together with their modified loss function, the model can reach a high mean average precision (mAP) of 85%. A mask of the lesion area can be generated at the pixel lever. The overall performance of the R-CNN series networks in the detection and classification of thyroid nodules is better than most of the other networks, but the use of R-CNN in the multi-scale detection task is a double-edged sword. Although it possesses high detection accuracy, the training and inference speed is limited due to the two-stage structure.

You only look once (YOLO) neural network series, first proposed by Joseph Red-mon and Ali Farhadi [24], [25], [26], is one of the state-of-art deep learning architectures. As a one-stage neural network, YOLO keeps a good balance between precision and efficiency. Based on YOLOv3, Ma et al. [27] proposed the YOLOv3-DMRF network, which comprised dense multi-receptive fields CNN and multiscale detection layers. The model can extract the edge and texture features of the thyroid nodules in different sizes. The experiments show that the network can correctly recognize the nodules with an mAP of 95.23% in 2.2 seconds.

Most machine learning methods require large-scale datasets, which are difficult to obtain for medical image tasks. In this paper, we propose a lightweight and data-efficient network for thyroid nodule detection from ultrasound images. Based on YOLOv5, we introduce the Label smoothing regularization (LSR) module to alleviate over-fitting and improve robustness when training on small datasets. We also incorporate the coordinate attention (CA) module into the network to enhance the texture and positional information extraction ability. The improved network sacrifices a small amount of inference speed (due to the increased model size) for higher accuracy and robustness, which is very suitable for medical image tasks.

The remainder of this paper is organized as follows: Section II introduces the detail of the workflow, including the patient data preparation, the network construction and the training settings. In section III, the experiment results are reported, this performance of the proposed method is compared with other state-of-the-art (SOTA) models. Section IV presents the discussion and provides a detailed analysis of the pros and cons of the network. Finally, the research is summarized in section V.

II. MATERIALS AND METHODS

The task of this study can be generally divided into 2 parts: the model training part and the model evaluation part. In the training part, as shown in Fig.1, the raw ultrasound images are collected and preprocessed as a dataset. We scale all the images to a size of 640 $\times$ 640$\times$3 (640 denotes the height and width of the image in pixel size, and 3 denotes the R, G, B tunnels). The dataset is, then, enlarged by flipping, rotating and cropping the images, which is called data augmentation. The model is trained with those augmented datasets to minimize its loss function. In the evaluation part, the ultrasound images are directly scaled and fed to the trained model and output with multi-head detection results. Some general post-processing will help make the output results more readable but won't go into detail in the paper. Below we will discuss the main steps of the proposed framework.

A. PATIENTS DATA PREPARING

Before training the model, a balanced sample of benign and malignant nodules is collected. The patient data includes 97 benign nodules from 85 patients (56 women and 29 men with average age of 51.12$\pm$12.19) and 97 malignant nodules from 86 patients (64 women and 22 men with average age of 48.18$\pm$11.28). All 171 patients were hospitalized at Wenzhou Central Hospital during the period of January 2011 to December 2014 and received ultrasonography. For those, whose nodules were highly suspected of malignancy ultrasonography-aided FNAB was conducted.

Ultrasound examination was performed by Acuson Sequoia 512 (Siemens Medical Solutions, MouView,

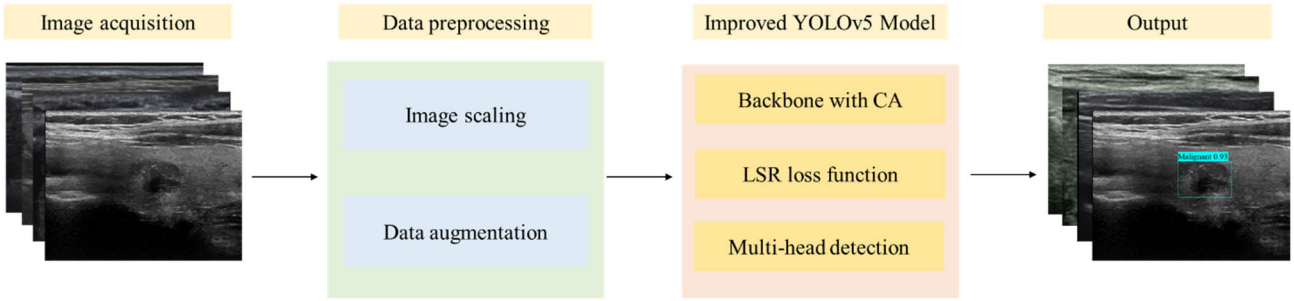

FIGURE 1. The model training workflow for the improved YOLOv5.

TABLE 1. Summary of demographic feature.

	Pathological findings	
	Benign	Malignant
No. of patients	85	86
Age, years, $\bar{x} \pm s$	51.22±12.09	48.18±11.28
Sex		
Male	29 (34.12%)	22 (25.58%)
Female	56 (65.88%)	64 (74.42%)
No. of nodules	97	97
Size, mm	21.24±12.00	9.73±4.23
< 5.0	7 (7.22%)	9 (9.28%)
5.0 ~ 10.0	17 (17.53%)	45 (46.39%)
10.0 ~ 20.0	26 (26.80%)	38 (39.18%)
⩾ 20.0	47 (48.45%)	5 (5.15%)

CA) and 128XP sonographic scanners (Siemens Medical Solutions, MouView, CA) with linear probes at the frequency of 10-12 MHz. The ultrasound images were then reviewed by two senior radiologists (with 20 and 6 years of sonography experience, respectively) independently, without being informed of the clinical information of the patients. A consensus on the type and location of nodules had been reached between the two radiologists. Finally, the ultrasound images were labeled by a radiologist with Labelme (a graphical image annotation tool) [28] for network training. There are 191 annotated ultrasound images in the dataset, each containing at least one thyroid nodule. The dataset is shuffled randomly and divided into three sets: training, validation and test. The training set contains 113 images (58 benign and 55 malignant), the validation set contains 42 images (19 benign and 23 malignant), and the test set contains 36 images (20 benign and 16 malignant). No data whatsoever from the training set is mixed into the test or validation sets.

To be noticed, all the benign lesions are nodular hyperplasias, and all the malignant tumors are papillary carcinoma. Table 1. summarizes the demographic features of the patient data.

B. NETWORK ARCHITECTURE

You only look once version 5 (YOLOv5) neural network [29] is a state-of-art single-stage object detector. The network size of YOLOv5 is only about a quarter of YOLOv3. Therefore, it's easy to train and predict, and capable of real-time tasks. In this study, an improved YOLOv5 network is proposed, in which the LSR module [30] and the CA module [31] are integrated to strengthen the robustness and positional extraction ability of the network respectively. The improved network structure is shown in Fig.2.

The YOLOv5 network consists of three parts: backbone network, neck network and head network. In the inference process, an image with the size of 640 × 640×3 is put into the net. The backbone part would extract the texture feature of the image into feature maps. Feature maps of three different sizes are generated in order to extract textures in different granularities. The neck of YOLOv5 adopts the feature pyramid network (FPN) [32] structure. The feature maps of different sizes are merged and enhanced in the FPN. Finally, feature maps with both high-resolution and strong semantics are generated and fed to the YOLO head for object detection. More detail about YOLOv5 can be found in [29].

Although YOLOv5 is enabled with quite efficient texture extraction capabilities, the positional information is neglected. We apply a novel attention module, which is called the CA module. The CA module factorizes channel attention into two feature vectors by x-average pooling and y-average pooling function to extract the long-range dependencies and precise positional information, respectively. The feature vectors are encoded separately into attention maps and then merged together to augment the representations of the objects of interest. In this research, the CA module is plugged into the BCSP module as a sub-residual part. The attention map is concatenated with the texture feature map. To be noticed, following the bottleneck setup of the original YOLOv5 network, the CA module used in the backbone net is slightly different from that in the neck net (CAM1 and CAM2 block in Fig.2).

Apart from that, the loss function set in the improved YOLOv5 is a highlighted design. The original loss function is given by:

$$L = \lambda_1 L_{bbax} + \lambda_2 L_{obj} + \lambda_3 L_{cls} \tag{1}$$

where, L_{bbax} indicates the loss between the ground truth bounding box and the predicted box, L_{obj} is the object loss training the net to tell whether there is an object in the bounding box, and L_{cls} is the classification loss, which illustrates how close the prediction is to the true class, L_n is the weight used to prioritize each loss term.

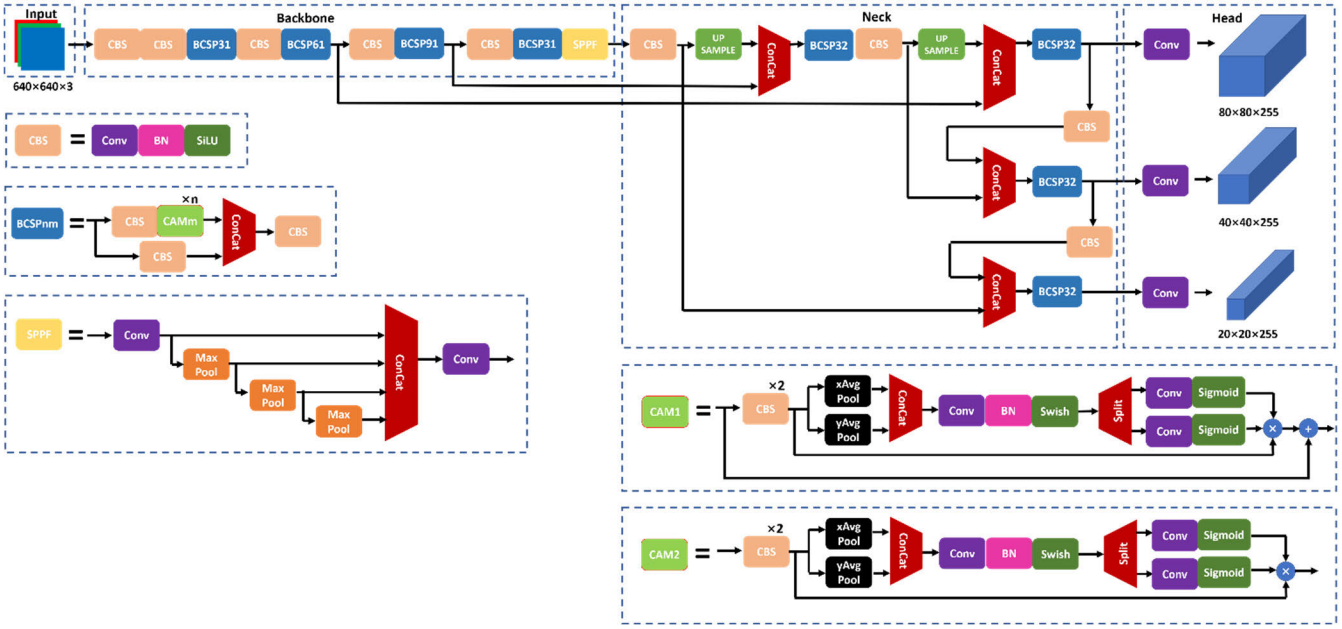

FIGURE 2. Improved YOLOv5 network architecture.

TABLE 2. Parameter setups for improved YOLOv5 network.

Parameters	Value
No. of anchor boxes	9
Batch size	12
Size of input image	640×640×3
Mosaic augmentation	True
No. of epochs	500
Early stop epoch (epoch without improvemrnt)	100
$[\lambda_1, \lambda_2, \lambda_3]$	[0.05, 1.0, 0.5]
Loss calculator of bounding box	CIoU [35]
Optimizer type	SGD
Initial learning rate	0.01
SGA momentum	0.937
Learning rate schedule	Cosine
Train BN layer	True
Pre-trained model	True
Soft factor ξ	0.05

Traditionally, the classification loss is calculated with the so-called "hard label". The true probability distribution (TPD) of each class is:

$$P_i = \begin{cases} 1, & i = y \\ 0, & i \neq y \end{cases} \tag{2}$$

so the class of an object should and only should either be this or that. To minimize the loss value of the classification term, the ideal probability distribution (IPD) for the network would be:

$$Z_i = \begin{cases} +\infty, & i = y \\ 0, & i \neq y \end{cases} \tag{3}$$

the training process would drive the network to fit the IPD as close as possible, which may lead to over-fitting.

TABLE 3. The ablation experiment result of YOLOv5 network.

Model	Class	mAP_{50}	Precision	Recall	Speed
Base	Benign	92.5	94.1	79.7	7.1 ms
	Malignant	89.4	80.9	90.9	
	All	**90.9**	87.5	85.3	
+LSR	Benign	94.3	83.5	95.0	7.1 ms
	Malignant	92.3	89.9	81.3	
	All	**93.3**	86.7	88.1	
+CAM	Benign	95.9	83.3	90.0	8.4 ms
	Malignant	89.7	90.1	86.4	
	All	**92.8**	86.7	88.2	
+LSR CAM	Benign	96.9	84.9	95.0	8.4 ms
	Malignant	98.3	95.2	90.6	
	All	**95.3**	90.1	92.8	

What's more, the training process would be sensitive, and the network can be misled by bug data, which is very common in practice.

The main idea of LSR is to soften the label of classes with a soft factor ξ. By adopting LSR, the TPD of each class transforms to:

$$P_i = \begin{cases} 1 - \xi, & i = y \\ \dfrac{\xi}{K-1}, & i \neq y \end{cases} \tag{4}$$

where K is the number of classes in the detection task. If we choose the cross-entropy function as the classification loss, the formula transforms:

$$L_{cls} = -\sum_{i=1}^{K} P_i \log q_i \Rightarrow L_i = \begin{cases} (1-\xi) \times L_{cls}, & i = y \\ \xi \times L_{cls}, & i \neq y \end{cases} \tag{5}$$

TABLE 4. Comparation of performance between the base model and the model with LSR module when 10 bug data are added into the training dataset. The values in the brackets denote the change with respect to Table 3.

Model	mAP_{50}	Precision	Recall
Base	87.3(-3.6)	72.1(-15.4)	92.1(+6.8)
+LSR	92.0(-1.3)	76.5(-10.2)	91.3(+3.2)

TABLE 5. Comparison of our method with SOTA object detection models.

Model	mAP_{50}	Precision	Recall	Speed
Improved YOLOv5 Neural Network	**95.3**	**90.1**	**92.8**	**8.4 ms**
Faster R-CNN	83.2	76.3	72.9	10.3 ms
Mask R-CNN	85.2	77.3	73.1	8.9 ms
SSD	84.8	**90.9**	82.8	7.6 ms
RetinaNet	80.1	80.2	78.2	8.1 ms

It's a very simple but useful technique to improve the robustness of the network. By revising the TPD, the probability distribution of the ideal model is considerably softened, which is given by:

$$Z_i = \begin{cases} \log \dfrac{(K-1)(1-\xi)}{\xi+\alpha}, & i = y \\ \alpha, & i \neq y \end{cases} \tag{6}$$

where is α a real number. Compared with formula (3), the new IPD have finite boundary and is much easier to train.

C. TRAINING SETUP

The training process is performed on an NVIDIA 3070 8GB GPU with an i7-10700F CPU. The batch size is set to 12 due to the small data size. Besides, the data augmentation technique is applied in the training to enrich the dataset, the images are rotated, scaled, cropped and stitched randomly. We use a stochastic gradient descent (SGD) optimizer with a cosine learning rate decay [33]. The initial learning rate is set to 0.01. An early stop is adopted, and the network is trained for 500 epochs. The model is pre-trained on the COCO [34] dataset to guarantee its generalization. The details configurations of the network are summarized in Table 2.

III. RESULTS

In this experiment, we use precision, recall, and mAP as evaluation metrics for our task. These metrics are widely used and appropriate for our task because they reflect the performance of our model across different classes of thyroid nodules, as well as the reliability and sensitivity of our model in detecting nodules of different sizes and shapes.

Precision indicates the ability to correctly identify the object of the desired class, which is given by:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

And recall indicates the ability to identify all the objects of the desired class without missing one. It can be considered as the omission rate of the model, which is given by:

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

where TP, FP and FN represent true positive, false positive and false negative respectively. True and false correspond to the inference result of the model being correct or wrong compared with the ground truth, while positive and negative represents whether the inference result of the model is the desired class or the others.

Furthermore, the precision and recall value can be considered as functions of the confidence threshold, whose value can be manually set from 0 to 1. A high confidence threshold value leads to less positive results from the model, so the precision will be high while the recall value is lower, and vice versa. By varying the confidence threshold, the Precision-Recall (PR) curve can be drawn. The area under the PR curve is called the average precision (AP), and the mAP is the average AP of each class.

Another sub-metric that would affect the value of precision and recall is the intersection over union (IoU) threshold. The definition of IoU is given by:

$$IOU = \frac{A_p \cap A_{gt}}{A_p \cup A_{gt}} \tag{9}$$

where, A_p denotes the area of the predicted bounding box from the model, A_{gt} represents the area of the ground truth bounding box. The IoU sub-metric indicates the similarity between the predicted bounding box and the ground truth. Following the Pascal VOC challenge [36] standard, we set it to 0.5 as a constant.

An ablation experiment is, firstly, carried out. The original YOLOv5 model is set as the base model, the LSR module and the CA module are plugged into the network in sequence. Table 3 summarizes the performance of each YOLOv5 model on the test set, the evaluation metrics includes the mAP at the IoU threshold of 0.5, precision and recall rate at the confidence threshold of 0.5 and detection time per image. In this experiment, the improved YOLOv5 network with the LSR and the CA modules achieves the highest average precision of 90.1% and the highest mAP of 95.3%. To be noticed, the LSR module only affects the loss calculation in the training process. It takes no extra cost in the inference process the PR-curve of each model is drawn in Fig. 3, the area under the PR-curve represents the mAP value of each model. The curve of the model with the LSR and the CA modules locates at the top of the figure.

The detection results of an ultrasound image of a malignant nodule are shown in Fig. 4. Fig. 4(a) is the ground truth, and Fig. 4(b)-(f) denotes the detection results of the four YOLOv5

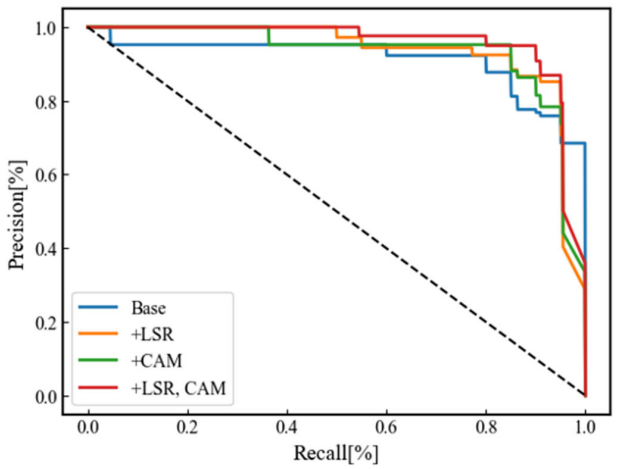

FIGURE 3. PR-curve of each YOLOv5 model in the ablation experiment.

models in Table 3. In Fig. 4(e), we collect the attention maps of the CA module at the entrance of the YOLO head. The attention maps are uniformed and visualized into a heatmap mask, which covers on the detection result.

Furthermore, we perform another test to prove the advantage of applying the LSR module by adding 10 ultrasound images (5 benign and 5 malignant) with wrong thyroid nodule types in the dataset. The soft factor ξ is still set to 0.05 in the test, and the training epoch was 500. Table 4 summarizes the performance of models with and without LSR module. The mAP of the base model decreases by about 3.6%, while the model with the LSR module only falls by 1.3%. The classification losses, as a function of the training epoch, are drawn in Fig. 5. The lines in dash represents the losses in validation, while the solid lines are the losses in training.

Finally, we conduct additional to compare our method with several state-of-the-art (SOTA) object detection models, including Faster R-CNN [37], Mask R-CNN [38], SSD [39], and RetinaNet [40]. We use the same dataset and evaluation metrics as in our original paper. The results are shown in Table 5. As can be seen from Table 5, our method is superior to the SOTA object detection model in terms of mAP and recall, but slightly lower in accuracy than the SSD.

IV. DISCUSSION

The concept of computer-aided diagnosis (CAD) for thyroid nodules was first suggested by Lim et al. [41] in 2008. The application of deep-learning has greatly improved the performance of CAD. The goal of this research is to propose a CAD tool for thyroid nodules on ultrasonography with both precision and efficiency. In Table 3, we compare the detection performance of the original YOLOv5 network and other networks with LSR and/or CA modules plugged in. The mAP and precision of the original network have lower scores on malignant than benign nodules, mostly because that the YOLOv5 network can hardly handle small targets, as shown in Fig. 4(b). The average size of the malignant nodules collected in this research is much smaller than that of the benign nodules. The application of the LSR and the CA modules lead to a considerable improvement in nearly all the metrics with respect to the original model. Fig.4 is a typical case in the experiment, comparing Fig.4(c)-(e), the LSR module will naturally lower the confidence level of detection, while the CA module boosts it. What's more, although the CA module costs 1.3 ms extra inference time, the predicted bounding box is much closer to the ground truth. The improved YOLOv5 network achieves a high average detection precision of 90.1% in the experiment, while the average diagnostics accuracy of FNAB is approximately 83%.

However, we recognize that the average size of the malignant nodules is less than half of the benign nodules' in our dataset. The imbalanced size distribution may introduce biases in the model. To address this concern, we selected extra 20 images of benign nodules with a diameter of less than 10 mm and 20 images of malignant nodules with a diameter of more than 10 mm. The prediction accuracy of our model achieved 90% on the large malignant nodules and 85% on the small benign nodules. Compared with the precision results in Table 3, the results indicate that the size of the nodules has a limited affection on the classification performance of our model. The proposed model should be further improved by collecting balanced nodule size data.

Considering a large database situation, it's evitable to have wrong labels in actual clinical practice, which will lower the precision of the network. Therefore, it's essential to apply the LSR module in the training process. By applying the LSR module, the network tends to learn the distribution characteristic of the dataset rather than fit specific labels. It provides robustness in the practical training process that radiologists don't need to worry about the flaws in the dataset. Moreover, the LSR efficiently improves the over-fitting in the training process. In Fig.5, the validation loss of the original model fluctuates wildly during the second half training process, which implies that the model has been misled by the bug data so that it can hardly tell the difference between benign and malignant nodules. While the model with the LSR model has a relatively smaller difference between training and validation loss.

With the proposed model, it's possible to diagnose hundreds of ultrasound images with high accuracy in just a few seconds. Most of the image cases, which have low malignant alert threshold can be just passed and defined as safe cases so that the workload of the radiologists can be greatly alleviated. Therefore, more effort can be put into the cases with risk, resulting in fair medical resource allocation. To sum up, our study can benefit both doctors and patients by providing a fast and reliable CAD tool for thyroid nodule screening and diagnosis. Combined with transfer learning techniques [42], the proposed model has the potential for other medical image detection tasks with just a small dataset collected.

Apart from that, there still are some shortcomings in this study: 1) Due to the limited dataset, the over-fitting problem is severe, despite that we utilize a pre-trained model and data augmentation strategy. The dataset shall be enriched in future work and more effective strategies should be applied to

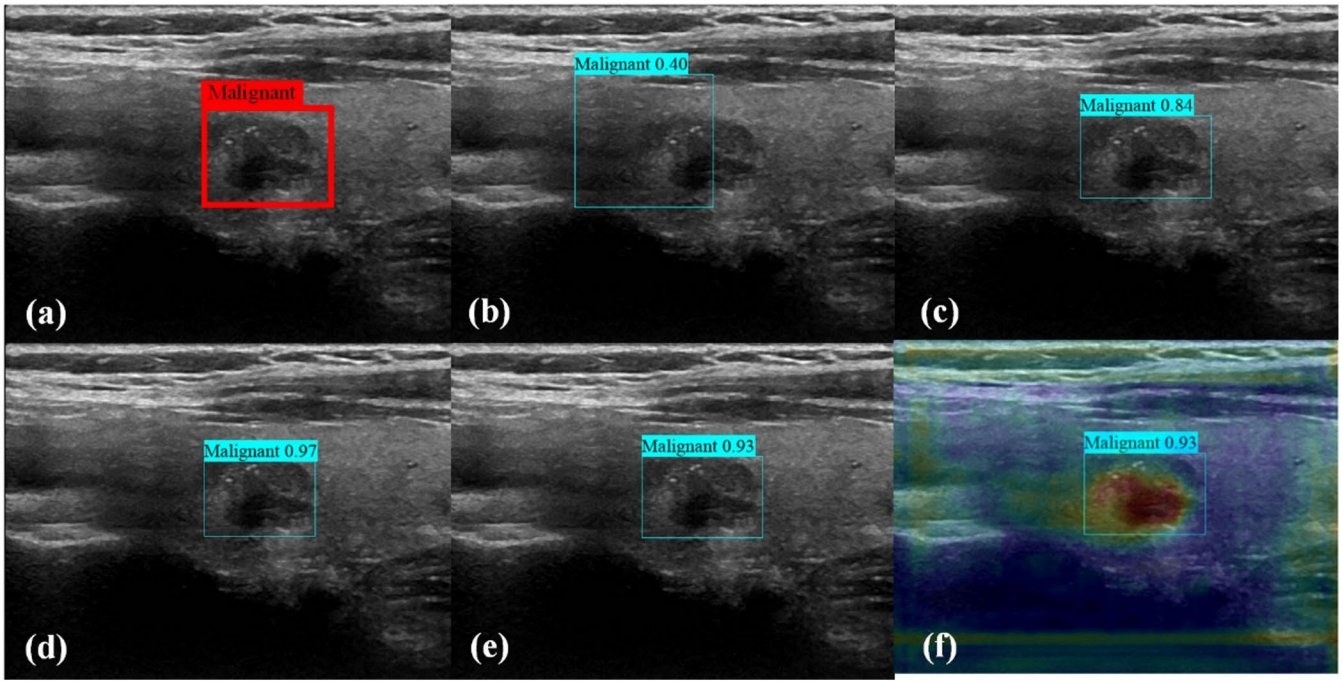

FIGURE 4. A detection example of malignant nodule image from different models: (a) the ground truth; (b) the base model prediction; (c) the LSR module added; (d) the CA module added; (e) added both the LSR and the CA module; (f) result(e) with the attention map mask on it. The flags above the bounding boxes represent the nodule type and confidence level results predicted by the models.

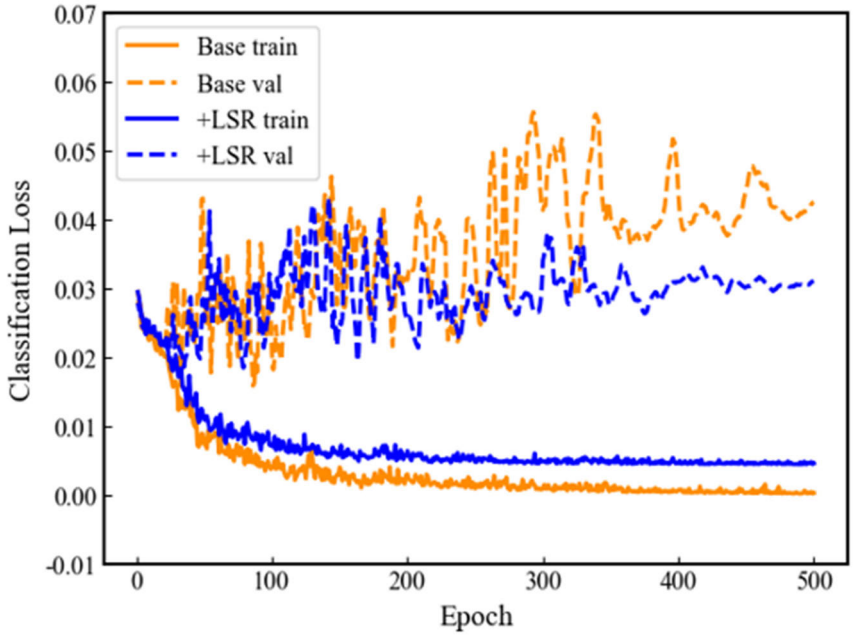

FIGURE 5. The classification loss function of the model in the training process with 10 bug data added.

overcome the over-fitting problem. 2) The lack of diversity in thyroid nodule types is another issue that may restrict the application potential of the proposed network. All the benign lesions were nodular hyperplasias, and all the malignant tumors were papillary carcinoma in this experiment. Although the acoustic characteristics of different nodule types are appreciably different [43], the YOLOv5 network is capable of distinguishing more than 80 classes of objects. A more specific classification of the thyroid nodule's type is of great help for the monitoring and surgical intervention planning afterward. 3) To deploy the network as a CAD tool in clinical practice, and to investigate its impact on the workflow and decision-making of radiologists and surgeons. To assess the user satisfaction, acceptance, and feedback of the network, and to optimize the network accordingly. 4) While the current study provides promising results within

the scope of our dataset, future work should focus on prospectively validating the model using additional images from diverse patients in various medical centers. This step is crucial to ensure the robustness and reliability of the model across different populations and healthcare settings.

V. CONCLUSION

Accurate automatic thyroid nodule diagnosis is a challenging task and is in great demand. In this study, we present an improved YOLOv5 network with the LSR module and the CA module plugged in. The network can correctly recognize the thyroid nodules in a millisecond-level detection time. The training process can be well conducted with only a small dataset, and the pre and post-processes of ultrasound images are barely needed. The experimental results show that the improved network has an mAP of 95.3%, 4.4% higher than the original network, and the detection time is only 8.7 ms per image. Furthermore, the improved network is more robust against the bug data and over-fitting problem, concerning the original one. The proposed network has significant superiority comparing with other SOTA model. It can be concluded that the proposed network is a promising CAD tool for thyroid nodule detection, with which the workload of the radiologists will be alleviated.

REFERENCES

[1] C. P. Wild, E. Weiderpass, and B. W. Stewart, "World cancer report: Cancer research for cancer prevention," in *International Agency for Research on Cancer*. Lyon, France: Avenue Tony Garnier, 2020.

[2] C. La Vecchia, M. Malvezzi, C. Bosetti, W. Garavello, P. Bertuccio, F. Levi, and E. Negri, "Thyroid cancer mortality and incidence: A global overview," *Int. J. Cancer*, vol. 136, no. 9, pp. 2187–2195, May 2015.

[3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[4] C. S. Park, S. H. Kim, S. L. Jung, B. J. Kang, J. Y. Kim, J. J. Choi, M. S. Sung, H. W. Yim, and S. H. Jeong, "Observer variability in the sonographic evaluation of thyroid nodules," *J. Clin. Ultrasound*, vol. 38, no. 6, pp. 287–293, Jul. 2010.

[5] B. R. Haugen, "2015 American thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: What is new and what has changed?" *Cancer*, vol. 123, no. 3, pp. 372–381, Feb. 2017.

[6] D. S. Cooper, G. M. Doherty, B. R. Haugen, R. T. Kloos, S. L. Lee, S. J. Mandel, E. L. Mazzaferri, B. McIver, F. Pacini, M. Schlumberger, S. I. Sherman, D. L. Steward, and R. M. Tuttle, "Revised American thyroid association management guidelines for patients with thyroid nodules and differentiated thyroid cancer," *Thyroid*, vol. 19, no. 11, pp. 1167–1214, Nov. 2009.

[7] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[8] X. Zhang, C. Xuan, J. Xue, B. Chen, and Y. Ma, "LSR-YOLO: A high-precision, lightweight model for sheep face recognition on the mobile end," *Animals*, vol. 13, no. 11, p. 1824, May 2023.

[9] J. X. Gu, Z. H. Wang, J. Kuen, L. Y. Ma, A. Shahroudy, B. Shuai, T. Liu, X. X. Wang, G. Wang, J. F. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Patten Recognit.*, vol. 77, pp. 354–377, May 2018.

[10] X. Wu, D. Sahoo, and S. C. H. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, Jul. 2020.

[11] A. Kunapinun, M. N. Dailey, D. Songsaeng, M. Parnichkun, C. Keatmanee, and M. Ekpanyapong, "Improving GAN learning dynamics for thyroid nodule segmentation," *Ultrasound Med. Biol.*, vol. 49, no. 2, pp. 416–430, Feb. 2023.

[12] C. F. Li, R. Q. Du, Q. Y. Luo, R. Wang, and X. H. Ding, "A novel model of thyroid nodule segmentation for ultrasound images," *Ultrasound Med. Biol.*, vol. 49, no. 2, pp. 489–496, 2023.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[16] T. Liu, S. Xie, J. Yu, L. Niu, and W. Sun, "Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 919–923.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[18] J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot, and M. Eramian, "Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network," *J. Digit. Imag.*, vol. 30, no. 4, pp. 477–486, Aug. 2017.

[19] 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 Med. Imag. Graph.*, vol. 71, pp. 9–18, Jan. 2019.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[21] H. Li, J. Weng, Y. Shi, W. Gu, Y. Mao, Y. Wang, W. Liu, and J. Zhang, "An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images," *Sci. Rep.*, vol. 8, no. 1, p. 6600, Apr. 2018.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[23] 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," *Comput. Biol. Med.*, vol. 122, Jul. 2020, Art. no. 103871.

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[25] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 7263–7271.

[26] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[27] J. Ma, S. Duan, Y. Zhang, J. Wang, Z. Wang, R. Li, Y. Li, L. Zhang, and H. Ma, "Efficient deep learning architecture for detection and recognition of thyroid nodules," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–15, Aug. 2020.

[28] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, May 2008.

[29] *YOLOv5*. Accessed: May 18, 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[30] C.-B. Zhang, P.-T. Jiang, Q. Hou, Y. Wei, Q. Han, Z. Li, and M.-M. Cheng, "Delving deep into label smoothing," *IEEE Trans. Image Process.*, vol. 30, pp. 5984–5996, 2021.

[31] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," 2021, *arXiv:2103.02907*.

[32] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[33] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

[34] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. Lawrence Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," 2014, *arXiv:1405.0312*.

[35] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," 2019, *arXiv:1911.08287*.

[36] *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*. Accessed: Nov. 6, 2007. [Online]. Available: http://www.pascalnetwork.org/challenges/VOC/voc2007/workshop/index.html
[37] S. Q. Ren, K. M. He, and R. Girshick, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[38] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020.
[39] W. Liu, D. Anguelov, and D. Erhan, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[41] K. J. Lim, C. S. Choi, D. Y. Yoon, S. K. Chang, K. K. Kim, H. Han, S. S. Kim, J. Lee, and Y. H. Jeon, "Computer-aided diagnosis for the differentiation of malignant from benign thyroid nodules on ultrasonography," *Academic Radiol.*, vol. 15, no. 7, pp. 853–858, Jul. 2008.
[42] L. Pratt, L. Pratt, and S. Thrun, *Machine Learning—Special Issue on Inductive Transfer*. Norwell, MA, USA: Kluwer, 1997.
[43] J. H. Yoon, H. J. Kwon, E.-K. Kim, H. J. Moon, and J. Y. Kwak, "The follicular variant of papillary thyroid carcinoma: Characteristics of preoperative ultrasonography and cytology," *Ultrasonography*, vol. 35, no. 1, pp. 47–54, Jan. 2016.

DAQING YANG received the bachelor's degree in clinical medicine and the master's degree in general surgery from Wenzhou Medical University, in 2004 and 2018, respectively. He is currently the Deputy Chief Physician of Colorectal and Anorectal Surgery with Wenzhou Central Hospital. He has published seven academic works in relevant professional fields. His research interests include the diagnosis and treatment of colorectal tumors, thyroid and breast tumors, and advanced digital technology.

JIANFU XIA received the master's (Medical) degree from the Department of Clinical Medicine, Wenzhou Medical University, China. Since 2010, he has been a General Surgeon with Wenzhou Central Hospital, China. In 2013, he became an Attending Surgeon. He has published more than ten papers in international and national journals and conference proceedings. His current research interests include colon tumor and thyroid cancer.

RIZENG LI received the bachelor's degree in clinical medicine and the master's degree in oncology from Wenzhou Medical University, in 2003 and 2018, respectively. He is currently the Deputy Chief Physician of Colorectal and Anorectal Surgery with Wenzhou Central Hospital. He has published eight academic works in relevant professional fields. His research interests include the diagnosis and treatment of colorectal tumors and thyroid and breast tumors.

WENCAI LI received the bachelor's degree in clinical medicine from Wenzhou Medical University, in 2010. He is currently the Attending Physician in colorectal and anal surgery with Wenzhou Central Hospital. He has published three academic works in related professional fields. His research interests include the diagnosis and treatment of colorectal tumors, thyroid and breast tumors, and surgical treatment of perianal benign diseases.

JISHENG LIU received the bachelor's degree in clinical medicine and the master's degree in general surgery from Hunan Normal University, in 2016 and 2020, respectively. He is currently a Resident Physician with the Colorectal and Anal Surgery Department, Wenzhou Central Hospital. He has published one academic treatise in the relevant professional field. His research interests include the diagnosis and treatment of colorectal tumors and thyroid and hepatobiliary tumors.

RONGJIAN WANG received the bachelor's degree in ophthalmology and the master's degree in general surgery from Wenzhou Medical University, in 2017 and 2021, respectively. He is currently a Resident Physician with the Colorectal and Anal Surgery Department, Wenzhou Central Hospital. He has published one academic treatise in the relevant professional field. His research interests include the diagnosis and treatment of colorectal tumors and thyroid and hepatobiliary tumors.

DONG QU received the bachelor's degree from Wenzhou Medical University, in 2004. He is currently an Attending Physician with Wenzhou Central Hospital. He has been working in medical imaging diagnosis for 19 years. He is also engaged in ultrasound, CT-guided percutaneous aspiration therapy, and artificial intelligence in medical imaging diagnosis.

JIE YOU received the master's degree from Wenzhou Medical University, in 2008. He is currently an Associate Chief Physician. He has been engaged in the diagnosis and treatment of thyroid tumors for 20 years. He is also engaged in the surgical treatment of thyroid malignant tumors and the correlation research of metabolism in thyroid malignant tumors.