Identifying Thyroid Nodules in Ultrasound Images Through Segmentation-Guided Discriminative Localization

Jintao Lu1,2, Xi Ouyang1, Tianjiao Liu3, and Dinggang Shen1(✉)

1 Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
Dinggang.Shen@gmail.com
2 Department of Control Science and Engineering, Zhejiang University, Hangzhou, China
3 Department of Electronic Engineering, Tsinghua University, Beijing, China

Abstract. In this paper, we propose a novel segmentation-guided network for thyroid nodule identification from ultrasound images. Accurate diagnosis of thyroid nodules through ultrasound images is significant for cancer detection at the early stage. Many Computer-Aided Diagnose (CAD) systems for this task ignore the inherent correlation between nodule segmentation task and classification task (i.e. cancer grading). Actually, segmentation results could be used as localization cues of thyroid nodules for facilitating their classifications as benign or malignant. Accordingly, we propose a two-stage thyroid nodule diagnosis method through 1) nodule segmentation and 2) segmentation-guided diagnosis. Specifically, in the segmentation stage, we use an ensemble strategy to integrate segmentations from diverse segmentation networks. Then, in the classification stage, the obtained segmentation result is integrated as additional information along with its corresponding original ultrasound images as the input of the classification network. Meanwhile, the segmentation result is further served as guidance to refine the attention map of the features used for classification. Our method is applied to the TN-SCUI 2020, a MICCAI 2020 Challenge, with the largest set of thyroid nodule ultrasound images according to our knowledge. Our method achieved the 2nd place in its classification challenge.

Keywords: Thyroid nodule · Ultrasound · Identification · Attention

1 Introduction

Thyroid nodule is a common disease with sometimes very rapid growth rates and bad outcomes. Convenient and non-invasive ultrasound has become a commonly used technique for detecting and diagnosing thyroid nodules. However, clinical
evaluation based on thyroid nodule ultrasound images for distinguishing malignant from benign nodules can be very tedious and challenging. For instance, Fig. 1(a) shows multiple suspicious nodules with varied locations. Their diagnosis process is time-consuming and highly relies on clinicians’ proficiency.

To alleviate the burden of clinicians, many Computer-Aided Diagnose (CAD) systems, including advanced deep learning methods, have been developed to identify and diagnose thyroid nodules. One typical type of these methods is to handle the diagnosis task as an instance segmentation task. Many models can be applied to localize and identify nodules [1,2]. Liu et al. [3] extracted diagnosis-oriented features with a multi-branch classification network after a multi-scale detector, while Ponugoti et al. [4] proposed a U-Net [5] based framework for thyroid nodule segmentation with an appended classification branch after the encoder. However, as all these methods extracted the same features for both nodule detection/segmentation and classification, it is not optimal for each task. Besides, different locations in the feature maps may contribute inconsistently to the segmentation and classification tasks. On the other hand, it is known that, by enforcing the model’s attention on salient locations, diagnosis/classification can be better achieved. For example, in natural image segmentation, attention to edges of objects could contribute more to segmentation performance. In this way, Song et al. [6] proposed decoupled classification head and detection head after RoI Pooling [7]. However, their framework is limited to natural images; directly application to detection and classification of much subtler and more complex thyroid nodules poses significant difficulty in generating decoupled attention.

In this paper, we utilize the localization cues derived from segmentation results and hereby propose a segmentation-guided attention network to achieve interpretable and accurate diagnosis results. Specifically, an online attention module is introduced to guide the network to focus on discriminative locations with nodules and thus generates a robust diagnosis. We apply our proposed novel network to the TN-SCUI 2020, a MICCAI 2020 Challenge, with the largest set of thyroid nodule ultrasound images to our knowledge, and achieved the 2nd rank in its classification task (F-1 score of 0.8541).

Specifically, we first ensemble different results from diverse segmentation algorithms to refine the segmentation mask by also suppressing false-positive results. Then we feed both the segmentation mask of nodules and the original images into the classification network, trained with an online attention strategy to ensure extracting features from nodule regions for classification, we use an efficient data augment strategy to simulate 1) variations of rotation, position, intensity of nodules in the ultrasound images and 2) variation of patch extraction by randomly cropping patches around the given location of each nodule.

2 Method

In this section, we first introduce the learning-based ensemble strategy for refining the nodule segmentation in Sect. 2.1. A cascaded U-Net (Fig. 2) is applied in the refinement stage after multiple trained networks to reduce segmentation
error. Then, in Sect. 2.2, we describe dilation of the refined mask as an input of the classifier while guiding the model’s attention onto the segmented nodules through generation and refinement of classification activation map (CAM) \[8\] in an online manner (Fig. 3).

**Fig. 1.** Examples of diverse thyroid nodule images. (a) A hard case with multiple suspicious nodules (Upper: original image; Bottom: segmentation). (b) Two high contrast images. (c) Two bright intensity images. (d) Two dark intensity images.

**Fig. 2.** Ensemble of segmentation results via a cascaded U-Net for generating a refined segmentation mask with precise and smooth nodule margin.

### 2.1 Cascaded U-Net for Segmentation Ensemble

We apply several high-performance segmentation models (e.g., LinkNet \[9\] and PSPNet \[10\]) for thyroid nodule segmentation. Since these semantic segmentation models may over-segment nodules or produce holes inside the segmented nodule (Fig. 2(a)), we add instance segmentation models, e.g., 1) Mask Scoring R-CNN \[11\] and 2) Deep Snake \[2\], to constrain the segmentation mask by its bounding box. In particular, the segmentation result from Mask Scoring R-CNN may lead to under-segmentation ( lesion outside the predicted bounding box was missed in Fig. 1(a)), which can be compensated by Deep Snake by deforming the boundary of segmentation mask (although Deep Snake may have over-smooth results (Fig. 2(b))).

To take advantage of multiple segmentation models, we propose a segmentation ensemble method via a cascaded U-Net. All the output segmentation masks
from multiple segmentation models are concatenated with the original image, and then fed into the U-Net. We use a Dice loss [12] to allow the network to learn from multi-style segmentation masks, thus finally obtaining the refined segmentation mask with suppressed error and also a more precise margin. Note that we can cascade more U-Nets for further improvement, similar to the refinement stages used in Hourglass [13]; alternatively, we can conduct the refinement recursively.

![Diagram of the segmentation-guided attention network](image)

**Fig. 3.** Overview of our segmentation-guided attention network. (a) The overall pipeline. We extract features from both original image and nodule segmentation mask. Then, we generate CAM online and compare it with the input segmentation mask to guide the locations for feature extraction. (b) Illustration of the foreground attention module. We fuse global spatial information into a channel attention sub-module. Then, the position attention sub-module forwards previous information to the convolution layer for generating the spatial attention.

### 2.2 Online Attention Module for Classification

We use ResNet101 [14] as a backbone for feature extraction. To keep high resolution feature maps, we change the stride from 2 to 1 in the last residual block; meanwhile, we apply dilated convolution [15] to expand the receptive field.

After the backbone, we add an attention module for both position attention and channel attention. Specifically, inspired by non-local [16] and DANet [17] for global spatial attention, we fuse non-local mechanism into the squeeze-and-excitation [18] attention module by replacing its Global Average Pooling (GAP) [19] layer. Therefore, this sub-module can merge dual attentions in only one step. After that, a lightweight attention sub-module is further applied to enhance the positional attention of our model. We use a $1 \times 1$ convolution kernel to reduce
the channels, and then apply a $3 \times 3$ convolution kernel with subsequent batch normalization and sigmoid activation to get a position attention map.

With CAM [8], the GAP layer can preserve spatial information in the feature maps and enable the classification network to show its attention on the discriminative locations. After propagating weights of the fully-connected layer to the convolutional feature maps, we generate class-specific discriminative regions as the attention map. Inspired by an online mechanism of 3D CAM for COVID-19 diagnosis [20], we leverage an online trainable CAM for this 2D ultrasound image task. Let $f$ represent the feature maps after the last convolutional layer and $w$ represent the weight matrix of the fully-connected layer. As the $1 \times 1$ convolutional layer is mathematically the same as fully-connected layer, we copy $w$ as the weight of a single $1 \times 1$ convolutional kernel with a ReLU activation function to generate the attention map $A$ as:

$$A = ReLU(\text{conv}(f, w)) = ReLU(\sum_{k=1}^{c} w_k f_k),$$

(1)

where $k$ stands for the index of channel with totally $c$ channel, and $A$ is of the same shape of $X$ and $Y$. Next, we post-process the attention map via upsampling to the original input image size and normalize the values to $(0,1)$. We then add the Attention Loss to maximize the overlap between activation map and the input segmentation mask. Together with the Binary Cross Entropy Loss of classification, we have the total loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Cls}} + \lambda \mathcal{L}_{\text{Attention}} = \text{BCE}(y_{\text{predict}}, y_{\text{gt}}) + \lambda \text{Dice}(\text{CAM}, \text{Mask}),$$

(2)

where $\lambda$ is the weight factor for attention and is set 0.6 in our experiments. Parameters in the $1 \times 1$ convolution layer are always copied from the fully-connected layer, so they will be updated only through the back-propagation of $\mathcal{L}_{\text{Cls}}$. $\mathcal{L}_{\text{Attention}}$ back-propagates before the GAP layer to guide the feature extractor and the attention module.

After building this mask-guided attention module, we expect the CAM to locate accurately inside or around the nodule segmentation masks, similar to clinicians who diagnose based on some important characteristics on the nodule areas. For instance, a nodule has a higher possibility to be malignant if calcification exists, or a nodule’s margin is unclear and hard to distinguish or in a complicated rather than ellipse shape. Also, since all these characteristics indicate aggressive growth, we dilate the ground-truth segmentation mask to include more neighborhood context, which is extremely important for small nodules with very limited internal features. Besides, replenishing neighborhood information from surrounding thyroid tissues is also vital for an accurate diagnosis.

### 2.3 Data Augmentation

We design an effective data augmentation strategy for robust nodule identification from a large set of ultrasound data with diverse imaging attributes and
appearance. As shown in Fig. 1, the majority of images are cross-sectional thyroid tissues with different locations and sizes. For example, an image may include the entire butterfly-shaped thyroid tissue (Fig. 1(a)), or only the left part or the right part (Fig. 1(d)) of thyroid tissue. Hence, we flip images in the horizontal direction (with a 50% chance) to imitate their counterparts. Besides, we randomly zoom out from 0.5× to 1.0× scale for a better view of small objects, and also crop the images covering the nodule areas in a range from the nodule size to the entire image size for simulating various uncertainties. Next, as commonly adopted in the field, we also create rotated and/or shifted images to mimic the real clinical scenario where different perspective images could be acquired when patients are scanned with varied positions and gestures.

Finally, we conduct random image deformations to simulate angles and forces of ultrasound probes. Since this dataset contains images from different ultrasonic machines, it includes different intensity scales and different tissue contrast levels (Fig. 1(b)(c)(d)). To address this, we not only augment image brightness, contrast, Gaussian noise, blurring, and sharpness to enhance the model’s robustness, but also normalize intensity distribution of each image to have zero mean and unit variance during pre-processing. In this way, we simulate many possible situations for better approaching real-world application scenarios and minimizing the differences of images acquired from different machines, thus further increasing the robustness of our model.

3 Experiments and Results

The provided training set of the TN-SCUI2020 contains ultrasound images of 1641 benign cases and 2003 malignant cases. To train our model, we conduct a 5-fold cross-validation strategy. Note that, each image is resized to 512×512 as the input to both the nodule segmentation and classification models. We also compare our classifier with other state-of-the-art models using the same cross-validation sets, and the result shows that our model outperforms other comparison methods. Ablation studies also show the effectiveness of each proposed component in our model. To further improve the performance, we finally use majority voting to combine the five trained models from five splits for obtaining robust and accurate results in the official testing dataset with 910 images.

3.1 Ablation Study on Segmentation Ensemble

We compare the refined outputs from the cascaded U-Net with those from the previously trained segmentation models. Using the initial testing split in one cross-validation case with 400 images, 1) LinkNet with ResNet101 backbone achieved 78.99% of mean IoU, 2) PSPNet with ResNet101 backbone achieved 78.50% of mean IoU, 3) Deep Snake with CenterNet detector and DLA34 [21] backbone achieved 76.71% of mean IoU, and 4) Mask Scoring R-CNN with the ResNet50 backbone and FPN [22] achieved 79.28% of mean IoU. Our ensemble method finally obtained 79.68% of mean IoU, which is 0.4% of mean IoU improvement compared to the best of these four segmentation models.
Table 1. Results of nodule classification obtained by six different models.

| Model                      | ACC  | Model                      | ACC  |
|----------------------------|------|----------------------------|------|
| Fuse-feature               | 0.7146 | Mask Scoring R-CNN (ResNet50 + FPN) | 0.7764 |
| ResNet50                   | 0.7523 | CenterNet (DLA34)         | 0.7804 |
| ResNeSt50                  | 0.7712 | Segmentation-guided ResNet101 (Ours) | 0.8224 |

Table 2. Ablation studies on our segmentation-guided classification model. SG stands for the segmentation guidance for CAM, FA stands for the usage of Foreground attention module. Mask-In indicates that we also feed the segmentation mask into the model. Dil stands for a dilation of 15 pixels to the segmentation mask. We show the mean IoU between the attention areas of CAM and the nodule segmentation masks.

| Model           | FA | Mask-In | Dil | F-1 Score | Accuracy | Mean IoU |
|-----------------|----|---------|-----|-----------|----------|----------|
| ResNet101       |    |         |     | 0.7589    | 0.7500   | 0.0410   |
| ResNet101       | ✓  |         |     | 0.7990    | 0.7896   | 0.0750   |
| SG-ResNet101    | ✓  |         |     | 0.8185    | 0.7910   | 0.5297   |
| SG-ResNet101    | ✓  | ✓       |     | 0.8249    | 0.7923   | 0.5504   |
| SG-ResNet101    | ✓  | ✓       | ✓   | 0.8331    | 0.7964   | 0.8398   |
| SG-ResNet101    | ✓  | ✓       | ✓   | 0.8337    | 0.8005   | 0.5943   |
| SG-ResNet101    | ✓  | ✓       | ✓   | **0.8467**| **0.8224**| **0.9140**|

3.2 Performance of Segmentation-Guided Classification

In the experiments shown in Table 1, we compared the accuracy of our model with other CAD models, including 1) Fuse-feature method [23] that combines HOG, LBP, and SIFT features together with the deep features extracted from a fine-tuned VGG Net [24], 2) ResNet50, 3) ResNeSt50 [25], 4) Mask Scoring R-CNN, and 5) CenterNet. For the classification models without a detector, we dilate the ground-truth segmentation masks and crop images around the nodules with context information. Results show that our proposed attention model yielded the best performance and brought an improvement of 4.2% accuracy compared to the best of five comparison models.

3.3 Ablation Study on Classifier

To evaluate the usefulness of each strategy in our classifier, we conducted experiments with different settings. All models were trained with an initial learning rate of 0.00003 and a learning rate decay strategy that decays the rate by a factor of 0.1 every 10 epochs. We set the batch size 20 on 4 NVIDIA TITAN X GPUs and iterated 20 epochs. Backbones were all pre-trained in ImageNet [26].

As shown in Table 2, without the segmentation guidance, the classifier cannot form attention on nodules as its CAM was overlapped poorly with the nodule segmentation mask, while this overlap increased quickly after the guidance of
CAM. Experiments also show that combining all our strategies together can achieve a significant boost to the final results, i.e., F-1 score increases and mIoU increases. Note that our model can still generate high accuracy and also locate discriminative nodule areas (Fig. 4) even without feeding any segmentation mask during the testing. These results indicate our model has the potential to be used without requiring the nodule segmentation in the final applications.

We randomly select 4 malignant cases and 4 benign cases, and visualize their attention maps generated by our models in Fig. 4. The second row shows the CAM from our original segmentation-guided attention model. Although the CAM was formed imperfectly with grid shape due to the dilated convolution layers, we can still see their coarse round-shaped attention on the nodule areas. The third row shows the improved results with detailed attention on the nodule areas when we added the foreground attention module. In the last row, we show the improved results when dilating the guided segmentation mask during the training for better attention distribution. Comparing the last two rows, we can see that the attention maps distributed less uniformly and focused more on the key diagnostic areas when dilating the segmentation masks. For example, for the nodules with calcification (e.g., Fig. 4(a)(d)), our model paid more attention on the possible calcification points near the margin of nodules after dilating the segmentation masks (e.g., point 1 in Fig. 4(a)). For the nodules with low and plain inner echo, our model paid more attention on the margin areas with discriminative information, indicating that the model can distinguish the malignant nodules with low margin sharpness and irregular margin shape (Fig. 4(c)) from the benign cystic thyroid nodules (Fig. 4(e)(f)). Instead of always distributing attention to the margins, our model with dilated segmentation masks may shrink its attention to concentrate more on the abnormal components inside a nodule when no useful context can be used (e.g., locations 2,3,4 in Fig. 4(b)(d)(g)).
4 Conclusion

In this study, we proposed a novel segmentation-guided attention network for nodule identification from ultrasound images and achieved a reasonable diagnosis. The use of dilated segmentation masks is able to provide more guidance for the classifier to localize nodules automatically, and finally form more accurate attention to the most informative parts in the nodules for better capturing subtle differences between benign and malignant cases. Experiments show the high performance of our model, indicating also the necessity of providing appropriate guidance to the classifier. Since all testing data were acquired from different machines under different situations, our model has the possibility to be used in real clinical scenarios.

References

1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6154–6162 (2018)
2. Peng, S., Jiang, W., Pi, H., Li, X., Bao, H., Zhou, X.: Deep snake for real-time instance segmentation. In: CVPR (2020)
3. 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)
4. Nikhila, P., Nathan, S., Ataide, E.J.G., Illanes, A., Friebe, D., Abbineni, S., et al.: Lightweight residual network for the classification of thyroid nodules. arXiv preprint arXiv:1911.08303 (2019)
5. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
6. Song, G., Liu, Y., Wang, X.: Revisiting the sibling head in object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11563–11572 (2020)
7. 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)
8. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)
9. Chaurasia, A., Culurciello, E.: Linknet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4. IEEE (2017)
10. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
11. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring R-CNN. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6409–6418 (2019)
12. Milletari, F., Navab, N., Ahmadi, S.-A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
13. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
14. 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 (2016)
15. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 472–480 (2017)
16. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
17. Fu, J., et al.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3146–3154 (2019)
18. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
19. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
20. Ouyang, X., et al.: Dual-sampling attention network for diagnosis of covid-19 from community acquired pneumonia. IEEE Trans. Med. Imaging 39, 2595–2605 (2020)
21. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2403–2412 (2018)
22. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
23. Gao, L., et al.: Computer-aided system for diagnosing thyroid nodules on ultrasound: a comparison with radiologist-based clinical assessments. Head Neck 40(4), 778–783 (2018)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
25. Zhang, H., et al.: Resnest: split-attention networks. arXiv preprint arXiv:2004.08955 (2020)
26. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)