2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)
April 13-16, 2021, Nice, France

2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-1246-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISBI48211.2021.9434087

MULTI-TASK LEARNING FOR THYROID NODULE SEGMENTATION WITH THYROID REGION PRIOR

Haifan Gong[1], *Guanqi Chen*[1] , *Ranran Wang*[2], *Xiang Xie*[1], *Mingzhi Mao*[1], *Yizhou Yu*[3], *Fei Chen*[4†], *Guanbin Li*[1†]

[1]Sun Yat-sen University, School of Computer Science and Engineering, China
[2] Nanchang University, School of Information Engineering, China
[3] Deepwise AI Lab, Beijing, China
[4] Southern Medical University, Zhujiang Hospital, China

ABSTRACT

Thyroid nodule segmentation in ultrasound images is a valuable and challenging task, and it is of great significance for the diagnosis of thyroid cancer. Due to the lack of the prior knowledge of thyroid region perception, the inherent low contrast of ultrasound images and the complex appearance changes between different frames of ultrasound video, existing automatic segmentation algorithms for thyroid nodules that directly apply semantic segmentation techniques can easily mistake non-thyroid areas as nodules. In this work, we propose a thyroid region prior guided feature enhancement network (TRFE-Net) for thyroid nodule segmentation. In order to facilitate the development of thyroid nodule segmentation, we have contributed TN3k: an open-access dataset of thyroid nodule images with high-quality nodule masks labeling. Our proposed method is evaluated on TN3k and shows outstanding performance compared with existing state-of-the-art algorithms. Source code and data are available [1].

Index Terms— Thyroid nodule, Ultrasound image, Segmentation, Multi-task learning, Attention modeling

1. INTRODUCTION

A thyroid nodule is an abnormal lump that grows in the thyroid gland. It is the early symptom of thyroid cancer that has become increasingly prevalent in the past 30 years. Based on a statistical sense, the vast majority of thyroid nodules are benign and the remaining small part are malignant [1], so the diagnose of malignant nodules is very dependent on the experience of clinicians. Considering that inexperienced clinicians can easily cause misdiagnosis, more and more computer-aided diagnosis systems are being developed for auxiliary diagnosis of thyroid diseases. The automatic segmentation of thyroid nodules is the basis for building intelligent diagnosis and a prerequisite for accurate diagnosis.

In recent years, deep convolutional neural networks have greatly facilitated the progress of visual semantic segmentation and have shown remarkable progress. A large amount of the advanced deep neural network structures have been proposed in the research area of semantic segmentation as well as medical image segmentation, e.g. FCN [2], U-Net [3], SegNet [4], Deeplab [5].

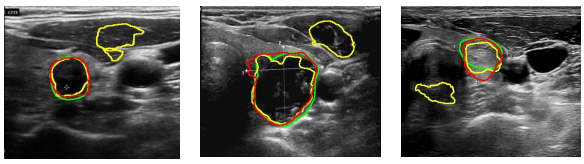


Fig. 1. Several visualization examples of our TRFE-Net comparing to U-Net. The curves in green, yellow and red represent the ground truth, U-Net segmentation, and TRFE-Net segmentation, respectively.

In the field of thyroid nodule segmentation, deep learning based models have significantly outperformed conventional methods[1]. Ma *et al.*[6] first reported to segment the thyroid nodule with convolutional neural networks. To accurately segment the thyroid nodules, Ying *et al.*[7] proposed a cascaded framework that first eliminates the influence of irrelevant regions with the help of an object detection network, and then uses the VGG network to accurately segment the thyroid nodules within detected region proposals. Considering the appearance shifts of the ultrasound image could result in performance degradation, Liu *et al.*[8] proposed a framework based on universal style transfer to make deep neural networks robust to appearance shift. Kumar *et al.*[9] proposed a framework that simultaneously segment the thyroid gland, thyroid nodule and the thyroid cystic. However, all previous research works are based on the end-to-end framework mapping from the overall ultrasound image to the localization of the nodule area, ignoring the preconditions that thyroid nodules must be located within the thyroid area, which results in

†These authors are the joint corresponding authors.
[1]https://github.com/haifangong/TRFE-Net-for-thyroid-nodule-segmentation

the algorithms generating incorrect location of thyroid nodules outside the thyroid gland region.

In this work, we propose a framework to segment the thyroid nodule precisely with the detected thyroid gland region as the incorporated prior information. There are two major differences compared with the existing works [9, 7]: (1) we incorporate separate training data labeled with either thyroid gland region or thyroid nodule to an integrated multi-task framework. (2) we use the result of thyroid gland region prediction to guide the realization of more accurate positioning and segmentation of thyroid nodules. Specifically, we develop a multi-task learning framework which contains a shared encoder backbone for feature representation learning and two separate decoders for thyroid gland region segmentation and thyroid nodule discovery, respectively. Furthermore, to make full use of the information learned from the branch of thyroid region segmentation, we design several feature enhancement modules with lightweight parameters. The contributions of this paper are summarized as follows.

1. A multi-task learning framework is proposed to learn to simultaneously infer the segmentation of thyroid regions and nodules, forcing the same backbone network to infer the localization of nodules within the thyroid region.

2. Three thyroid region prior guided feature enhancement modules are designed to better embed the thyroid region prior into the branch of the thyroid nodule segmentation and improve its positioning accuracy.

3. An open-access dataset of thyroid nodule images with high-quality nodule masks labeling are proposed to facilitate the research of thyroid nodule segmentation.

2. METHOD

2.1. Multi-task Learning Framework

Based on the observation that existing single-branch neural network for thyroid nodules segmentation tends to regard the non-thyroid areas of similar nodule appearance as thyroid nodules due to the lack of prior knowledge of glandular region, we propose a multi-task learning framework called thyroid region prior guided feature enhancement network(TRFE-Net) for precise thyroid nodule segmentation from ultrasound image. The structure of the TRFE-Net is shown in Fig. 2. It contains a shared encoder backbone for feature representation learning and two separate decoders for thyroid gland region segmentation and thyroid nodule discovery, respectively. During the training process, we select image from the thyroid nodule segmentation dataset and the thyroid gland segmentation dataset one by one to construct a batch, which is shown in the upper of Fig.2. The former contains the mask label of the thyroid nodule and is used to train the nodule segmentation branch and the latter is used to supervise the branch of thyroid gland segmentation. It should be noted that the image containing the thyroid nodule

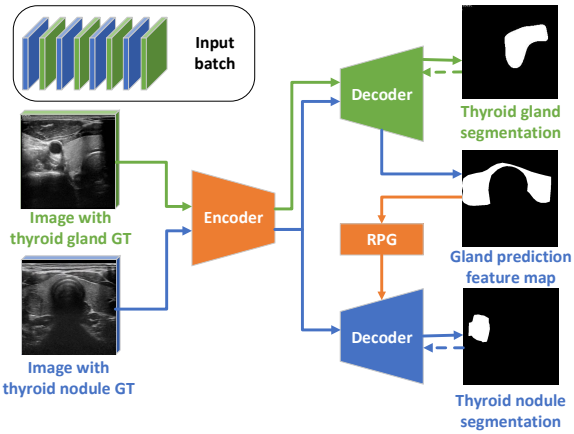


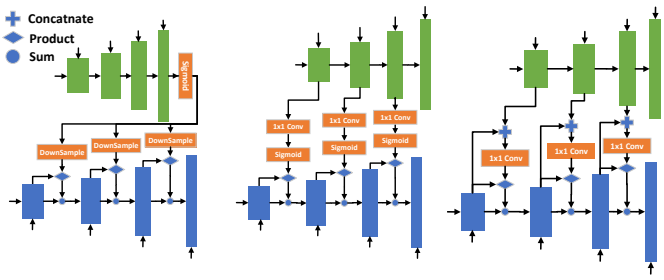
Fig. 2. The structure of the thyroid region prior guided feature enhancement network (TRFE-Net). The TRFE-Net learns to segment the thyroid nodule region and the thyroid gland region simultaneously. The solid lines in green and blue with arrow represent the information stream of the thyroid gland region segmentation and nodule segmentation, respectively. The dotted arrows represent the corresponding gradient back-propagation process. The input batch is shown in the top of the figure, the blue block denotes the nodule image and the green block denotes the gland image.

annotation will be simultaneously input to the upper branch for thyroid region inference, and the obtained features is further fed to the region prior guidance module to enhance the precise positioning of the gland during thyroid nodule segmentation. The encoder consists of five layers including two repeated 3×3 convolutional layers with padding zero, each is followed by a ReLU and a 2×2 max pooling operation with stride 2. Each decoder is composed of five layers by repeating the 2×2 transposed convolutions, and a concatenation of the corresponding feature map in the encoder and the current feature map using skip connections. The following items are two 3×3 convolutions each followed by a ReLU. Finally, a 1×1 convolution layer is applied to predict the thyroid gland or the thyroid nodule area independently. The details of the region prior guide modules are introduced as follows.

2.2. Region Prior Guidance Module

To make full use of the prior knowledge acquired from the thyroid gland segmentation branch, we design several region prior guidance modules (RPG) to enhance the learned feature of the thyroid nodule segmentation branch, which are shown in Fig. 3. To give a formulaic definition, we denote "$\cdot$" as the dot-product operation, $\sigma(\cdot)$ as the Sigmoid operator, B_{tg} as the thyroid gland segmentation branch, and B_{tn} as the thyroid nodule segmentation branch.

The first RPG variant is shown in Fig. 3(a). We utilize the



(a) RPG-A (b) RPG-B (c) RPG-C

Fig. 3. Three different kinds of region prior guidance (RPG) modules. The thyroid gland segmentation branch , the thyroid nodule segmentation branch, and the feature fusion modules are represented by blocks in green, blue, and orange, respectively. The top and bottom arrows in the figure represent the features from the corresponding layer of the encoder.

predictions of B_{tg} as the guidance. The RPG module is added to the last 2, 3, 4 layers in B_{tn} by downsampling the feature map to the corresponding shape. This process is defined as:

$$y = x_{tn} \cdot I[\sigma(x_{tg})] + x_{tn} \tag{1}$$

where $I(\cdot)$ denotes the downsampling operation. x_{tn} represents the feature map in B_{tn} of size $h \times w \times c$, while x_{tg} denotes the output of the last layer in B_{tg} of size $h \times w \times 1$. It is worth noting that this approach does not require additional parameters in the neural network.

The second variant of RPG module is shown in Fig. 3(b). We use the output feature maps of B_{tg} branch to enhance the corresponding feature maps of B_{tn} branch, which is formulated as:

$$y = x_{tn} \cdot \sigma[g(x_{tg})] + x_{tn} \tag{2}$$

where $g(\cdot)$ denotes the 1×1 convolution layer with 1 output channel. x_{tn} represents the feature vector in B_{tn} of size $h \times w \times c$, and x_{tg} denotes the corresponding feature map of the B_{tg} branch of shape $h \times w \times c$.

The third feature fusion approach is shown in Fig. 3(c). We use the same layers as the second method, but replace the 1×1 convolution layer and sigmoid layer with the concatenate operation, which is formulated as:

$$y = x_{tn} \cdot f([x_{tn}, x_{tg}]) + x_{tn} \tag{3}$$

where $[\cdot, \cdot]$ denotes the concatenate operation. The definition of x_{tg} and x_{tn} is the same as with method 2. $f(\cdot)$ denotes a 1×1 convolution layer with c output channels.

2.3. Loss Function

To train the proposed TRFE-Net, we design a multi-task loss $\mathcal{L}_{total}$ with the fine-grained mask supervision of thyroid nodule and thyroid gland regions, which is shown below:

$$\mathcal{L}_{total} = \sum_{i=1}^{N_{nodule}} \mathcal{L}_{nodule\ i} + 0.5 * \sum_{j=1}^{N_{gland}} \mathcal{L}_{gland\ j} \tag{4}$$

where N_{gland} and N_{nodule} are set to half of the batchsize. $\mathcal{L}_{nodule}$ and $\mathcal{L}_{gland}$ represent the dice loss of the nodule segmentation and the gland segmentation, respectively.

Table 1. Thyroid ultrasonic image statistics.

	Thyroid Nodule	Thyroid Gland[10]
Training set	2879	3226
Test set	614	359

3. A NEW DATASET

In order to facilitate the development of thyroid nodule segmentation, we construct a thyroid nodule region segmentation dataset called TN3k, which includes 3493 ultrasound images taken from 2421 patients within the date period from January 2016 to August 2020. These images are selected from over 30,000 images provided by our partner hospital based on the following criteria: (1) Each image contains at least one thyroid nodule area; (2) Lymphatic images or images containing a large number of colored areas are excluded; (3) Only one representative image among the images of the same area or the same perspective of the patient is preserved.

We ask three volunteers to label the image under the guidance of the experienced radiologist. Each image is converted to grayscale and the non-ultrasound image area is cropped. To verify the performance of the algorithms, we divide the dataset into the training set and test set by ensuring that the images from the same patient only appear in a certain subset. The statistical information of TN3k is shown in the first column of Table 1.

4. EXPERIMENTS

4.1. Implementation Details

We train the models on a single NVIDIA GTX TITAN GPU. The framework is implemented in PyTorch 1.5. Unless specified, the weight of the model in the ablation study is initialized by Xavier. Stochastic gradient descent is applied to optimize the model at a learning rate of 0.001 for 60 epochs. The batchsize is set to 8, and all the images are resized to 224×224. In order to prevent the model from overfitting to the test set, the evaluated model is obtained by performing five-fold cross-validation on the training set. In other words, we use 2303 pictures to train the model, leaving 576 ultrasound images as the validation set to select the best model,

Table 2. Ablation study. The best result is shown in **bold**.

models	Jaccard(%)	Dice(%)
Unet [3]	63.51± 0.44	77.68± 0.61
Pretrain-U-Net	63.89± 0.55	77.97± 0.71
MT-Net	67.91± 0.35	80.89± 0.52
TRFE-Net-A	**68.44± 0.38**	**81.26± 0.55**
TRFE-Net-B	68.18± 0.19	81.08± 0.32
TRFE-Net-C	67.78± 0.34	80.80± 0.51

and finally use the test set to evaluate the performance of the best model. To quantitatively measure the models' performance, we apply the Dice coefficient and Jaccard index as the metrics.

To construct the thyroid gland segmentation set, we extract 3585 images with the thyroid gland ratio greater than 0.06 from 16 videos [10]. For U-net pre-training, we divide the thyroid gland segmentation images into training set and test set with the proportion of 9:1, which is shown in Table 1. For the multi-task learning task, we hybridize 2303 images of thyroid gland segmentation set and the training set of TN3k as the multi-task training dataset which contains 4606 images.

4.2. Ablation Study

As shown in Table 2, we verify the effectiveness of the proposed methods separately. The Pretrain-U-Net denotes the U-net pre-trained with thyroid gland images, which outperforms the U-net directly trained with thyroid nodule images by 0.38% w.r.t Jaccard index. The multi-task learning framework without the RPG module is represented by MT-Net. By taking advantage of the independently labeled data and the shared backbone structure, it significantly exceeds the Pretrain-U-Net by 4.02% Jaccard index. We evaluate three types of RPG modules in section 2.2. The TRFE-Net-A outperforms the MT-Net by 0.53% Jaccard index and achieves the best result. We guess that the reason for the better performance of TRFE-Net-A is that the segmentation mask is clearer than the inner feature map, contains less noise, and has more accurate contour boundary information.

4.3. Comparison with the State-of-the-art Methods

As shown in Table 3, the proposed TRFE-Net-A is compared with other advanced models including FCN [2], SegNet [4], Deeplabv3+ [5], and U-Net [3]. The FCN and the SegNet are trained with the ImageNet pre-trained VGG-Net backbone, while the Deeplabv3+ is trained with the ImageNet pre-trained ResNet-50 backbone. The proposed TRFE-Net-A considerably exceeds the second-best FCN by 3.04% Jaccard score and significantly outperforms the U-Net by 4.93% Jaccard index. Some visualization examples are shown in Fig. 1.

Table 3. Comparisons with the state-of-the-art semantic segmentation models. The best result is shown in **bold**.

models	Jaccard(%)	Dice(%)
FCN [2]	65.40±0.56	79.08±0.72
SegNet [4]	61.96±0.43	76.51±0.60
Deeplabv3+ [5]	53.89±0.44	70.03±0.61
U-Net [3]	63.51±0.44	77.68±0.61
TRFE-Net-A	**68.44±0.38**	**81.26±0.55**

5. CONCLUSION

This paper presents a multi-task learning framework for thyroid nodule segmentation from ultrasound images, which uses inferred thyroid region prior to enhance the feature representation for thyroid nodule segmentation. To the best of our knowledge, the proposed TRFE-Net is the first to successfully utilize the thyroid gland region prior to boost the thyroid nodule segmentation performance. Specifically, TRFE-Net contains a shared encoder backbone for feature representation learning and two separate decoders for thyroid gland region segmentation and thyroid nodule discovery, respectively. Besides, the RPG modules are proposed to enhance the thyroid nodule segmentation performance by utilizing the feature of the thyroid gland segmentation branch. With the help of these efforts, TRFE-Net achieves superior performance for thyroid nodule segmentation comparing w.r.t other state-of-the-art methods. Last but not least, a challenging dataset called TN3k is proposed to facilitate the future development of thyroid nodule segmentation.

6. REFERENCES

[1] Junying Chen, Haijun You, and K. Li, "A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images," *Computer methods and programs in biomedicine*, vol. 185, pp. 105329, 2020.

[2] J. Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[4] Vijay Badrinarayanan, Alex Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. PAMI*, vol. 39, pp. 2481–2495, 2017.

[5] Liang-Chieh Chen, Y. Zhu, G. Papandreou, Florian Schroff, and H. Adam, "Encoder-decoder with atrous

separable convolution for semantic image segmentation," in *ECCV*, 2018, pp. 801–818.

[6] Jinlian Ma, Fa Wu, Tian'an Jiang, Q. Zhao, and Dexing Kong, "Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, pp. 1895–1910, 2017.

[7] Xiang Ying, Zhihui Yu, Ruiguo Yu, Xuewei Li, M. Yu, Mankun Zhao, and Kai Liu, "Thyroid nodule segmentation in ultrasound images based on cascaded convolutional neural network," in *ICONIP*, 2018, pp. 373–384.

[8] Zhendong Liu, X. Yang, R. Gao, Shengfeng Liu, Haoran Dou, S. He, Yu-Hao Huang, Yankai Huang, Huanjia Luo, Y. Zhang, Y. Xiong, and Dong Ni, "Remove appearance shift for ultrasound image segmentation via fast and universal style transfer," in *ISBI*, 2020, pp. 1824–1828.

[9] Viksit Kumar, Jeremy M Webb, A. Gregory, Duane D. Meixner, J. Knudsen, M. Callstrom, M. Fatemi, and Azra Alizad, "Automated segmentation of thyroid nodule, gland, and cystic components from ultrasound images using deep learning," *IEEE Access*, vol. 8, pp. 63482–63496, 2020.

[10] Tom Wunderling, B. Golla, Prabal Poudel, C. Arens, M. Friebe, and Christian Hansen, "Comparison of thyroid segmentation techniques for 3d ultrasound," in *Medical Imaging*, 2017, p. 1013317.

Compliance with Ethical Standards

This is a numerical simulation study for which no ethical approval was required.

Acknowledgments

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation(No.2020B1515020048), in part by the National Natural Science Foundation of China (No.61976250, No.61702565).