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

SGUNET: SEMANTIC GUIDED UNET FOR THYROID NODULE SEGMENTATION

Huitong Pan[1], *Quan Zhou*[2] *and Longin Jan Latecki*[1]

[1] Department of Computer and Information Sciences, Temple University, USA
[2] National Engineering Research Center of Communications and Networking, Nanjing University of Posts Telecommunications, China.

2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-1246-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISBI48211.2021.9434051

ABSTRACT

Thyroid nodule segmentation from ultrasound images is an important step for early diagnosis of thyroid diseases. This paper introduces a novel encoder-decoder network architecture, called Semantic Guided UNet (SGUNet), for automatic thyroid nodule segmentation. In contrast to previous UNet architecture that only utilizes high-dimensional features on the up-sampling paths, our SGUNet further abstracts a single-channel pixel-wise semantic map from the high-dimensional features in each decoding step, which serves as a high-level semantic guidance to low-level features for obtaining more accurate nodule representation. We evaluate our SGUNet on Thyroid Digital Image Database (TDID) with high noise, blurry nodule boundaries and no embedded calipers, which marks the extremes of nodules. The 5-fold cross validation experiments show that our SGUNet achieves 72.9% in terms of Dice Coefficient, yielding 2.0% and 2.4% improvements with respect to traditional UNet and its variant UNet++.

Index Terms— Deep Convolutional Neural Networks, Thyroid nodule, Segmentation, Semantic guidance.

1. INTRODUCTION

Thyroid nodules are among the most common endocrine complaints in the adult population and they are clinically important primarily due to their malignant potential [1]. For detection and evaluation of thyroid nodules, ultrasound technology is widely employed to provide information with regard to nodule dimensions, structure, and thyroid parenchymal changes [2]. Segmentation is an essential step to detect and produce region of interest (ROI) of nodules, which is beneficial for analysis in nodule characteristics and forthcoming diagnosis. However, due to the inherent characteristics of ultrasound images, such as attenuation, speckles, shadows, low contrast, and signal loss, it is very challenging to segment thyroid nodules from these images.

The traditional thyroid nodule segmentation methods have been reviewed in [2], which can be roughly divided into three categories: contour and shape based methods, region based methods, and machine learning methods [3, 4]. Due to the powerful ability to abstract high-level semantics from raw images, there is a remarkable progress for thyroid nodule segmentation using deep convolutional neural networks (CNNs). Fully convolutional network (FCN) was first adopted to ensure 2D segmentation output [5]. Due to FCN's consecutive spatial pooling and convolution stride, there is a significant loss in spatial details, which may be harmful segmentation. As a result, UNet based models were proposed [6, 7, 8], where UNet's encoder-decoder architecture allows a gradual recovery of high-resolution output. For other types of images, attention UNet [13] and its extension UNet++ [14] are proposed to generate more sophisticated features from encoding layers to input into decoding layers. In spite of achieving promising results in their proposing datasets, the abstracted features in decoder, especially in the shallow layers, are often subject to noise interference from ultrasound images.

This paper introduces Semantic Guided UNet (SGUNet) to address the above problems. The idea is intuitive and simple. In contrast to the convolution features that can be easily corrupted by noisy pixels, SGUNet is more robust to abstract features from the semantic perspective, which is not apt to be influenced by visual variety of ultrasound images. More specifically, in the decoder, we produce a single-channel pixel-wise semantic map from the high-level features, which can be considered as semantic guidance to help low-level features to obtain more accurate semantic representations. We evaluate our SGUNet on a very challenging thyroid nodule ultrasound dataset, Thyroid Digital Image Database (TDID) [9]. In summary, the main contributions of our paper are three-fold: (1) We develop a semantic guidance module that complements the noise-sensitive high-dimensional decoding features, and allows incremental semantic learning following the layer-wise decoder structure. (2) For ultrasound images without additional annotations from calipers, we provide an additional model choice other than FCN for segmenting thyroid nodule. (3) We validate our method on TDID dataset and achieve the state-of-the-art performance.

2. METHOD

This section first briefly introduces the entire network architecture of SGUNet, and then elaborates on the details of SGM.

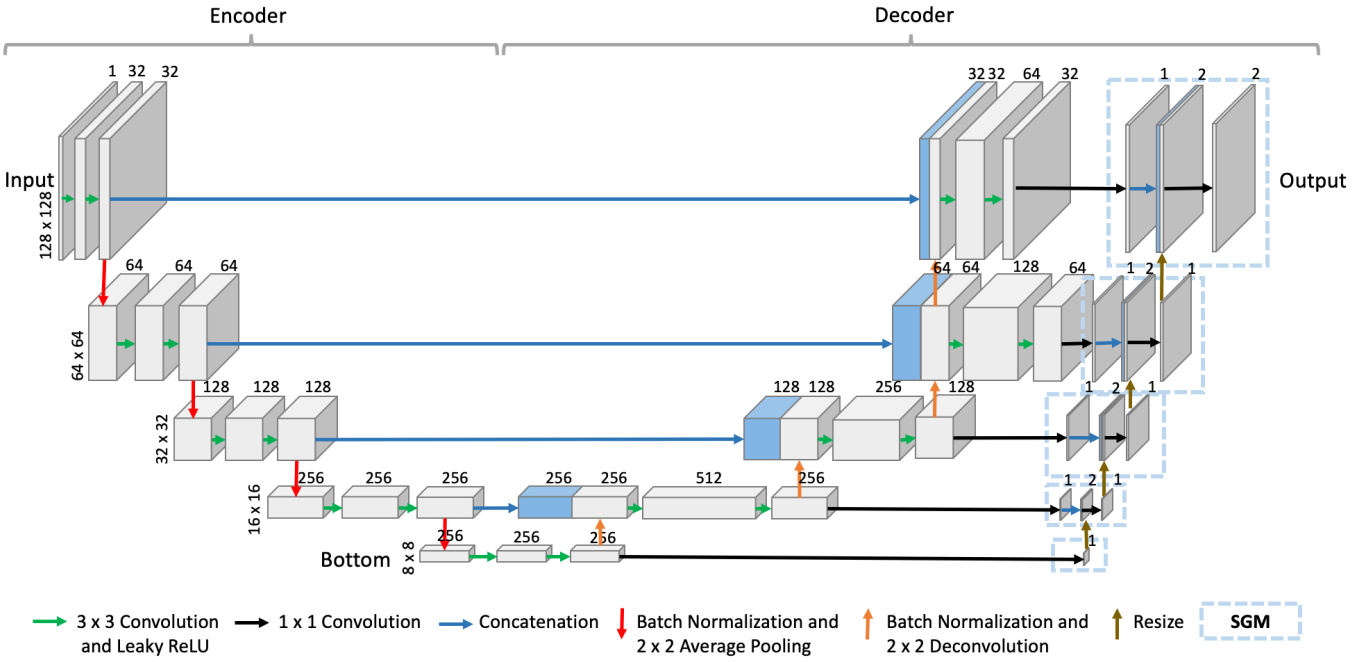

Fig. 1. The overall network architecture of our method. (Best viewed in color)

2.1. Overall Network Architecture

The overall structure is shown in Fig. 1. In order to obtain accurate high-resolution segmentation outputs, we adopt UNet [10] as backbone following encoder-decoder architecture, which has been proven very effective and successful for medical image segmentation [10, 11, 12]. As can be seen in Fig. 1, there are five convolutional stages in encoder, leading to the resolution of 1, $\frac{1}{2}$, $\frac{1}{4}$,$\frac{1}{8}$, and $\frac{1}{16}$ with respect to the input image. Correspondingly, in decoder, a series of deconvolutions is performed to gradually recover the resolution of feature maps. Due to a more powerful capacity to represent visual data, one may prefer to employ UNet++ [14] as backbone. This method, however, is at the cost of a complicated network structure, thus requiring more expensive computation. Note we make two major modifications with respect to the original UNet [10]. Firstly, the downsampling operation adopts average-pooling instead of max-pooling, as it is helpful to suppress noise in the features. Secondly, rather than ReLU, we use the leaky ReLU, helping activation to escape from the negative filter responses.

After gathering features from the backbone, we project abstracted high-dimensional features into one-channel pixel-wise feature map using SGM. It serves as a high-level guidance that integrates the semantic information from two adjacent deconvolutional layers step-be-step. Finally, a softmax layer is adopted to produce channel-wise probability maps, which later receive their supervisions from the ground truths. Immediately below, we introduce the details of the proposed SGM.

2.2. SGM

As can be seen from Fig. 1, due to the mirror architecture of UNet [10], the noise interference from ultrasound images can not only be transferred from multiple stages of spatial pooling and strided convolution in encoder, but can also be directly duplicated to the decoder through skipped-connections. Feeding these potentially noisy features to softmax classifier is inadequate for dense estimation problems, especially for thyroid nodule segmentation. To address this problem, as illustrated in the rectangles with blue dashed borders in Fig. 1, a set of SGMs is adopted to transfers high-dimensional features to one channel semantic feature. This low-dimensional embedding can be viewed as a high-level guidance, which highlights the semantically meaningful region and thus reduces noise interference.

SGM consists of three steps. Let $F_l^0 \in R^{W\times H\times C}$ denote the input of SGM at current layer l, where W, H, and C are width, height, and channel numbers, respectively. (1) An 1×1 convolution is applied to F_l^0 to produce one channel semantic feature $F_l^1 \in R^{W\times H\times 1}$. (2) From the previous layer of SGM, we obtain the final sematic feature $F_{l-1}^3 \in R^{\frac{W}{2}\times\frac{H}{2}\times 1}$. Note, F_{l-1}^3 is a high-level feature with half resolution with respect to the low-level feature at current layer. F_{l-1}^3 is upsampled and combined with F_l^1 using concatenation:

$$F_l^2 = F_l^1 \bigodot \mathcal{U}(F_{l-1}^3) \tag{1}$$

where $\bigodot$ denotes the concatenation operation and $\mathcal{U}$ denotes the bilinear upsampling. (3) The stacked two-channel feature

tensor $F_l^2 \in R^{W\times H\times 2}$ is fed into an 1×1 convolution to produce one channel pixel-wise semantic map $F_l^3 \in R^{W\times H\times 1}$. SGM is applied to all decoding layers. For the bottom layer, SGM only has the first step and its F_l^3 equals to F_l^1.

For example, at the layer of 16×16 resolution, we first compress the decoding features with the dimension of 16× 16×256 into a 16×16×1 semantic feature. Then, we upsample the lower layer's semantic feature with dimension 8×8×1 to 16×16×1, concatenate it with the single-channel feature from the first step and obtain a stacked output with dimension 16×16×2. Lastly, we use a 1×1 convolution to obtain the final semantic feature with the dimension of 16×16×1.

One main merit of our SGM lies in the fact that it introduces very small number of additional model parameters, yet yielding more powerful and robust feature representations for the task of thyroid nodule segmentation. Another advantage of SGM is that, compared with UNet [10] and its variant UNet++ [14], it provides a relatively shorter path between prediction output and the deep layers, which is beneficial for gradient propagation. Comparing to attention UNet [13] and UNet++ [14], SGM engineers on the output of decoding layers instead of the input of decoding layers. In this way, SGM provides a relatively shorter path between prediction output and the deep layers, which is beneficial for gradient propagation and closing the semantic gap between high and low dimension decoding layers.

3. EXPERIMENTS

3.1. Dataset

In this work, we evaluate our SGUNet on Thyroid Digital Image Database (TDID), which is a public dataset for thyroid nodule segmentation created by Universidad Nacional de Colombia [9]. The dataset consists of 400 sets of B-mode Ultrasound images, including a complete annotation and diagnostic description of suspicious thyroid lesions by expert radiologists. In total, the TDID dataset contains ultrasound thyroid images of 298 patients. For each patient, one or more ultrasound thyroid images were captured. All the original images have a shape of 560×360 pixels.

In terms of data cleaning, we first removed the samples without proper annotation. Then, we split the images with two views into two samples, and crop out the side annotation from the original ultrasound images, resulting in total 420 images, where 358 of them are malign cases and 62 are benign cases. Some cleaned images are shown in the first row of Fig. 2.

3.2. Evaluation metrics and baselines

For all our experiments, we conduct 5-fold cross-validation where the data is split into training 80% and validation 20%, and report the average validation results. The evaluation metrics we adopt are Recall, Specificity, Precision, Dice Score, Intersection-over-Union (IoU), and Accuracy.

Table 1. Average validation performance for all cases

Model	Parm.	Rec	Spec	Prec	Dice	IoU	Acc
UNet	29,956K	**0.784**	0.954	0.718	0.709	0.580	0.924
UNet++	44,931K	0.767	0.960	0.733	0.705	0.574	0.924
SGUNet	29,959K	0.783	**0.962**	**0.753**	**0.729**	**0.604**	**0.931**

In order to show the advantages of SGUNet, we selected 2 state-of-the-art networks as baselines for comparison, including UNet [10] and UNet++ [14]. Results for attention UNet [13] is not shown as its results is worse than the baseline UNet in most perspectives.

3.3. Implementation

In all experiments, the batch size equals to the total number of training sample, which is 336. The input images are all resized to 128×128 pixels. The Adam optimizer is used with a learning rate of 0.0001. All models are trained for 3000 epochs. The models are implemented based on the Tensorflow framework and trained with a NVIDIA GeForce GTX 1080 GPU. We use a loss combining Dice loss (L_{Dice}) and cross entropy loss(L_{CE}). It is defined as:

$$L = 0.5L_{Dice} + 0.5L_{CE} \tag{2}$$

For data augmentation, we randomly apply one or a combination of the following methods in training: left-right flipping, shifting, noise addition, contrast adjusting, rotating, and elastic distortion. Code for SGUNet can be viewed through https://github.com/Jo-Pan/SGUNet.

3.4. Quantitative results

Table 1 reports the quantitative results and the parameter size in bytes for UNet [10], UNet++ [14] and our SGUNet. SGUNet outperforms UNet and UNet++ across all evaluation metrics except recall. Comparing to UNet, which is the baseline, SGUNet improves 2.0% in Dice score and 2.4% in IoU, with only small increase in parameter size.

Comparing UNet++ with UNet, UNet++ under-performs in terms of Dice, IoU and recall. By having additional dense connection, UNet++ does not improve performance while increasing number of parameters significantly. There are two possible causes for UNet++'s under-performance. First, TDID may be more challenging than the evaluated datasets by UNet++ [14], as TDID images often have high amount of noise and unclear boundaries between foreground and background. Second, as there is a limited amount of useful features from the input image, the high-dimensional computation from dense connections make UNet++ vulnerable to noise interference.

Table 2. Average validation performance for malign and benign cases

Group	Model	Rec	Spec	Prec	Dice	IoU	Acc
Malign	UNet [10]	**0.792**	0.956	0.717	0.714	0.585	0.929
	UNet++ [14]	0.772	0.962	0.730	0.708	0.577	0.930
	SGUNet	0.798	**0.965**	**0.753**	**0.736**	**0.611**	**0.936**
Benign	UNet [10]	0.736	0.941	0.719	0.678	0.550	0.895
	UNet++ [14]	0.736	0.946	0.746	0.686	0.553	0.895
	SGUNet	**0.737**	**0.947**	**0.749**	**0.687**	**0.564**	**0.901**

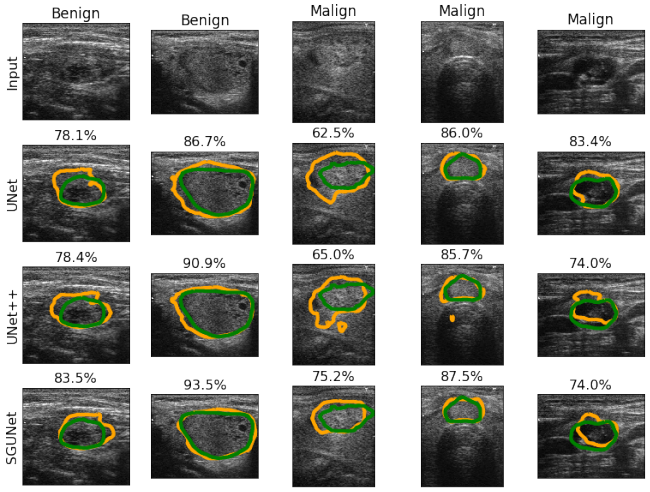

Fig. 2. Sample test predictions. Orange contour is the predicted nodule segmentation. Green contour is the ground truth (GT) segmentation. (Best viewed in color)

For both groups of benign and malign cases, as shown in Table 2, SGUNet consistently outperform the baseline UNet and UNet++ as well. For malign group, SGUNet outperforms UNet by 2.2% and 2.6% in terms of Dice Coefficient and IoU, respectively. For benign group, SGUNet outperforms UNet by 0.9% and 1.4% in terms of Dice Coefficient and IoU. As the dataset is heavily biased toward malign cases, it is reasonable to see a better performance on the malign group.

3.5. Qualitative results

We visualize prediction results for both malign and benign cases in Fig. 2. SGUNet's prediction tends to be more focus around the nodule while both UNet and UNet++ have cases with small disconnected prediction regions or uneven contour edges. To demonstrate the challenging nature of the chosen dataset, the last column of Fig. 2 shows one of the most challenging samples where SGUNet's prediction can be a reasonable segmentation from non-expert perspective as the predicted contour follows one of the apparent contours in the image.

In order to gain a better understanding of the proposed semantic guidance module, we also visualize the output semantic features from each SGM at all decoding layers for

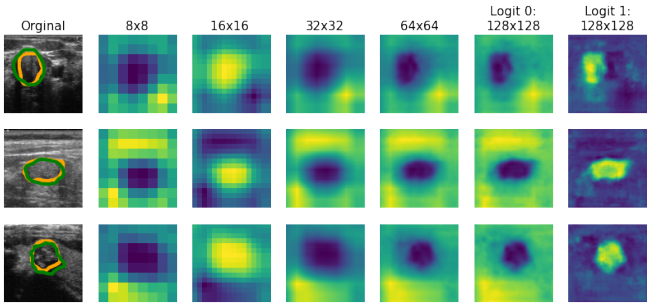

Fig. 3. All SGMs' output semantic features from resolution 8×8 pixels to 128×128 pixels. In the first column, input images, the orange contours are the final prediction from the model. The green contours are the ground truth segmentation. Logit 0 and Logit 1 represents the two channels of the predicted logits outputs from the last SGM. (Best viewed in color)

two testing examples in Fig. 3. As expected, the semantic feature gains more details incrementally through the upsampling process. On the other hand, we observe that semantic features are not consistently representing either the nodule or the background. We also notice that, for all testing samples, except the final SGM outputting the logit output, it is always the case that only resolution 16×16 has a semantic feature representing the nodule while the rest of the semantic features at other resolutions representing the background. This may mean that the model finds features from resolution 16×16 provide most indicative information about the nodule, while at other resolutions, the model finds more reliable features representing the background.

4. CONCLUSIONS

We have presented SGUNet, a novel segmentation network that makes use of semantic features to solve a challenging thyroid nodule segmentation problem. To incorporate semantic features, we proposed a SGM to abstract semantic features at each decoding layer and then use them as a high-level guidance to low-level features. We evaluated our method on a challenging thyroid nodule dataset, TDID, with high noise, blurry boundaries and no calipers. The Experimental results demonstrate that our proposed method consistently outperforms state-of-the-art approaches, such as UNet and UNet++. We also proved the validity of SGM through visualization of the learnt semantic features. In the future, in addition to achieving promising results for thyroid nodule segmentation, we believe that our method can be easily transferred to any existing network architectures that is used for other medical image segmentation tasks.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by University Nacional de Colombia [9]. Ethical approval was not required as confirmed by the license attached with the open access data.

6. ACKNOWLEDGMENTS

This work is in part supported by the NSF under Grant No. IIS-1814745.

7. REFERENCES

[1] Popoveniuc, G., Jonklaas, J.: Thyroid Nodules. The Medical clinics of North America, 96(2), 329–349 (2012). 10.1016/j.mcna.2012.02.002

[2] Chen, J., You, H., Li, K.: A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images. Computer Methods and Programs in Biomedicine, Vol. 185 (2020).

[3] Chang, C., Huang, H., Chen, S.: Thyroid nodule segmentation and component analysis in ultrasound images, in: Proceedings of APSIPA Annual Summit and Conference, pp. 910–917 (2009).

[4] Keramidas, E.G., Maroulis, D. Iakovidis, D.K.: ND: A Thyroid Nodule Detection System for Analysis of Ultrasound Images and Videos. J Med Syst 36, 1271–1281 (2012).

[5] Ma, J., Wu, F., Jiang, T. Zhao, Q., Kong, D.: Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. Int J CARS 12, 1895–1910 (2017).10.1007/s11548-017-1649-7

[6] Ying X., Yu, Z., Yu, R.: Thyroid Nodule Segmentation in Ultrasound Images Based on Cascaded Convolutional Neural Network. Neural Information Processing, pp. 373–384 (2018).10.1007/978-3-030-04224-0_32

[7] Buda, M., Tobriner B., Castor, K., Hoang, J., Mazurowski, M.: Deep Learning-Based Segmentation of Nodules in Thyroid Ultrasound: Improving Performance by Utilizing Markers Present in the Images. Ultrasound in Medicine Biology, vol. 46, issue 2, pp. 415-421 (2020).

[8] Yu, R., Liu, K., Wei, X., Zhu, J., Li, X., Wang, J., Ying, X., Yu, Z.: Localization of thyroid nodules in ultrasonic images, in: Proceedings of International Conference on Wireless Algorithms, Systems, and Applications, pp. 635–646 (2018).

[9] Pedraza L., Vargas C., Narvaez F., Duran O., Munoz E., Romero E.: An open access thyroid ultrasound-image database. Proceedings of the 10th International Symposium on Medical Information Processing and Analysis, pp. 1–6 (2015). http://cimalab.intec.co/?lang=enmod=programid=5.

[10] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI 2015, LNCS, vol. 9351, pp. 234–241. (2015).10.1007/978-3-319-24574-4_28

[11] Dong, H., Yang, G., Liu, F., Mo, Y., Guo, Y.: Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks. In: Valdés Hernández M., González-Castro V. (eds) Medical Image Understanding and Analysis. Communications in Computer and Information Science, vol 723. Springer, Cham (2017).

[12] Norman, B., Pedoia, V., Majumdar, S.: Use of 2D U-Net Convolutional Neural Networks for Automated Cartilage and Meniscus Segmentation of Knee MR Imaging Data to Determine Relaxometry and Morphometry. Radiology, 288:1, pp. 177-185 (2018).

[13] Oktay, O., Schlemper, J., Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N., Kainz, B., Glocker, B., Rueckert, D.: Attention U-Net: Learning Where to Look for the Pancreas. The Medical Imaging with Deep Learning conference MIDL (2018).

[14] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018).10.1007/978-3-030-00889-5_1.

[15] 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).