ET-Net: A Generic Edge-aTtention Guidance Network for Medical Image Segmentation

Zhijie Zhang[1], Huazhu Fu[2(✉)], Hang Dai[2], Jianbing Shen[2], Yanwei Pang[1], and Ling Shao[2]

[1] School of Electrical and Information Engineering, Tianjin University, Tianjin, China
[2] Inception Institute of Artificial Intelligence, Abu Dhabi, UAE
huazhu.fu@inceptioniai.org

Abstract. Segmentation is a fundamental task in medical image analysis. However, most existing methods focus on primary region extraction and ignore edge information, which is useful for obtaining accurate segmentation. In this paper, we propose a generic medical segmentation method, called Edge-aTtention guidance Network (ET-Net), which embeds edge-attention representations to guide the segmentation network. Specifically, an edge guidance module is utilized to learn the edge-attention representations in the early encoding layers, which are then transferred to the multi-scale decoding layers, fused using a weighted aggregation module. The experimental results on four segmentation tasks (*i.e.*, optic disc/cup and vessel segmentation in retinal images, and lung segmentation in chest X-Ray and CT images) demonstrate that preserving edge-attention representations contributes to the final segmentation accuracy, and our proposed method outperforms current state-of-the-art segmentation methods. The source code of our method is available at https://github.com/ZzzJzzZ/ETNet.

1 Introduction

Medical image segmentation is an important procedure in medical image analysis. The shapes, size measurements and total areas of segmentation outcomes can provide significant insight into early manifestations of life-threatening diseases. As a result, designing an efficient general segmentation model deserves further attention. Existing medical image segmentation methods can mainly be divided into two categories: edge detection and object segmentation. The edge detection methods first identify object boundaries utilizing local gradient representations, and then separate the closed loop regions as the objects. These methods, which aim to obtain highly localized image information, can achieve high accuracy in boundary segmentation, and are adequate for simple structures. For example, the level-set technique is employed to minimize an objective function for estimating tumor segmentation based on shape priors [16]. The

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11764, pp. 442–450, 2019.
https://doi.org/10.1007/978-3-030-32239-7_49

template matching method is proposed to obtain optic disc boundary approximations with the Circular Hough Transform in retinal images [1]. Other edge detection methods are employed to extract blood vessels in retinal images [6,11]. However, these edge detection methods depend on local edge representations and lack object-level information, which leads to trivial segmentation regions and discontinuous boundaries. By contrast, object segmentation methods [7,18] utilize global appearance models of foregrounds and backgrounds to identify the target regions, which preserves the homogeneity and semantic characteristics of objects, and reduces the uncertainties in detecting the boundary positions. A common way of doing this is to classify each pixel/patch in an image as foreground or background. For example, a superpixel classification method was proposed to segment the optic disc and cup regions for glaucoma screening [4]. However, without utilizing edge information, several object segmentation methods need to refine the initial coarse segmentation results using additional post-processing technologies (*e.g.*, Conditional Random Field and shape fitting), which is time-consuming and less related to previous segmentation representations.

Recently, the success of U-Net has significantly promoted widespread applications of segmentation on medical images, such as cell detection from 2D image [12], vessel segmentation from retinal images [6] and lung region extraction from chest X-Ray and CT images [10]. However, there are still several limitations when applying Deep Convolutional Neural Networks (DCNNs) based on a U-Net structure. In medical image segmentation, different targets sometimes have similar appearances, making it difficult to segment them using a U-Net based DCNN. Besides, inconspicuous objects are sometimes over-shadowed by irrelevant salient objects, which can confuse the DCNNs, since it cannot extract discriminative context features, leading to false predictions. In addition, target shapes and scale variations are difficult for DCNNs to predict. Although U-Net proposes to aggregate high-level and low-level features to address this problem, it only slightly alleviates it, since it aggregates features of different scale without considering their different contributes. Herein, we propose a novel method to extract discriminative context features and selectively aggregate multi-scale information for efficient medical image segmentation.

In this paper, we integrate both edge detection and object segmentation in one deep learning network. To do so, we propose a novel general medical segmentation method, called Edge-aTtention guidance Network (ET-Net), which embeds edge-attention representations to guide the process of segmentation. In our ET-Net, an edge guidance module (EGM) is provided to learn the edge-attention representations and preserve the local edge characteristics in the early encoding layers, while a weighted aggregation module (WAM) is designed to aggregate the multi-scale side-outputs from the decoding layers and transfer the edge-attention representations to high-level layers to improve the final results. We evaluate the proposed method on four segmentation tasks, including optic disc/cup segmentation and vessel detection in retinal images, and lung segmentation in Chest X-Ray and CT images. Results demonstrate that the proposed ET-Net outperforms the state-of-the-art methods in all tasks.

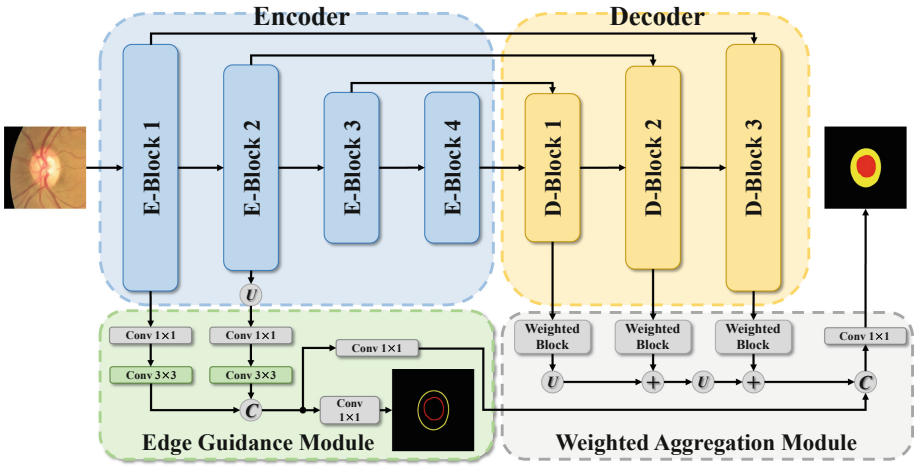

Fig. 1. Illustration of our ET-Net architecture, which includes the main encoder-decoder network with an edge guidance module and weighted aggregation module. 'Conv' denotes the convolutional layer, while 'U', 'C', and '+' denote the upsampling, concatenation, and addition layers, respectively.

2 Method

Figure 1 illustrates the architecture of our ET-Net, which is primarily based on an encoder-decoder network, with the EGM and WAM modules appended on the end. The ResNet-50 [8] is utilized as the encoder network, which comprises of four Encoding-Blocks (E-Blocks), one for each different feature map resolution. For each E-Block, the inputs first go through a feature extraction stream, which consists of a stack of $1 \times 1 - 3 \times 3 - 1 \times 1$ convolutional layers, and are then summed with the shortcut of inputs to generate the final outputs. With this residual connection, the model can generate class-specific high-level features. The decoder path is formed from three cascaded Decoding-Blocks (D-Blocks), which are used to maintain the characteristics of the high-level features from the E-Blocks and enhance their representation ability. As shown in Fig. 2(a), the D-Block first adopts a depth-wise convolution to enhance the representation of the fused low-level and high-level features. Then, a 1×1 convolution is used to unify the number of channels.

2.1 Edge Guidance Module

As stated in Sect. 1, edge information provides useful fine-grained constraints to guide feature extraction during segmentation. However, only low-level features preserve sufficient edge information. As such, we only apply the EGM at the top of early layers, *i.e.*, E-Block 1 and 2, of the decoding path, as shown in Fig. 1. The EGM has two main fashions: (1) it provides an edge-attention representation to

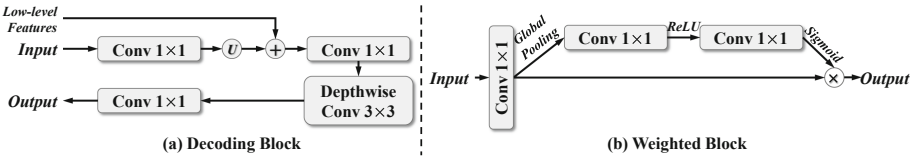

Fig. 2. Illustration of the E-Block and Weighted Block. 'U', '+' and '×' denote upsampling, addition, and multiplication layers, respectively.

guide the process of segmentation in the decoding path; (2) it supervises the early convolutional layers using the edge detection loss.

In our EGM, the outputs of E-Block 2 are upsampled to the same resolution as the outputs of E-Block 1, and then fed into the $1\times1-3\times3$ convolutional layers and concatenated together. After that, the concatenated features go through one of two branches: a 1×1 convolutional layer to act as the edge guidance features in the decoding path, or another 1×1 convolutional layer to predict the edge detection results for early supervision. The Lovász-Softmax loss [2] is used in our EGM, since it performs better than cross entropy loss for class imbalanced problems. It can be formulated as:

$$\mathcal{L} = \frac{1}{C}\sum_{c\in C}\overline{\Delta_{J_c}}(m(c)), \quad \text{and} \quad m_i(c) = \begin{cases} 1 - p_i(c) & \text{if } c = y_i(c), \\ p_i(c) & \text{otherwise,} \end{cases} \tag{1}$$

where C denotes the class number, and $y_i(c) \in \{-1, 1\}$ and $p_i(c) \in [0, 1]$ are the ground truth label and predicted probability of pixel i for class c, respectively. $\overline{\Delta_{J_c}}$ is the Lovász extension of the Jaccard index [2]. With the edge supervision, the transmitted edge features are better able to guide the extraction of discriminative features in high-level layers.

2.2 Weighted Aggregation Module

In order to adapt to the shape and size variations of objects, existing methods tend to sum up multi-scale outputs along the channel dimension for final predictions (*e.g.*, [5,19]). However, not all features in high-level layers are activated and benefit the recovery of objects. Aiming to address this, we develop the WAM to emphasize the valuable features, and aggregate multi-scale information and edge-attention representations to improve the segmentation performance. As shown in Fig. 1, outputs of each D-Block are fed into the Weighted Blocks to highlight the valuable information. The structure of a Weighted Block is shown in Fig. 2(b). In this block, global average pooling is first employed to aggregate the global context information of inputs, and then two 1×1 convolutional layers with different non-linearity activation functions, *i.e.*, ReLU and Sigmoid, are applied to estimate the layer relevance and generate the weights along the channel dimension. After that, the generated weights are multiplied with the outputs to yield more representative features.

Our WAM integrates the features of different scales via a bottom-up pathway, which generates a feature hierarchy consisting of feature maps of different sizes. Finally, the WAM also concatenates edge-attention representations from the EGM, and then applies a 1×1 convolution to extract features under edge-guided conditions. As with the edge detection in the EGM, our WAM also utilizes the Lovász-Softmax loss as the segmentation loss function. Thus, the total loss function of our ET-Net is defined as: $\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{seg} + (1 - \alpha) \cdot \mathcal{L}_{edge}$, where $\mathcal{L}_{edge}$ and $\mathcal{L}_{seg}$ denote the losses for edge detection in EGM and segmentation in WAM, respectively. In our experiments, weight α is empirically set to 0.3.

2.3 Implementation Details

For data augmentation, we apply a *random mirror*, *random scale*, which ranges from 0.5 to 2, and *random rotation* between -10 and $10°$, for all datasets. *random color jitters* with a probability of 0.5 are also applied to the data. All input images are randomly cropped to 512×512.

The initial weights of the encoder network come from ResNet-50 [8] pre-trained on ImageNet, and the parameters of the other layers are randomly initialized. A dilation strategy is used in E-Block 4, with an output stride of 1/16. During training, we set the *batch_size* to 16 with synchronized batch normalization, and adopt 'poly' learning rate scheduling $lr = base\_lr \times (1 - \frac{iters}{total\_iters})^{power}$, in which the power is set to 0.9 and *base_lr* is 0.005. The *total_iters* is calculated by the $num\_images \times epochs/batch\_size$, where *epochs* is set to 300 for all datasets. Our deep models are optimized using the Adam optimizer with a momentum of 0.9 and a weight decay of 0.0005. The whole ET-Net framework is implemented using PyTorch. Training (300 epochs) requires approximately 2.5 h on one NVIDIA Titan Xp GPU. During testing, the segmentation results, including edge detection and object segmentation, are produced within 0.015 s. per image.

3 Experiments

We evaluate our approach on three major types of medical images: retinal images, X-Ray and CT images. For convenience comparison, we select the evaluation metrics that are highly related to generic segmentation.

Optic Disc and Cup Segmentation in Retinal Images: We evaluate our method on optic disc and cup segmentation in retinal images, which is a common task in glaucoma detection. Two public datasets are used in this experiment: the REFUGE[1] dataset, which consists of 400 training images and 400 validation images; the Drishti-GS [14] dataset, which contains 50 training images and 51 validation images. Considering the negative influence of non-target areas in fundus images, we first localize the disc centers following the existing automatic

[1] https://refuge.grand-challenge.org/.

Table 1. Optic disc/cup segmentation results on retinal fundus images.

Method	REFUGE			Drishti-GS		
	Dice_{OC} (%)	Dice_{OD} (%)	mIoU (%)	Dice_{OC} (%)	Dice_{OD} (%)	mIoU (%)
FCN [13]	84.67	92.56	82.47	87.95	95.69	83.92
U-Net [12]	85.44	93.08	83.12	88.06	96.43	84.87
M-Net [5]	86.48	93.59	84.02	88.60	96.58	85.88
Multi-task [3]	86.74	94.01	84.36	88.96	96.55	85.94
pOSAL [17]	87.50	94.60	-	90.10	97.40	-
Our ET-Net	**89.12**	**95.29**	**86.70**	**93.14**	**97.52**	**87.92**

Table 2. Segmentation results on retinal fundus, X-Ray and CT images.

Method	DRIVE		MC		LUNA	
	Acc. (%)	mIoU (%)	Acc. (%)	mIoU (%)	Acc. (%)	mIoU (%)
FCN [13]	94.13	74.55	97.35	90.53	96.18	93.82
U-Net [12]	94.45	75.46	97.82	91.64	96.63	94.79
M-Net [5]	94.58	75.81	97.96	91.95	97.27	94.92
Multi-task [3]	94.97	76.21	98.13	92.24	97.82	94.96
Our ET-Net	**95.60**	**77.44**	**98.65**	**94.20**	**98.68**	**96.23**

disc detection method [5], and then transmit the localized images into our network. The proposed approach is compared with the classic segmentation methods (*i.e.*, FCN [13] U-Net [12], M-Net [5], and Multi-task [3] (it predicts edge and object predictions on the same features.), and the state-of-the-art segmentation method pOSAL [17], which achieved first place for the optic disc and cup segmentation tasks in the REFUGE challenge. The dice coefficients of optic disc (Dice_{OD}) and cup (Dice_{OD}), as well as mean intersection-over-union (mIoU), are employed as evaluation metrics. As shown in Table 1, our ET-Net achieves the best performance on both the REFUGE and Drishti-GS datasets. Our model achieves particularly impressive results for optic cup segmentation, which is an especially difficult task, achieving 2% improvement of Dice_{OC} over the next best method.

Vessel Segmentation in Retinal Images: We evaluate our method on vessel segmentation in retinal images. DRIVE [15], which contains 20 images for training and 20 for testing, is adopted in our experiments. The statistics in Table 2 show that our proposed method achieves the best performance, with 77.44% mIoU and 95.60% accuracy, when compared with classical methods (*i.e.*, U-Net [12], M-Net [5], FCN [13] and Multi-task [3]).

Lung Segmentation in X-Ray Images: We conduct lung segmentation experiments on Chest X-Rays, which is an important component for computer-aided diagnosis of lung health. We use the Montgomery County (MC) [9] dataset,

which contains 80 training images and 58 testing images. We compare our ET-Net with FCN [13], U-Net [12], M-Net [5] and Multi-task [3], in terms of mIoU and accuracy (Acc.) scores. Table 2 shows the results, where our method achieves the state-of-the-art performance, with an Acc. of 98.65% and mIoU of 94.20%.

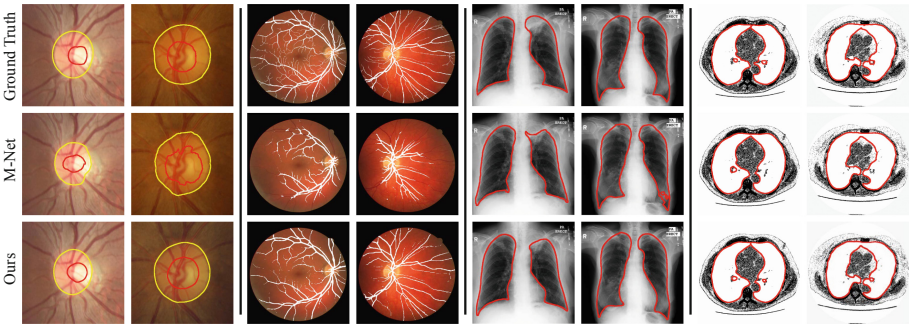

Fig. 3. Visualization of segmentation results. From left to right: optic disc/cup, and vessel segmentation in retinal fundus images, lung segmentation in Chest X-Ray and CT images.

Table 3. Ablation study of optic disc/cup segmentation on the Drishti-GS *test* set

Method	Dice_{OC} (%)	Dice_{OD} (%)	mIoU (%)
Base network	90.11	95.77	84.41
Base network + EGM	91.24	97.17	86.49
Base network + WAM	91.49	97.07	86.62
Base network + EGM + WAM	**93.14**	**97.52**	**87.92**

Lung Segmentation in CT Images: We evaluate our method on lung segmentation from CT images, which is fundamental for further lung nodule disease diagnosis. The Lung Nodule Analysis (LUNA) competition dataset[2] is employed, which is divided into 214 images for training and 53 images for testing. As with the lung segmentation from Chest X-Ray images, we compare our method with FCN [13], U-Net [12], M-Net [5], and Multi-task [3], in terms of mIoU and Acc. scores. The randomly cropped images are fed into the proposed network. As shown in Table 2, our ET-Net outperforms previous state-of-the-art methods, obtaining an Acc. of 98.68% and mIoU of 96.23%.

In addition to quantitative results, we provide qualitative segmentation results, shown in Fig. 3. As can be seen, our results are close to the ground truth. When compared with the predictions of other methods, it is clear that our results are better, especially, in the edge regions.

[2] https://www.kaggle.com/kmader/finding-lungs-in-ct-data/data.

3.1 Ablation Study

To evaluate the contributions of each component of the proposed method, we conduct experiments with different settings on the Drishti-GS dataset. As shown in Table 3, we choose the encoder-decoder network shown in Fig. 1 as the base network, which achieves 90.11%, 95.77% and 84.41% in terms of Dice_{OC}, Dice_{OD} and mIoU, respectively. When we append the proposed EGM, it yields results of 91.24%/97.17%/86.49% (Dice_{OC}/Dice_{OD}/mIoU). This dramatically outperforms the base network, with only a small addition to the computational cost, proving that edge information is of vital importance for segmentation. To study the effect of the WAM, we append the WAM to the base network, without concatenating the edge features to base network. With the same training settings, this approach achieves performances of 91.49%/97.07%/86.62%, compared to the base network. The obvious performance gains for all three metrics illustrate the efficiency of the proposed the WAM. Finally, our whole ET-Net, with both EGM and WAM, obtains the best performance on the Drishti-GS *test* set.

4 Conclusion

In this paper, we propose a novel Edge-aTtention Guidance network (ET-Net) for general medical image segmentation. By assuming that edge detection and region segmentation are mutually beneficial, we have proposed the Edge Guidance Module to detect object edges and generate edge-attention representations that contain sufficient edge information. Moreover, a Weighted Aggregation Module has been employed to highlight the valuable features of high-level layers, which are combined with the edge representations, to guide the final segmentation. Experiments on various medical imaging tasks have demonstrated the superiority of our proposed ET-Net compared to other state-of-the-art methods. In future work, we will extend our approach to 3D segmentation on CT and MRI volumes.

References

1. Aquino, A., Gegundez-Arias, M.E., Marin, D.: Detecting the optic disc boundary in digital fundus images using morphological, edge detection, and feature extraction techniques. IEEE TMI **29**(11), 1860–1869 (2010)
2. Berman, M., Rannen Triki, A., Blaschko, M.B.: The Lovász-Softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: CVPR (2018)
3. Chen, H., Qi, X., et al.: DCAN: deep contour-aware networks for accurate gland segmentation. In: CVPR (2016)
4. Cheng, J., Liu, J., et al.: Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. IEEE TMI **32**(6), 1019–1032 (2013)
5. Fu, H., Cheng, J., et al.: Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE TMI **37**(7), 1597–1605 (2018)

6. Fu, H., Xu, Y., Lin, S., Kee Wong, D.W., Liu, J.: DeepVessel: retinal vessel segmentation via deep learning and conditional random field. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 132–139. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_16
7. Gu, Z., et al.: CE-Net: context encoder network for 2D medical image segmentation. IEEE TMI (2019, in press). https://doi.org/10.1109/TMI.2019.2903562
8. He, K., Zhang, X., et al.: Deep residual learning for image recognition. In: CVPR (2016)
9. Jaeger, S., Candemir, S., et al.: Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. QIMS **4**(6), 475–477 (2014)
10. Mansoor, A., Bagci, U., et al.: Segmentation and image analysis of abnormal lungs at CT: current approaches, challenges, and future trends. Radiographics **35**(4), 1056–1076 (2015)
11. Moccia, S., Momi, E.D., et al.: Blood vessel segmentation algorithms - review of methods, datasets and evaluation metrics. CMPB **158**, 71–91 (2018)
12. 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
13. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. TPAMI **39**(4), 640–651 (2017)
14. Sivaswamy, J., Krishnadas, S.R., et al.: Drishti-GS: retinal image dataset for optic nerve head (ONH) segmentation. In: IEEE ISBI (2014)
15. Staal, J., Abràmoff, M.D., et al.: Ridge-based vessel segmentation in color images of the retina. IEEE TMI **23**(4), 501–509 (2004)
16. Tsai, A., Yezzi, A., et al.: A shape-based approach to the segmentation of medical imagery using level sets. IEEE TMI **22**(2), 137–154 (2003)
17. Wang, S., Yu, L., et al.: Patch-based output space adversarial learning for joint optic disc and cup segmentation. IEEE TMI (2019, in press). https://doi.org/10.1109/TMI.2019.2899910
18. Wang, W., Lai, Q., et al.: Salient object detection in the deep learning era: an in-depth survey. arXiv:1904.09146 (2019)
19. Wang, W., Shen, J., Ling, H.: A deep network solution for attention and aesthetics aware photo cropping. IEEE PAMI **41**(7), 1531–1544 (2019)