AEC-Net: Attention and Edge Constraint Network for Medical Image Segmentation

Jingyi Wang[1], Xu Zhao[2], Qingtian Ning[1] and Dahong Qian[3]

***Abstract*— Semantic segmentation is a fundamental and challenging problem in medical image analysis. At present, deep convolutional neural network plays a dominant role in medical image segmentation. The existing problems of this field are making less use of image information and learning few edge features, which may lead to the ambiguous boundary and inhomogeneous intensity distribution of the result. Since the characteristics of different stages are highly inconsistent, these two cannot be directly combined. In this paper, we proposed the Attention and Edge Constraint Network (AEC-Net) to optimize features by introducing attention mechanisms in the lower-level features, so that it can be better combined with higher-level features. Meanwhile, an edge branch is added to the network which can learn edge and texture features simultaneously. We evaluated this model on three datasets, including skin cancer segmentation, vessel segmentation, and lung segmentation. Results demonstrate that the proposed model has achieved state-of-the-art performance on all datasets.**

I. INTRODUCTION

Currently, an increasing number of excellent solutions based on deep learning have been proposed in medical image segmentation. U-Net [1] is undoubtedly one of the most successful methods. In U-Net, however, there is a lot of redundant information in the low-level features, which is greatly different from the high-level features. Hence, it is not appropriate to combine these two without modification. Moreover, U-Net does not explicitly extract edge features. While in [5], the author concludes that convolutional neural network(CNN) actually learns texture features rather than shape features. The robustness of network can be enhanced by improving the shape learning ability of network.

Recently, attention mechanism is proposed to integrate two different characteristics in a better way, which has increasingly become a powerful tool for deep neural networks. The Context Encoding Module is introduced in EncNet [2] to capture global context information. Similarly, the attention concept is leveraged into medical image segmentation in [12], [13] as well. Attention U-Net [3] proposes the attention mechanism in U-Net. Before splicing the corresponding features in the encoder and decoder modules, an attention block is used to readjust the output characteristics of the encoder. The module generates a gating signal that controls the importance of features at different spatial locations. Although it has brought about an improvement in the results, there are still some problems of importing a large amount of redundant information and increasing the parameters of the network. Similar to Parsenet [4], they all make use of the rich semantic information of high-level features to adjust low-level features. Nevertheless, high-level features inherently lack spatial information, using them to filter low-level features can have unsatisfactory results.

ET-Net [6] combines edge detection and semantic segmentation, which can use edge information to monitor and guide the process of segmentation. Nonetheless, it utilizes too few features, and the network structure for extracting edge information is quite simple. Gated-SCNN [7] designs a new two-stream CNN architecture. It adds a separate branch to explicitly process shape information, which is in parallel with classic streams. However, it doesn't have skip connection and U-shaped structure, which are extremely significant for medical images.

To solve the above-mentioned problems, we propose the Distillation Attention Module (DAM), which only uses the information of low-level features to complete the information screening. Without convolution operation, we reduce the amount of parameters. Meanwhile, we are capable of achieving the same or even better performance than the above methods. Moreover, we introduce an edge branch based on the U-Net network structure, and design an edge detection module to optimize the edge learning ability of it through the attention mechanism.

The contribution of this paper are three folds:

- We introduce an attention mechanism in the network, which can boost the valid low-level features, and reduce the parameters in the model to make the network lightweight.
- We propose an edge feature structure for shape learning, in order to make sure that the current mainstream network structure fully utilize existing information.
- We introduce style transfer [8] to further strengthen the edge learning ability.

Our proposed model has been evaluated on three datasets. They all outperform the state-of-the-art methods, indicating that our network has good generalization capability.

*This research has been supported in part by the funding from NSFC programs (61673269, 61273285) and in part by the project funding from Institute of Medical Robotics, Shanghai Jiao Tong University.

[1]Jingyi Wang and Qingtian Ning are with the Department of Automation, Shanghai Jiao Tong University, Shanghai, 200240, China.

[2]Xu Zhao is with the Department of Automation, Shanghai Jiao Tong University, also with the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, 200240, China. zhaoxu@sjtu.edu.cn

[3]Dahong Qian is with Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai 200240, China. dahong.qian@sjtu.edu.cn

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

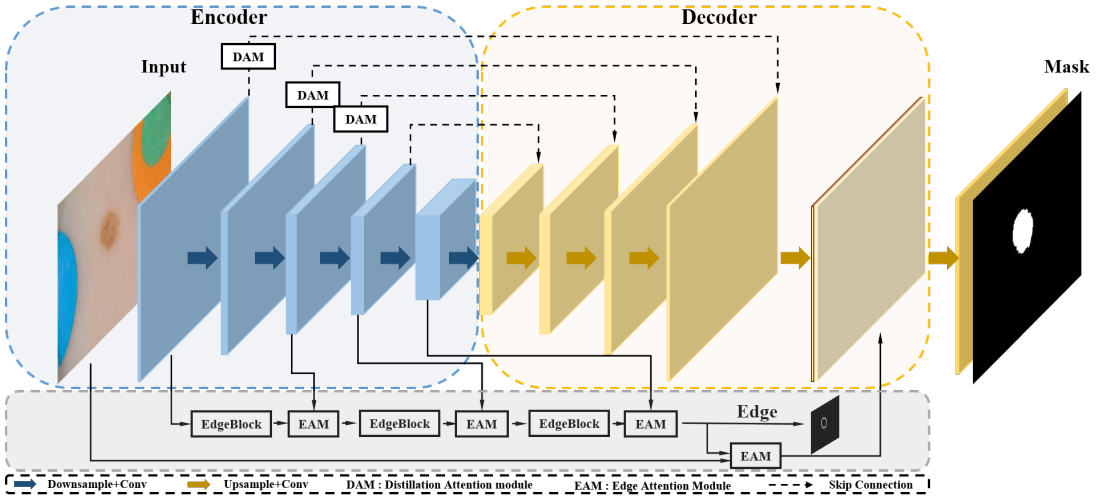

Fig. 1. An overview of the AEC-Net.

II. METHOD

A. Overview

The architecture of our model is illustrated in Fig. 1. In the encoder network, we employ ResNeXt101 [21] to extract the feature map with different resolutions. Then, in the decoder network, we use bilinear upsampling operations to restore the resolution of the feature maps. In order to further enhance the feature representations and model long-range dependencies, we attempt to combine the features of corresponding stages to improve the discriminative ability of feature representations for pixel-level recognition. Instead of combining high-level features with low-level features directly, we first use DAM to filter redundant information and extract more discriminating features from shallow layers. Furthermore, we add an edge branch in the early encoding layers and force the network to learn the shape information of the object better. The edge branch is composed of several EdgeBlocks and Edge Attention Modules(EAM). In the EdgeBlock, we use BasicBlock from ResNet and bilinear interpolation, followed by a convolution operation to obtain the features. In order to guide the edge branch learn more important information, EAM is designed to utilize the consistency of higher stages. Finally, the segmentation map and the edge map are combined together, and then a convolution operation is carried out to achieve the best prediction.

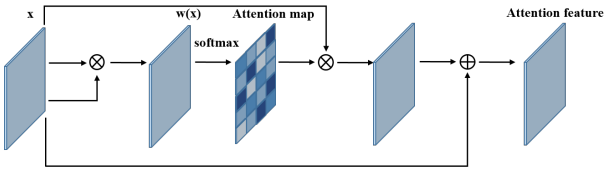

Fig. 2. Illustration of the distillation attention module (DAM)

B. Distillation Attention Module

Shallow layers encode finer spatial features and have more detailed information, while they are weak in semantic information. Deeper layers have larger receptive view and richer semantic information, but they lose a lot of details. In order to make better use of their respective advantages, we utilize low stage information to help high stage information to refine the spatial information and restore image details. Nevertheless, due to the different scales of receptive views, combining them directly will lead to inconsistent results. As a consequence, we design a distillation attention module to accomplish information filtering, which can select the discriminative and effective features. As shown in Fig. 2, we first multiply the low-level features by themselves to get the dependence between pixels, and then obtain the attention map through a softmax layer. The larger the softmax value, the more reliable and stronger the relative dependence. Then we perform a matrix multiplication between the attention map and low-level features. Eventually, we perform an element-wise sum operation on the new features with the low-level features to obtain the attention feature:

$$f(x) = \sigma(w(x)) \times x + x, \tag{1}$$

where x is the input feature of DAM, $w(x)$ represents the weight map and σ denotes a softmax function.

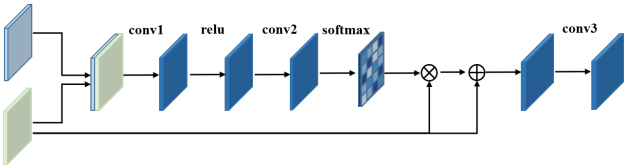

Fig. 3. Illustration of the edge attention module (EAM)

C. Edge Attention Module

In medical images, different tissues and organs have different shapes, as well as normal and diseased parts. In

order to make the features of higher layers guide the extraction of edge features better, we propose the edge attention module. As illustrated in Fig. 3, the output of EdgeBlock and the higher features are concatenated first. Then, we use a 1×1 convolutional layer to integrate features. A ReLU activation function is used to introduce nonlinearity, followed by another convolution operation to unify the number of feature channels. After that, we use a softmax layer to get the attention map. Then, we multiply it by the output of EdgeBlock and add it to the output of EdgeBlock. At the end of EAM, a convolutional layer is used to produce the final attentional features. Since we input characteristics of different stages, multi-scale information is retained.

D. Loss Function

Standard Dice loss is used for each output of predicted semantic segmentation network and predicted boundary network, which is defined as:

$$L_{dice}(y, \hat{y}) = 1 - \frac{2|y \cap \hat{y}|}{|y| + |\hat{y}|}, \tag{2}$$

where y and $\hat{y}$ are the ground truth and predicted image.

The total loss L_t can be formulated as:

$$L_t = \lambda_1 L_{mask} + \lambda_2 L_{edge}, \tag{3}$$

where λ_1 and λ_2 are the weight of two terms, satisfied with $\lambda_1 = 0.3$ and $\lambda_2 = 0.7$ in this paper.

III. EXPERIMENTS AND RESULTS

To evaluate our method, we have used three different public medical imaging datasets: ISIC2017 [9], DRIVE [10], and LUNA [11]. The final results on these datasets are shown in Fig. 4, which prove that our method has good generalization performance. It can be seen that our results are very close to the ground truth. The edges are smooth and there are no artificially marked burrs. Even in edge regions, our method can perform well.

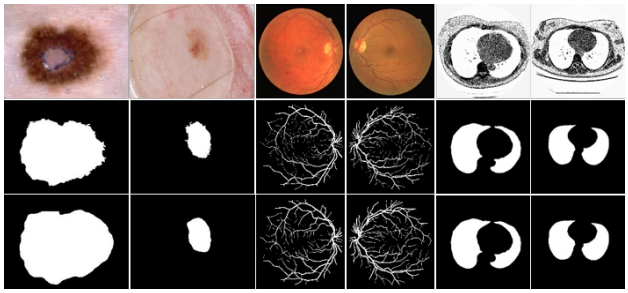

Fig. 4. Visualization of segmentation results on three datasets. From left to right: ISIC2017, DRIVE and LUNA. From top to bottom: input image, ground truth and predictions.

A. Datasets and Evaluation Metrics

ISIC-2017: The 2017 International Skin Imaging Collaboration (ISIC) skin lesion segmentation challenge dataset is a well-known skin lesion dataset, which contains 2000 training images, 150 validation images, and 600 test images. In our experiment, the images and their corresponding segmentation masks are resized to 192×256 pixels. In order to accelerate the convergence of training, the initial weights of the encoder network come from ResNeXt101 pretrained on ImageNet.

DRIVE: The DRIVE dataset is one of the most commonly used retinal databases, which consists of 40 color images. There are 20 images for training and 20 images for testing. We resize each image to 576×544 pixels. The weights of the encoder network are initialized using He initialization [10].

LUNA: The Lung Nodule Analysis (LUNA) competition dataset consists of 2D and 3D CT images, and the size of each image is 512×512. We use 267 2D samples and divided them into 214 images for training and 53 images for testing. As with the previous dataset, we use He initialization [10] to initialize the encoder network.

Evaluation metrics: Accuracy (AC), sensitivity (SE), specificity (SP), F1-score, dice coefficient (DI), Jaccard similarity (JA), and the area under receiver operating characteristics curve (AUC) are used to evaluate the performance of the models.

B. Implementation Details

For data augmentation, random rotations between -10 and 10 degrees, random color jitters with a probability of 0.5, horizontal and vertical flips are applied to the data. Moreover, we also subtract the mean value of image from every RGB channels.

Our whole framework is implement on PyTorch. We use the Adam optimizer to train it and the initial learning rate is set to 0.0002. When the absolute improvement of the evaluation index is smaller than 10^{-3} during the last 20 epochs, the learning rate is gradually reduced by a factor of 0.9 to improve the network performance. As for the batch size, we set it to 16 in the ISIC2017 dataset and set it to 1 in the other two datasets.

TABLE I

RESULTS OF AEC-NET AND OTHER METHODS ON ISIC2017 DATASET.

Method	year	AC	SE	SP	JA	DI
UNet [1]	2015	0.923	0.808	0.978	0.739	0.828
ECDN [14]	2017	0.934	0.825	0.975	0.765	0.849
SSPFCN [15]	2018	0.938	0.855	0.973	0.773	0.857
SLSDeep [16]	2018	0.936	0.816	**0.983**	0.782	0.878
ours	2019	**0.941**	**0.861**	0.979	**0.791**	**0.881**

C. Performance evaluation of medical image segmentation

The proposed approach is compared with several recently published skin lesion segmentation methods. As shown in Table I, our model surpasses the ISIC-2017 challenge winner (ECDN) by 2.6% and 3.2% with respect to JA and DI. We also carry out the best performance and improve JA by 0.9% when comparing to SLSDeep[16]. The quantitative results on DRIVE and LUNA datasets are shown in Table II and Table III. From Table II, it can be observed that our model is superior to the performance of representative state-of-the-art vessel segmentations method, with 82.88% F1-score and

96.74% accuracy. Table III shows that the proposed approach achieves the best performance on LUNA dataset. Compared with ET-Net[6], our method improves F1-score by 0.67% and accuracy by 0.81%.

TABLE II

RESULTS OF AEC-NET AND OTHER METHODS ON DRIVE DATASET.

Method	Year	F1-score	SE	SP	AC	AUC
R2U-Net [17]	2018	0.8171	0.7792	0.9813	0.9556	0.9784
LadderNet [18]	2018	0.8202	0.7856	0.9810	0.9561	**0.9793**
Unsupervised Ensemble [19]	2019	0.8225	0.8072	0.9780	0.9559	0.9779
Dual Encoding U-Net [20]	2019	0.8270	0.7940	0.9816	0.9567	0.9772
Ours	2020	**0.8288**	**0.8173**	**0.9821**	**0.9674**	0.9776

TABLE III

RESULTS OF AEC-NET AND OTHER METHODS ON LUNA DATASET.

Method	Year	SE	SP	JA	F1-score	AC
Residual UNet [17]	2018	0.9555	0.9945	0.9850	0.9690	0.9849
Recurrent UNet [17]	2018	0.9734	0.9866	0.9836	0.9638	0.9836
R2U-Net [17]	2018	0.9832	0.9944	0.9918	0.9823	0.9918
ET-Net [6]	2019	0.9811	0.9887	0.9922	0.9799	0.9868
Ours	2020	**0.9898**	**0.9954**	**0.9926**	**0.9866**	**0.9949**

D. Ablation Study

To verify the contributions of each component of our method, we perform an ablation study with different settings on the ISIC2017 dataset. The result is shown in Table IV. We can observe that DAM achieves 1.47% improvement in performance in terms of JA and 1.05% improvement in terms of DI, which confirms the effectiveness of DAM. The edge branch also significantly improves the performance, which reaches 78.62% and 87.53% in relation to JA and DI, respectively. Finally, we take advantage of style transfer [8] for data augmentation, which can generate images with conicting shape and texture information. In our experiment, ten different styles of images are generated for each original image. Ultimately, we improve on several metrics.

TABLE IV

ABLATION STUDY EXPERIMENT

Method	SE	SP	JA	DI
Baseline	0.8222	0.9703	0.7516	0.8389
Baseline + DAM	0.8168	0.9796	0.7663	0.8494
Baseline + DAM + edge	0.8337	0.9851	0.7862	0.8753
Baseline + DAM + edge + style transfer [8]	0.8612	0.9793	0.7914	0.8814

IV. CONCLUSIONS

In this paper, we present an inventively network model that can learn both texture and edge features. We introduce a creatively low-level feature attention mechanism in the encoder stage to optimize the network, which has solved some of the problems existing in current mainstream medical images. The results show that the new attention model, as well as the learning of both edges and textures, have performed the desired results. Furthermore, We do not use any post-processing techniques and implement an end-to-end approach. In future work, we will make better use of prior information and apply the attention mechanism to semi-supervised learning.

REFERENCES

[1] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[2] Zhang H, Dana K, Shi J, et al. Context encoding for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7151-7160.

[3] Oktay O, Schlemper J, Folgoc L L, et al. Attention u-net: Learning where to look for the pancreas[J]. arXiv preprint arXiv:1804.03999, 2018.

[4] Liu W, Rabinovich A, Berg A C. Parsenet: Looking wider to see better[J]. arXiv preprint arXiv:1506.04579, 2015.

[5] Geirhos R, Rubisch P, Michaelis C, et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness[J]. arXiv preprint arXiv:1811.12231, 2018.

[6] Zhang Z, Fu H, Dai H, et al. ET-Net: A Generic Edge-aTtention Guidance Network for Medical Image Segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019: 442-450.

[7] Takikawa T, Acuna D, Jampani V, et al. Gated-scnn: Gated shape cnns for semantic segmentation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 5229-5238.

[8] Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1501-1510.

[9] Codella N C F, Gutman D, Celebi M E, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)[C]//2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, 2018: 168-172.

[10] Staal J, Abràmoff M D, Niemeijer M, et al. Ridge-based vessel segmentation in color images of the retina[J]. IEEE transactions on medical imaging, 2004, 23(4): 501-509.

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

[12] Isensee F, Petersen J, Klein A, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation[J]. arXiv preprint arXiv:1809.10486, 2018.

[13] Nie D, Gao Y, Wang L, et al. ASDNet: Attention based semi-supervised deep networks for medical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2018: 370-378.

[14] Yuan Y, Lo Y C. Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks[J]. IEEE journal of biomedical and health informatics, 2017, 23(2): 519-526.

[15] Mirikharaji Z, Hamarneh G. Star shape prior in fully convolutional networks for skin lesion segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2018: 737-745.

[16] Sarker M M K, Rashwan H A, Akram F, et al. SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2018: 21-29.

[17] Alom M Z, Hasan M, Yakopcic C, et al. Recurrent residual convolutional neural network based on u-net (R2U-net) for medical image segmentation[J]. arXiv preprint arXiv:1802.06955, 2018.

[18] Zhuang J. LadderNet: Multi-path networks based on U-Net for medical image segmentation[J]. arXiv preprint arXiv:1810.07810, 2018.

[19] Liu B, Gu L, Lu F. Unsupervised Ensemble Strategy for Retinal Vessel Segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019: 111-119.

[20] Wang B, Qiu S, He H. Dual Encoding U-Net for Retinal Vessel Segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019: 84-92.

[21] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492-1500.