Cascaded multi-scale feature interaction for choroidal atrophy segmentation

Song, Jiahuan, Chen, Xinjian, Zhu, Weifang

Jiahuan Song, Xinjian Chen, Weifang Zhu, "Cascaded multi-scale feature interaction for choroidal atrophy segmentation," Proc. SPIE 11596, Medical Imaging 2021: Image Processing, 115960J (15 February 2021); doi: 10.1117/12.2580652

SPIE.

Event: SPIE Medical Imaging, 2021, Online Only

Cascaded Multi-Scale Feature Interaction for choroidal atrophy segmentation

Jiahuan Song[1], Xinjian Chen[1,3], and Weifang Zhu[1,2,*]

[1]School of Electronics and Information Engineering, Soochow University, Suzhou, 215006, China
[2]Collaborative Innovation Center of IoT Industrialization and Intelligent Production, Minjiang University, Fuzhou 350108, China
[3]State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou, 215123, China

ABSTRACT

The recent work has achieved great success in utilizing multi-scale feature ensembling for medical image segmentation. In this paper, we propose a new module called cascaded multi-scale feature interaction (CMSI) for choroidal atrophy segmentation in fundus images. The proposed CMSI module makes full use of multi-scale features, including using cascaded pooling and convolution to implement feature interactions at different scales and using strip pooling to capture long-distance features, which makes it more flexible than traditional convolution on the choroidal atrophy region with various scales in fundus image. Based on the U-shape network, we use the ResNet as the backbone to extract hierarchical feature representations. The proposed CMSI module is added at the top of the encoder path. In summary, our main contributions are summarized in two aspects as follows: (1) The CMSI module is proposed for multi-scale feature ensembling by cascading multi-scale pooling and strip pooling. (2) The Dice coefficients of our proposed network on choroidal atrophy segmentation increased by 4.15% compared to U-Net.

Keywords: Choroidal Atrophy Segmentation, Deep Learning, Cascaded Multi-Scale Feature Interaction, Medical Image Processing

1. INTRODUCTION

The complications from pathologic myopia are a major cause of visual impairment and blindness.[1] Choroidal atrophy is one of the earliest pathological changes of pathologic myopia. Automatic segmentation of choroidal atrophy is important for early diagnosis. Choroidal atrophy segmentation in fundus images is a challenge task due to the following reasons: (1) The shapes and areas of choroidal atrophy are various, e.g., the area of optic disc atrophy in patients with common myopia is small while the area of optic disc atrophy in patients with high myopia is mostly circular arc. (2) Atrophy adjacent to the optic disc is usually blurred and difficult to identify. (3) Fundus of patients with pathologic myopia in the early stage is similar to that of patients with high myopia, but shows large area of atrophy in the later stage.Some examples are shown in Fig.1.

In the framework of convolutional network, there are operations of multiple convolution and downsampling. As the number of layers increases, the receptive field gradually increases and the semantic information becomes richer. Therefore, many new structures based on fully convolutional networks are applied to semantic segmentation tasks. U-Net[2] uses skip-connections to make the structure information from the shallow layer can be reused, and has achieved remarkable performance in medical image segmentation. But U-Net can not extract and utilize multi-scale context information effectively.

Recently, some networks have been proposed to explore and integrate multi-scale context information. PSPNet,[3] DeepLabV3[4] and CE-Net[5] adopt multiple parallel poolings or atrous convolutions to process high-level feature maps. CPFNet[6] adopts the atrous convolution with shared weights to obtain multi-scale features, in which a pixel-level soft weight is learned for each scale feature. CCNet[7] and EMANet[8] utilize attention mechanism to allow the network to obtain a global receptive field and aggregate the features of each pixel.

Medical Imaging 2021: Image Processing, edited by Ivana Išgum, Bennett A. Landman, Proc. of SPIE Vol. 11596, 115960J · © 2021 SPIE · CCC code: 1605-7422/21/$21 · doi: 10.1117/12.2580652

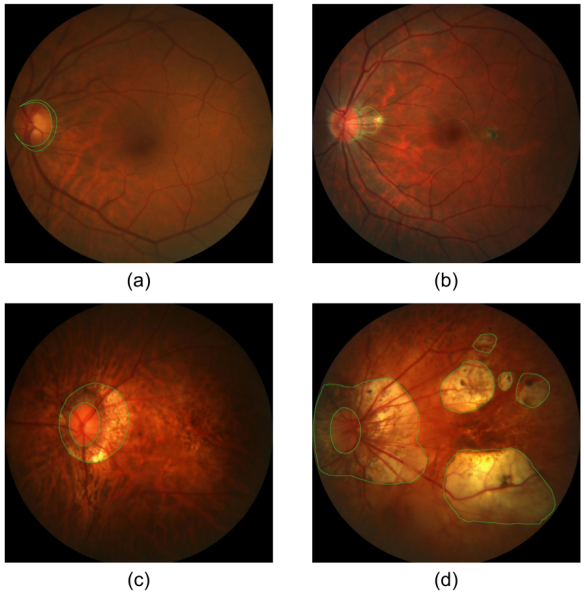

Figure 1. Examples of different fundus images.(a)Normal fundus. (b)High myopia fundus. (c)Pathologic myopia fundus with choroidal atrophy in early stage. (d)Pathologic myopia fundus with severe choroidal atrophy. The region with green boundary represents choroidal atrophy.

In this paper, we propose a U-shape based network with the novel cascaded multi-scale feature interaction (CMSI) module for choroidal atrophy segmentation in fundus images. By the CMSI module, multi-scale features can be extracted by feature interactions between sub-networks and long distance features can be extracted by strip pooling. The experimental results show that our proposed method can solve the above problems well.

2. METHODS

2.1 Overall structure of the network

Inspired by U-Net,[2] we use the encoder-decoder architecture in our network. The ResNet-34[9] pre-trained over the ImageNet dataset is adopted as the backbone, and the average pooling layer and fully connected layers are removed from the original ResNet-34. In the encoder path, the input image is downsampled to 1/32 to obtain hierarchical feature representations. The proposed CMSI module is inserted at the top of the encoder path to extract multi-scale semantic information. The decoder path contains four double convolutions with skip connection. At last, upsample the score map to the original image size. The overall architecture of our network is shown in Fig.2.

2.2 Cascaded multi-scale feature interaction module

The size of the receptive field determines how much context information the network can use. Although the theoretical receptive field of ResNet is already larger than the input image, the empirical receptive field of CNN is much smaller than the theoretical one especially on high-level layers.[10] Previous work including PSPNet[3] and DeepLabV3[4] both proposed effective methods to solve this problem. But these methods deal with features of different scales separately, without considering the relationship between features.

The structure of the proposed CMSI module is shown in Fig.3 (a), which contains two branches. The results of the two branches are concatenated with the input, followed by a 3×3 convolution to transform the channel dimension to the original input channel dimension. Different from previous work, our CMSI module can use features of other scales to enhance semantic information at a specific scale via cascaded pooling and 3×3 convolution. The structure of Branch1 in CMSI is shown in Fig.3 (b), which uses different pooling sizes to get multi-scale features and interact information between different scales simultaneously. Branch1 fuses features with

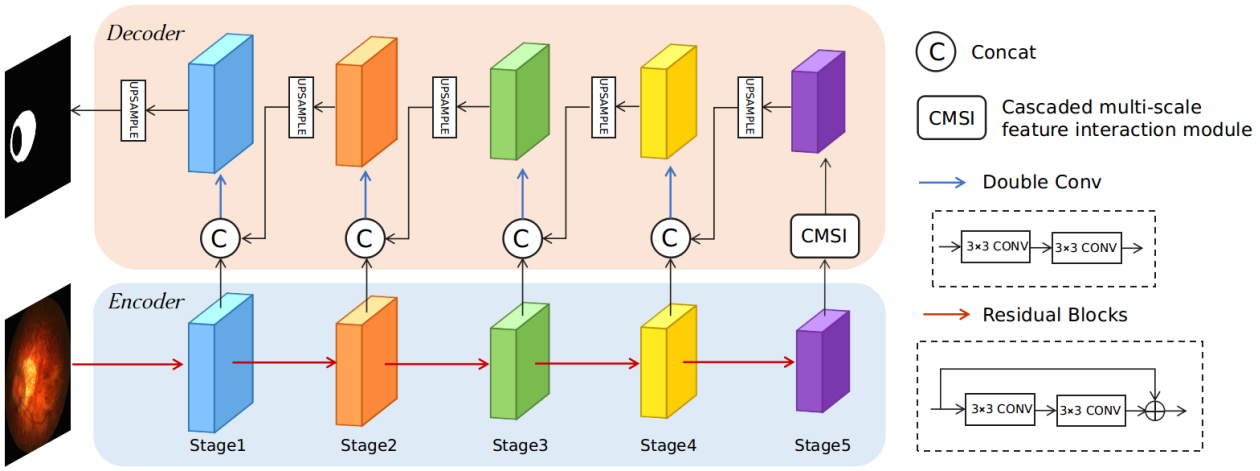

Figure 2. Overall architecture of the proposed network.

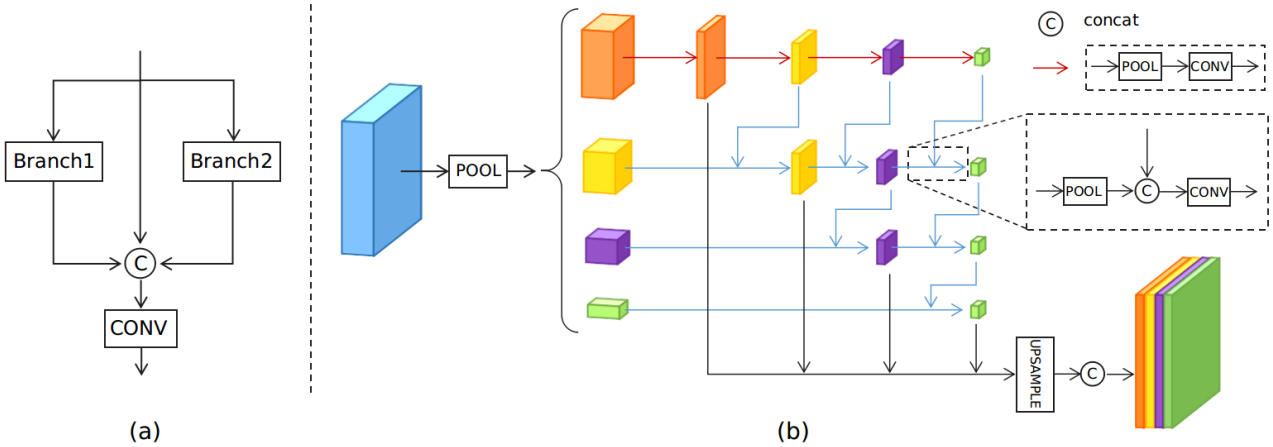

Figure 3. (a)The structure of the proposed CMSI module. (b) The structure of Branch1 in CMSI, which uses cascaded pooling and convolution to interact with different scale features.

four different scales. Each branch is a sub-network, in which the feature maps are downsampled step by step until the spatial resolution is 1 and the information of the feature map with the same resolution is transmitted to each other between each sub-network. Low-dimensional feature maps gradually concatenate pooled high-dimensional feature maps to fuse features. Then the enhanced feature maps are directly upsampled to get the final feature map with the same size as the original feature map via bilinear interpolation. Finally, different levels of features are concatenated.

In Branch2 shown in Fig.4, as we think that the square receptive field may not be suitable for all objects, especially for slender objects, the strip pooling is adopted to capture long-distance features. We consider an input feature map with size $H \times W \times C$ from a single sample. First, the input feature map is pooled into $H \times 1 \times C$ and $1 \times W \times C$, respectively. Then, we use 3×1 and 1×3 convolutions to reduce channel dimension. Finally, upsample the feature maps and element-wise add the two feature maps.

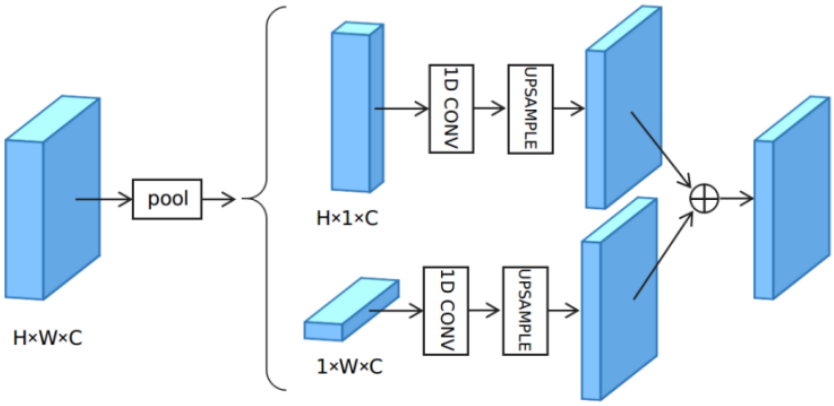

Figure 4. The structure of Branch2 in CMSI, which uses strip pooling to capture long-distance features.

3. RESULTS

3.1 Datasets

Our segmentation approach was evaluated based on the dataset which is from ISBI 2019 Pathologic Myopia Challenge.[11] The training dataset contains 161 normal images, 26 high myopia images, and 213 pathological myopia images. As the images in the dataset have two different resolutions including 1444×1444 and 2124×2056, the bottom of the images with size 2124×2056 are padded to get 1:1 aspect ratio, and then all the images are resized to 448×448. The training data are randomly splitted into four parts and the four-fold cross validation strategy is adopted in our experiments.

3.2 Implementation Details

We conduct experiments based on PyTorch. The training and testing bed is Ubuntu 16.04 system with a Nvidia GeForce 1660ti graphics card, which has 6 Gigabyte memory. In training, we employ Adam optimization with poly learning rate policy. The initial learning rate is set to 0.0001. Batch size and weight decay coefficients are set to 4 and 0.0001, respectively. For data augmentation, we apply rotation, horizontal and vertical flipping, contrast and brightness transformation of the image to augment the training data. The loss function contains two items: cross-entropy loss function and Dice loss function expressed in Eq.(1). The coefficient of Dice loss function λ is set to 0.5.

$$L_{Total} = L_{CE} + \lambda L_{Dice} \tag{1}$$

3.3 Metrics

To quantitatively evaluate the segmentation performance, we use five common segmentation evaluation metrics: Dice similarity coefficient (DSC), Intersection over union (IoU), Accuracy (Acc), Sensitivity (Sens), Specificity (Spec). The definitions of these metrics are shown as follows:

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{2}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{3}$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$Sens = \frac{TP}{TP + FN} \tag{5}$$

$$Spec = \frac{TN}{FP + TN} \tag{6}$$

where TP denotes true positive, FP denotes false positive, FN denotes false negative, TN denotes the true negative.

Metrics include Dice similarity coefficient and F1 score are adopted for the online comparison with other methods in the ISBI challenge leaderboard. F1 score is defined as:

$$F1 = 2\frac{precision * recall}{precision + recall} \tag{7}$$

3.4 Results

The proposed method is compared with four other state-of-art networks including U-Net,[2] Attention U-Net,[12] CE-Net[5] and CPFNet.[6] As shown in Table 1, our method outperforms other methods by a large margin, especially in Dice similarity coefficient and Sensitivity.

Table 1. Comparison experiments and ablation experiments (w/o means without the following component).

Method	DSC(%)	IoU(%)	Acc(%)	Sens(%)	Spec(%)
U-Net[2]	78.33	70.81	98.05	81.35	99.24
Attention U-Net[12]	78.86	71.29	97.94	81.58	99.14
CE-Net[5]	80.48	73.25	98.25	83.66	99.24
CPFNet[6]	81.26	73.95	98.38	83.70	99.30
Ours (w/o CMSI)	80.88	73.88	98.30	83.64	99.20
Ours (CMSI w/o Branch1)	81.26	74.16	98.22	84.75	99.19
Ours	82.49	75.12	98.33	86.06	99.24

The ablation experiments illustrate the effectiveness of our proposed CMSI module, which are shown in the last three rows of Table 1. The performance of the proposed network will decrease if Branch1 in CMSI module is removed (Ours (CMSI w/o Branch1)), but it will be still higher than that of the network without CMSI module (Ours (w/o CMSI)), indicating that both branches in the module are effective.

The online comparison with other methods in the ISBI challenge leaderboard are shown in Table 2. In terms of Dice similarity coefficient, our proposed network without pre-processing and post-processing , exceeds the top 5 teams in the leaderboard. However, it does not show a great advantage in F1 score, which mainly due to the false positive detection of atrophy in normal or high myopia fundus images. Post-processing such as small target exclusion in the segmentation results may further improve the F1 score.

Table 2. Performance comparison with the methods in the challenge leaderboard online.

Team Name/Method	DSC(%)	F1 Score
LAIS	79.42	0.8842
KUL_VITO	80.68	0.9336
VistaLab	81.40	0.8540
PingAn Smart Health	82.77	0.9389
CUHK	83.68	0.9197
Ours	84.13	0.9134

Fig.5 shows some segmentation results of choroidal atrophy. It can be seen from Fig.5 that our method improves the segmentation performance of choroidal atrophy with various scales, which shows that our proposed CMSI module can extract richer global contextual semantic information.

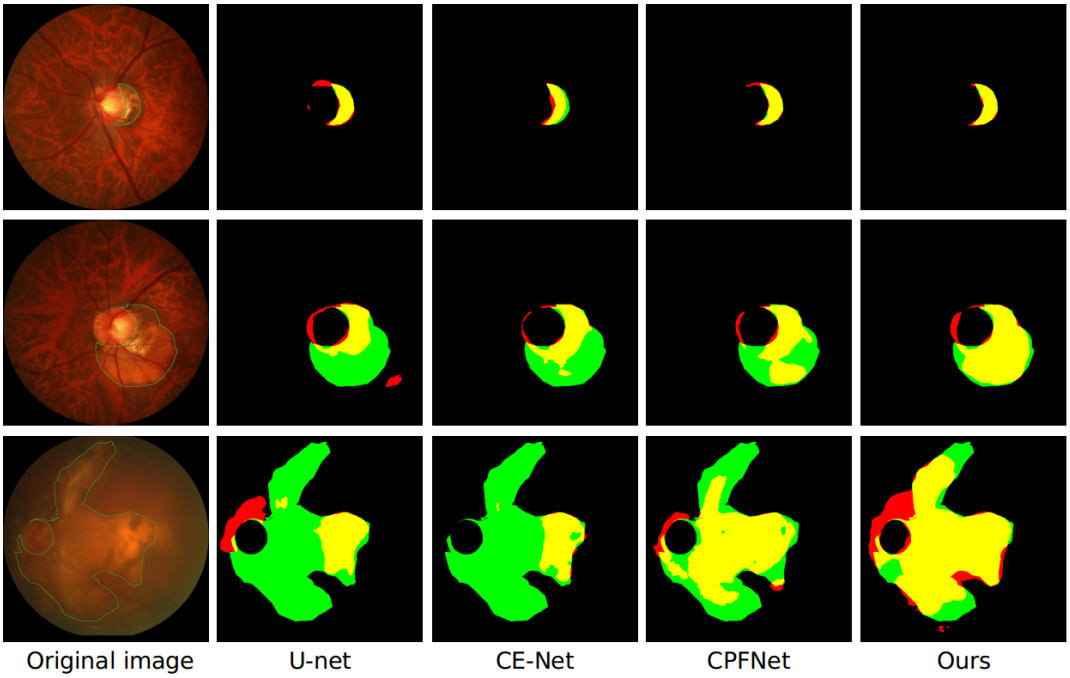

Figure 5. Visualization of segmentation results. The green lines in original images represent the boundaries of the ground truth. Green, red, and yellow regions represent the false negative, false positive and true positive, respectively.

4. CONCLUSIONS

In this paper, we propose a U-shape based network with the novel cascaded multi-scale feature interaction (CMSI) module for choroidal atrophy segmentation in fundus images. The proposed CMSI module can effectively increase the ability of multi-scale feature extraction of the network for choroidal atrophy with various scales in fundus image. The primary experimental results show the effectiveness of our proposed network.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0701700, in part by the National Nature Science Foundation of China under Grant 61622114, in part by the National Basic Research Program of China under Grant 2014CB748600, and in part by Collaborative Innovation Center of IoT Industrialization and Intelligent Production, Minjiang University under Grant IIC1702.

REFERENCES

[1] Ohno-Matsui, K., Lai, T. Y. Y., Lai, C. C., and Cheung, C. M. G., "Updates of pathologic myopia," *Progress in Retinal and Eye Research* , 156–187 (2016).

[2] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*], Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., eds., 234–241, Springer International Publishing, Cham (2015).

[3] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J., "Pyramid scene parsing network," in [*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 6230–6239, IEEE Computer Society, Los Alamitos, CA, USA (jul 2017).

[4] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H., "Rethinking atrous convolution for semantic image segmentation," (2017).

[5] Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., and Liu, J., "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE Transactions on Medical Imaging* **38**(10), 2281–2292 (2019).

[6] Feng, S., Zhao, H., Shi, F., Cheng, X., Wang, M., Ma, Y., Xiang, D., Zhu, W., and Chen, X., "Cpfnet: Context pyramid fusion network for medical image segmentation," *IEEE Transactions on Medical Imaging* **39**(10), 3008–3018 (2020).

[7] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., and Liu, W., "Ccnet: Criss-cross attention for semantic segmentation," in [*2019 IEEE/CVF International Conference on Computer Vision (ICCV)*], 603–612 (2019).

[8] Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., and Liu, H., "Expectation-maximization attention networks for semantic segmentation," in [*2019 IEEE/CVF International Conference on Computer Vision (ICCV)*], 9166–9175 (2019).

[9] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 770–778, IEEE Computer Society, Los Alamitos, CA, USA (jun 2016).

[10] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A., "Object detectors emerge in deep scene cnns," (2015).

[11] Zhang, H. F. F. L. J. I. O. H. B. X. S. J. L. Y. X. S. Z. X., "Palm: Pathologic myopia challenge," (2019).

[12] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., and Rueckert, D., "Attention u-net: Learning where to look for the pancreas," (2018).