**ABSTRACT**

Caused by new blood vessels in chroid glowing through the Bruch’s membrane, choroidal neovascularization (CNV) is a main reason for vision loss in retinal diseases. In clinic treatment, quantitative analysis is essential to help ophthalmologists better diagnose and evaluate CNV. Recently, many effective deep learning methods have been proposed and have achieved good results in medical image segmentation. However, lots of CNVs are small in size and have many kinds of morphological characteristics, making accurate segmentation in OCT images a big challenge. In addition, it is difficult for deep learning models to detect the edges of CNV regions due to its blurry or missing boundary. In order to solve the problems above, we creatively proposed a convolutional neural network with the guide of classification for CNV segmentation in OCT images. By the approach of wavelet transform, prior knowledge can be introduced into the model learning. A new pyramid mixed pooling module (PMP) is designed to help the network obtain the context and detailed information of the image. With the supervision of classification, the model can effectively distinguish CNV from other retinal diseases. The proposed method is evaluated on our own dataset consisting of 1216 OCT images. The experimental results demonstrate the advantage of our proposed algorithm.

**Keywords**

optical coherence tomography (OCT), choroidal neovascularization (CNV), classification guided convolutional neural network

# INTRODUCTION

Choroid neovascularization (CNV) refers to the proliferating vessels from choroid capillaries, expanding through the tear of the Bruch membrane. These blood vessels will underneath the retina and leak which is a characteristic feature of age-related macular degeneration (AMD) , usually resulting in vision loss, visual distortion or even blindness. CNV is common in adults over 60 years old and ought to receive treatment as soon as possible.

Recently, lots of advanced methods using for medical diagnosis have been raised. Li et al. generated a random forest model to detect CNV region by extracting the 3d-HOG feature. Zhu et al. proposed a CNV growth prediction with treatment based on reaction-diﬀusion model in 3-D OCT images. Feng et al. designed a novel context pyramid fusion network(CPFNet) for sementic image segmentation to fuse global and multi-scale context information. Meng et al. proposed a multi-scale information fusion network(MF-Net) for CNV segmentation to catch multi-scale deformation of the targets and aggregate contextual information.

Though deep learning techniques have made great achievements in CNV medical diagonosis in the last few years. However, it’s hard for existing methods to get satisfactory results due to the specificity of CNV lesions. The difficulty of segmentation are as follows. (1)Structurally, lots of CNVs are small while some methods are disabled to capture features of small object, which may got lost in the process of multiple convolution and pooling layers. In addition, various shapes of CNV are also a big challenge which requires model to have good robustness. (2)Visually, as shown in Fig. 1, CNV has similar contrast with background, making it difficult to achieve accurate classification between CNV foreground and backgound. Besides, some vessels are often blurred by fluid exudates and hemorrhage, resulting in few quantitative basis for CNV segmentation.

In order to tackle these issues above, we proposed a convolutional neural network with the guide of classification, which is capable of automatically achieving accurate CNV segmentation results in OCT images. In the input phase of network, we use the way of wavelet transform to enrich detailed information in image. What’s more, a pyramid mixed pooling module (PMP) referred to PSPNet is designed and integrated to the network to get the context and detailed information within OCT images. Considering other diseases and the background have similar contrast with CNV, we added a classification supervised branch, which effectively prevents the model from misjudgment. The results on our own dataset demonstrate the advantage of our proposed algorithm.

|  |
| --- |
|  |
| Figure 1: Examples of CNV in OCT image. The first column is the origin image and the red boundaries in second column is the ground truth of CNV. |

# METHOD

|  |
| --- |
|  |
| **Figure 3: Overview of the proposed network** |

The whole structure of our proposed network is based on U-Net which is depicted in Fig. 3. The main modules of our network contain dilated resnet blocks, pyramid mixed pooling module (PMP) and bilinear interpolation upsampling layers. Specifically, encoder composed of ResUNet with dilated convolution and decoder including several upsample layers is adopted as our network’s backbone while PMP module is integrated to the top of the encoder to obtain the multi-scale and global information of the high-level feature map. Two output branches are designed to obtain classification results and segmentation masks of the origin image.

## Wavelet transform process

Before our network, the origin OCT image is firstly processed by wavelet transform (WT) . In a nutshell, WT is a process of separating low-frequency (HF) and high-frequency (LF) signal like filiters. Based on the idea of introducing prior knowledge into our model and above, haar wavenet is used to get LF and HF information within the image. HF information represents the detailed message while the LF one contains the global message. With respect to HF information, we only used the horizontal one which is consistent with retinal structure. These two outputs are concatenated with the original image and then fed to the model. The detail of WT process is shown in Fig. 3.

|  |
| --- |
|  |
| **Figure 2: Output of wavelet transform process. The left colume is the origin image. The middle colume is the HF filiter’s output which contains horizontal details. The right colume is the LF filter’s output representing the global message from the image.** |

## Backbone

In encoder path, we use a pretrained ResNet-34 as the feature extractor. The average pooling layer and fully connected layers are removed for the task of segmentation. Considering CNVs are commonly small objects and easily got lost in deep CNN, we replaced the normal convolution with dilation convolution in each basic block in order to get a greater recetive field. By introducing dilation convolution, pooling operation can be avoided and internal data structure of the feature map can be retained as much as possible so that only three downsampling operations are performed during feature extracting process.

Corresponding to the encoder, decoder is composed of three upsampling layers using the way of bilinear interpolation, which aims to up sample the informative feature map and restore the high-level semantic features. The skip-connection between the same resolution of encoder and decoder is adopted to bring detailed information from the encoder to the decoder. In this way, information loss due to pooling and strided convolution can be remedy.

## Pyramid Mixed Pooling Module (PMP)

In a deep CNN, receptive field has always been a concern in the process of continual convolution. As is discussed in previous study, the empirical receptive field is much smaller than the theoretical one in high-level stage of the network. Recently, transformer may be a state-of-the-art method to solve long-range dependencies in feature extraction. From another perspective, transformer is handling the problem that receptive field is too small to incorporate the global scenery prior. With that in mind, we designed a pyramid mixed pooling module (PMP) referred to PSPNet, while it has better performance in dealing with image details than PSPNet.

Fig. 4 demonstates the structure of PMP module. The module fuses features from four pyramid scales (1×1, 2×2, 3×3 and 6×6), which highlights different kinds of information in high-level feature map from global to local scope. In each input branch, global average pooling and global max pooling are performed. As is well known, max pooling preserves texture features, while average pooling preserves global features. Although max pooling is widely used in retaining details and reducing the computation, it eaily results in feature loss. So we add the outputs of the two pooling operations and use 1×1 convolution layer to reduce the dimension of context representation to 1/4 of the origin one. Then we upsample each feature map to get the same resolution as the origin feature map. Finally, different levels of features including the origin one are concatenated as the final output feature.

|  |
| --- |
|  |
| **Fig. 4: The illustration of pyramid mixed pooling module (PMP)** |

## Classification Guided Branch and Loss Function

Due to the particularity of lesions, other diseases and the background are close to CNV in color and contrast within OCT images. To prevent the model from misjudging, we add a classification guided branch for supervising the presence of lesions that need to be segmented in origin image. In this way, FP metric which denotes false positive will get smaller, enabling the model to optimize the learning process and detect class information within the image. In addition, adding a extra classification branch hardly brings FLOPs to the model backpropagation. In the testing phase, the branch we add will be ignored and not participate in the testing process.

In our data set, the images without CNV lesions are nearly twice as many as the images with CNV leisions. In view of unbalance between positive and negative samples in our classification task, we use focal loss to reduce the weight of simple negative samples in training, which is a strategy for difficult sample mining. The focal loss is formulated as

where X and Y denote the output of the classification branch and corresponding ground truth. is the factor which influence the rate of weight reduction for simple samples.

The cross entropy (CE) loss is commonly used in semantic segmentation tasks. To segment the single lesion, we adopt binary cross-entropy (BCE) loss to measure the similarity between the model output mask and the ground truth. Dice loss is also considered to improve model’s ability to capture small objects. Therefore, a joint loss function consisting of BCE loss and Dice loss is used to get segmentation result.

Finally, we integrate the focal loss in classification task and the joint loss in segmentation task as the final loss in the training process. Due to the main status of segmentation, we add weight to balance the focal loss. The formulation are as follows:

Where and are the ground truth and predicted mask. C is the sum of the pixels of the output results. is the weight of classification loss, which is set to 0.1 in our paper.

# EXPERIMENT AND RESULTS

## Dataset

Our own dataset was collected by Shanghai First Municipal Hospital. The dataset contains 63 eyes from 63 patients which was acquired using Heidelberg Spectralis OCT (Heidelberg Engineering, Germany) and annotated by professional ophthalmologists. Each patient has 4 3D-OCT volumes during treatment from 4 time points and each volume contains 512×496×19 voxels. In our experiment, 16 patients (64 3D volumes, 1216 B-scans) were selected without exclusion criterion based on age, gender or race. All the slices with and without CNV are included.

## Implementation Details and Evaluation Metric

We split data into trainng and test set by patients and the strategy of 4-fold cross validation is adopted. The origin image is 512x496 and we resize it to 512x512. For data augmentaton, we apply a random horizontal flip, random vertical flip, random rotation between -10 and 10 degrees, random smoothing method including Gaussian blur, median blur and average blur. In the training process, a joint loss consisting of binary cross-entropy loss, dice loss and focal loss is used for backpropagation. SGD is used as the optimizer. Batch size is set to 4 and 60 epochs are trained per fold. The initial learning rate is 0.01 and the poly learning rate scheduling is adopted while decay coefficient is set to 0.9. All the network are trained based on pytorch framework and NVIDIA RTX3090 GPU with 24G memory.

To evalute the segmentation performance, the metrics including dice similarity coeﬃcient (DSC), Jaccard similarity coefficient (JAC), sensitivity (SEN) and specificity(SPE) are adopted, among which, Dice and JAC are the most important evaluate metric.

## Results

We compare our proposed network with the state-of-the-art method including UNet, PSPNet, CPFNet, DeeplabV3+, CENet and MFNet. It should be noted that all the models were trained and tested under the same hyperparameters and data set. The details of the comparison experiments are shown in Table 1.

Table 1. The performance of different segmentation models on CNV

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Methods** | **DSC** | **JAC** | **SEN** | **SPE** |
| UNet | 0.8225±0.06 | 0.7888±0.07 | 0.8955±0.03 | 0.9992±0.00 |
| PSPNet | 0.8829±0.02 | 0.8504±0.02 | 0.9083±0.02 | 0.9995±0.00 |
| CPFNet | 0.8874±0.01 | 0.8545±0.02 | 0.9138±0.01 | 0.9994±0.00 |
| DeeplabV3+ | 0.8583±0.02 | 0.8244±0.03 | 0.9043±0.01 | 0.9994±0.00 |
| CENet | 0.8478±0.06 | 0.8156±0.06 | 0.9231±0.02 | 0.9991±0.00 |
| MFNet | 0.8684±0.02 | 0.8360±0.02 | 0.9114±0.02 | 0.9994±0.00 |
| Backbone | 0.8826±0.01 | 0.8499±0.01 | 0.9191±0.01 | 0.9992±0.00 |
| Backbone  +ppm | 0.8863±0.01 | 0.8548±0.00 | 0.9195±0.02 | 0.9992±0.00 |
| **Proposed method** | **0.8951±0.01** | **0.8639±0.01** | 0.9190±0.01 | 0.9994±0.00 |

It can be seen from Table 1 that our proposed method achieves outperformance in DSC and JAC compared with other models, such as DeepLabV3+, PSPNet and CENet, which are popular in semantic segmentation task. Our method also performs better than models using specially for medical image segmentation such as UNet, CPFNet and MFNet. Due to the nature of CNV, our backbone using dilated resnet blocks as decoder gets a relatively good result, whose DSC and JAC achieves 88.26% and 84.99% respectively. With the PPM module and our extra classification branch, our method improves average DSC by 1.25% and JAC by 1.4% compared with backbone, demonstrating that the addition of PPM module and extra branch effectively improves the model’s performance to solve the CNV segmentation task. In addition, SPE of the proposed network has a slight increase which proves that the extra branch has an effect in reducing the FP metric.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Image | Ground Truth | DeepLabV3+ | MFNet | Backbone | **Proposed** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| **Figure 6: CNV segmentation reults via different models. From left to right are origin image, ground truth, DeepLabV3+, MFNet, backbone and our proposed method.** | | | | | |

# CONCLUSION

In this paper, we propose a novel encoder-decoder network for CNV segmentation in retinal OCT images. Before the network, we consider a method of wavelet transform to introduce prior knowledge of CNV structure. Then a pyramidal module named pyramid mixed pooling (PMP) which is capable of capturing detailed information and global message is designed and integrated to our network. In addition, an extra classification branch and focal loss is added to the process of training in order to optimize the network, promoting it to obtain class information within the image. Proved by experiment, our method achieves excellent performance and has great potential in CNV segmentation.