BMM-Net: Automatic Segmentation of Edema in Optical Coherence Tomography Based on Boundary Detection and Multi-Scale Network

Ruru Zhang
Beijing University of Posts and Telecommunications
Haidian District, Beijing, China
zrr@bupt.edu.cn

Jiawen He
Beijing University of Posts and Telecommunications
Haidian District, Beijing, China
euphy@bupt.edu.cn

Shenda Shi
Beijing University of Posts and Telecommunications
Haidian District, Beijing, China
cy_z_feng@bupt.edu.cn

Haihong E
Beijing University of Posts and Telecommunications
Haidian District, Beijing, China
ehaihong@bupt.edu.cn

Zhonghong Ou
Beijing University of Posts and Telecommunications
Haidian District, Beijing, China
zhonghong.ou@bupt.edu.cn

Meina Song
Beijing University of Posts and Telecommunications
Haidian District, Beijing, China
mnsong@bupt.edu.cn

ABSTRACT

Retinal effusions and cysts caused by the leakage of damaged macular vessels and choroid neovascularization are symptoms of many ophthalmic diseases. Optical coherence tomography (OCT), which provides clear 10-layer cross-sectional images of the retina, is widely used to screen various ophthalmic diseases. A large number of researchers have carried out relevant studies on deep learning technology to realize the semantic segmentation of lesion areas, such as effusion on OCT images, and achieved good results. However, in this field, problems of the low contrast of the lesion area and unevenness of lesion size limit the accuracy of the deep learning semantic segmentation model. In this paper, we propose a boundary multi-scale multi-task OCT segmentation network (BMM-Net) for these two challenges to segment the retinal edema area, subretinal fluid, and pigment epithelial detachment in OCT images. We propose a boundary extraction module, a multi-scale information perception module, and a classification module to capture accurate position and semantic information and collaboratively extract meaningful features. We train and verify on the AI Challenger competition dataset. The average Dice coefficient of the three lesion areas is 3.058% higher than the most commonly used model in the field of medical image segmentation and reaches 0.8222.

CCS CONCEPTS

• **Applied computing** → **Imaging**; • **Computing methodologies** → **Image segmentation**; *Biometrics.*

ACM CHIL '20, April 2–4, 2020, Toronto, ON, Canada

ACM ISBN 978-1-4503-7046-2/20/04…$15.00
https://doi.org/10.1145/3368555.3384447

ACM Reference Format:
Ruru Zhang, Jiawen He, Shenda Shi, Haihong E, Zhonghong Ou, and Meina Song. 2020. BMM-Net: Automatic Segmentation of Edema in Optical Coherence Tomography Based on Boundary Detection and Multi-Scale Network. In *ACM Conference on Health, Inference, and Learning (ACM CHIL '20), April 2–4, 2020, Toronto, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3368555.3384447

1 INTRODUCTION

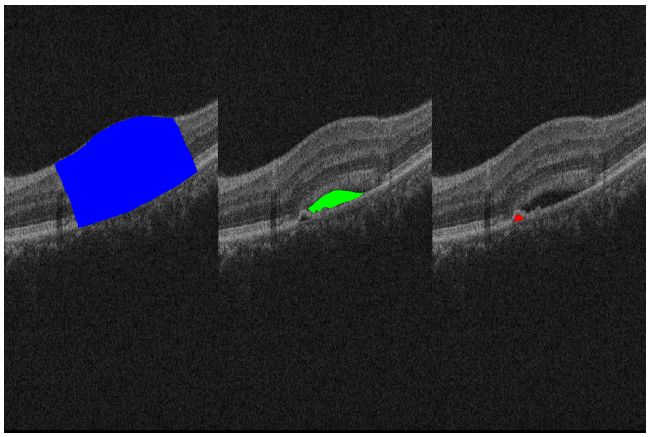

Figure 1: A comparison of the size of the three lesion areas in the AI Challenger competition dataset. The blue area is REA, the green area is SRF, and the red area is PED. Best viewed in color.

The human eye is one of the most important sensory organs, and the eye tissue structure is fine. Even minor damage can cause visual impairment or vision loss [21]. Optical coherence tomography (OCT) is a non-contact, non-invasive imaging technology that can rapidly visualize ocular structures at high resolution and provide clear pathological cross-section imaging [17]. Compared with B-mode, computed tomography, magnetic resonance imaging, and other imaging methods, OCT offers more than 10 times

higher resolution. Nowadays, OCT technology has become the gold standard for the diagnosis and imaging of age-related macular degeneration (AMD), diabetic macular edema, and other ophthalmic diseases [18].

Intraretinal fluid (IRC), subretinal fluid (SRF), and pigment epithelium detachment (PED) caused by the leakage of damaged macular vessels and choroid neovascularization are symptoms of many ophthalmic diseases such as retinal vein occlusion and AMD [13]. Most doctors provide a diagnosis of fundus disease based on the area of effusions and cysts, such as IRC and SRF. In addition, IRC represents one of the most important variables related to vision loss, and SRF may enhance visual prognosis [16]. Therefore, diagnosis and treatment of many diseases are important to accurately identify the lesion areas, such as effusions and cysts, on OCT images. Deep learning technology integrates excellent expert knowledge, which can quantify pathological features in medical images on a pixel-by-pixel basis [4], capture lesion features, and assess the type of disease quickly and accurately. It can improve the accuracy and stability of the diagnosis while reducing the doctor's diagnosis time and provide clinical assistance for doctors. To date, a large number of researchers have carried out relevant studies on deep learning technology to realize the semantic segmentation of lesion areas such as effusion on OCT images, and they have achieved good results.

Lee H. et al. [8] from the Department of Ophthalmology, Konkuk University Medical Center, Seoul, South Korea, developed the AMD intelligent assisted diagnostic system, which can segment multiple lesions, such as IRC, SRF, PED, and subretinal hyperreflective material of neovascular AMD, with an average Dice coefficient of 0.7875. Lu, D. et al. [12] from Simon Fraser University segmented the internal limiting membrane, Bruchs membrane, and IRC, SRF, and PED in OCT images based on the FCN model and used the random forest algorithm to classify the lesion area and remove the incorrectly marked lesion area. The model has an average Dice coefficient of 0.7667; it won first place in the 2017 MICCAI RETOUCH Challenge. Hu J et al. [7] from the College of Computer Science of Sichuan University proposed the automatic segmentation of SRF and PED using deep neural networks and stochastic atrous spatial pyramid pooling (sASPP). The Dice coefficients of SRF and PED reached 0.8759 and 0.7371, respectively, and the model achieved second place in the macular edema segmentation competition in AI Challenger.

However, some insurmountable challenges in the research of OCT lesion region segmentation persist. First, OCT images can show more than ten layers of retinal structures, and lesions at different layers belong to different types of diseases. However, the OCT images are usually blurred, because they are black and white images, and some artifacts are often present during the shooting process [27]. This results in a relatively low contrast between the lesion area and the normal area, and between different lesion areas, which may cause blurring or loss of the lesion boundaries and incorrect labeling of the lesion type. Second, the size of the lesion area is imbalanced. For example, Figure 1 shows a comparison of the size of the three lesion areas in the AI Challenger competition dataset. The retinal edema area (REA) and SRF focus area are often large, while the PED is often small. The uneven size of the lesion area will greatly affect the performance of the model. Compared with the detection of REA and SRF, the detection of PED is more difficult.

In this paper, we build a deep learning model to achieve semantic segmentation of REA, SRF, and PED lesions in OCT images. The model is trained and tested on the AI Challenger competition dataset. Our contributions are as follows:

- Construct a boundary extraction module (BEM) to achieve accurate boundary positioning and alleviate the problems of low contrast and blurred boundaries between the lesion areas.
- Construct a multi-scale information perception module (MIPM) to address the imbalanced size of the lesion area through the multi-scale pyramid. At the same time, the attention mechanism is added to each layer of the pyramid structure to better focus on the relationship between the retinal layers and the distribution of the lesions, so as to guide the model to correctly mark the lesion types.
- Design and implement joint multi-task supervised learning. During the training process, three-task joint supervised is performed and supervised learning is carried out on the BEM, MIPM, and classification module at the same time.
- The effectiveness of our proposed model is proved by multiple comparative experiments, and our model achieves better performance than previous advanced methods.

2 RELATED WORKS

2.1 Boundary Detection

Given the ambiguity of the OCT image, accurate boundary positioning will provide useful fine-grained constraints that can effectively guide feature extraction in the segmentation process. However, the extraction of boundaries and contours is a complex task. The texture of the image is a weak boundary distribution mode, and artifacts exist in the image, so some details are easily covered.

Su J et al. [19] proposed a novel boundary-aware network with successive dilation for salient object detection. In this network, the feature selectivity at boundaries is enhanced by incorporating a boundary localization stream, while the feature invariance at interiors is guaranteed with a complex interior perception stream. Moreover, a transition compensation stream is adopted to amend the probable failures in transitional regions between interiors and boundaries. He J et al. [6] proposed a bi-directional cascade network structure to improve the boundary detection for objects at different scales. In this structure, an individual layer is supervised by labeled boundaries at its specific scale, and a scale enhancement module is introduced to generate multi-scale features using dilated convolution. Ruan T et al. [15] developed a simple yet effective context embedding with boundary perceiving framework and then integrated feature resolution, global context information, and boundary details into a unified network to achieve human segmentation in images. They won first prize on all three human parsing tracks in the "Second Look into Person Challenge." Liu Y et al. [11] proposed a novel fully convolutional neural network architecture using diverse deep supervision within a multitask framework where lower layers aim at generating category-agnostic boundaries, while higher layers are responsible for the detection of category-aware semantic boundaries.

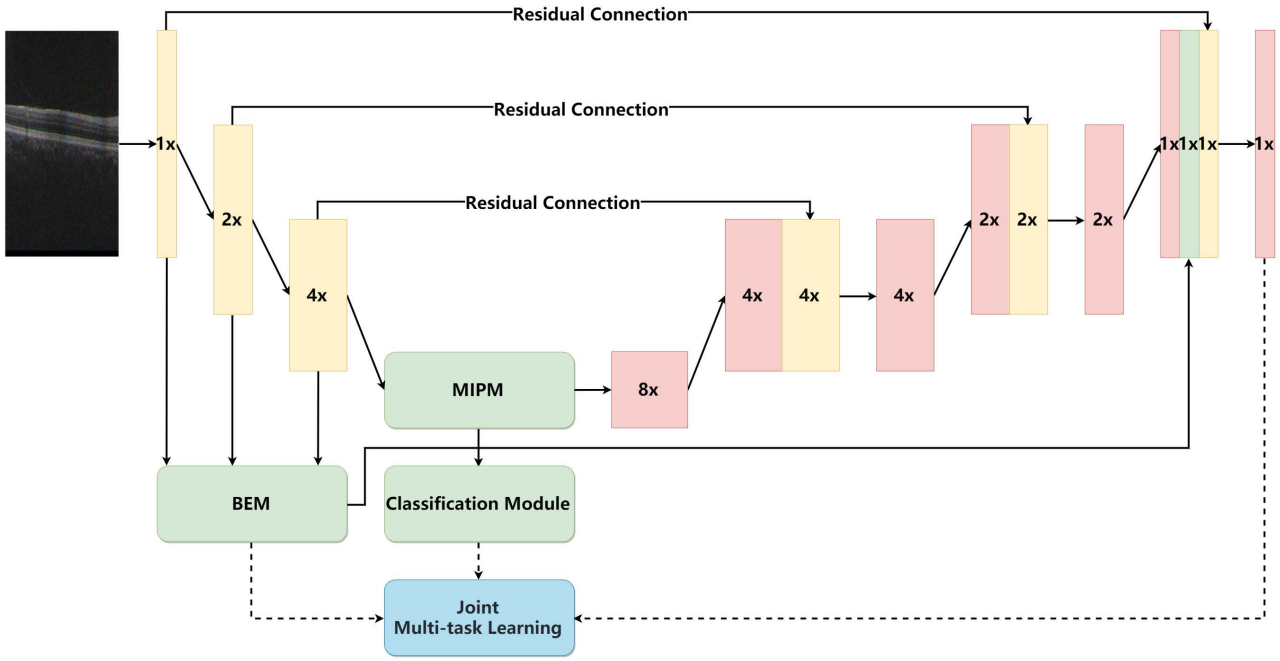

Figure 2: The overall framework of BMM-Net. The basic network is a U-Net model. Features are extracted at the encoder stage and inputted into BEM, MIPM, and classification module. At the decoder stage, feature maps from BEM and MIPM are fused. 1x, 2x, 4x and 8x represent the model downsampling multiples at this layer, respectively. We perform joint multi-task learning on three modules. Best viewed in color.

2.2 Multi-scale and Contextual Aggregation

Traditional full convolutional neural networks reduce the resolution of the final feature map due to continuous pooling and downsampling. Lead to the lack of ability to describe the segmentation results in detail and to accurately segment different scale targets. In the context of semantic segmentation networks, especially in the study of segmentation of target regions of different sizes, multi-scale feature fusion shows remarkable performance.

In RefineNet [9], a multi-path refinement block is utilized to combine multi-scale features in the downsampling process to generate high-resolution predictions. The U-Net++ network [28] is designed with an encoder-decoder architecture; the low-level feature map of the encoder module and the high-level feature map of the decoder module are combined to capture clear object boundaries by gradually recovering spatial information. Feature pyramid networks [10] progressively upsample feature maps of different scales in a bottom-up fashion and aggregate them in a top-down fashion to obtain a multi-scale feature map. PSPNet [25] uses pyramid pooling module blocks to perform pooling operations at different grid scales and fuses context information of different scales between various sub-regions through an aggregation network, thereby improving global information acquisition. In the DeepLab [2] model, ASPP probes an incoming convolutional feature layer with filters at multiple sampling rates and effective fields-of-views, thereby capturing objects and image context at multiple scales. Adaptive pyramid context network [5] introduces adaptive context modules that generate local affinity coefficients with a global-guided local affinity to capture different scales objects. Although context fusion helps capture target objects at different scales, it fails to take advantage of the relationship between targets in the global image, which is also crucial to semantic segmentation.

2.3 Attention Modules

Traditional convolutional networks can only model local interactions, resulting in insufficient semantic information and affecting the final segmentation performance. The attention mechanism can obtain larger receptive fields and contextual information by capturing global information, so as to better utilize the relationship between targets in the global image to emphasize useful features, which is beneficial to the improvement of segmentation performance.

Non-local network (NLNet) [22] presents a pioneering approach for capturing long-range dependencies via aggregating query-specific global context to each query position. SENet adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels. GCNet [1] absorbs the advantages of NLNet and SENet and introduces the channel space of SENet on the basis of non-local spatial attention. The point-wise spatial attention network proposed in [26] connects each location in the feature map to all other locations through a self-adaptively learned

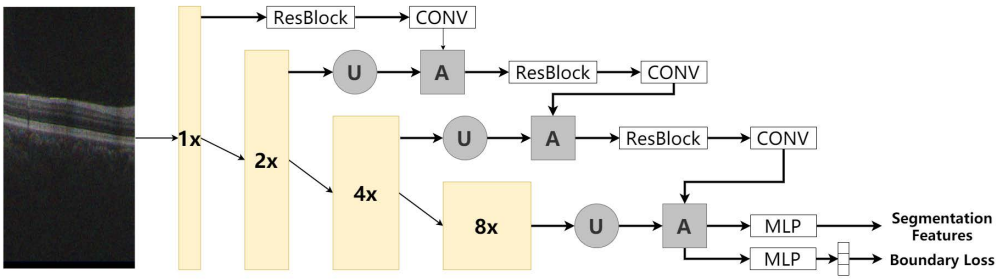

Figure 3: Framework of BEM. Extracting boundary information for feature maps obtained from the first four layers of downsampling. In this figure, U stands for upsampling, A stands for attention mechanism module, and MLP stands for multilayer perceptron. 1x, 2x, 4x and 8x represent the model downsampling multiples at this layer, respectively. Best viewed in color.

attention mask, enabling flexible and dynamic aggregation of remote contextual information. The DANet model [3] introduces position attention module and channel attention module to capture rich contextual relationships for improved feature representations with intra-class compactness. Although attention mechanisms are becoming increasingly popular on many visual problems, the literature on attention in medical image segmentation is scarce and focuses on simple attention modules or applications to other related research.

3 METHODS

In view of the diversity of the shape and size of the lesion area in the OCT image, we propose the boundary multi-scale multi-task OCT segmentation network (BMM-Net) based on the U-Net model. The model architecture is shown in Figure 2. The model first extracts image features through an encoder and then inputs them to the BEM, MIPM, and classification module for joint multi-task supervised learning.

To address the problems of low contrast, blurriness, and vanishing boundaries in OCT images, we construct the BEM to enhance the boundary features of the lesion area. To solve the uneven size of different lesions in OCT images, we construct the MIPM to effectively improve the segmentation accuracy of lesions at different scales. We also build classification modules for collaborative learning to further improve the performance of segmented networks.

The decoder module implements the upsampling of multi-scale information perception feature maps. The high-resolution features of the downsampling path are combined with the upsampled output to obtain rich detailed information. We combine the semantic region features obtained by the MIPM with the boundary features of the BEM to obtain an accurate segmentation result.

3.1 BEM

The low contrast of the image causes the boundaries of the tissue or lesion area to blur or even disappear. The existing method to solve this problem is to combine low-level features with high-level features through skip connections. This strategy works well in high-contrast, clear, and consistent images. However, when it is applied to low-contrast images, the local appearance features extracted from the lower layer may not suppress surrounding artifacts and fail to identify disappearing boundaries [27]. These features will greatly hinder the successful completion of OCT image automatic analysis by computer algorithms.

To achieve accurate boundary location in OCT images, we introduce a boundary detection algorithm to obtain additional feature information related to the boundary. The boundary information provides useful fine-grained constraints to guide the feature extraction process of the MIPM. We use the natural image boundary detection method recently proposed in [20], as shown in Figure 3. Given that only the low-level features of the semantic segmentation model can retain sufficient boundary information, the boundaries of medical images are more blurred compared with natural images. Therefore, the original model extracts the boundary information of the feature map obtained by the first, third, fourth, and fifth layers of downsampling. We then modify the model to extract the boundary information of the feature map obtained by the first four layers of downsampling. (The fourth layer is a feature map with eightfold reduced resolution in the MIPM, which is not reflected in Figure 2.) The feature maps of the second to fourth layers are enlarged to the same resolution as the input image by the upsampling module. All four layers of feature maps are fused by alternately connecting the residual module, 1×1 convolution module, and attention mechanism module. Using the attention mechanism can effectively focus on the information related to the boundaries and achieve accurate extraction of the boundary regions.

3.2 MIPM

To obtain global information further, we add the MIPM between the encoder and decoder. This module is used to accurately extract feature information at different scales and alleviate the problem of the uneven size of the lesion area in OCT images. This module is illustrated in Figure 4.

To improve the resolution of feature maps during model downsampling and alleviate the disappearance of boundaries or small lesions, we construct a real-time processing and highly responsive model suitable for the medical field. We consider using the joint pyramid upsampling (JPU) module proposed in the FastFCN model [24] to aggregate high-resolution features. In addition, compared

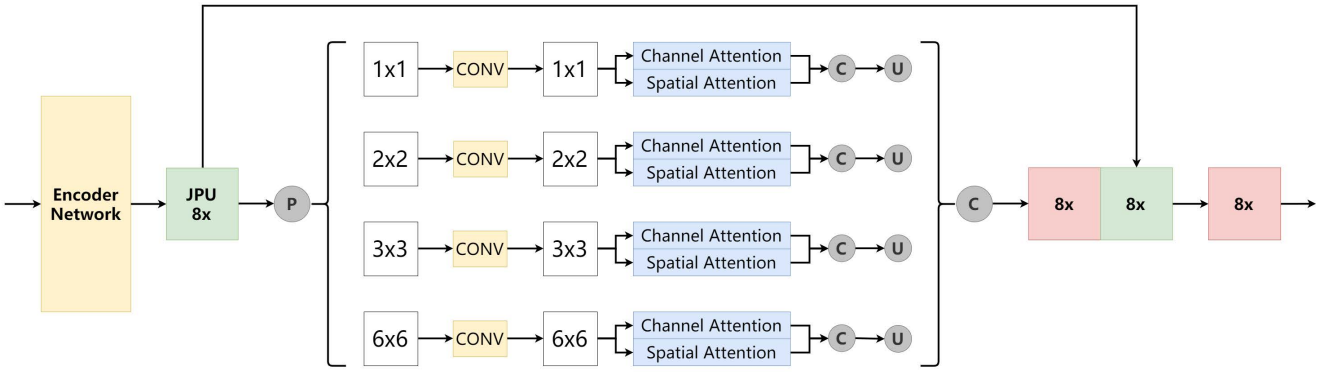

Figure 4: The framework of MIPM. In this figure, the encoder network represents the first three layers of the overall model, P represents pooling, C represents concatenate, and U represents upsampling. 1x1, 2x2, 3x3 and 6x6 respectively represent the size of the feature map after pooling. The detailed introduction of the JPU module is shown in Figure 5. Best viewed in color.

with the dilated convolution method commonly used to solve the resolution problem, this module reduces the computational complexity by more than three times without performance loss. After the JPU module, we use the pyramid pooling module(PPM) to obtain context information of four different scales in the OCT image. However, the information concerned by the network is different at varying scales. If the feature maps of different scales are simply fused, there may be redundant information, which is not conducive to the subsequent recovery of valuable information. To make each layer of pyramids pay close attention to key contextual information of targets at different scales, we add a spatial attention module and a channel attention module [3] at each scale, as shown in Figure 4. The spatial attention module and the channel attention module construct a context vector for each size of the pyramid model. Simultaneously, the spatial attention feature map and the channel attention feature map are extracted and fused on four scales and then inputted to the subsequent decoder part.

The multi-scale semantic feature map output by the MIPM is fused with the boundary feature map output by the BEM to obtain accurate segmentation results. The boundary feature map is also used to calculate boundary detection errors for multi-task supervised learning. The highest dimensional feature map in MIPM is used as input to the classification module. The feature map obtains the probability that the OCT image contains three kinds of lesion areas through a maximum pooling layer. The classification loss for joint multi-task learning is calculated using the output of the classification module. See Figure 5 for details.

3.3 Joint Multi-Task Learning

During the training process, we perform three-task joint supervised learning on the BEM, MIPM, and classification module. The boundary feature map is a binary representation of the boundaries of all the lesions in the OCT image. We use the standard binary cross-entropy loss on the predicted boundary feature maps and classification results. The formula is as follows:

$$L_{BCE}(\hat{s}, s) = \frac{1}{N_m} \sum_{i}^{N_m} [s_i \log \hat{s}_i + (1 - s_i) \log \hat{s}_i] \tag{1}$$

$$L_{BCE}(\hat{y}, y) = \frac{1}{N_L} \sum_{j=0}^{N_L - 1} (I\{\hat{y}_j = 1\} \log y_j) \tag{2}$$

where $L_{BCE}(\hat{s}, s)$ represents the cross-entropy loss of the boundary extraction graph. $L_{BCE}(\hat{y}, y)$ represents the cross-entropy loss of the classification result. $\hat{s}$ represents the true boundary map label. N_m represents the number of all pixels in the boundary feature map. $\hat{s}_i$ and s_i are the truth value and output value of the i-th pixel, respectively. N_L represents the number of category tags; here, it is 3 (REA, SRF, and PED). $\hat{y}$ and y are a 1×3 matrix (if REA exists in the OCT image, then the number in the corresponding position in the matrix is 1, otherwise 0). $\hat{y}_j$ and y_j correspond to the truth value and output value of the index j, respectively. $I\{\}$ is a binary indicator. $I\{\hat{y}_j = 1\}$ is 1 only when $\hat{y}_j = 1$; otherwise, it is 0.

We fuse the predicted semantic segmentation feature map and boundary feature map into the final feature map f. To obtain high-quality region segmentation and clear boundaries, we use Dice loss and structural similarity index (SSIM) [23] mixed loss functions. Among them, Dice loss is a loss function commonly used in medical image segmentation and is related to the Dice coefficient. The Dice coefficient is used to calculate the similarity of the two samples. Dice loss is used to measure the loss value using the Dice coefficient as an indicator. The specific formula is as follows:

$$L_{Dice}(\hat{f}, f) = 1 - \frac{1}{N_L} \sum_{j=0}^{N_L - 1} \frac{2 \left\| \hat{f}_j \cap f_j \right\|}{\left\| \hat{f}_j \right\| + \left\| f_j \right\|} \tag{3}$$

Among them, $\hat{f}$ and f are ground truth and model prediction segmentation maps, respectively, which are the superposition of N_L images. $\hat{f}_j$ and f_j are ground truth and model prediction segmentation map of the j-type lesion, respectively.

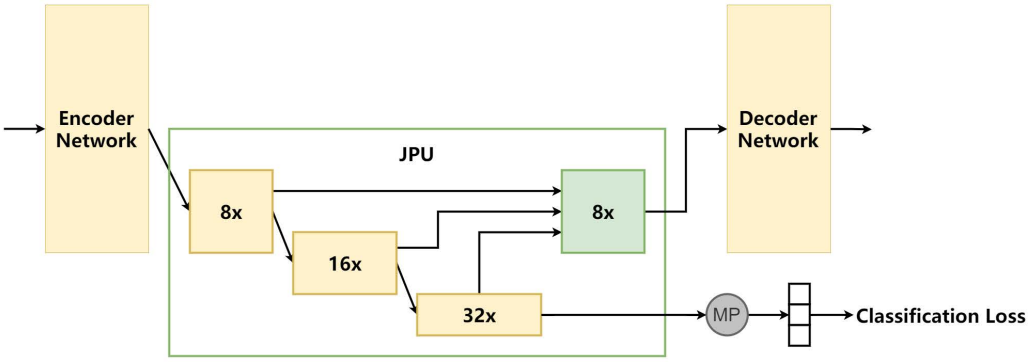

Figure 5: The framework of the classification module. We use the 32x downsampling feature map in the JPU module for classification. In this figure, MP stands for maximum pooling. Best viewed in color.

SSIM loss is a block-level measure, which is sensitive to the perception of local structural changes and usually gives high weights in the boundary area. We crop on $\hat{f}_j$ and f_j to obtain two pixels blocks of size N × N (N = 5 in this article), which are represented by n and m, respectively. The structural similarity between n and m is calculated and traversed over the entire f and $\hat{f}$ to obtain the overall SSIM loss:

$$L_{ssim}(\hat{f}, f) = 1 - \frac{(2\mu_n\mu_m + C_1)(2\sigma_{nm} + C_2)}{(\mu_n^2 + \mu_m^2 + C_1)(\sigma_n^2 + \sigma_m^2 + C_2)} \quad (4)$$

Among them, μ_n, μ_m, σ_n^2, σ_m^2, and σ_{nm} are the mean, variance, and covariance of n and m, respectively.

The multi-task learning joint loss function is:

$$L = \lambda_1 L_{BCE}(\hat{s}, s) + \lambda_2 L_{BCE}(\hat{y}, y) + \lambda_3 L_{Dice}(\hat{f}, f) + \lambda_4 L_{ssim}(\hat{f}, f) \quad (5)$$

Among them, λ_1, λ_2, λ_3 and λ_4 are four hyperparameters that control the weight between losses.

4 EXPERIMENTS

In this section, we demonstrate the effectiveness of our proposed BMM-Net on the OCT image dataset. In the first part of the experiment, we compare the experiment with advanced models commonly used in the medical field. In the second part of the experiment, we perform ablation experiments to verify the effectiveness of each component in our BMM-Net.

4.1 Dataset

We use the AI Challenger competition dataset. The dataset has a total of 100 OCT volumes, of which the training set contains 70 volumes, and the validation and test sets both contain 15 volumes. Each volume contains 128 slices with a resolution of 512 × 1024, and pixel-level annotations are performed on the REA, SRF, and PED lesion areas on each slice. Given that the label of the test set is not public, we only conduct comparison experiments on the validation set.

During the experiment, we adopted the following preprocessing method for the above data set: we superimposed the adjacent 3 OCT slices in the channel dimension and trained on this unit. Use the labeled result of the middle slice as the ground truth. And we only use a single OCT slice for testing. Other than that, we did not perform any other preprocessing.

4.2 Implementation Details

Our BMM-Net is based on PyTorch. The network trains a total of 60 epochs. The learning rate is initially set to 0.001, and every 10 epochs drop by one order. The momentum is set to 0.9, and the minimum batch size is eight. All experiments are optimized using Adam optimizer until the verification loss converges. All models and training settings are consistent with the original implementation. All these experiments are performed on two NVIDIA GeForce GTX GPUs with 12GB of memory.

We conducted several sets of comparative experiments on the choice of hyperparameters, and selected the value with the best effect. λ_1, λ_2, λ_3 and λ_4 are all set to 0.25.

4.3 Comparative experiments with other methods

To prove the advantages of our proposed method, we conduct comparison experiments with two other deep learning models which are widely used in the medical field. All models have added classification modules, and other operating conditions are exactly the same.

(1) U-Net [14]: U-Net is a pioneering work that introduces full convolutional neural networks to medical image analysis. The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. This network achieves the best performance on the ISBI 2012 EM challenge dataset.

(2) U-Net++ [28]: U-net++ was proposed at the DLMIA2018 conference, and some improvements were made based on U-net. The main improvement is the reduced semantic gap between the feature maps of the encoder and decoder subnets through a series of nested, dense skip pathways to extract detailed information.

Table 1 quantitatively compares the performance of our BMM-Net with the two deep learning segmentation models introduced above through Dice coefficients. We add The same classification

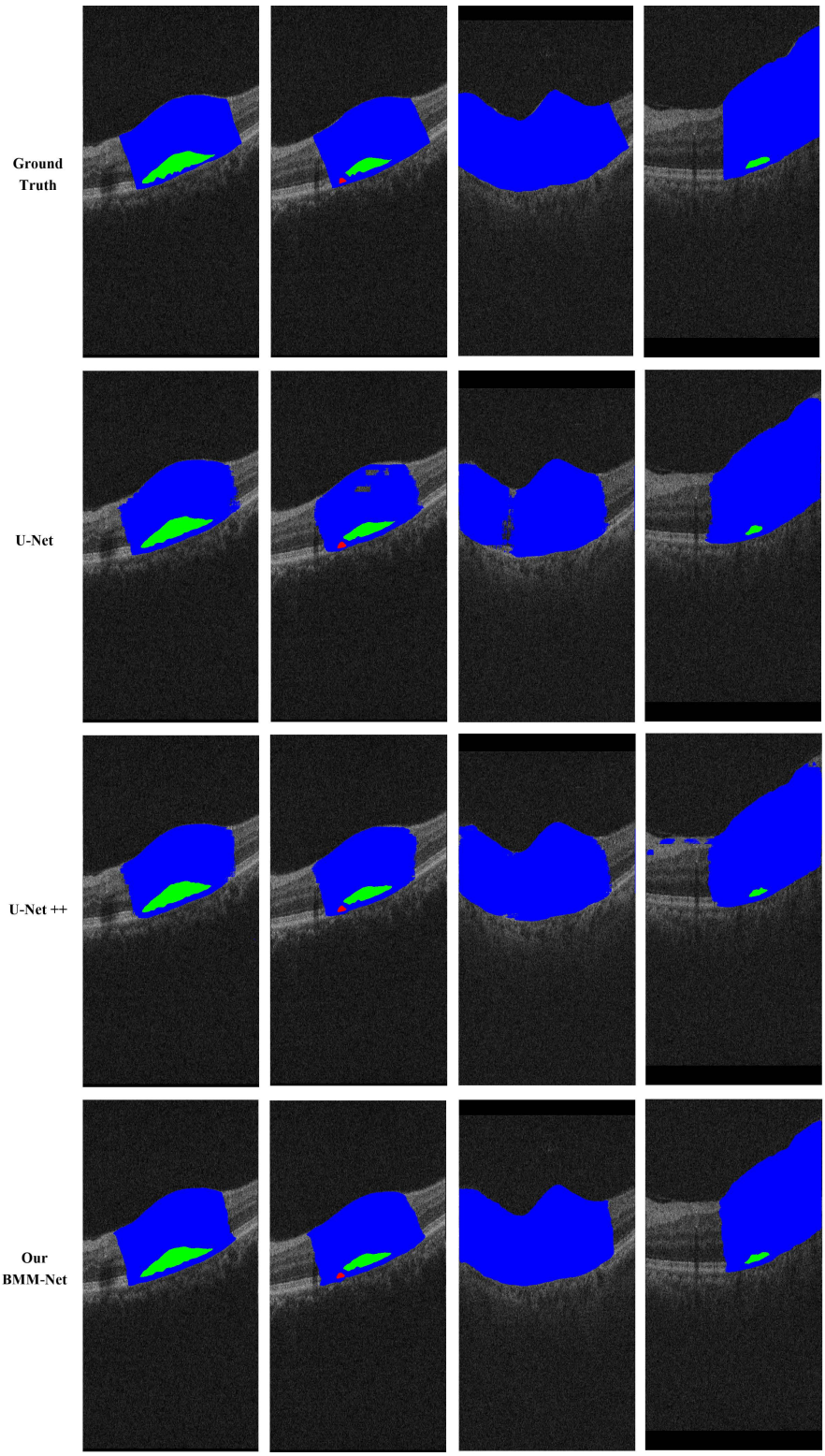

Figure 6: Qualitative comparison between U-Net, U-Net++ and our BMM-Net. The four rows denote the ground truth, prediction results of U-Net, U-Net++ and BMM-Net. Each column represents a specific slice in OCT. In these figures, different colors indicate different types of lesions: SRF (green), PED (red), REA (blue).

Table 1: Comparison with state-of-the-art models.

Model	Dice				AUC			
	Mean	REA	SRF	PED	Mean	REA	SRF	PED
U-Net	0.7889	0.7637	0.8378	0.7653	0.9860	0.9752	0.9990	0.9838
U-Net ++	0.7978	0.7784	0.8336	0.7814	0.9761	0.9825	0.9987	0.9472
Our BMM-Net	**0.8222**	**0.7916**	**0.8767**	**0.7985**	0.9845	0.9783	**0.9993**	0.9758

Table 2: Ablation experiments

Model	Dice				AUC			
	Mean	REA	SRF	PED	Mean	REA	SRF	PED
U-Net	0.7889	0.7637	0.8378	0.7653	0.9860	0.9752	0.9990	0.9838
U-Net + MIPM	0.7955	0.7495	0.8714	0.7657	0.9684	0.9822	0.9985	0.9244
U-Net + BEM	0.7985	0.7493	0.8531	0.7931	0.9834	0.9787	0.9987	0.9728
U-Net + BEM + JPU + PPM	0.8160	0.7778	0.8944	0.7759	0.9810	0.9676	0.9985	0.9769
U-Net + MIPM + BEM (ours)	**0.8222**	**0.7916**	0.8767	**0.7985**	0.9845	0.9783	**0.9993**	0.9758

module into all models and calculate the classification AUC scores. Our model achieves better performance compared with the two advanced models commonly used in the medical field. Figure 6 shows some typical segmentation results. The results show that our proposed model is 3.085% higher in the average Dice coefficient than the current advanced methods. At the same time, the performance of our model improved compared to the team that achieved second place in the macular edema segmentation competition in AI Challenger [7]. From the visual segmentation results of representative samples in Figure 6, we find that our model does not only achieve more accurate segmentation on large lesion areas but also obtains more robust results on small lesion areas. The advantages of our model are mainly reflected in two aspects: (1) compared with other models, the segmentation of lesion region boundary of our model is clearer and smoother; and (2) multi-scale lesions are appropriately handled and the detailed areas are treated more accurately.

4.4 Ablation Experiments

To prove the validity of the proposed model in this paper, we perform ablation experiments on each component of the model. Except for different combinations of the BEM and MIPM, all other configurations are the same. Table 2 shows the comparison results of the ablation experiments. Compared with the benchmark model U-Net, the constructed BEM and MIPM both bring performance gains, especially when the two modules are trained in collaboration, the performance is significantly improved.

U-net+MIPM. To study the performance of the MIPM, we append the MIPM to the benchmark model, without BEM. Compared with U-Net, the addition of the MIPM improves the average segmentation performance of REA, SRF, and PED lesion regions.

U-Net+BEM. When we only append BEM to the benchmark model, the overall effect of the model does not improve significantly. In order to verify the collaborative training promotion effect between modules, the next set of ablative experiments were carried out.

U-net+BEM+JPU+PPM. When we append BEM, JPU, and PPM to the benchmark model, the segmentation performance of REA, SRF, and PED lesion regions was significantly improved. This model dramatically outperforms the benchmark model, proving that the collaborative learning between boundary detection, JPU model and PPM model is crucial for segmentation.

BMM-Net. Our whole BMM-Net, obtains the best performance through the collaborative training of MIPM and BEM.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose BMM-Net to segment OCT images. In this model, the BEM, MIPM, and classification module learn cooperatively to capture accurate position and semantic information and extract meaningful features. The BEM uses residual convolution and attention mechanisms to extract accurate location semantic information and perform boundary positioning. The MIPM uses pyramid structure and attention mechanisms to obtain comprehensive multi-scale information. The classification module determines the existence of lesion types through maximum pooling. We integrate the BEM, MIPM, and classification module into the UNet network and conduct multi-task joint training in an end-to-end learning manner. We achieve accurate segmentation of the lesion area in OCT images and find better performance than previous advanced methods. In the future, we will work with ophthalmologists to label OCT image segmentation data sets for real clinical scenarios, including labeling retinal layers and multiple lesion types. More importantly, we will further refine our experiments and try to add more expert prior knowledge.

6 ACKNOWLEDGE

This work was supported by the National Natural Science Foundation of China (Grant No. 61902034), the National Natural Science Foundation of China (Grant No. 61702046) and National Key R&D Program of China (Grant No. 2017YFB1401500).

REFERENCES

[1] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. 2019. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic

image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

[3] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3146–3154.

[4] Mrinal Haloi. 2018. Towards Ophthalmologist Level Accurate Deep Learning System for OCT Screening and Diagnosis. *arXiv preprint arXiv:1812.07105* (2018). https://arxiv.org/abs/1812.07105

[5] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. 2019. Adaptive Pyramid Context Network for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7519–7528.

[6] Jianzhong He, Shiliang Zhang, Ming Yang, Yanhu Shan, and Tiejun Huang. 2019. Bi-Directional Cascade Network for Perceptual Edge Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3828–3837.

[7] Junjie Hu, Yuanyuan Chen, and Zhang Yi. 2019. Automated segmentation of macular edema in OCT using deep neural networks. *Medical image analysis* 55 (2019), 216–227.

[8] Hyungwoo Lee, Kyung Eun Kang, Hyewon Chung, and Hyung Chan Kim. 2018. Automated segmentation of lesions including subretinal hyperreflective material in neovascular age-related macular degeneration. *American journal of ophthalmology* 191 (2018), 64–75.

[9] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1925–1934.

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[11] Yun Liu, Ming-Ming Cheng, Deng-Ping Fan, Le Zhang, JiaWang Bian, and Dacheng Tao. 2018. Semantic edge detection with diverse deep supervision. *arXiv preprint arXiv:1804.02864* (2018). https://arxiv.org/abs/1804.02864

[12] Donghuan Lu, Morgan Heisler, Sieun Lee, Gavin Weiguang Ding, Eduardo Navajas, Marinko V Sarunic, and Mirza Faisal Beg. 2019. Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network. *Medical image analysis* 54 (2019), 100–110.

[13] Abdolreza Rashno, Dara D Koozekanani, and Keshab K Parhi. 2018. Oct fluid segmentation using graph shortest path and convolutional neural network. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 3426–3429.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[15] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. 2019. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4814–4821.

[16] Thomas Schlegl, Sebastian M Waldstein, Hrvoje Bogunovic, Franz Endstraßer, Amir Sadeghipour, Ana-Maria Philip, Dominika Podkowinski, Bianca S Gerendas, Georg Langs, and Ursula Schmidt-Erfurth. 2018. Fully automated detection and quantification of macular fluid in OCT using deep learning. *Ophthalmology* 125, 4 (2018), 549–558.

[17] Ursula Schmidt-Erfurth, Amir Sadeghipour, Bianca S Gerendas, Sebastian M Waldstein, and Hrvoje Bogunović. 2018. Artificial intelligence in retina. *Progress in retinal and eye research* (2018).

[18] Ursula Schmidt-Erfurth and Sebastian M Waldstein. 2016. A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. *Progress in Retinal and eye Research* 50 (2016), 1–24.

[19] Jinming Su, Jia Li, Yu Zhang, Changqun Xia, and Yonghong Tian. 2019. Selectivity or invariance: Boundary-aware salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 3799–3808.

[20] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. 2019. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 5229–5238.

[21] Gerard J Tortora and Bryan H Derrickson. 2018. *Principles of anatomy and physiology*. John Wiley and Sons.

[22] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7794–7803.

[23] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[24] Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yizhou Yu. 2019. FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation. *arXiv preprint arXiv:1903.11816* (2019). https://arxiv.org/abs/1903.11816

[25] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2881–2890.

[26] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. 2018. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 267–283.

[27] Sihang Zhou, Dong Nie, Ehsan Adeli, Jianping Yin, Jun Lian, and Dinggang Shen. 2019. High-Resolution Encoder–Decoder Networks for Low-Contrast Medical Image Segmentation. *IEEE Transactions on Image Processing* 29 (2019), 461–475.

[28] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 3–11.