IPEM
Institute of Physics and
Engineering in Medicine

PAPER

Multi-class retinal fluid joint segmentation based on cascaded convolutional neural networks

To cite this article: Wei Tang *et al* 2022 *Phys. Med. Biol.* **67** 125018

View the article online for updates and enhancements.

Physics in Medicine & Biology

IPEM
Institute of Physics and Engineering in Medicine

CrossMark

PAPER

Multi-class retinal fluid joint segmentation based on cascaded convolutional neural networks

Wei Tang[1] , Yanqing Ye[1], Xinjian Chen[2] , Fei Shi[1] , Dehui Xiang[1] , Zhongyue Chen[1] and Weifang Zhu[1,*]

[1] MIPAV Lab, School of Electronic and Information Engineering, Soochow University, Jiangsu 215006, People's Republic of China
[2] MIPAV Lab, School of Electronic and Information Engineering and the State Key Laboratory of Radiation Medicine and Protection, Soochow University, Jiangsu 215006, People's Republic of China
[*] Author to whom any correspondence should be addressed.

E-mail: wfzhu@suda.edu.cn

Abstract

Objective. Retinal fluid mainly includes intra-retinal fluid (IRF), sub-retinal fluid (SRF) and pigment epithelial detachment (PED), whose accurate segmentation in optical coherence tomography (OCT) image is of great importance to the diagnosis and treatment of the relative fundus diseases. *Approach.* In this paper, a novel two-stage multi-class retinal fluid joint segmentation framework based on cascaded convolutional neural networks is proposed. In the pre-segmentation stage, a U-shape encoder–decoder network is adopted to acquire the retinal mask and generate a retinal relative distance map, which can provide the spatial prior information for the next fluid segmentation. In the fluid segmentation stage, an improved context attention and fusion network based on context shrinkage encode module and multi-scale and multi-category semantic supervision module (named as ICAF-Net) is proposed to jointly segment IRF, SRF and PED. *Main results.* the proposed segmentation framework was evaluated on the dataset of RETOUCH challenge. The average Dice similarity coefficient, intersection over union and accuracy (Acc) reach 76.39%, 64.03% and 99.32% respectively. *Significance.* The proposed framework can achieve good performance in the joint segmentation of multi-class fluid in retinal OCT images and outperforms some state-of-the-art segmentation networks.

1. Introduction

MACULAR edema is the accumulation of fluid in the macular area of the retina due to the destruction of the blood-retinal barrier. It occurs secondary to diabetic retinopathy (DR), retinal vein occlusion and other retinal diseases, which could lead to serious impairment in vision (Bringmann *et al* 2004). Retinal fluids mainly include three types: intra-retinal fluid (IRF), sub-retinal fluid (SRF) and pigment epithelial detachment (PED). Figure 1 shows three examples of optical coherence tomography (OCT) (Huang *et al* 1991) B-scan images with IRF, SRF and PED. As can be seen from figure 1, IRF is usually located between the inner and outer nuclear layers and appears as low-reflective cystic edema, which increases the overall thickness of the retina. SRF is an exudate that accumulates between the photoreceptor cell layer and the retinal pigment epithelium (RPE), which is also a low-reflective area caused by retinal detachment. PED is caused by the separation of RPE layer and Bruch's membrane, which can be further divided into serous type (low-reflective and dome-shaped area), fibrovascular type (irregular-shaped area with non-uniform reflection) and drusen type (medium to high reflection area with smooth boundary) (Marmor 1999). In clinical cases, IRF, SRF and PED may appear simultaneously in the same eye. Studies have shown that retinal fluid is an effective biomarker, whose size, location and shape can provide accurate information for the diagnosis and treatment of macular edema (Ristau *et al* 2013, Waldstein *et al* 2016). Therefore, accurate segmentation of multi-class fluid in the retinal OCT image is of great importance, which is also challenging due to the various shapes, locations and blurred boundaries.
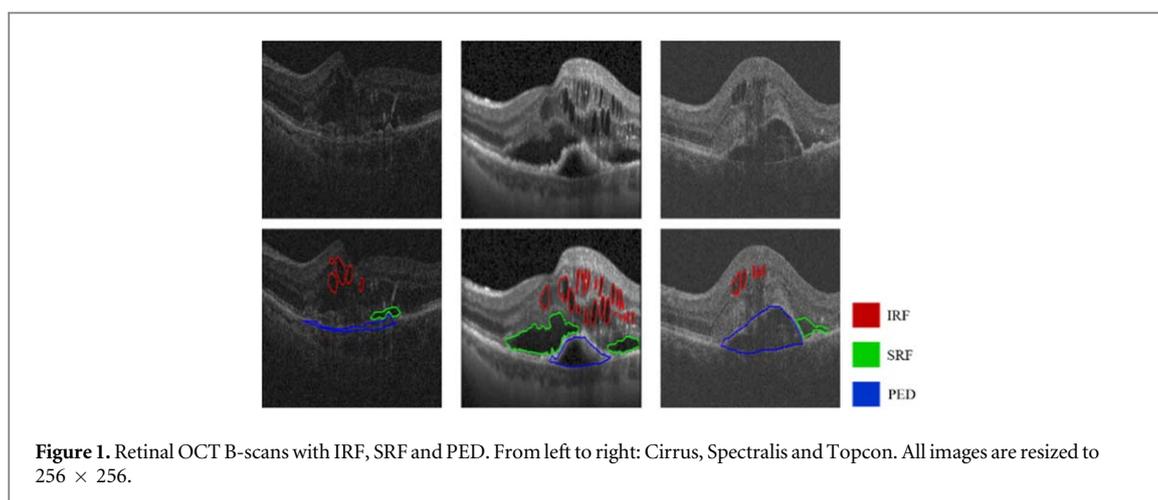
Figure 1. Retinal OCT B-scans with IRF, SRF and PED. From left to right: Cirrus, Spectralis and Topcon. All images are resized to 256 × 256.

In recent years, many researches have been proposed for the segmentation of retinal fluid in OCT images, which can be classified as traditional machine learning based methods, deep learning based methods and methods based on the combination of these two methods. In traditional machine learning based methods, Wilkins *et al* proposed a thresholding and boundary tracking based method to segment retinal fluid (Wilkins *et al* 2012). Xu *et al* obtained the fluid segmentation result by stratified sampling voxel classification (Xu *et al* 2015). Wu *et al* proposed a continuous max flow optimization random forest classification based method to segment serous retinal detachment (Wu *et al* 2017). Montuoro *et al* combined unsupervised feature representation and heterogeneous spatial context to realize the joint segmentation of retinal layer and fluid (Montuoro *et al* 2017). Wu *et al* proposed a Gaussian mixture model based intra-retinal cystoid macular edema segmentation method (Wu *et al* 2020). Rashno *et al* used a neutrosophic transformation and graph-based shortest path method to segment fluid/cyst regions in OCT images of subjects with diabetic macular edema (Rashno *et al* 2017). Rashno *et al* proposed a neutrosophic C-means clustering method for fluid segmentation in retinal OCT images (Rashno *et al* 2019). Esmaeili *et al* proposed a three-dimensional curvelet transform based dictionary learning for the automatic segmentation of intraretinal cysts in OCT images (Esmaeili *et al* 2016). The traditional methods were mainly well-designed for a certain type of fluid, whose segmentation efficiency is generally low.

There are some studies which tried to combine traditional machine learning based algorithms with convolutional neural network (CNN) to segment the retinal fluid. Lu *et al* proposed a framework for the multi-class fluid segmentation including IRF, SRF and PED in retinal OCT images, in which a fully convolutional neural network was trained to label the fluid pixels based on the intensity of OCT images and retinal layer segmentation provided by a graph-cut algorithm. Random forest classification was performed as post-processing to exclude the false positives (Lu *et al* 2019). Rashno *et al* proposed a graph shortest path algorithm and CNN based multi-class fluid segmentation method, in which the graph shortest path algorithm was used to segment internal limiting membrane (ILM) and RPE layers, PED was segmented by RPE layer flattening, and IRF and SRF were segmented by a CNN trained with supervision (Rashno *et al*). Gopinath *et al* designed a CNN to locate retinal cysts in three-dimensional OCT images by focusing on the axial movement of the OCT B-scan image to selectively enhance the lesion area. The final segmentation of cysts is obtained via K-means clustering of the detected cyst locations (Gopinath and Sivaswamy 2018). Compared with the traditional methods, the segmentation performance of the tradition and deep learning combination based methods is generally improved, but there is still room for improvement in segmentation efficiency.

With the rapid development of convolutional neural networks, more and more studies have focused on CNN based networks to segment retinal fluid in OCT images. Lee *et al* used a U-shape encoder–decoder structure based convolutional neural network to segment IRF (Lee *et al* 2017). Roy *et al* proposed a fully convolutional deep architecture based ReLayNet for end-to-end segmentation of retinal layers and fluid masses (Roy *et al* 2017). Schlegl *et al* proposed an encoder–decoder structure based CNN for IRF and SRF segmentation (Schlegl *et al* 2018). Venhuizen *et al* realized the automatic segmentation and quantification of intraretinal cystoid fluid by using the cascaded U-shape convolutional neural network and retinal anatomical prior information (Venhuizen *et al* 2018). These segmentation networks realized the end-to-end retinal fluid segmentation by using the relatively primitive U-shape encoder–decoder structure based networks, whose segmentation performances can be further improved by introducing strategies such as multi-scale information fusion and attention mechanism. Gao *et al* proposed a fully convolutional network DA-FCN based on dual-branch and region constraint to achieve automatic SRF segmentation, in which the dual branches structure

Figure 2. The overall architecture of the proposed segmentation framework. Skip connections for both branches are omitted for clarity.
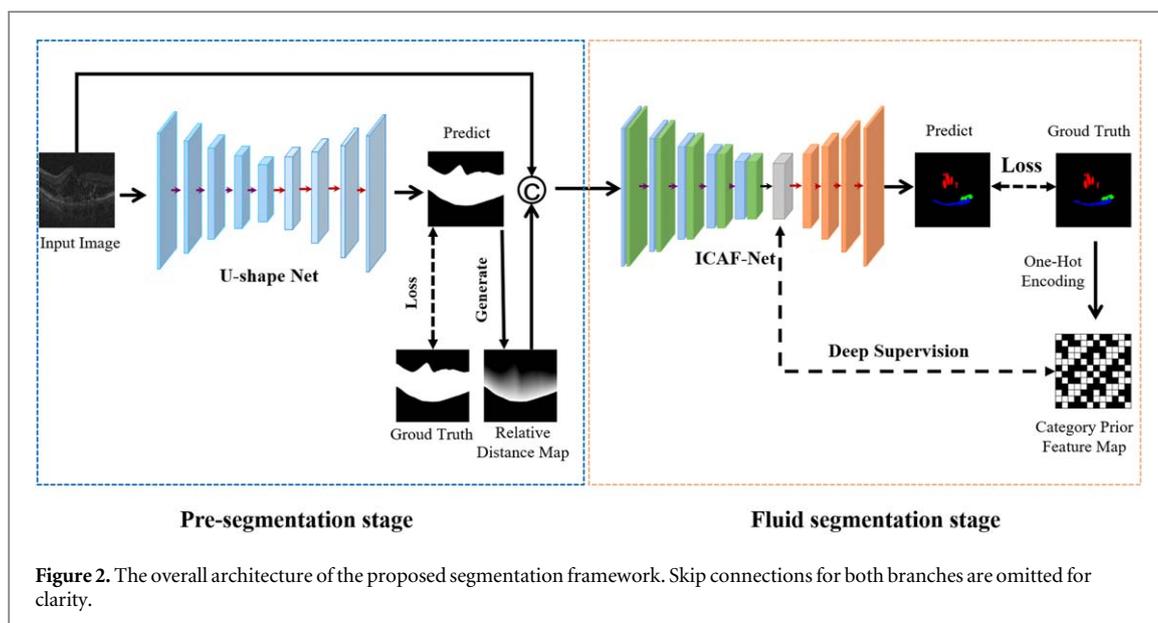
enables the network learn multi-level feature representations (Gao *et al* 2019). Yang *et al* proposed a residual multiple pyramid pooling network (RMPPNet) to realize the segmentation of retinal neurosensory layer detachment in OCT images. With residual architectures and pyramid pooling modules, RMPPNet can deal with the receptive field and multi-scale features well (Yang *et al* 2020). Chen *et al* used the improved convolutional neural network SEUNet based on the channel attention mechanism to segment the retinal fluid (Chen *et al* 2020). Feng *et al* designed CPFNet equipped with two pyramid modules to extract and fuse multi-scale information and achieved good performance in multi-class fluid segmentation including retinal edema area (REA), SRF and PED, in which the pyramid mechanism can effectively strengthen the segmentation ability of the network for targets with various sizes (Feng *et al* 2020). Hassan *et al* proposed a new multi-scale feature extractor module and demonstrated superior results in macular fluids segmentation including IRF, SRF and PED (San *et al* 2020). In our previous work, we proposed a multi-class retinal fluid automatic segmentation network CAF-Net based on context shrinkage encode (CSE) module and context pyramid guide module (Ye *et al* 2021), which can be further improved by considering the fluid prior information both in spatial location and fluid category.

Although there have been many methods for retinal fluid segmentation in OCT images as described above, most of these methods are designed for single-class fluid segmentation. Due to the various shapes, random locations and blurred boundaries of IRF, SRF and PED, the joint segmentation of multi-class fluid in OCT images is very challenging. In this paper, based on our earlier proposed CAF-Net and making full use of the prior information including the category information and the relative spatial location information of fluid, we propose a new multi-class retinal fluid joint segmentation framework based on the cascaded convolutional neural networks.

The main contributions of this paper are as follows:

(1) A two-stage multi-class retinal fluid segmentation framework based on ICAF-Net (Improved CAF-Net,) is proposed, which combines the prior information including the category information and relative spatial location of fluid to improve the multi-class retinal fluid joint segmentation performance.

(2) In the pre-segmentation stage, a U-shape network is used to obtain the retinal mask and a retinal relative distance map is generated, which can provide the prior information about the relative spatial location of the fluid.

(3) In the fluid segmentation stage, a novel multi-scale and multi-category semantic supervision (MSMC) module is proposed, in which the fluid category prior information generated by MSMC is fully used as the deep supervision to make ICAF-Net learn more intra-class consistency and inter-class difference in the advanced semantic features.

(4) The proposed framework was evaluated on the training and validation dataset of MICCAI2017 RETOUCH Challenge. The results of ablation experiments and comparison experiments show that the proposed method can effectively achieve the joint segmentation of multi-class retinal fluid in OCT images.
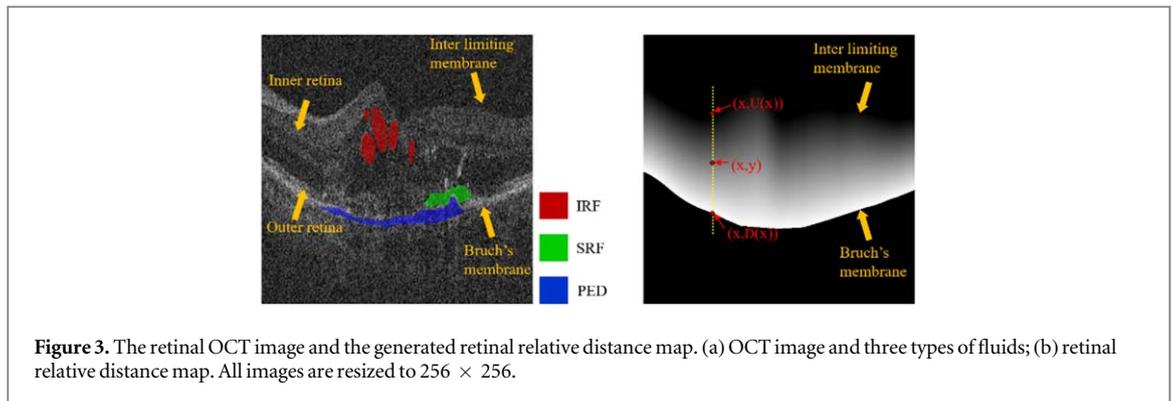
Figure 3. The retinal OCT image and the generated retinal relative distance map. (a) OCT image and three types of fluids; (b) retinal relative distance map. All images are resized to 256 × 256.
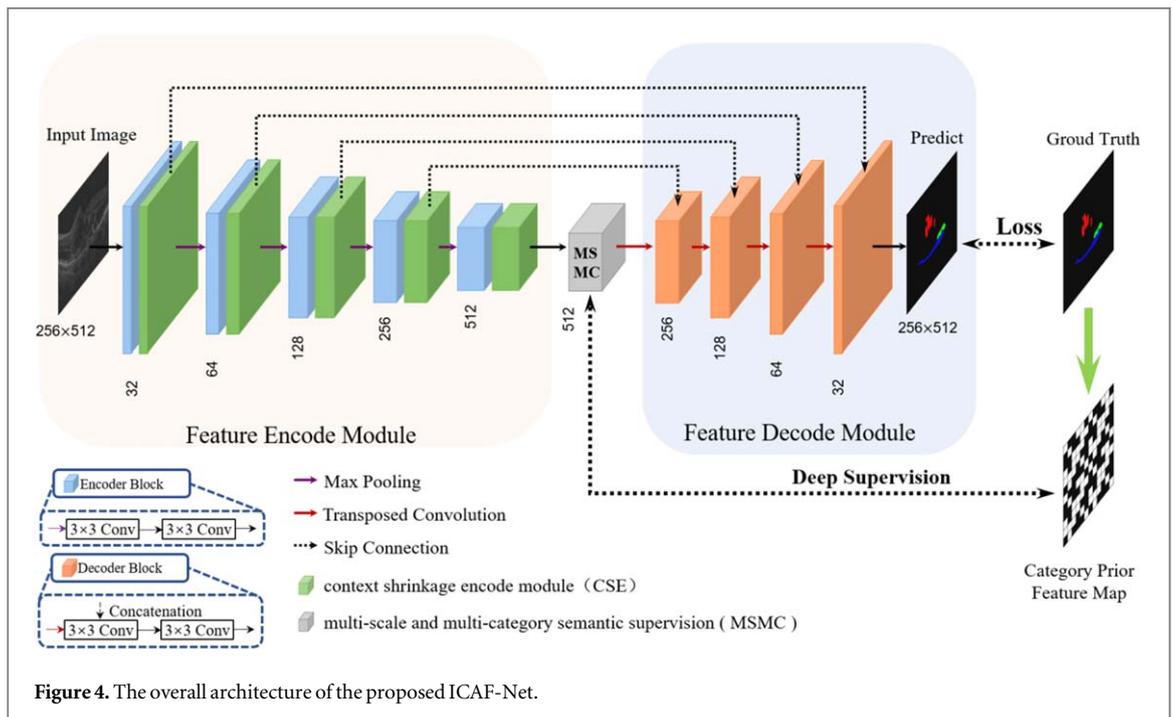
Figure 4. The overall architecture of the proposed ICAF-Net.

The rest of this paper is arranged as follows: in section 2, we describe the proposed multi-class retinal fluid segmentation framework in details. The experimental results are presented and analyzed in section 3, followed by the conclusions and discussions in section 4.

2. Methods

2.1. Overview of the framework architecture

Figure 2 is the overall illustration of the proposed multi-class retinal fluid joint segmentation framework based on the cascaded convolutional neural networks, which includes pre-segmentation stage and fluid segmentation stage. In the pre-segmentation stage, a U-shape encoder–decoder network is adopted to acquire the retinal mask and generate the retinal relative distance map, which can provide the spatial prior information for the next fluid segmentation. The retinal relative distance map is concatenated to the original retinal OCT image in the channel dimension and taken as the input of the next fluid segmentation stage. In the fluid segmentation stage, the ICAF-Net based on CSE module and multi-scale and MSMC module is adopted to jointly segment IRF, SRF and PED.

2.2. Pre-segmentation stage

According to the clinical knowledge, although the quantity, shape and spatial location of IRF, SRF and PED inter-eyes are various, the three generally maintain a relatively stable relationship in the depth direction, that is, from the inner retina to the outer retina (from top to bottom in figure 3), they usually are IRF, SRF and PED. This spatial prior information is useful to the joint fluid segmentation.

First, the retinal region (from the inner limiting membrane to the Bruch's membrane in figure 3(a)) is segmented by a U-shape network with pre-trained ResNet34 (He *et al* 2016) as backbone. A joint loss function

Figure 5. The structure of multi-scale and multi-category semantic supervision (MSMC) module.
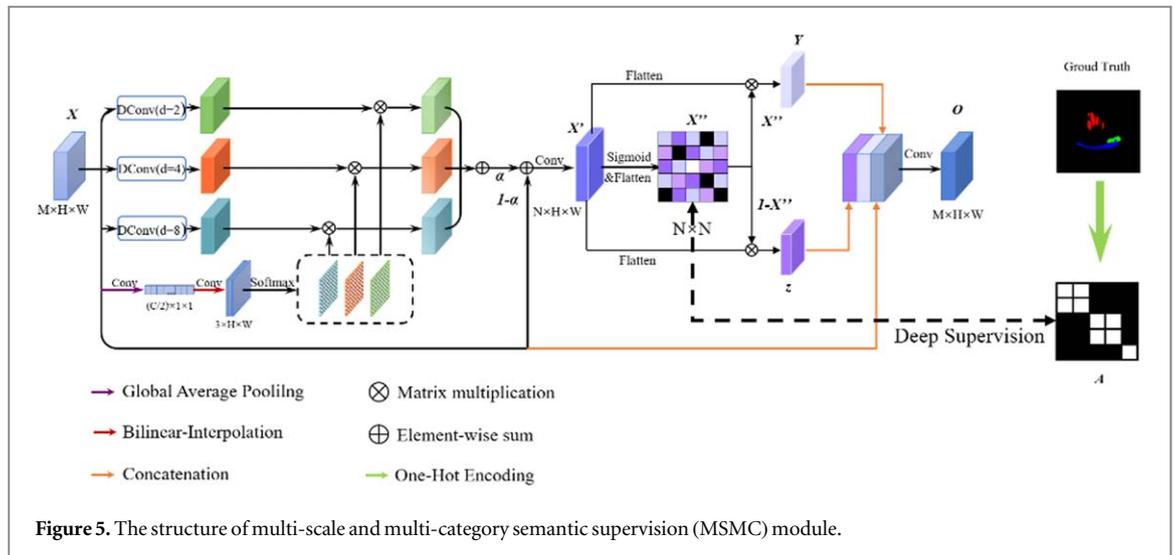
consisting of cross entropy loss and exponential logarithmic Dice loss is used during the network training. Due to the retinal boundary deformation caused by the fluid, the retinal region segmentation result is post-processed including morphological open and close operation, maximum connected area extraction and curve fitting to obtain a more accurate retinal mask. Then, the relative distance map of the retina is generated based on the retinal mask. Specifically, the value $I(x, y)$ of point (x, y) in the retinal relative distance map can be calculated as follows:

$$I(x, y) = \frac{y - U(x)}{D(x) - U(x)}, \tag{1}$$

where $U(x)$ represents the vertical ordinate of the upper retinal boundary of point (x, y) in column, that is, the vertical ordinate of the inner limiting membrane. $D(x)$ represents the vertical ordinate of the lower boundary of point (x, y) in column, that is, the vertical ordinate of the Bruch's membrane. Figure 3(b) shows the retinal relative distance map generated according to Formula (1), which can be used as the spatial prior information to guide the specific target feature extraction and enhancement in the subsequent retinal fluid segmentation network.

2.3. Fluid segmentation stage

Figure 4 shows the overall structure of ICAF-Net, which mainly composes of feature encoder, CSE module, MSMC module and feature decoder. The channel attention mechanism and soft thresholding based CSE module is the same as that is proposed in our previous CAF-Net (Ye *et al* 2021) (shown in figure 5) and embedded after each feature encoder. The MSMC module is embedded in the top of the feature encoder path.

2.3.1. Feature encoder and decoder

In order to get representative feature maps, we adopt the same feature encoder path as the original U-Net (Ronneberger 2015). The number of feature maps of each encoder is halved to reduce the risk of overfitting. In order to efficiently reconstruct high-resolution feature maps, each feature decoder of ICAF-Net includes a 3×3 convolution and a transposed convolution. The number of feature maps in the decoder path is also halved compared to the original U-Net. Specifically, the decoder fuses the feature maps from the corresponding CSE module (Donoho 1995) through skip connection and 3×3 convolution, and then upsamples the fused feature map through transposed convolution.

2.3.2. Multi-scale and MSMC module

Images generally contain category prior information, which determines whether there is a target of a certain class in the image and is the basis for pixel-level classification (Yu *et al* 2020). We propose a multi-scale and multi-class semantic supervision (MSMC) module (shown in figure 5) to aggregate long-range and multi-scale semantic information, then deep supervision will be applied to directly supervise the aggregation of high-level features and distinguish the intra-class and inter-class context information. As shown in figure 5, assuming the size of the input feature map X is $M \times H \times W$ (M is the number of channels, H and W are the height and the width of the feature map respectively). First, we use three parallel dilated convolutions to aggregate multi-scale information and global average pooling to capture long-range information. Second, we use 3×3 convolution with learnable parameters to resize $X \in \mathbb{R}^{M \times H \times W}$ to $X' \in \mathbb{R}^{N \times H \times W}$ ($N = H \times W$). The category prior affinity map A, which

Figure 6. Generation of the category priori affinity map *A*.
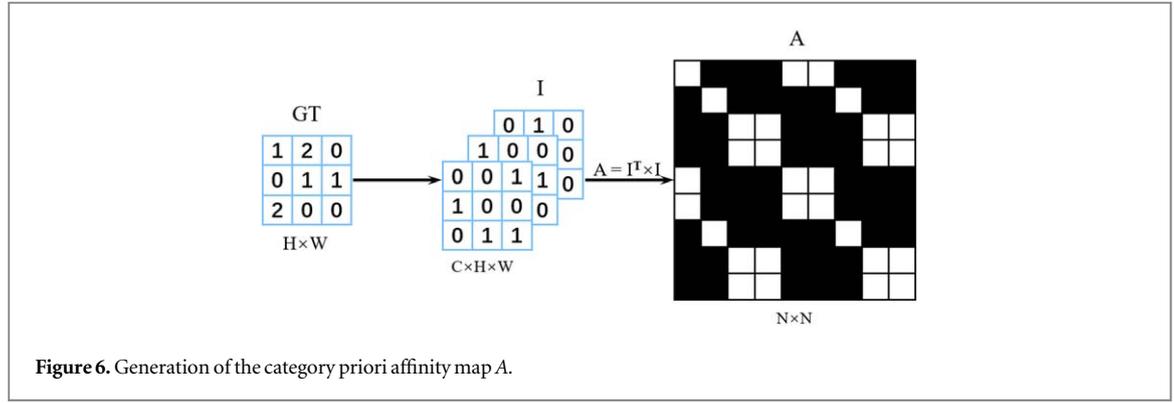
is generated based on the ground truth and one-hot encoding (Yu *et al* 2020, Karthiga *et al* 2021), is adopted to guide the category prior feature map X'' to learn the intra-class and inter-class relationships between pixels. The feature map X' is reshaped to $N \times N (N = H \times W)$ and followed by sigmoid activation function to generate category prior feature map X''. Third, matrix multiplication is performed between reshaped X' and X'' to generate intra-class context information feature Y. And the inter-class context information feature Z is adaptively learned by multiplying X' with $1 - X''$. Finally, the input feature map X, intra-class context information feature map Y and inter-class context information feature map Z are concatenated in the channel dimension, and a 1×1 convolution is used to fuse and restore the dimension of the output feature map O to $M \times H \times W$. The overall MSMC module can be summarized as:

$$X'' = \sigma(Re@X') \in \mathbb{R}^{N \times N} \tag{2}$$

$$Y = X' \otimes X'' \in \mathbb{R}^{N \times H \times W} \tag{3}$$

$$Z = X' \otimes (1 - X'') \in \mathbb{R}^{N \times H \times W} \tag{4}$$

$$O = Conv(X' © Y © X) \in \mathbb{R}^{M \times H \times W}, \tag{5}$$

where σ represents the sigmoid activation, $Re@$ represents the operation of reshape, $\otimes$ represents the matrix multiplication, © represents the operation of concatenation, $Conv$ represents the 1×1 convolution.

In order to make full use of the pixel-wise category prior information contained in the ground truth, the category prior affinity map A is generated based on the ground truth and one-hot encoding (Yu *et al* 2020, Karthiga *et al* 2021) and adopted as the deep supervision of the network training, which can guide the category prior feature map X'' to learn the intra-class and inter-class relationships between pixels. Figure 6 shows the generation of the category prior affinity map A. First, the ground truth is resized to $GT \in \mathbb{R}^{H \times W}$. Second, one-hot encoding is adopted to encode each category label in GT to get $I \in \mathbb{R}^{C \times H \times W}$, where C represents the number of categories. Third, $I \in \mathbb{R}^{C \times H \times W}$ is reshape to $I \in \mathbb{R}^{C \times N}$, where $N = H \times W$. At last, the category prior affinity map A is obtained through $A = I^T \times I \in \mathbb{R}^{N \times N}$.

2.4. Loss function

To overcome the class distribution imbalance problem, the joint of multi-class cross entropy loss L_{CE} and exponential logarithmic Dice loss (Wong *et al* 2018). L_{ELDice} is used as the joint segmentation loss function L_{Seg} to supervise the prediction results of the network. The definition of L_{CE}, L_{ELDice} and L_{Seg} are as follows:

$$L_{Seg} = L_{CE} + \lambda_1 L_{ELDice} \tag{6}$$

$$L_{CE} = -\frac{1}{T} \sum_{i=1}^{T} \sum_{c=1}^{C} y_{i,c} \log(p_{i,c}) \tag{7}$$

$$L_{ELDice} = \frac{1}{C} \sum_{c=1}^{C} \left\{ -ln \left[\frac{2\sum_{i=1}^{T} y_{i,c} p_{i,c} + \varepsilon}{\sum_{i=1}^{T}(y_{i,c} + p_{i,c}) + \varepsilon} \right]^{\gamma} \right\}$$
$$\times (\gamma = 0.5), \tag{8}$$

where λ_1 is a trade-off between L_{CE} and L_{ELDice} and is set to 1 in all experiments. $y_{i,c} \in \{0, 1\}$ represents the probability that the pixel i belongs to class c in ground truth. $p_{i,c} \in [0, 1]$ represents the probability that the pixel i is predicted to be class c. T represents the total number of pixels in the image, and C represents the total number of categories. ε is a smooth factor and is set to 1e-6.

In order to better guide the network to model the intra-class and inter-class relationships, the category prior feature map X' is deeply supervised during the network training by adopting the auxiliary loss function L_{Aux}, which combines the unstructured loss L_{BCE} and the structured loss L_G:

$$L_{Aux} = L_{BCE} + \lambda_2 L_G, \tag{9}$$

where L_{BCE} is the binary cross-entropy loss and L_G means global-based structured loss. λ_2 is a hyperparameter to balance the two losses and is set to 1 in all experiments.

The definition of L_{BCE} is as follows:

$$L_{BCE} = -\frac{1}{N^2}\sum_{i=1}^{N^2}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i)), \tag{10}$$

where $y_i \in \{0, 1\}$ represents the probability that the pixel i belongs to the target in the category prior affinity map A, $p_{i,c} \in [0, 1]$ represents the probability that the pixel i is predicted to be the target in the category prior feature map A. N represents the rows of A, N^2 represents the total number of pixels in A.

The definition of L_G is as follows:

$$L_G = -\frac{1}{N}\sum_{k=1}^{N}(\Phi_k^p + \Phi_k^r + \Phi_k^s) \tag{11}$$

$$\Phi_k^p = \log\frac{\sum_{i=1}^{N} y_k^i p_k^i}{\sum_{i=1}^{N} p_k^i} \tag{12}$$

$$\Phi_k^r = \log\frac{\sum_{i=1}^{N} y_k^i p_k^i}{\sum_{i=1}^{N} y_k^i} \tag{13}$$

$$\Phi_k^s = \log\frac{\sum_{i=1}^{N}(1 - y_k^i)(1 - p_k^i)}{\sum_{i=1}^{N}(1 - y_k^i)}, \tag{14}$$

where Φ_k^p, Φ_k^r and Φ_k^s represent the intra-class accuracy, true intra-class rate and true inter-class rate of the kth row in the category prior feature map X'', respectively. N represents the rows of A and X''. $y_k^i \in \{0, 1\}$ represents the probability that the pixel i belongs to the target in the kth row of the category prior affinity map A, $p_k^i \in [0, 1]$ represents the probability that the pixel i is predicted to be the target in the kth row of the category prior feature map X''.

The joint segmentation loss function L_{seg} and auxiliary loss function L_{Aux} are combined as the total loss function of the proposed ICAF-Net:

$$L_{Total} = L_{Seg} + \lambda_3 L_{Aux}, \tag{15}$$

where λ_3 is a trade-off to balance L_{Seg} and L_{Aux}, and is set to 1 in all our experiments.

3. Experiment settings

3.1. Dataset
The dataset used in this paper was acquired from a public competition: RETOUCH Challenge in MICCAI2017 (only training set and validation set are available) (Bogunović *et al* 2019), including 70 three-dimensional retinal OCT images collected by 3 different types of devices: Cirrus (24 volumes with size $1024 \times 512 \times 128$), Spectralis (24 volumes with $496 \times 512 \times 49$) and Topcon (11 volumes with $885 \times 512 \times 128$ and 11 volumes with $650 \times 512 \times 128$). The ground truth is obtained from manual voxel-wise annotations of the fluid lesions, which were completed by 4 or 2 graders under the supervision of retinal specialist. Figure 1 shows some OCT B-scans acquired by these three devices. It can be seen from figure 1. that there are obvious differences in noise level and contrast between the images from different devices. Three-fold cross-validation strategy is used to objectively evaluate the proposed multi-class fluid segmentation framework and each fold contains images from different types of devices. Data augmentation strategies including random horizontal flip, random rotation ($\pm10^{\circ}$) and Gaussian noise addition are applied during the training process.

3.2. Evaluation metrics
In order to objectively evaluate the proposed multi-class retinal fluid joint segmentation framework, three evaluation indicators including Dice similarity coefficient (DSC), intersection over union (IoU) and accuracy (Acc) are adopted. These three metrics are also used in the pre-segmentation stage to evaluate the retinal region segmentation performance. Metrics are calculated as follows:

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{16}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{17}$$
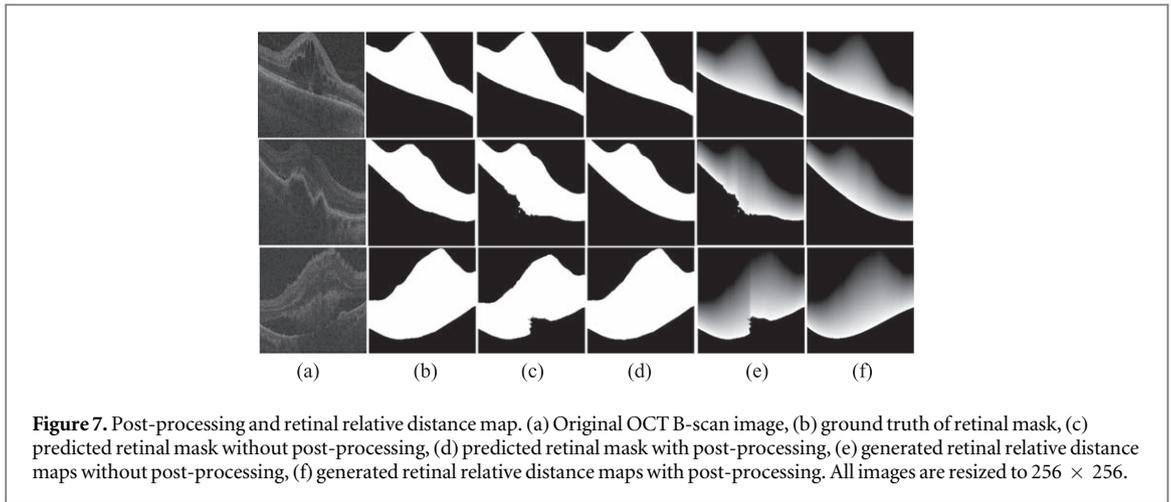
Figure 7. Post-processing and retinal relative distance map. (a) Original OCT B-scan image, (b) ground truth of retinal mask, (c) predicted retinal mask without post-processing, (d) predicted retinal mask with post-processing, (e) generated retinal relative distance maps without post-processing, (f) generated retinal relative distance maps with post-processing. All images are resized to 256 × 256.

Table 1. The ablation experiment results for retinal region segmentation (%).

Methods	DSC	IoU	Acc
U-shape network	99.01 ± 0.75	98.04 ± 1.41	99.35 ± 0.47
U-shape network + pre-trained ResNet34	99.08 ± 0.75	98.20 ± 0.14	99.41 ± 0.50
U-shape network + pre-trained ResNet34 + post-processing	99.09 ± 0.74	98.21 ± 0.14	99.42 ± 0.46

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}, \tag{18}$$

where TP, FP, TN and FN represent true positive, false positive, true negative and false negative, respectively.

3.3. Implementation details

The proposed method is implemented based on the public platform Pytorch and NVIDIA RTX 2080Ti with 11 GB memory. In the training process of both the pre-segmentation U-shape network and the fluid segmentation ICAF-Net, stochastic gradient descent (SGD) algorithm with poly learning rate policy is used to optimize the weights of the network with learning rate lr:

$$lr = lr_b \times \left(1 - \frac{Iter}{Iter_t}\right)^p, \tag{19}$$

where $Iter$ and $Iter_t$ represent the current number of iterations and the total number of iterations respectively. In all ablation and comparison experiments, the basic learning rate lr_b is set to 0.01 and the declining index p is set to 0.9. The batch size is set to 8 and the epochs is set to 50.

4. Results

4.1. Experimental results of the pre-segmentation stage

To evaluate the retinal region segmentation performance in the pre-segmentation stage, three ablation experiments are conducted, including basic U-shape network, U-shape network with pre-trained ResNet34 (U-shape network + pre-trained ResNet34) and U-shape network with pre-trained ResNet34 and curve fitting based post-processing (U-shape network + pre-trained ResNet34 + post-processing). Table 1 shows the results of these three ablation experiments. As shown in figure 1, because the contrast between the retinal region and the background is relatively large, the segmentation of retinal region in OCT images is a relatively easy task and good segmentation performance can be achieved by using the basic U-shape network with all evaluation metrics over 98%. It can be seen from table 1 that the loading of the pre-trained ResNet34 model can further improve the retinal region segmentation performance, indicating that loading the pre-trained weights can not only promote the convergence of the network, but also enhance the network's ability to capture useful features and finally optimize the segmentation performance. In a few B-scans, the retinal region segmentation results are affected by fluid (mainly PED), resulting in severe errors as shown in the last row of figure 7(c). For most of the B-scans, the retinal region segmentation results are either the correct one as shown in the first row of figure 7(c), or the one with a few false negatives and/or false positives s as shown in the second row of figure 7(c). But as shown in the second and last rows of figure 7(e), the corresponding retinal relative distance maps may provide

Figure 8. The segmentation results of ablation experiments on multi-class fluid joint segmentation. (a) Original OCT image, (b) ground truth, (c) result of Baseline, (d) result of Baseline + CSE, (e) result of Baseline + CSE + MSMC, (f) result of the proposed method. Red represents IRF, green represents SRF and blue represents PED. All images are resized to 256 × 256.
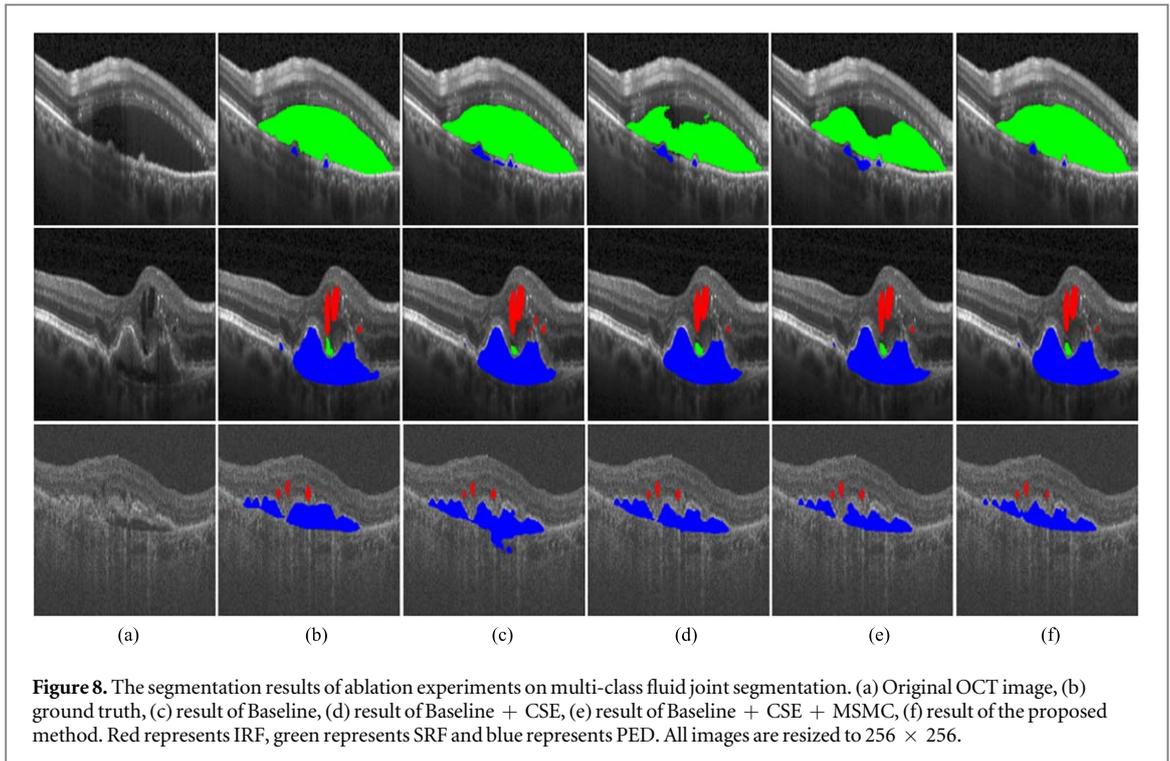
negative spatial prior information for the next fluid segmentation stage. Some post-processing operations are used to correct the retinal boundaries. First, morphological operations including open and close operations with circular element (radius of 5) are adopted to fill the small holes and make the retinal boundary smoother. Second, the maximum connected area is extracted as the retinal region. Third, cubic curve fitting is performed to correct the retinal lower boundary. Then, the maximum ordinate error between the fitted retinal lower boundary and the original lower boundary is calculated. If the error is greater than 15 pixels, the original lower boundary will be replaced with the fitted one. Figure 8 shows the comparison of the results of the retinal relative distance map before and after post-processing. As can be seen from the comparison of the relative distance maps shown in the last two columns of the first row and third row of figure 7 (figures 7(e) and (f)), post-processing is of great significance for the correction of the relative distance map, which in turn helps to improve the subsequent fluid segmentation performance.

4.2. Experimental results of the fluid segmentation stage

4.2.1. Ablation experiments

Five ablation experiments are conducted, including Baseline (U-Net with half number of feature map channels), Baseline with CSE module (Baseline +CSE), Baseline with CSE and MSMC modules (Baseline + CSE + MSMC, ICAF-Net), Baseline with CSE and MSMC modules and the retinal relative distance map without post-processing (proposed w/o post-processing) and the proposed method. Table 2 shows the results of the ablation experiments. The DSC metrics of the proposed framework for IRF, SRF and PED are 73.62%, 80.92% and 74.64% respectively, and the average DCS reaches 76.39%, which achieves a significant improvement compared with the Baseline and means that the use of the context information and prior information can achieve robust feature expression and improve the multi-target recognition ability.

Table 2 shows that, compared with Baseline, the addition of CSE module improves the segmentation performance of all three types of fluid, especially for IRF and SRF. The possible reason is that the feature information of the retinal region is more abundant than the background region, which may be also redundant for fluid segmentation. Compared with feature extraction only by stacking convolutional layers in the Baseline, the CSE module can distinguish features according to the global context information and make the semantic information of output feature map have a stronger correlation with the target fluid. That is, the CSE module can suppress redundant information and retain useful information, thereby effectively extract key information from noisy feature maps.

It can be seen from table 2 that the category prior information introduced by MSMC has significantly improved the fluid segmentation performance, with average DSC and average IoU increased by 2.12% and 2.05% respectively. In addition, it can also be found that the segmentation performance improvements of IRF, SRF and PED are relatively close. The possible reason is that the category prior information is equally important

Table 2. The ablation experiment results for multi-class fluid joint segmentation (%).

Methods	DSC				IoU				Acc
	IRF	SRF	PED	Average[a]	IRF	SRF	PED	Average[a]	Overall[b]
Baseline	71.09 ± 13.95	75.26 ± 22.98	68.50 ± 22.08	71.62 ± 19.67	56.74 ± 15.15	64.41 ± 23.24	55.63 ± 22.08	58.93 ± 20.16	99.13 ± 1.03
Baseline +CSE	72.76 ± 13.50	78.12 ± 18.12	68.60 ± 22.89	73.16 ± 18.17	58.72 ± 14.86	66.92 ± 19.97	56.11 ± 23.38	60.58 ± 19.40	99.22 ± 0.88
Baseline+CSE + MSMC(ICAF-Net)	**73.78** ± 15.03	80.10 ± 20.12	71.96 ± 22.38	75.28 ± 19.18	59.63 ± 15.46	69.03 ± 21.00	59.22 ± 23.27	62.63 ± 19.91	99.26 ± 0.89
Proposed + w/o post-processing[c]	72.92 ± 12.99	79.80 ± 16.67	73.91 ± 17.63	75.54 ± 15.76	58.84 ± 14.67	68.88 ± 18.81	61.26 ± 19.72	62.99 ± 17.73	99.24 ± 0.88
Proposed	73.62 ± 13.19	**80.92** ± 12.68	**74.64** ± 19.08	**76.39** ± 14.98	**59.72** ± 14.54	**69.65** ± 16.43	**62.72** ± 21.78	**64.03** ± 17.58	**99.32** ± 0.78

[a] Means the average value of the corresponding indexes of IRF, SRF and PED respectively.

[b] The overall segmentation accuracy of IRF, SRF and PED.

[c] Means the proposed two-stage segmentation framework without the post-processing in the retinal region segmentation stage.
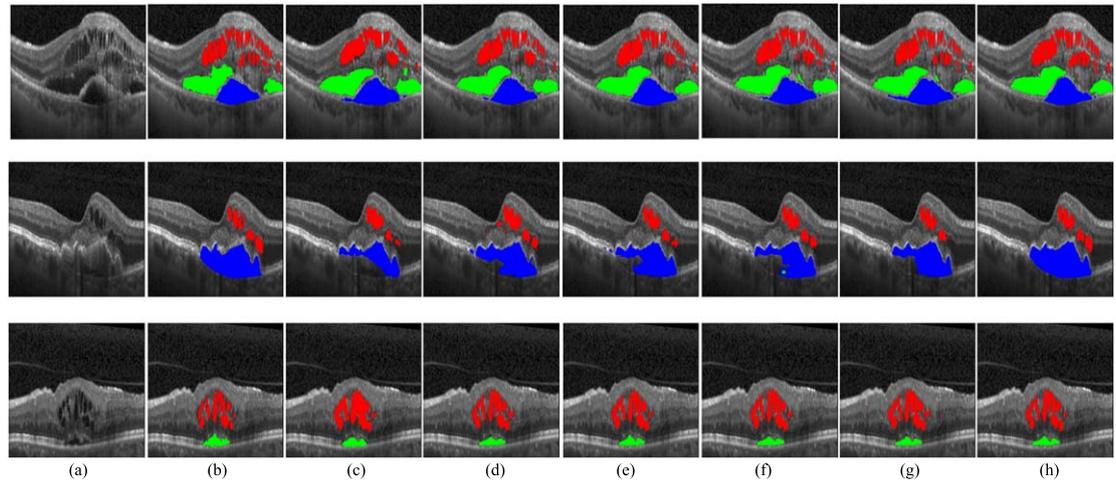
Figure 9. The segmentation results of different methods. (a) Original OCT images, (b) ground truth, (c) the results of FCN, (d) the results of CE-Net, (e) the results of CPFNet, (f) the results of U-Net, (g) the results of CAF-Net, (h) the results of the proposed method. Red represents IRF, green represents SRF and blue represents PED. All images are resized to 256 × 256.

for all the three types of fluids and MSMC can effectively aggregate long-range and multi-scale semantic information and selectively capture the intra-class and inter-class dependencies between pixels, which makes the network learn more intra-class consistency and inter-class difference. It can be seen from table 2 that the spatial prior information introduced by the retinal relative distance map can effectively improve the segmentation performance of PED, increasing its average DSC, average IoU and Acc by 1.11%, 1.40% and 0.06% respectively. The retinal relative distance map based spatial prior information has little effect on the segmentation of IRF. We think the possible reason is that the spatial position of IRF in the retina is relatively fixed and stable in depth direction, which can be learned well by ICAF-Net during network training. While the distributions of SRF and PED are relatively close, it is necessary to use the retina relative distance map to distinguish them effectively. Comparing the fourth and fifth rows in table 2, it can be seen that if there is no post-processing in the retinal region segmentation stage, it may cause errors in the relative distance map generated subsequently (as shown in figure 7(e)), which in turn leads to a decline in the overall segmentation performance of the proposed framework.

    Figure 8 shows some multi-class retinal fluid joint segmentation results of ablation experiments, which shows that the use of CSE module, MSMC and the retinal relative distance map information can improve the segmentation performance of the proposed network effectively.

4.2.2. Comparison experiments
In order to evaluate its performance, the proposed multi-class fluid joint segmentation framework is compared with some other state-of-the-art deep learning based image semantic segmentation networks, including U-Net (Ronneberger *et al* 2015), Attention U-Net (Oktay *et al* 2018), CE-Net (Gu *et al* 2019), FCN (Shelhamer *et al* 2017), SegNet (Badrinarayanan *et al* 2017), PSPNet (Zhao *et al* 2017), CPFNet (Feng *et al* 2020), DeepLab-V3 (Chen *et al* 2017a, 2017b), DANet (Fu *et al* 2019) and our previous CAF-Net (Ye *et al* 2021). For fair comparison, the experimental implementation details of all methods are kept consistent.

    Table 3 shows the comparison results of different methods. As can be seen from table 3, the proposed framework achieves the best performance. The segmentation performance of FCN is bad, which may because that FCN loses feature information for the small targets during deconvolution based feature map upsampling without reasonable skip connections. SegNet uses the max-pooling indices to upsample the feature maps without learning and also fails to outperform U-Net with classical skip connections, indicating that the full fusion of high-level and low-level feature information is important in restoring the local details of the feature map. Although DANet introduces spatial and channel attention modules, its segmentation performance is not good, which may because its decoder structure is too simple to be suitable for the multi-class fluid segmentation task. Attention U-Net introduces the attention gate module in the skip connection and achieves some improvement compared with U-Net, indicating attention mechanism plays an important role in image semantic segmentation. CE-Net combines pooling operation and dilated convolution to design a multi-scale contextual information extraction module, which outperforms PSPNet and DeepLab-V3. CPFNet combines two pyramidal modules to fuse global and multi-scale context information and outperforms CE-Net. CAF-Net introduces CSE module and context pyramid guide (CPG) module and outperforms all the networks mentioned

Table 3. Comparisons of different methods (%).

Methods	DSC				IoU				Acc Overall
	IRF	SRF	PED	Average	IRF	SRF	PED	Average	
DANet (Fu te al 2019)	58.34 ± 15.64	68.97 ± 23.16	61.62 ± 23.91	62.98 ± 20.90	42.73 ± 14.39	56.61 ± 23.57	48.21 ± 22.23	49.18 ± 20.06	98.92 ± 1.14
DeepLab-V3 (Chen *et al* 2017a)	59.71 ± 14.47	69.37 ± 24.21	65.08 ± 22.29	64.72 ± 20.32	43.97 ± 14.06	57.41 ± 24.34	51.63 ± 21.52	51.01 ± 19.96	98.96 ± 1.07
FCN (Shelhamer *et al* 2017)	65.24 ± 14.03	72.95 ± 22.88	64.19 ± 21.17	67.46 ± 19.36	49.87 ± 14.06	61.41 ± 22.63	50.42 ± 20.45	53.90 ± 19.05	98.98 ± 1.10
PSPNet (Zhao *et al* 2017)	66.28 ± 14.74	74.09 ± 22.43	66.40 ± 22.37	68.92 ± 19.85	51.14 ± 14.72	62.68 ± 22.57	53.25 ± 22.13	55.69 ± 19.86	99.08 ± 1.02
CE-Net (Gu *et al* 2019)	69.02 ± 16.31	75.12 ± 22.35	68.34 ± 23.99	70.83 ± 20.88	54.70 ± 16.57	64.07 ± 22.95	55.98 ± 23.46	58.25 ± 20.99	99.16 ± 0.96
SegNet (Badrinarayanan *et al* 2017)	69.31 ± 15.58	76.31 ± 20.78	68.06 ± 24.02	71.23 ± 20.12	54.89 ± 16.06	65.26 ± 22.20	55.75 ± 23.89	58.63 ± 20.72	99.17 ± 1.00
CPFNet (Feng *et al* 2020)	69.55 ± 14.27	76.25 ± 18.69	69.11 ± 21.93	71.64 ± 18.30	54.94 ± 15.17	64.66 ± 20.94	56.36 ± 22.34	58.66 ± 19.48	99.19 ± 0.90
U-Net (Ronneberger *et al* 2015)	72.06 ± 12.63	77.22 ± 19.11	70.34 ± 20.21	73.21 ± 17.32	57.69 ± 14.15	66.06 ± 21.19	57.33 ± 20.72	60.36 ± 18.69	99.19 ± 0.92
Attention U-Net (Oktay *et al* 2018)	71.95 ± 14.46	77.87 ± 18.84	70.56 ± 20.98	73.46 ± 18.10	57.90 ± 15.53	66.80 ± 20.59	57.96 ± 22.10	60.89 ± 19.41	99.19 ± 1.00
CAF-Net (Ye *et al* 2021)	73.17 ± 13.49	79.70 ± 13.80	71.06 ± 22.12	74.64 ± 16.47	59.21 ± 14.68	68.12 ± 16.92	58.92 ± 23.20	62.08 ± 18.27	99.26 ± 0.84
Proposed	73.62 ± 13.19	80.92 ± 12.68	74.64 ± 19.08	76.39 ± 14.98	59.72 ± 14.54	69.65 ± 16.43	62.72 ± 21.78	64.03 ± 17.58	99.32 ± 0.78

Table 4. Comparison of parameters and training time of different methods (M: Mbyte, H: hour, Sec/image: second per image).

Methods	DANet	DeepLab V3	FCN	PSPNet	CE-Net	CPFNet	SegNet	U-Net	Attention U-Net	CAF-Net	proposed
Parameters (M)	66.42	58.16	30.32	27.49	38.96	21.05	29.44	34.52	34.88	22.89	23.18
Training time (H)	4.42	4.88	4.68	3.77	3.82	3.93	6.43	10.03	11.28	4.90	5.63
Prediction time (Sec/image)	0.022	0.017	0.019	0.019	0.021	0.021	0.018	0.022	0.022	0.019	0.021

above. Figure 9 shows some example of comparison experiments, which indicates that the proposed cascaded segmentation framework outperforms other state-of-the-art networks.

In addition, in order to further verify the performance of the proposed method, we compare the parameters and training time of the different methods mentioned in comparison experiments, which is shown in table 4. As can be seen from table 4, the proposed method achieves the best performance with relatively less parameters and training time.

5. Conclusion and discussion

In this paper, we propose a two-stage multi-class retinal fluid joint segmentation framework based on cascaded convolutional neural networks. In the pre-segmentation stage, the retinal relative distance map is generated according to the retinal mask segmented by a simple U-shape network, which can provide spatial prior information for the next fluid segmentation stage. In the fluid segmentation stage, an ICAF-Net is proposed based on CSE modules and MSMC, which can fully use the context information and category prior information to guide the feature extraction and fusion. The RETOUCH challenge dataset (training and validation set) is used to evaluate the proposed framework. The quantitative and qualitative analysis of the experimental results shows the proposed segmentation framework achieves good performance in the multi-class retinal fluid segmentation task.

Although the proposed framework performs well in the multi-class fluid segmentation task, the dataset used in this paper is only composed of 70 three-dimensional retinal OCT images, which is far from enough to ensure the generalization of the model. Even if the dataset can be augmented by collecting more OCT images in the future, the labeling of ground truth will be time-consuming. To further improve the segmentation performance and generalization of the proposed framework, we will focus on utilizing the domain adaptive ability of the transfer learning technology (Long *et al* 2015, Chen *et al* 2019, Sahu *et al* 2021), trying to use a small amount of labeled data to guide and enhance the network training based on a large amount of unlabeled data. In addition, with the subsequent collection of 3D OCT volumes and the annotation of the corresponding ground truth, we will try to extent our network for 3D fluid segmentation, which can make full use of the spatial structure prior information of 3D OCT volumes.

Acknowledgments

ORCID iDs

Wei Tang ● https://orcid.org/0000-0001-7580-4132
Xinjian Chen ● https://orcid.org/0000-0002-0871-293X
Fei Shi ● https://orcid.org/0000-0002-8878-6655
Dehui Xiang ● https://orcid.org/0000-0001-7873-9778
Weifang Zhu ● https://orcid.org/0000-0001-9540-4101

References

Badrinarayanan V, Kendall A and Cipolla R 2017 SegNet: A deep convolutional encoder-decoder architecture for image segmentation *IEEE Trans. Pattern Anal. Mach. Intell.* **39** 2481–95

Bogunović H *et al* 2019 RETOUCH: the retinal OCT fluid detection and segmentation benchmark and challenge *IEEE Trans. Med. Imaging* **38** 1858–74

Bringmann A, Reichenbach A and Wiedemann P 2004 Pathomechanisms of cystoid macular edema *Ophthalmic Res.* **36** 241–9

Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L 2017 Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution and fully connected CRFs *IEEE Trans. Pattern Anal. Mach. Intell.* **40** 834–48

Chen L, Papandreou G, Schroff F and Adam H 2017 Rethinking atrous convolution for semantic image segmentation arXiv:1706.05587

Chen M, Xue H and Cai D 2019 Domain adaptation for semantic segmentation with maximum squares loss *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)* pp 2090–9

Chen Z, Li D, Shen H, Mo H, Zeng Z and Wei H 2020 Automated segmentation of fluid regions in optical coherence tomography B-scan images of age-related macular degeneration *Opt. Laser Technol.* **122** 105830

Donoho D L 1995 De-noising by soft-thresholding *IEEE Trans. Inf. Theory* **41** 613–27

Esmaeili M, Dehnavi A M, Rabbani H and Hajizadeh F 2016 Three-dimensional segmentation of retinal cysts from spectral-domain optical coherence tomography images by the use of three-dimensional curvelet based K-SVD *J. Med. Signals Sens.* **6** 166–71

Feng S, Zhao H, Shi F, Cheng X and Chen X 2020 CPFNet: Context pyramid fusion network for medical image segmentation *IEEE Trans. Med. Imaging* **39** 3008–18

Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z and Lu H 2019 Dual attention network for scene segmentation *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* pp 3146–54

Gao K, Niu S, Ji Z, Wu M, Cheng Q, Xu R, Yuan S, Fan W, Chen Y and Dong J 2019 Double-branched and area-constraint fully convolutional networks for automated serous retinal detachment segmentation in SD-OCT images *Comput. Methods Prog. Biomed.* **176** 69–80

Gopinath K and Sivaswamy J 2018 Segmentation of retinal cysts from optical coherence tomography volumes via selective enhancement *IEEE J. Biomed. Health Inform.* **23** 273–82

Gu Z, Cheng J, Fu H, Zhou K, Hao H, Zhao Y, Zhang T, Gao S and Liu J 2019 Ce-net: Context encoder network for 2d medical image segmentation *IEEE Trans. Med. Imaging* **38** 2281–92

He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* pp 770–8

Huang D *et al* 1991 Optical coherence tomography *Science* **247** 1178–81

Karthiga R, Usha G, Raju N and Narasimhan K 2021 Transfer learning based breast cancer classification using one-hot encoding technique *Proc. of 2021 IEEE Int. Conf. on Artificial Intelligence and Smart Systems* pp 115–20

Lee C S, Tyring A J, Deruyter N P, Wu Y, Rokem A and Lee A Y 2017 Deep-learning based, automated segmentation of macular edema in optical coherence tomography *Biomed. Opt. Express* **8** 3440–8

Long M, Cao Y and Wang J 2015 *Michael Jordan Proc. of the 32nd Int. Conf. on Machine Learning, PMLR* pp 97–105

Lu D, Heisler M, Lee S, Ding G W, Navajas E, Sarunic M V and Beg M F 2019 Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network *Med. Image Anal.* **54** 100–10

Marmor M F 1999 Mechanisms of fluid accumulation in retinal edema *Doc. Ophthalmol.* **2000** 239–49

Montuoro A, Waldstein S M, Gerendas B S, Schmidt-Erfurth U and Bogunovi H 2017 Joint retinal layer and fluid segmentation in OCT scans of eyes with severe macular edema using unsupervised representation and auto-context *Biomed. Opt. Express* **8** 1874–88

Oktay O, Schlemper J, Folgoc L, Lee M, Heinrich M, Misawa K and Rueckert D 2018 Attention u-net: learning where to look for the pancreas arXiv:1804.03999

Rashno A, Koozekanani D D, Drayna P M, Sadri B, Rabbani H S and Parhi K K 2017 Fully automated segmentation of fluid/cyst regions in optical coherence tomography images with diabetic macular edema using neutrosophic sets and graph algorithms *IEEE Trans. Biomed. Eng.* **65** 989–1001

Rashno A, Koozekanani D D and Parhi K K 2018 OCT fluid segmentation using graph shortest path and convolutional neural network *Int. Conf. of the IEEE Engineering in Medicine and Biology Society* pp 3426–9

Rashno E, Rashno A and Fadaei S 2019 Fluid segmentation in neutrosophic domain *5th Iranian Conf. on Signal Processing and Intelligent Systems (ICSPIS)* pp 1–5

Ristau T, Keane P A, Walsh A C, Engin A, Mokwa N, Kirchhof B, Sadda S R and Liakopoulos S 2013 Relationship between Visual Acuity and Spectral Domain Optical Coherence Tomography Retinal Parameters in Neovascular Age-Related Macular Degeneration *Ophthalmologica* **231** 37–44

Ronneberger O, Fischer P and Brox T 2015 U-net: Convolutional networks for biomedical image segmentation *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer) pp 234–41

Roy A G, Conjeti S, Karri S P K, Sheet D, Katouzian A, Wachinger C and Navab N 2017 ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks *Biomed. Opt. Express* **8** 3627–42

Sahu M, Mukhopadhyay A and Zachow S 2021 Simulation-to-Real domain adaptation with teacher-student learning for endoscopic instrument segmentation *Int. J. Comput. Assist. Radiol. Surg.* **16** 849–59

San B, Qin S and Ahmed R 2020 SEADNet: Deep learning driven segmentation and extraction of macular fluids in 3D retinal OCT scans *Int. Symp. on Signal Processing and Information Technology (ISSPIT)* (IEEE) 51521.1

Schlegl T, Waldstein S M, Bogunovic H, Endstraßer F, Sadeghipour A, Philip A M, Podkowinski D, Gerendas B S, Langs G and Schmidt-Erfurth U 2018 Fully automated detection and quantification of macular fluid in OCT using deep learning *Ophthalmology* **125** 549–58

Shelhamer E, Long J and Darrell T 2015 Fully convolutional networks for semantic segmentation *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 3431–40

Venhuizen F G, Ginneken B V, Liefers B, Asten F V, Schreur V, Fauser S, Hoyng C, Theelen T and Sánchez C I 2018 Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography *Biomed. Opt. Express* **9** 1545–69

Waldstein S M, Philip A, Leitner R, Simader C, Langs G, Gerendas B S and Schmidt-Erfurth U 2016 Correlation of 3-dimensionally quantified intraretinal and subretinal fluid with visual acuity in neovascular age-related macular degeneration *JAMA Ophthalmol.* **134** 182–90

Wilkins G R, Houghton O M and Oldenburg A L 2012 Automated segmentation of intraretinal cystoid fluid in optical coherence tomography *IEEE Trans. Biomed. Eng.* **59** 1109–14

Wong K C L, Moradi M, Tang H and Syeda-Mahmood T 2018 3D segmentation with exponential logarithmic loss for highly unbalanced object sizes *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer) 612–9

Wu J, Niu S, Chen Q, Fan W and Li D 2020 Automated segmentation of intraretinal cystoid macular edema based on Gaussian mixture model *J. Innovative Opt. Health Sci.* **13** 1950020

Wu M, Fan W, Chen Q, Du Z and Park H 2017 Three-dimensional continuous max flow optimization-based serous retinal detachment segmentation in SD-OCT for central serous chorioretinopathy *Biomed. Opt. Express* **8** 4257–74

Xu X, Lee K, Zhang L, Sonka M and Abràmoff M D 2015 Stratified sampling voxel classification for segmentation of intraretinal and subretinal fluid in longitudinal clinical OCT data *IEEE Trans. Med. Imaging* **34** 1616–23

Yang J, Ji Z, Niu S, Cheng Q, Yuan S and Fan W 2020 RMPPNet: Residual Multiple Pyramid Pooling Network for Subretinal Fluid Segmentation in SD-OCT Images *OSA Continuum* **3** 1751–69

Ye Y, Chen X, Shi F, Xiang D, Pan L and Zhu W 2021 Context attention-and-fusion network for multiclass retinal fluid segmentation in OCT image *SPIE Medical Imaging 2021: Image Processing* 1159622

Yu C, Wang J, Gao C, Yu G, Shen C and Sang N 2020 Context prior for scene segmentation *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* pp 12416–25

Zhao H, Shi J, Qi X, Wang X and Jia J 2017 Pyramid scene parsing network *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* pp 2881–90