IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 June 2022, accepted 5 August 2022, date of publication 16 August 2022, date of current version 19 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3198657

RESEARCH ARTICLE

Multiscale Attention Gated Network (MAGNet) for Retinal Layer and Macular Cystoid Edema Segmentation

ALEX CAZAÑAS-GORDÓN AND LUÍS A. DA SILVA CRUZ, (Senior Member, IEEE)
Department of Electrical and Computer Engineering, University of Coimbra, 3030-290 Coimbra, Portugal
Instituto de Telecomunicações, University of Coimbra, 3030-290 Coimbra, Portugal

Corresponding author: Alex Cazañas-Gordón (acazanas@deec.uc.pt)

This work was supported in part by the Secretariat of Higher Education, Science, Technology and Innovation, Ecuador; and in part by the Fundação para a Ciência e a Tecnologia [FCT/Ministério da Ciência, Tecnologia e Ensino Superior (MCTES)], Portugal, through National Funds co-funded by the European Union (EU) Funds under Project UIDB/EEA/50008/2020, Project UIDP/50008/2020, and Project LA/P/0109/2020.

ABSTRACT Retinal optical coherence tomography (OCT) imaging is a mainstay in the clinical diagnosis of several sight-threatening diseases. Due to the wide variability in shape and orientation of retinal structures, analyzing and interpreting OCT images are complex tasks that require domain knowledge. Within the analysis process, delineating anatomical landmarks and pathological formations, i.e., segmenting OCT scans, is a labor-intensive task usually carried out by expert graders. Recently, several studies have proposed methods based on fully convolutional neural networks (FCN) to alleviate the burden of manual OCT segmentation. Despite the promising performance of FCN-based methods, the negative impact of the class imbalance problem on the segmentation of small foreground targets such as macular cystoid edemas remains a challenge. This article proposes a novel end-to-end automatic method for segmenting retinal layers and macular cystoid edema in OCT images. The proposed method introduces a novel FCN architecture that leverages spatial and channel-attention gates at multiple scales for fine-grained segmentation and a weighting loss approach to handle class imbalance. Results on a benchmark dataset that includes cases of severe retinal edema show the robustness of the proposed algorithm, which achieved state-of-the-art performance with a mean Dice score of 0.92 ± 0.03.

INDEX TERMS Deep learning, semantic segmentation, attention gates, fully convolutional networks, optical coherence tomography.

I. INTRODUCTION

Due to its non-invasive nature and high resolution, optical coherence tomography (OCT) imaging is a widely used diagnostic tool in the characterization of retinal pathologies such as diabetic macular edema (DME), age-related macular degeneration (AMD), and retinal vein occlusion (RVO). The quantification of diagnostic findings in OCT images is central to detecting sight-threatening conditions and relies on the accurate segmentation of anatomical landmarks and abnormal structures. As a highly complex task, OCT segmentation requires specialized knowledge, which is not always readily available because of labor shortages and increased demand. Moreover, OCT-manual segmentation is labor-intensive and prone to interobserver variability. Following the success of deep learning in computer vision tasks, several authors proposed deep learning-based methods for automatic OCT segmentation with promising results [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12].

This article introduces MAGNet: Multiscale Attention Gated Network, a novel lightweight fully convolutional neural network (FCN) for end-to-end automatic segmentation of retinal layers and macular cystoid edema (MCE) in OCT images. The proposed network builds upon the encoder-decoder architecture and attention-gated framework to exploit pixel-wise spatial information for fine-grained

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

segmentation. Unlike related work, we model spatial and channel attention independently to serve complementary objectives. Channel-attention gates are placed on the encoder to learn which feature maps are the most informative and emphasize them before reaching the decoder. On the other hand, spatial-attention gates at the decoder capture pixel-wise contextual information to focus learning on the most relevant regions within feature maps. To train the segmentation model, we propose a novel loss-weighting approach based on the Euclidean distance transform [13] tailored to handle class imbalance. The proposed method derives pixel-wise loss weightings directly from the data without adding training hyperparameters. Furthermore, to support the Green AI initiative [14], we take advantage of Fire modules [15] to reduce the number of parameters of the proposed FCN.

This research aims to advance the state of the art of medical-imaging processing by developing a deep-learning application for end-to-end automatic segmentation of diagnostic markers in retinal OCT B-scans. The main contributions of this study are summarized as follows:

- We propose two attention gates designed to capture channel and spatial correlations within intermediate feature maps. These gates provide a mechanism for improving the allocation of computational resources toward the most informative features for the segmentation task.
- The proposed attention gates are integrated into a novel fully-convolutional network, which delivers state-of-the-art results with fewer parameters than comparative networks.
- We introduce a novel adaptive-weighting scheme for the loss function to circumvent the negative effect of the class-imbalance problem. The weighting approach is pixel-wise, parameter-free, and easily adaptable to any cost function.

II. RELATED WORK

Owing to their ability to learn complex hierarchical representations directly from data, convolutional neural networks (CNN) are the preferred approach for computer vision tasks such as image classification, object detection, and image segmentation. CNNs have set the benchmark for image segmentation outperforming methods based on graph theory, dynamic programming, and energy minimization, such as graph-cut, shortest path, and active contours [16]. Within the context of medical data, CNNs are increasingly used for a wide variety of classification tasks across several imaging modalities, including MRI [17], [18], CT/X-rays [19], [20], fundus photography [21], [22], and ultra-widefield retinal imaging [23].

The development of CNN architectures for image segmentation led to several innovations in network connectivity to aid gradient flow. The seminal work of Long *et al.* [24] on fully convolutional networks for semantic segmentation introduced the notion of *skip connections*. These connections combine coarse-high-level semantic information from deeper layers with fine-grained information from shallow layers to refine segmentation results. Later developments improved the FCN framework with the encoder-decoder architecture, unpooling operations [25], and atrous convolutions [26]. Encoder-decoder architectures leveraging skip connections, such as U-net [27], are extensively used for medical-image segmentation.

Even though architectural developments have been instrumental in enhancing the representational capacity of FCNs, some shortcomings remain unaddressed. Namely, the interleaved coding of spatial and channel information in feature maps generated at each convolutional layer and the wasteful computation of redundant, low-level features brought across via skip connections. Prior work has proposed to improve the joint encoding of spatial and channel information in CNNs via explicitly embedded learning mechanisms that capture spatial interdependencies between the feature-map channels. Recently, attention and gating mechanisms have been incorporated into the FCN workflow to improve the allocation of computational resources towards the most informative features of the input signal [28], [29]. Attention mechanisms have been demonstrated to enhance network performance across several computer vision tasks, from classification [30] and localization [31], to image captioning [32]. For image segmentation, attention and gating mechanisms leverage global information to highlight relevant features while suppressing less informative ones.

Besides architectural innovations, a considerable body of knowledge places the loss function at the center of numerous approaches to overcome the negative effect of class imbalance on CNN-based image segmentation [33], [34], [35], [36]. According to the optimization objective, most loss functions fall under two categories: distribution-based and region-based [34]. Distribution-based loss functions measure the difference between the probability distribution of the segmentation target and that of the predictions. The cross-entropy function is a distribution-based loss widely used in image segmentation [37]. Region-based loss functions measure the similarity of the predicted segmentation and the target ground truth. The most extensively-used loss in this category is the Dice loss function [38]. Apart from pure distribution and region-based losses, compound loss functions combine terms derived from the cross-entropy and the Dice loss functions. Compound loss functions typically improve the segmentation performance but add extra training parameters, increasing the search space of the hyperparameter optimization.

III. METHOD

A. PROBLEM DEFINITION

Given an OCT B-scan, the segmentation task is approached as a multiclass classification problem where every pixel $\boldsymbol{x}$ in a given input image $\boldsymbol{I}$ is mapped to a class label $\hat{y}$ in the label space $\hat{\boldsymbol{Y}} = \{c_1, c_2, \ldots, c_K\}$. Fig. 1 shows a preview of the proposed MAGNet segmentation results, including seven retinal layers and macular cystoid edema.

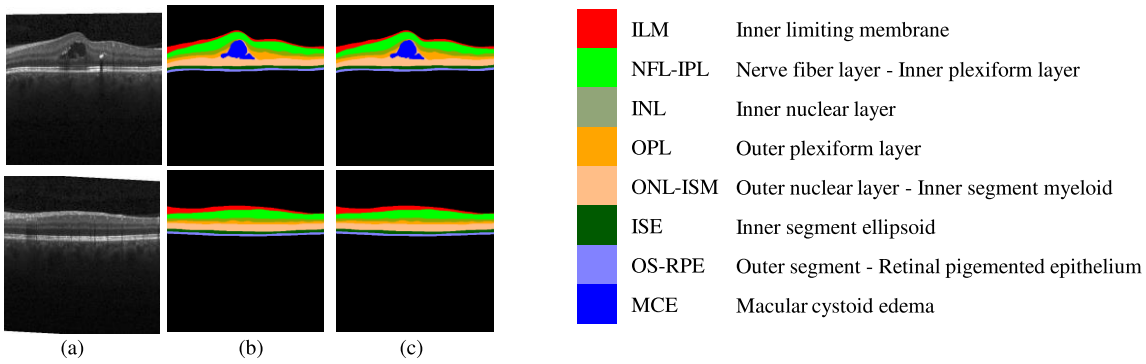

FIGURE 1. Preview of the segmentation results of the proposed MAGNet. Input OCT B-scans (a), Human grader markings (b), proposed-approach segmentation maps (c). Foreground class-label color code is shown to the right. Background classes above and below the retina are shown in black.

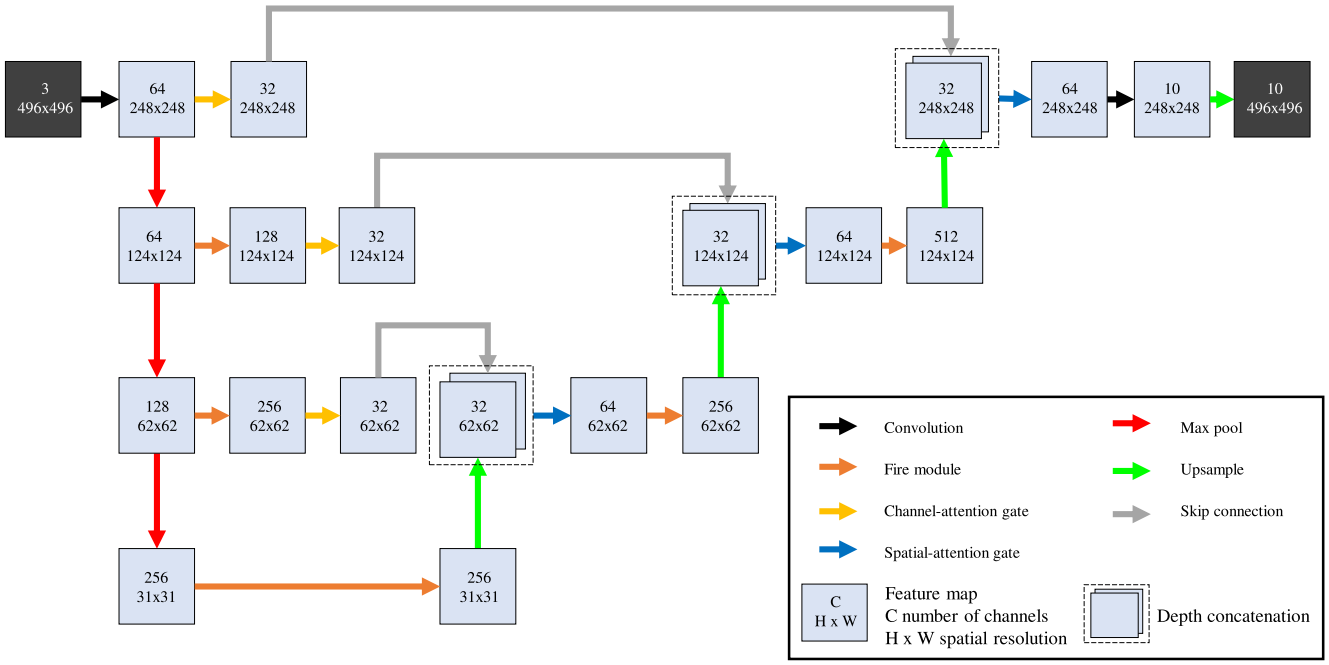

FIGURE 2. Overview of the proposed fully convolutional MAGNet architecture. The spatial resolution of the feature maps are indicated inside corresponding boxes.

B. NETWORK ARCHITECTURE

The proposed FCN is based on the widely-used encoder-decoder architecture, which has proven to be a suitable framework for image segmentation. As shown in Fig. 2, the architecture comprises three major components: Fire modules, channel-attention gates, and spatial-attention gates. The network encoder extracts feature maps through Fire modules applied at multiple scales. On the other hand, the decoder concatenates encoder features with low-level features extracted from the network backbone to refine the segmentation results. Attention gates are strategically placed at every level of the encoding and decoding branches to focus the learning effort on the most relevant features for the segmentation task.

Based on prior work findings, we adopted skip connections to retain global information that otherwise is lost at the encoding stage due to pooling operations. Skip connections ease the flow of gradients through the network and improve the segmentation results by supplementing contextual information at the decoding stage. However, this approach overlooks correlations between feature maps which causes redundant, irrelevant features to be processed repeatedly at multiple scales. To address this shortcoming, we use channel-attention gates to learn which feature maps are worth passing to the decoder. At the decoding stage, spatial-attention gates learn pixel-wise attention maps to weight the features within a given feature map. As a result, features are emphasized based on their relevance to the segmentation or de-emphasized

otherwise. The detail of the main components of the proposed architecture is described below.

1) FIRE MODULE

Drawing inspiration from the efficiency-driven network design of SqueezeNet [15], we adopted the Fire module as the building block of the proposed architecture's backbone. Using Fire modules as its basic unit, the SqueezeNet achieves state-of-the-art classification performance with 50x fewer parameters than the AlexNet [39] architecture. The Fire module processes a given feature map in three steps, as shown in Fig. 3. First, the feature map is compressed along the channel dimension using one convolution layer with a 1×1 filter. The number of channels after the compressing operation is $C/4$, where C denotes the number of input channels. Then, two convolutions layers expand the compressed feature map producing two feature maps with C channels each. Lastly, the output feature map is obtained by concatenating the outputs of the expansive operations along the channel dimension. In general, the Fire module establishes a mapping $\mathbf{F}(.) : \mathbf{X} \rightarrow \mathbf{U}$, where $\mathbf{X} \in \mathbb{R}^{H \times W \times C'}$ is the input feature map, $\mathbf{U} \in \mathbb{R}^{H \times W \times C}$ is the output feature map, H and W are spatial dimensions height and width, C is the number of input channels and C' is the number of output channels.

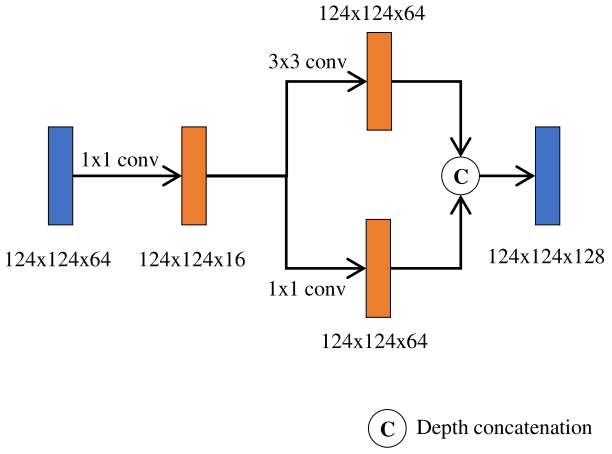

FIGURE 3. Architecture of the Fire module. The dimension of the feature maps are shown next to corresponding boxes.

2) CHANNEL-ATTENTION GATE

Channel attention gates are applied to the output of Fire modules in the encoder to increase the model sensitivity to informative features and, consequently, suppress less useful ones. This attention mechanism applies a series of transformations to a given input to identify which channels within the feature map are the most relevant for the segmentation objectives. As shown in Fig. 4, an encoder feature map $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_C]$, here represented as a collection of channels $\mathbf{u}_i \in \mathbb{R}^{H \times W}$, is first spatially summarized by a global average pooling layer. The output of this operation is an embedding $\mathbf{z} \in \mathbb{R}^{1 \times 1 \times C}$ of the global information in $\mathbf{U}$, where the c-th element in $\mathbf{z}$ is computed by:

$$z_c = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} u_c(i, j) \tag{1}$$

The vector $\mathbf{z}$ is further transformed using a gating mechanism designed to capture channel-wise dependencies. The gating mechanism is enforced through a sigmoid activation that brings the activation range to the interval [0,1]. The output of the channel-attention gate $\mathbf{s}$ is given by:

$$\mathbf{s} = \sigma(\mathbf{W}_2 \psi(\mathbf{W}_1 \mathbf{z})) \tag{2}$$

where $\sigma(.)$ denotes the sigmoid activation function, $\psi(.)$ the ReLU activation function, and $\mathbf{W}_1 \in \mathbb{R}^{C \times \frac{C}{2}}$ and $\mathbf{W}_2 \in \mathbb{R}^{\frac{C}{2} \times C}$ represent the weights of two 1×1 convolutional layers. The vector $\mathbf{s} = [s_1, s_2, \ldots, s_c]$ is then used to obtain a weighted feature map $\hat{\mathbf{U}} = [s_1 \mathbf{u}_1, s_2 \mathbf{u}_2, \ldots, s_C \mathbf{u}_C]$. During training, the network progressively tunes the activations s_i to emphasize or ignore channels according to their importance for the segmentation task. Weighted feature maps are passed through skip connections and concatenated with intermediate feature maps of similar resolution to add contextual information at every stage of the upsampling path.

3) SPATIAL-ATTENTION GATE

Previous work demonstrated that spatial-attention gates are effective in focusing the network onto the target regions automatically without additional supervision [16], [40]. Based on these works, we introduce spatial-attention gates at the decoder to learn an attention map $\hat{\mathbf{A}}$ that defines the pixel-wise relevance of each location (i, j) in a given input feature map $\mathbf{U}$. In contrast with the channel-attention gate, the spatial-attention gate summarizes the feature map $\mathbf{U}$ along the channel dimension and scales spatially, according to the importance of the spatial location. As illustrated in Fig. 5, the input feature map $\mathbf{U} = [\mathbf{u}^{1,1}, \mathbf{u}^{1,2}, \ldots \mathbf{u}^{i,j}, \ldots, \mathbf{u}^{H,W}]$, with $\mathbf{u}^{i,j} \in \mathbb{R}^{1 \times 1 \times C}$ representing the spatial location (i, j), $i \in [1, 2, \ldots, H]$, and $j \in [1, 2, \ldots, W]$ is compressed along the channel dimension twice, first with a 1×1 convolutional that reduces its channel dimension by half and then, a second time to obtain a projection tensor $\mathbf{q} \in \mathbb{R}^{H \times W}$, that contains linear combinations $q_{i,j}$ representing the channel-wise expression at each location(i, j). The projection $\mathbf{q}$ is passed through a sigmoid operation $\sigma(.)$ to obtain the attention map which is used to scale the input feature map. Lastly, a residual connection was added as a safeguard to prevent vanishing gradients. The output of the spatial-attention gate $\hat{\mathbf{U}} = [(\sigma(q_{1,1}) + 1)\mathbf{u}^{1,1}, (\sigma(q_{1,2}) + 1)\mathbf{u}^{1,2}, \ldots, (\sigma(q_{i,j}) + 1)\mathbf{u}^{i,j}, \ldots, (\sigma(q_{H,W}) + 1)\mathbf{u}^{H,W}]$ is a feature map that either emphasizes or de-emphasizes what is relevant or irrelevant for fine-grained segmentation.

4) MODEL COMPLEXITY

As shown in Fig. 2, the proposed MAGNet follows an encoder-decoder design comprising three encoder blocks,

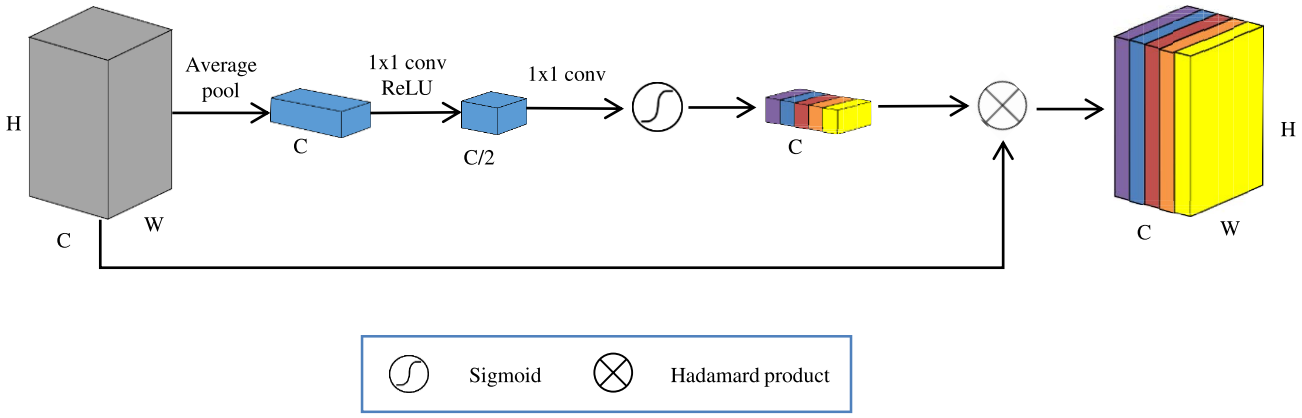

FIGURE 4. Architecture of the channel-attention gate. H and W denote spatial dimensions height and width. C denotes the channel dimension.

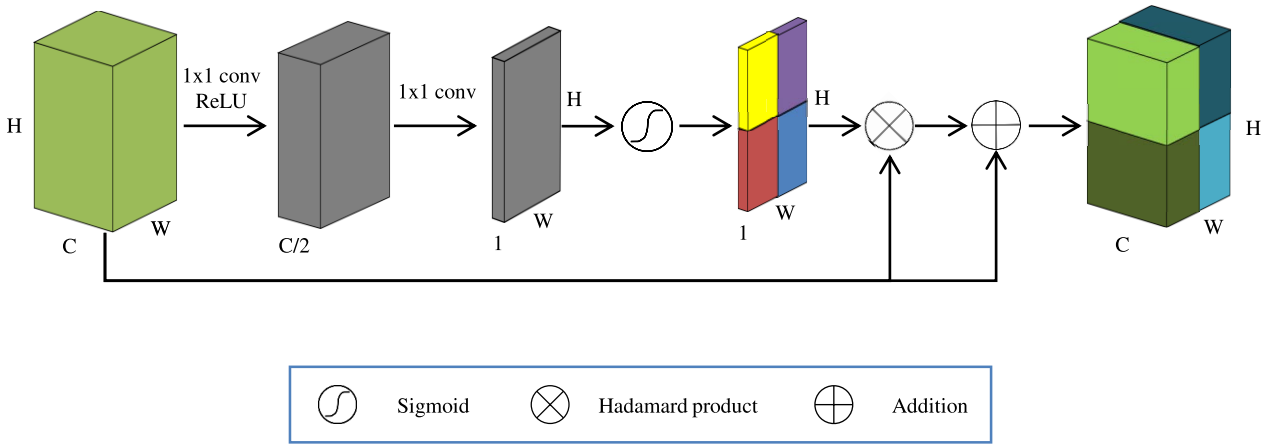

FIGURE 5. Architecture of the spatial-attention gate. H and W denote spatial dimensions height and width. C denotes the channel dimension.

one bottleneck, and three decoder blocks. In contrast to related work, our design leverages attention gates and Fire modules to obtain a deep model while keeping the number of parameters low. As shown in Table 1, the proposed architecture has significantly fewer parameters than state-of-the-art FCNs for image segmentation.

C. LOSS FUNCTION

Because of the dominance of background classes in diagnostic images, highly imbalanced datasets are commonplace in the medical field. Tuning FCNs with such imbalanced datasets can negatively affect the segmentation performance since the learning process focuses on the classes that weigh the most to the loss. Weighting the loss function counters the unwanted effects of the class imbalance problem by balancing the contribution of individual classes to the loss value. Ordinarily, weightings emphasize the importance of underrepresented classes over others in the loss during training. This study introduces a novel pixel-wise weighting approach based on the Euclidean distance transform to handle class imbalance. Our loss weighting method uses distance maps computed from the boundary of the segmentation targets to assign pixel-wise weightings. Fig. 6(c) shows an example of distance maps computed from the ground truth of an MCE-class sample. To formally define the distance-transform weightings, let be G_j the set of pixels representing the ground truth of class j. The distance transform of a pixel i in G_j is determined as follows:

$$DT_{ij} = \min_{i' \in G_j} (\|i - i'\|_2 + \mathbb{1}_G^j(i)) \tag{3}$$

TABLE 1. Model complexity comparison.

Network	Size in disk (MB)	Parameters (Millions)	Layers
FCN-8	537	134.3	42
DeepLabV3+	176.86	43.98	205
SqueezeSeg	39.8	9.89	88
Unet	30.89	7.69	45
PointSegNet	15.6	3.8	198
Proposed	2.76	1.45	126

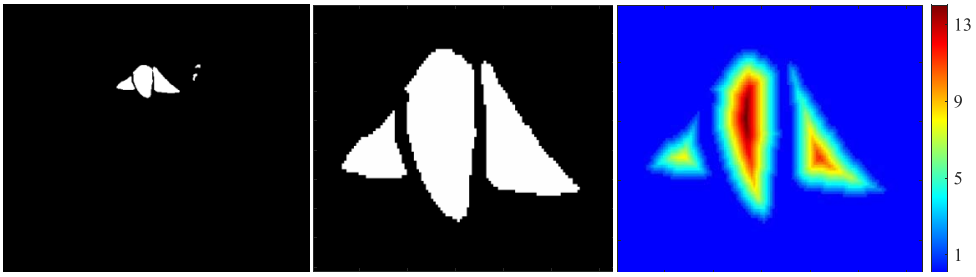

FIGURE 6. Illustration of the pixel-wise weighting scheme used to balance the loss function. The loss weightings are obtained based on the distance transform of segmentation targets. A sample's ground truth of the class MCE is shown in (a), a close-up of the ground truth and corresponding distance map are shown in (b) and (c), respectively.

where $\|.\|_2$ represents the Euclidean distance, and $\mathbb{1}_G^j(i)$ is an indicator function defined as:

$$\mathbb{1}_G^j(i) = \begin{cases} 0 & \text{if } i \in G_j \\ \infty & \text{otherwise} \end{cases} \tag{4}$$

Distance-based weightings highlight boundary errors in proportion to the distance to the ground-truth. The farther a misclassified pixel is from the target boundary, the higher its assigned weight and contribution to the loss. The distance-transform weighting for pixel i of class j is given by:

$$\mathcal{W}_{ij}^{DMT} = 1 + DT_{ij} \tag{5}$$

The formulation of the proposed weighting scheme is task agnostic, which makes weightings applicable to any base cost function. This study uses the cross-entropy loss as the base cost function. The definition of the cross-entropy loss is as follows:

$$CE = \frac{1}{N} \sum_{j=1}^{K} \sum_{i=1}^{N} y_{ij} \log \hat{y}_{ij} \tag{6}$$

where $y_{ij} \in \{0, 1\}$ is the one-hot encoding of the ith pixel for class label j, $\hat{y}_{ij}$ is the predicted class probability of the ith pixel for class label j, K is the number of classes, and N is the number of pixels. The distance-map-transform weighted cross-entropy is given by:

$$DWCE = \frac{1}{N} \sum_{j=1}^{K} \sum_{i=1}^{N} \mathcal{W}_{ij}^{DMT} y_{ij} \log \hat{y}_{ij} \tag{7}$$

IV. EXPERIMENTAL DESIGN

A. DATASET

We used 110 annotated B-scans sourced from the publicly available Duke SD-OCT dataset [1]. The dataset consists of 10 volumes acquired from a cohort of ten DME patients using a Spectralis HRA+OCT scanner (Heidelberg Engineering, Heidelberg, Germany). All OCT volumes are centered at the fovea and have 61 B-scans each. The B-scans are 768 × 496 pixels in size and have an axial resolution of 3.87 μm. Each volume includes annotations from two retina specialists on eleven non-consecutive scans where expert annotations delineate eight retinal boundaries and fluid masses. Henceforth, we refer to Duke's retina-specialist annotations as Expert1 and Expert2.

In addition to the Duke dataset, we used the publicly available HCMS dataset [41] to validate the proposed segmenting algorithm. The data comprise scans of the right eye of 35 subjects, including healthy controls (14) and multiple sclerosis patients (21). The dataset includes 35 OCT volumes containing 49 B-scans each. The acquisition device was a Spectralis OCT system (Heidelberg Engineering, Heidelberg, Germany). All B-scans are 1024 × 496 pixels in size, with a lateral resolution of 5.8 μm and axial resolution of 3.9 μm. Besides the OCT images, the dataset provides manual delineations for nine layer boundaries in every B-scan.

We defined ground-truth class labels for each dataset based on the corresponding expert annotations. The class-label definition for the Duke dataset included eight foreground classes corresponding to seven retinal layers and fluid (CME) and two background classes for the regions above and below the retina. Similarly, we defined ten class labels for the HCMS data, including eight retinal layers and the same two background classes defined for the Duke dataset.

As a pre-processing step, we center-cropped all B-scans along their longest axis to make them square with a size of 496 pixels. Then, to ease gradient computations, we rescaled the pixel-intensity values to the range [0 1]. Finally, we split the data into disjointed partitions for training and testing purposes. Considering the amount of the available ground-truth labels, we separated the Duke data into two sets, 8 OCT volumes for training and 2 OCT volumes for testing. Similarly, we split the HCMS dataset into two partitions having 18 OCT volumes (training set) and 17 OCT volumes (test set).

B. NETWORK TRAINING PROTOCOL

We trained the proposed FCN-segmentation model to optimize the weighted cross-entropy loss (DWCE) defined in (7). The network parameters were updated with the stochastic

gradient descent algorithm [42] using a variable learning rate and minibatch size of two. Preliminary experiments showed an inverse correlation between the segmentation performance and the minibatch size hyperparameter. This observation coincides with previous research that verified a higher performance attained from training regimes with small minibatch sizes rather than large ones [43], [44], [45]. Optimal values for the initial-learning rate, decay drop factor, and decay-drop period hyperparameters were determined by grid search. Table 2 lists the training hyperparameters and the intervals of the search space. Two training rounds were conducted with the Duke dataset, each using one of the two expert annotations as reference. The development and testing environment was MATLAB® release 2021b and CUDA® library version 9.0. Model training and evaluation was conducted on a Windows 10 PC (CPU: Intel i7 8700K CPU @ 3.7 GHz - 6 cores, RAM: 32 GB) with a GPU NVIDIA® GeForce GTX® 1080 Ti with 11 GB RAM.

TABLE 2. Network training hyperparameters.

Hyperparameter	Value
Optimizer	Stochastic gradient descent
Momentum	0.9
Max. Epochs	[10 , 100]
Minibatch size	{2, 4, 8, 16}
Initial learning rate	$[10^{-3}, 10^{-1}]$
Learning rate decay	$[10^{-2}, 10^{-1}]$

C. PERFORMANCE EVALUATION

To evaluate the segmentation performance, we used the Dice-similarity score. This metric measures the overlap between the regions of the model prediction and the ground truth. The definition of the Dice similarity score for class j is as follows:

$$DSC_j(P_j, G_j) = \frac{2|P_j \cap G_j|}{|P_j| + |G_j|} \tag{8}$$

where $|.|$ represents set cardinality, P_j is the predicted segmentation of class j and G_j the corresponding ground truth.

D. COMPARISON WITH STATE OF THE ART

The proposed method was contrasted with state-of-the-art algorithms for layer and fluid segmentation in OCT images. Specifically, FCN-based methods including FCN-8 [24], FCN with conditional random fields [46] (FCN-CRF), U-Net [27], and U-Net-based networks: RelayNet [47], U-Net with shape-based regression [48] (U-Net-SR), and DMP Net [49]. Besides FCN architectures, the comparison included methods based on graph theory and dynamic programming (GDP), namely, Kernel regression with GDP [1] (GDP-KR), Neutrosophic-set and GDP [3] (GDP-NS). In line with the most common setup in related work, we used a MAGNet trained and evaluated with Expert1 annotations for the performance comparison on the Duke dataset. In addition, we evaluated MAGNet with Expert2 annotations and compared it to competing methods using said setup. Similarly, the proposed MAGNet was evaluated on the HCMS dataset and contrasted with published works using said data.

E. ABLATION STUDY

To evaluate the impact of the proposed contributions, we present an ablation study with plausible variations of the proposed algorithm. Besides the proposed approach, we conducted five additional experiments to evaluate the importance of the attention mechanisms and the loss function on the overall results. Regarding the influence of the attention gates, three network variants were considered for evaluation. The variant N1 corresponds to the proposed architecture with the channel-attention gates removed. Conversely, the variant N2 keeps the channel-attention gates and removes the spatial ones from the complete model. On the other hand, the variant N3 is obtained by removing all attention gates from the proposed architecture. As for the loss function, we devised two experiments (N4 and N5) to contrast the performance of models trained with the proposed DWCE loss against that of corresponding models trained with the cross-entropy. In both experiments the models were optimized with the cross-entropy loss function instead of the DWCE loss. Table 3 summarizes the configuration of the experiments in the ablation study.

TABLE 3. Configuration of the salient components evaluated in the ablation study.

Networks	Channel gate	Spatial gate	Loss function
Proposed	×	×	DWCE
N1		×	DWCE
N2	×		DWCE
N3			DWCE
N4			Cross-entropy
N5	×	×	Cross-entropy

V. RESULTS AND DISCUSSION

Table 4 presents the results of the proposed and state-of-the-art methods on the Duke dataset. The results correspond to two experimental setups: a) using Expert1 annotations as ground truth for training and testing, and b) using Expert1 annotations as a reference for training and Expert2 for testing. We report the Dice score for seven retinal layers and the cystoid edema class. In addition, we report the overlap between human-grader annotations (denoted as Expert1 and Expert2).

Looking at the overlap between grader markings, it is noticeable that the Dice score is low in most classes and particularly low in the MCE class. This observation illustrates the high degree of difficulty that the segmentation task entails. The proposed approach obtained the highest segmentation performance in layers ILM, NFL-IPL, INL, OPL, ONL-ISM, and ISE, with Dice scores values over 0.9. Moreover, the proposed method achieved the highest Dice score in the class MCE. Our deep-learning approach improved the performance of graph-based algorithms, with substantial improvements in layers NFL, OPL, and the MCE class. These

TABLE 4. Dice score of the proposed and comparative methods on the Duke dataset evaluated with: (a) Expert1 annotations as ground truth and (b) Expert2 annotations as ground truth. Best results in group columns are shown in bold.

Method	ILM	NFL-IPL	INL	OPL	ONL-ISM	ISE	OS-RPE	MCE
a) Ground truth: Expert1								
Expert2	0.86	0.90	0.79	0.74	0.94	0.86	0.82	0.58
Proposed	**0.92**	**0.96**	**0.92**	**0.90**	0.94	**0.95**	0.88	**0.88**
DMP Net	0.91	0.95	0.88	0.86	**0.95**	0.92	0.88	0.81
RelayNet	0.90	0.94	0.87	0.84	0.93	0.92	**0.90**	0.77
U-Net	0.86	0.91	0.83	0.81	0.91	0.9	0.83	0.67
GDP-KR	0.85	0.89	0.75	0.74	0.93	0.87	0.82	0.53
U-Net-SR	0.81	0.85	0.72	0.71	0.87	0.84	0.83	0.46
FCN-8	0.81	0.84	0.72	0.71	0.88	0.89	0.86	0.28
GDP-NS	-	-	-	-	-	-	-	0.62
FCN-CRF	-	-	-	-	-	-	-	0.61
b) Ground truth: Expert2								
Expert1	0.86	0.90	0.79	0.74	0.94	0.86	0.82	0.58
Proposed	**0.87**	**0.93**	**0.88**	**0.86**	**0.90**	**0.91**	**0.83**	**0.83**
U-Net-SR	0.79	0.84	0.72	0.66	0.85	0.82	0.81	0.4
GDP-NS	-	-	-	-	-	-	-	0.53

TABLE 5. Dice score of the proposed MAGNet models trained using annotations from Expert1 (MAGNet1) and Expert 2 (MAGNet2). Highest scores in column are shown in bold.

Method	Ground truth	ILM	NFL-IPL	INL	OPL	ONL-ISM	ISE	OS-RPE	MCE
MAGNet1	Expert1	**0.92**	**0.96**	**0.92**	**0.90**	**0.94**	**0.95**	**0.88**	0.88
MAGNet1	Expert2	0.87	0.93	0.88	0.86	0.90	0.91	0.83	0.83
MAGNet2	Expert1	0.89	0.95	0.89	0.87	0.91	0.91	0.86	0.84
MAGNet2	Expert2	0.91	**0.96**	0.92	**0.90**	0.93	0.94	0.87	**0.89**
Expert2	Expert1	0.86	0.90	0.79	0.74	**0.94**	0.86	0.82	0.58

TABLE 6. Dice score of the proposed and comparative methods on the HCMS dataset. Best results in column are shown in bold.

Method	ILM	NFL-IPL	INL	OPL	ONL	IS	OS	RPE	Mean
Proposed	**0.89**	**0.95**	**0.87**	**0.90**	0.94	**0.92**	**0.89**	**0.93**	**0.91**
DMP Net	0.88	0.87	0.76	0.84	**0.95**	0.85	0.87	**0.93**	0.87
RelayNet	0.85	0.82	0.74	0.85	0.93	0.83	0.87	**0.93**	0.85
U-Net	0.87	0.85	0.72	0.81	0.94	0.85	0.86	0.92	0.85

classes are challenging segmentation targets, as they have the lowest overlap scores between human-grader annotations. Contrasting the proposed network with state-of-the-art FCNs, we remark that our network achieved competitive performance with a much smaller architecture (see Table 1). Furthermore, we highlight the improvement in the segmentation performance of imbalanced classes INL, OPL, ISE, OS-RPE, and MCE. We stress that this improvement is a direct result of the proposed pixel-wise weighting approach, which, unlike comparative methods, does not add training hyperparameters.

Regarding the results using Expert2 annotations as ground truth, the performance of the proposed and competing methods drops relative to the experiment using Expert1 markings. This decrease in performance is not unexpected, considering the high disagreement between expert annotations. However, we remark that the proposed method outperformed comparative methods and achieved a higher Dice score than Expert1. We also report results for all possible experimental configurations with MAGNet and each set of expert annotations for completeness (Table 5). Consistent with prior results, training and testing on the same set of annotations delivered higher performance than other experimental configurations. Furthermore, the proposed method improved the human-grader overlap in all experimental settings with substantial improvements in classes with a low inter-observer agreement.

Furthermore, we evaluated the proposed MAGNet on 833 unseen images from the HCMS dataset. Table 6 shows the results of the proposed and comparative methods on this set of images, where the proposed MAGNet attained the best overall performance with a mean Dice similarity score of 0.91%. Moreover, MAGNet achieved the highest segmentation performance in seven of eight layers with substantial improvements relative to competing methods in the NFL-IPL, INL, and IS layers. These results confirm the robustness of the proposed method and its performance consistency.

A. QUALITATIVE EVALUATION

Fig. 7 shows a qualitative comparison between top performing FCNs and the proposed approach on the Duke dataset. The algorithm predictions correspond to

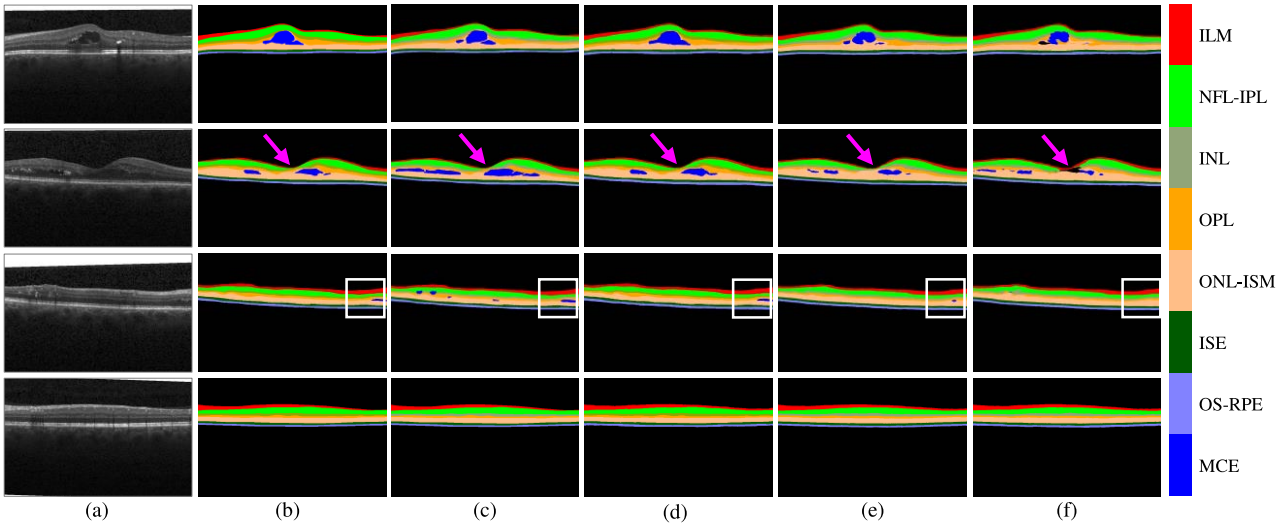

FIGURE 7. Segmentation results of the proposed MAGNet and comparative methods on the Duke dataset. Models were trained and evaluated on the Duke dataset using Expert1 annotations as reference. (a) OCT B-scan input, (b) Expert1 annotations (ground truth), (c) Expert2 annotations, (d) MAGNet, (e) DMPNet, (f) RelayNet. Magenta arrows in the second row indicate the location of the fovea. White boxes in the third row indicate a small fluid pocket.

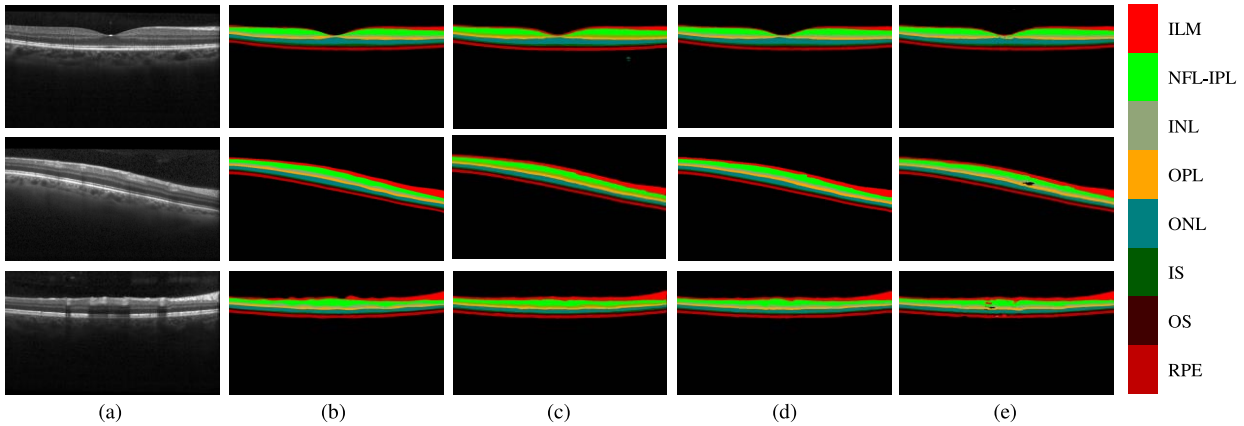

FIGURE 8. Segmentation results of the proposed MAGNet and comparative U-Net-based networks on the HCMS dataset. (a) OCT inputs, (b) Ground truth, (c) MAGNet, (d) DMPNet, (e) RelayNet.

a representative sample of the benchmark dataset that includes one case of severe cystoid edema disrupting the fovea (top row), two scans showing diffused fluid accumulations (second and third rows), and one scan without edema distal from the fovea (bottom row). We remark quality differences in the pathological scans regarding the segmentation of fluid accumulation and the retinal layers at the fovea. Fluid accumulation is a hard-to-segment target as it can be seen in the marked disagreement between human-manual annotations (Fig. 7(b) and Fig. 7(c)). State-of-the-art methods under-segment fluid masses in Fig. 7(e) and Fig. 7(f). Moreover, a small fluid pocket at the right of the scan in Fig. 7(f) (indicated by a white box) is absent in the prediction of one of the comparative methods. Another challenging target is the foveal region (indicated by a magenta arrow in 7(b)), where retinal layers are ordinarily at their thinnest. As it can be seen, the proposed-method segmentation is of high quality and comparable to human grader markings. Similarly, Fig. 8 presents a representative sample of segmentation results obtained in the HCMS dataset. Consistent with prior evaluation, the proposed method show quality segmentation results, on par with human grader performance. By contrast, competing FCNs results are affected by blood-vessel shadows that occur in B-scans distal from the fovea (middle and bottom rows) where segmentation errors brake the layer delineation producing a jagged effect (Fig. 8(d) and Fig. 8(e)).

Overall, the proposed MAGNet attained high segmentation performance in both test sets. Nevertheless, few B-scans in the Duke dataset showed minor segmentation errors in

TABLE 7. Dice score of the proposed MAGNet variants evaluated in the ablation study. Models were trained and evaluated on the Duke dataset taking: (a) Expert1 annotations as reference and (b) Expert2 annotations as reference. Highest scores in group columns are shown in bold.

Networks	ILM	NFL-IPL	INL	OPL	ONL-ISM	ISE	OS-RPE	MCE	Mean
a) Ground truth: Expert1									
Proposed	**0.92**	**0.96**	**0.92**	**0.90**	**0.94**	**0.95**	**0.88**	**0.88**	**0.92**
N1	0.91	**0.96**	0.91	0.89	0.93	0.93	0.86	0.86	0.90
N2	0.89	0.95	0.89	0.87	0.92	0.93	0.86	0.86	0.90
N3	0.89	0.94	0.88	0.85	0.90	0.91	0.84	0.83	0.88
N4	0.84	0.92	0.84	0.80	0.90	0.91	0.80	0.78	0.85
N5	0.88	0.94	0.86	0.84	0.90	0.90	0.82	0.82	0.87
b) Ground truth: Expert2									
Proposed	**0.91**	**0.96**	**0.92**	**0.90**	**0.93**	**0.94**	**0.87**	**0.89**	**0.92**
N1	0.89	0.95	0.90	0.86	0.91	0.93	0.85	0.87	0.90
N2	0.89	0.95	0.89	0.86	0.92	0.92	0.82	0.87	0.89
N3	0.88	0.94	0.87	0.86	0.91	0.90	0.81	0.85	0.88
N4	0.86	0.93	0.85	0.84	0.89	0.89	0.81	0.81	0.86
N5	0.87	0.92	0.85	0.84	0.89	0.90	0.83	0.83	0.87

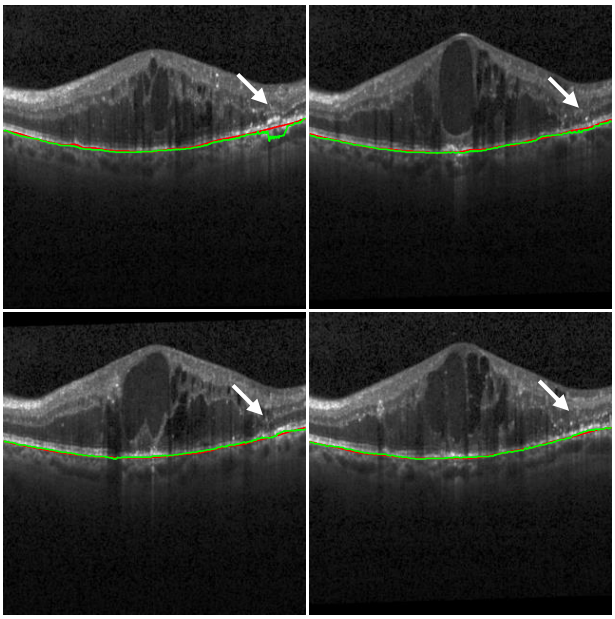

FIGURE 9. Segmentation errors in the ERP boundary of retinas showing hyperreflective foci (white arrows). Automatic and manual delineations are shown in green and red, respectively.

the RPE boundary. Upon visual inspection of said B-scans, we found that errors concentrate around hyperreflective foci (HRF) near the RPE. HRF are small, punctiform hyperreflective lesions that appear brighter than surrounding tissue. When HRF groupings occur close to the RPE, the proposed algorithm label HRF pixels as RPE, which in turn results in the predicted segmentation overstepping the target boundary (see Fig. 9). The low incidence of samples showing HRF near the RPE in the dataset (less than 4%) might explain the occurrence of these segmentation errors. Thus, adding more examples of these abnormal formations to the training data might enhance the discriminative capability of the segmentation model.

B. ABLATION STUDY

We evaluated plausible variations of the proposed method to observe the impact of the main components of the framework on the segmentation performance. Five experimental configurations were evaluated twice on the Duke dataset, each using either Expert1 or Expert2 annotations as reference. Table 7 summarizes the results of the ablation experiments grouped by ground truth.

1) ATTENTION GATES

Comparing the variant without attention gates (N3) with the proposed architecture, we observe a substantial improvement across all classes due to the combined effect of the two attention mechanisms (see Table 7). Notably, the proposed network improved the segmentation performance of challenging classes by 5% on average. We also evaluated the effect of the two types of attention gates separately by removing the channel-attention gates in N2 and the spatial-attention gates in N1. Contrasting N2 and N1 with the variant without any attention mechanism, we observed a slightly better improvement with N1 than with N2. Although this difference might be attributed to the increase in parameters due to the attention blocks, it is worth noting that said increase was less than 2% in both variants. We also note the complementary nature of the two attention mechanisms, demonstrated by the higher improvement margins of the proposed approach relative to those of the variants with only one type of attention. Fig. 10 shows qualitative results of the network variants evaluated in this ablation analysis.

2) LOSS WEIGHTING

We introduced a novel weighted loss to counter the negative effect of the class imbalance on the segmentation performance. To evaluate the impact of the proposed loss function while factoring out the effect of the attention mechanism, we trained the model without attention gates (N3) with the cross-entropy loss. As shown in Table 7, the

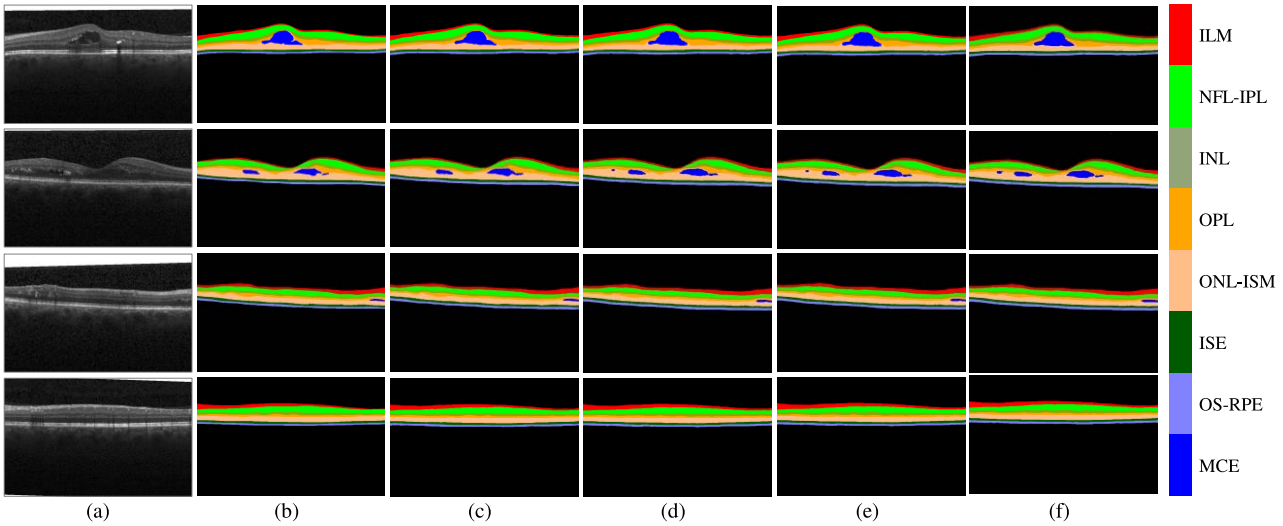

FIGURE 10. Segmentation results of the MAGNet variants evaluated in the ablation study. Models were trained and evaluated on the Duke dataset using Expert1 annotations as reference. (a) Input OCT scans, (b) ground truth, (c) proposed MAGNet, (d) variant N1: spatial-attention gate, (e) variant N2: channel-attention gate, (f) variant N3: sans attention gates.

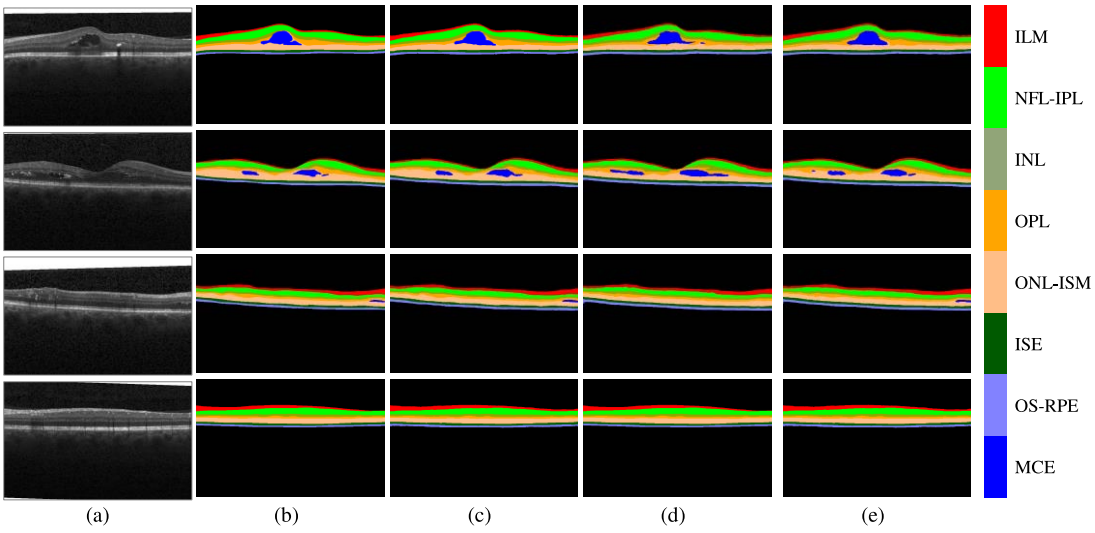

FIGURE 11. Segmentation results of the MAGNet variants evaluated in the ablation study. Models were trained and evaluated on the Duke dataset using Expert1 annotations as reference. (a) Input OCT scans, (b) ground truth, (c) proposed MAGNet trained with the DWCE loss, (d) experiment N4: variant N3 trained with the cross-entropy loss, (e) experiment N5: proposed MAGNet trained with the cross-entropy loss.

segmentation performance dropped 3% on average across all classes. In addition, we trained the proposed full model with the cross-entropy loss (experiment N5) and contrasted its performance against that of the same model trained with the DWCE loss. As shown in Table 7, the model trained with the proposed weighted loss improves the mean Dice score of the model trained with the cross-entropy loss by 5%. These observations confirm the effectiveness of the proposed loss weighting approach in handling class imbalance, with the added advantage that our weighting scheme is data-driven and parameter-free. A qualitative comparison of the loss function experiments is presented in Fig. 11.

VI. CONCLUSION

This study proposes a novel end-to-end automatic method for segmenting retinal layers and macular cystoid edema in OCT B-scans. The proposed method addresses the need for automating retinal OCT image analysis, which is a labor-intensive task, prone to human error and inter-observer variability. Results of the evaluation on two publicly available benchmark datasets showed that the proposed method reached competitive performance on par with state-of-the-art FCNs. The proposed method was evaluated on a benchmark dataset of OCT B-scans from DME patients where it achieved a mean Dice score of 0.92 ± 0.03 and improved the

state-of-the-art performance of the CME class by 7%. In addition, the proposed algorithm attained a mean Dice score of 0.91 ± 0.03 in a OCT dataset acquired from healthy controls and multiple sclerosis patients. Furthermore, the proposed approach introduces an effective, parameter-free loss weighting scheme to handle the class imbalance problem. In future work, we hope to extend the proposed method to other retinal pathology diagnosed with OCT, such as macular holes and drusen, as well as to other imaging modalities. In addition, the interplay between contradicting expert annotations warrants further research conducing to leverage such a differential knowledge concurrently in a single deep learning pipeline.

REFERENCES

[1] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," *Biomed. Opt. Exp.*, vol. 6, no. 4, pp. 1172–1194, 2015, doi: 10.1364/BOE.6.001172.

[2] S. P. K. Karri, D. Chakraborthi, and J. Chatterjee, "Learning layer-specific edges for segmenting retinal layers with large deformations," *Biomed. Opt. Exp.*, vol. 7, no. 7, pp. 2888–2901, 2016.

[3] A. Rashno, D. D. Koozekanani, P. M. Drayna, B. Nazari, S. Sadri, H. Rabbani, and K. K. Parhi, "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.*, vol. 65, no. 5, pp. 989–1001, May 2018.

[4] J. Oliveira, S. Pereira, L. Gonçalves, M. Ferreira, and C. A. Silva, "Multi-surface segmentation of OCT images with AMD using sparse high order potentials," *Biomed. Opt. Exp.*, vol. 8, no. 1, pp. 281–297, 2017.

[5] E. Parra-Mora, A. Cazanas-Gordon, R. Proenca, and L. A. da Silva Cruz, "Epiretinal membrane detection in optical coherence tomography retinal images using deep learning," *IEEE Access*, vol. 9, pp. 99201–99219, 2021, doi: 10.1109/ACCESS.2021.3095655.

[6] D. Alonso-Caneiro, J. Kugelman, J. Hamwood, S. A. Read, S. J. Vincent, F. K. Chen, and M. J. Collins, "Automatic retinal and choroidal boundary segmentation in OCT images using patch-based supervised machine learning methods," in *Proc. Comput. Vis. ACCV Workshops*, G. Carneiro and S. You, Eds. Cham, Switzerland: Springer, 2019, pp. 215–228, doi: 10.1007/978-3-030-21074-8_17.

[7] K. Hu, B. Shen, Y. Zhang, C. Cao, F. Xiao, and X. Gao, "Automatic segmentation of retinal layer boundaries in OCT images using multi-scale convolutional neural network and graph search," *Neurocomputing*, vol. 365, pp. 302–313, Nov. 2019, doi: 10.1016/j.neucom.2019.07.079.

[8] M. Pekala, N. Joshi, T. Y. A. Liu, N. M. Bressler, D. C. DeBuc, and P. Burlina, "Deep learning based retinal OCT segmentation," *Comput. Biol. Med.*, vol. 114, Nov. 2019, Art. no. 103445, doi: 10.1016/j.compbiomed.2019.103445.

[9] P. Zang, J. Wang, T. T. Hormel, L. Liu, D. Huang, and Y. Jia, "Automated segmentation of peripapillary retinal boundaries in OCT combining a convolutional neural network and a multi-weights graph search," *Biomed. Opt. Exp.*, vol. 10, no. 8, pp. 4340–4352, 2019.

[10] J. Kugelman, D. Alonso-Caneiro, Y. Chen, S. Arunachalam, D. Huang, N. Vallis, M. J. Collins, and F. K. Chen, "Retinal boundary segmentation in stargardt disease optical coherence tomography images using automated deep learning," *Transl. Vis. Sci. Technol.*, vol. 9, no. 11, p. 12, Oct. 2020, doi: 10.1167/tvst.9.11.12.

[11] A. Cazanas-Gordon, E. Parra-Mora, and L. A. D. S. Cruz, "Ensemble learning approach to retinal thickness assessment in optical coherence tomography," *IEEE Access*, vol. 9, pp. 67349–67363, 2021, doi: 10.1109/ACCESS.2021.3076427.

[12] B. Hassan, S. Qin, R. Ahmed, T. Hassan, A. H. Taguri, S. Hashmi, and N. Werghi, "Deep learning based joint segmentation and characterization of multi-class retinal fluid lesions on OCT scans for clinical use in anti-VEGF therapy," *Comput. Biol. Med.*, vol. 136, Sep. 2021, Art. no. 104727, doi: 10.1016/j.compbiomed.2021.104727.

[13] C. R. Maurer, R. Qi, and V. Raghavan, "A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 265–270, Feb. 2003.

[14] R. Schwartz, J. Dodge, N. Smith, and O. Etzioni, "Green AI," *Commun. ACM*, vol. 63, no. 12, pp. 54–63, Dec. 2020.

[15] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," 2016, *arXiv:1602.07360*.

[16] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019.

[17] B. Khagi and G. R. Kwon, "Convolutional neural network-based natural image and MRI classification using Gaussian activated parametric (GAP) layer," *IEEE Access*, vol. 9, pp. 96930–96947, 2021.

[18] V. Jain, O. Nankar, D. J. Jerrish, S. Gite, S. Patil, and K. Kotecha, "A novel AI-based system for detection and severity prediction of dementia using MRI," *IEEE Access*, vol. 9, pp. 154324–154346, 2021.

[19] F. Jiao, Z. Gui, K. Li, H. Shangguang, Y. Wang, Y. Liu, and P. Zhang, "A dual-domain CNN-based network for CT reconstruction," *IEEE Access*, vol. 9, pp. 71091–71103, 2021.

[20] T. Iqbal, A. Shaukat, M. U. Akram, A. W. Muzaffar, Z. Mustansar, and Y.-C. Byun, "A hybrid VDV model for automatic diagnosis of pneumothorax using class-imbalanced chest X-rays dataset," *IEEE Access*, vol. 10, pp. 27670–27683, 2022.

[21] Y. Li, Z. Song, S. Kang, S. Jung, and W. Kang, "Semi-supervised auto-encoder graph network for diabetic retinopathy grading," *IEEE Access*, vol. 9, pp. 140759–140767, 2021.

[22] A. Cazanas-Gordon, E. Parra-Mora, and L. A. da Silva Cruz, "3D modeling of the optic nerve head of glaucomatous eyes using fundus stereo images," in *Proc. Telecoms Conf. (ConfTELE)*, Feb. 2021, pp. 1–5.

[23] E. Parra-Mora, A. Cazanas-Gordon, and L. A. da Silva Cruz, "Detection of peripheral retinal breaks in ultra-widefield images using deep learning," in *Proc. Telecoms Conf. (ConfTELE)*, Feb. 2021, pp. 1–5.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[25] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[26] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 833–851, doi: 10.1007/978-3-030-01234-2_49.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[28] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[29] Z. Xu, S. Wang, L. V. Stanislawski, Z. Jiang, N. Jaroenchai, A. M. Sainju, E. Shavers, E. L. Usery, L. Chen, Z. Li, and B. Su, "An attention U-Net model for detection of fine-scale hydrologic streamlines," *Environ. Model. Softw.*, vol. 140, Jun. 2021, Art. no. 104992.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[31] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, D. Ramanan, and T. S. Huang, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2956–2964.

[32] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.

[33] S. Jadon, "A survey of loss functions for semantic segmentation," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2020, pp. 1–7, doi: 10.1109/CIBCB48159.2020.9277638.

[34] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. L. Martel, "Loss Odyssey in medical image segmentation," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102035, doi: 10.1016/j.media.2021.102035.

[35] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, "A comprehensive survey of loss functions in machine learning," *Ann. Data Sci.*, pp. 1–26, 2020, doi: 10.1007/s40745-020-00253-5.
[36] Y. Prajna and M. K. Nath, "A survey of semantic segmentation on biomedical images using deep learning," in *Advances in VLSI, Communication, and Signal Processing*, D. Harvey, H. Kar, S. Verma, and V. Bhadauria, Eds. Singapore: Springer, 2021, pp. 347–357, doi: 10.1007/978-981-15-6840-4_27.
[37] D. Oliva, S. Hinojosa, V. Osuna-Enciso, E. Cuevas, M. Pérez-Cisneros, and G. Sanchez-Ante, "Image segmentation by minimum cross entropy using evolutionary methods," *Soft Comput.*, vol. 23, no. 3, pp. 431–450, 2019, doi: 10.1007/s00500-017-2794-1.
[38] A. Reinke *et al.*, "Common limitations of image processing metrics: A picture story," 2021, *arXiv:2104.05642*.
[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9.
[40] T. L. B. Khanh, D.-P. Dao, N.-H. Ho, H.-J. Yang, E.-T. Baek, G. Lee, S.-H. Kim, and S. B. Yoo, "Enhancing U-Net with spatial-channel attention gate for abnormal tissue segmentation in medical imaging," *Appl. Sci.*, vol. 10, no. 17, p. 5729, Aug. 2020.
[41] Y. He, A. Carass, S. D. Solomon, S. Saidha, P. A. Calabresi, and J. L. Prince, "Retinal layer parcellation of optical coherence tomography images: Data resource for multiple sclerosis and healthy controls," *Data Brief*, vol. 22, pp. 601–604, Feb. 2019.
[42] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006, pp. 206–212, ch. 4.
[43] A. Cazanas-Gordon, E. Parra-Mora, and L. A. da Silva Cruz, "Evaluating transfer learning for macular fluid detection with limited data," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 1348–1352, doi: 10.23919/Eusipco47968.2020.9287859.
[44] I. Kandel and M. Castelli, "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset," *ICT Exp.*, vol. 6, no. 4, pp. 312–315, Jan. 2020, doi: 10.1016/j.icte.2020.04.010.
[45] J. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks Trade* (Lecture Notes in Computer Science), G. Montavon and G. B. Orr, Eds. Cham, Switzerland: Springer, 2012, pp. 437–478, doi: 10.1007/978-3-642-35289-8_26.
[46] F. Bai, M. J. Marques, and S. J. Gibson, "Cystoid macular edema segmentation of optical coherence tomography images using fully convolutional neural networks and fully connected CRFs," 2017, *arXiv:1709.05324*.
[47] A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "ReLayNet: Retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomed. Opt. Exp.*, vol. 8, no. 8, pp. 3627–3642, 2017, doi: 10.1364/BOE.8.003627.
[48] T. Kepp, J. Ehrhardt, M. P. Heinrich, G. Huttmann, and H. Handels, "Topology-preserving shape-based regression of retinal layers in oct image data using convolutional neural networks," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1437–1440.
[49] H. Wei and P. Peng, "The segmentation of retinal layer and fluid in SD-OCT images using mutex dice loss based fully convolutional networks," *IEEE Access*, vol. 8, pp. 60929–60939, 2020, doi: 10.1109/ACCESS.2020.2983818.

ALEX CAZAÑAS-GORDÓN received the B.E. degree in electrical engineering from the National Polytechnic School, Quito, Ecuador, in 2003, and the M.Sc. degree in information technology from The University of Queensland, Brisbane, Australia, in 2015. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Coimbra, Coimbra, Portugal.

Since 2018, he has been a Researcher with the Multimedia Signal Processing Laboratory, Department of Electrical and Computer Engineering, University of Coimbra. His research interests include signal processing, deep learning, optical coherence tomography, scanning laser ophthalmoscopy, and fundus photography.

LUÍS A. DA SILVA CRUZ (Senior Member, IEEE) received the Licenciado and M.Sc. degrees in electrical engineering from the University of Coimbra, Portugal, in 1989 and 1993, respectively, and the M.Sc. degree in mathematics and the Ph.D. degree in electrical, computer, and systems engineering from the Rensselaer Polytechnic Institute (RPI), Troy, NY, USA, in 1997 and 2000, respectively. He has been with the Department of Electrical and Computer Engineering, University of Coimbra, since 1990, as a Teaching Assistant and as an Assistant Professor, since 2000. He is currently a Researcher with the Institute for Telecommunications of Coimbra, where he works on image and video processing and coding and medical image processing. He is a member of the EURASIP, SPIE, and IEEE technical societies.