

Received January 16, 2020, accepted January 31, 2020, date of publication February 7, 2020, date of current version February 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2972318

Improving the Performance of Convolutional Neural Network for the Segmentation of Optic Disc in Fundus Images Using Attention Gates and Conditional Random Fields

BHARGAV J. BHATKALKAR¹, DHEERAJ R. REDDY¹, SRIKANTH PRABHU¹,
AND SULATHA V. BHANDARY²

¹Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education (MAHE), Manipal 576104, India

²Department of Ophthalmology, Kasturba Medical College, Manipal Academy of Higher Education (MAHE), Manipal 576104, India

Corresponding author: Bhargav J. Bhatkalkar (bhargav.jb@manipal.edu)

ABSTRACT The localization and segmentation of optic disc (OD) in fundus images is a crucial step in the pipeline for detecting the early onset of retinal diseases, such as macular degeneration, diabetic retinopathy, glaucoma, etc. In this paper, we are proposing a novel convolutional neural network architecture for the precise segmentation of the OD in fundus images. We modify the basic architectures of DeepLab v3+ and U-Net models by integrating a novel attention module between the encoder and decoder to attain the finest accuracy. We also use fully-connected conditional random fields to further boost the performance of these architectures. We compare the results of our best proposed architecture against other established architectures for optic disc segmentation on our private dataset, as well as on publicly available datasets, namely, DRIONS-DB, RIM-ONE v.3, and DRISHTI-GS. The results obtained with the proposed method outperforms the existing methods in the literature.

INDEX TERMS Optic disc, attention network, conditional random fields, deep learning, biomedical imaging.

I. INTRODUCTION

Recent advancements in deep learning and computer vision have shown the effectiveness of Convolutional Neural Networks (CNNs) in solving challenging tasks such as image classification, image segmentation, image captioning, object detection and tracking, surpassing traditional algorithms, and achieving state-of-the-art results. The success of CNNs is attributed to their ability to progressively learn the abstract representations from the raw input domain, without hand-labeled features. The hierarchical architecture of CNNs allows for shallower layers to grasp local information, whereas deeper layers with larger receptive fields capture the global information.

Segmentation is a significant medical imaging task, as the automatic delineation of biological structures of importance is required for the automatic detection of disease, computer-

assisted diagnosis, and interventions. Since a vast majority of automated diagnostic data consists of 2D images, being able to perform segmentation by taking the entirety of the image content at once has an important relevance.

The precise segmentation of OD in fundus images is a challenging problem primarily owing to the retinal diseases bring in the pathological changes in the anatomy of the retina. Diseases like optic disc edema, optic disc hemorrhages, and glaucoma sometimes make the segmentation of OD very hard [1]. Also, in many cases, the quality of fundus images is not good enough to detect the OD precisely. The reasons like image distortions, noise introduced, and lack of technical expertise of the technician are the major causes of degraded fundus image quality. Figure [something] shows three fundus images having different levels of visibility of the OD with respect to its appearance.

Accurate labeling in medical images is both an expensive and time-consuming process. It requires skilled labor, and being completely error-free isn't possible in dense labels such

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

as segmentation masks. Automating this process is desired to increase the clinical work efficiency and to help with diagnostic decisions through faster and automatic extraction of regions of interest in the images.

Due to the difficulty in obtaining the medical images and annotating them, medical image datasets often tend to be quite small in numbers as compared to other computer vision datasets [2], [3]. The same situation is exaggerated for segmentation datasets due to more complex annotations as compared to the classification annotations.

The U-Net [4] is a benchmark architecture for biomedical image segmentation. It uses an encoder-decoder architecture with skip-connections [5] from the encoder stage to the corresponding decoder stage. To overcome the scarcity of large annotated medical datasets, U-Net uses extensive image augmentation to artificially increase the size of the dataset with the use of affine transforms and elastic deformations to mimic the natural biological variations found in biological structures in the image.

The DeepLabv3+ [6] is an architecture that uses spatial pyramid pooling module along with an encoder-decoder structure that achieves excellent results in various image segmentation tasks. The spatial pyramid pooling encodes multi-scale information at differing rates and effective fields-of-view. It is an extension of DeepLabv3 [7] with the addition of a decoder network to refine and sharpen object boundaries in the final output. The general modules of DeepLabv3+ can be incorporated with various other popular deep learning networks like ResNet, MobileNet, etc. which act as backbone networks.

Attention is one of the most influential ideas in the Deep Learning community. Processing the entire image during the subsequent training steps is time consuming. Attention mechanisms are efficiently employed to focus only on the target features in the image during training. Attention gates help the neural network to learn faster by highlighting the salient features useful for a specific task.

In medical imaging, fully-connected conditional random fields (CRFs) are used in image post-processing to refine the predictions. The usage of CRF results in much smoother boundaries between different class predictions, as well as correcting any noisy segmentation outputs of the CNN.

Although CRFs have been extensively used in various biomedical image segmentation tasks, recent advances in deep learning models and mechanisms have made them obsolete due to their higher training and inference time, with the gain in performance being rather minimal. We have compared our modified U-Net architecture with and without CRFs to examine both the accuracy of prediction and the complexity of the model (in terms of time taken).

The contributions of this paper are two-fold. Firstly, with our novel CNN, we achieve state-of-the-art OD segmentation results on public fundus datasets (DRIONS-DB, RIME-ONE v.3, DRISHTI-GS). Secondly, for the segmentation of OD, we compare the results and performance of our CNN

architecture with other state-of-art CNN architectures which are commonly used in biomedical image segmentation.

II. RELATED WORK

An automated segmentation of OD is a well-researched problem due to its very significance in detecting the other anatomical structures present in retinal images. An OD in the color retinal image has a defined boundary and is the brightest region within the scope of the image.

Application of CNN has become a standard approach for the OD segmentation in recent years. In 2014, Fully Convolutional Networks by [8] popularized CNN architectures for dense predictions without the need for any fully-connected layers, which allowed segmentation maps to be generated for image of any size and was also much faster compared to the patchwise training [9]

Since we use semantic segmentation with CNNs combined with attention-gates in deep learning for the segmentation of OD, we are focusing on these related works in the literature review.

A. SEMANTIC SEGMENTATION USING CNN

Semantic segmentation [10] is one of the vital problem areas in the field of computer vision. It is a high-level task in the image processing extensively used for complete scene understanding. Due to the popularity and the advancements in deep-learning algorithms in recent years, most of the semantic segmentation problems are addressed by the deep-learning architectures [11]. Convolutional Neural Networks are the leading deep-learning architectures in terms of their efficiency and accuracy [12].

The first of a kind CNN for the segmentation of OD and optic cup (OC) in color fundus images is proposed by [13]. The authors used a unique entropy sampling method which uses entropy filtering to calculate the entropy maps for each color channel. As a step in the pre-processing, they are converting the RGB fundus images into $L * a * b$ color space as it is closer to the human perseverance. As analyzing every pixel in the image is time-consuming, they use entropy maps to select only the dominant features in each channel and are used for training their CNN model. The usage of entropy maps during the training boosts the convolutional filters in their network. As a part of the post-processing, the probability map produced by the classifier is fed into the Graph Cut algorithm [14] for smoothing it. A convex haul function is applied to the smoothed mask to connect the disjoint points in the OD and OC regions. The authors have come up with their revised paper [15] explaining in detail about the ensemble learning approach followed by their CNN to boost the learning of filters.

The transfer learning technique was utilized by authors in [16] with pre-trained VGG [17] network for the image classification. In their modified VGG network, the fully connected layers at the end are removed, and only the pre-trained convolutional layers are retained for the generation of segmentation mask. To deal with the feature maps of different

size produced by the convolutional layers, they are using the inception architecture of GoogLeNet [18]. Their results have set a new benchmark for OD and optic nerve segmentation in color fundus images.

The automated, simultaneous segmentation of OD, blood vessels and fovea using a single 7-layer CNN is proposed in [19]. The authors have considered a pixel as an effective point in the fundus image if it lies in any of the three features to be segmented. The network is trained on DRIVE public database and evaluated with sensitivity, specificity and overlap metrics.

The U-net architecture [4] has been a benchmark model in the biomedical image segmentation. A modified U-net architecture is proposed in [20] for the segmentation of both OD and OC to detect the presence of glaucoma. As a preprocessing step, the input image is first passed to the CLAHE [21] method which improves the contrast across the image. The enhanced image is then passed to the modified U-Net network. The proposed method has very few trainable parameters in the network and is very much light-weighted in terms of memory requirement compared to the original U-Net model.

The authors of M-Net [22] jointly segment OD and OC using a four-layered architecture. The authors use the method given in [23] to detect the OD and its center. The first layer of the network performs the polar transformation of the fundus image by considering the center of the OD and sends it to the second layer which is a modified U-Net network. The second layer produces the multi-label prediction maps for OD and OC regions in the transformed image. The third layer is a slide-output layer that acts as a classifier and assigns each output instance to multiple binary labels. The inverse polar transformation is finally applied to recover the segmentation map in its original Cartesian form.

The recent work by authors in [24] used full-resolution residual networks (FRRN) [25] and atrous convolutions [7] in their deep CNN for the segmentation of OD. They call their unique network Fine-net. FRRN architectures are memory intensive and to overcome this, the authors are using atrous convolutions in FRRN units instead of conventional convolutions. They have assessed their network with five-fold cross-validation on three commonly used public databases, and the segmentation result is one of the best compared to the existing methods.

Another popular architecture for image segmentation is DeepLabV3+ [6] which uses a atrous spatial pyramid pooling module along with a decoder module to refine the segmentation masks along the boundary between objects of different classes. This achieves accurate segmentation maps, and the underlying backbone architecture can be changed based on the trade-off between speed of inference and accuracy.

B. ATTENTION IN DEEP LEARNING

The use of attention, especially soft-attention, has shown great promise in the field of deep learning. The authors [26] applied attention network in the domain of NLP for

machine translation to achieve state-of-the-art results in English-French translation. They used an attention module in the LSTM encoder-decoder for the alignment between the English and French words introducing the concept of trainable soft-attention.

Soft-attention allows a neural network to learn a feature-selector based on some external gating information or in contrast; it uses its own feature maps to learn a gating mechanism [27]. The benefit of soft-attention lies in its ability to train the feature selection mechanism in an end-to-end manner using backpropagation, and without the need for Monte-Carlo sampling [28] as in the case of hard-attention mechanism. Soft-attention module is both computationally and memory efficient compared to a hard-attention module making it easily integrated with the pre-existing deep learning models.

Following the success of [26], the authors of [29] utilized attention in a CNN along with an LSTM encoder-decoder for improving the accuracy of image captioning. They classified attention modules into three separate domains: spatial, channel-wise, and multi-layer attention. They used spatial and channel-based attention on the feature map of each CNN layer to create an attention vector. This attention vector is used to create an “attended” output, which is obtained by the weighted average of the attention vector and the feature maps of the same layer. This output is fed as the input to the next CNN layer.

The spatial and channel-wise attention is used in a residual network for image classification [30]. In this work, the attention mechanism is used as a feature selector in the forward propagation, and as a gradient filter in the backpropagation for an end-to-end training model. The attention residual learning introduced in this work is used to train a very deep attention network.

Attention module that requires no external gating information and which uses its own feature mapped input to generate a contextual attention vector is called self-attention module. Self-attention is a very powerful mechanism that is used as a standalone attention module or in conjunction with other attention modules to extract the significant features during training a deep network. The authors in [31] have used only the self-attention layers in their network and excluded the expensive recurrent and convolutional layers to train their sequence-to-sequence translation model. The result of their model outperforms traditional sequence-to-sequence models in different translation tasks. In the field of medical image segmentation, self-attention is used to capture contextual information across multi-scales without the need for explicit multi-scale training. The better contextual information allows the network to better correlate the features of interest from local-level to the global level.

III. METHODOLOGY

In this section, we will outline our architecture, dataset, and training procedure in detail.

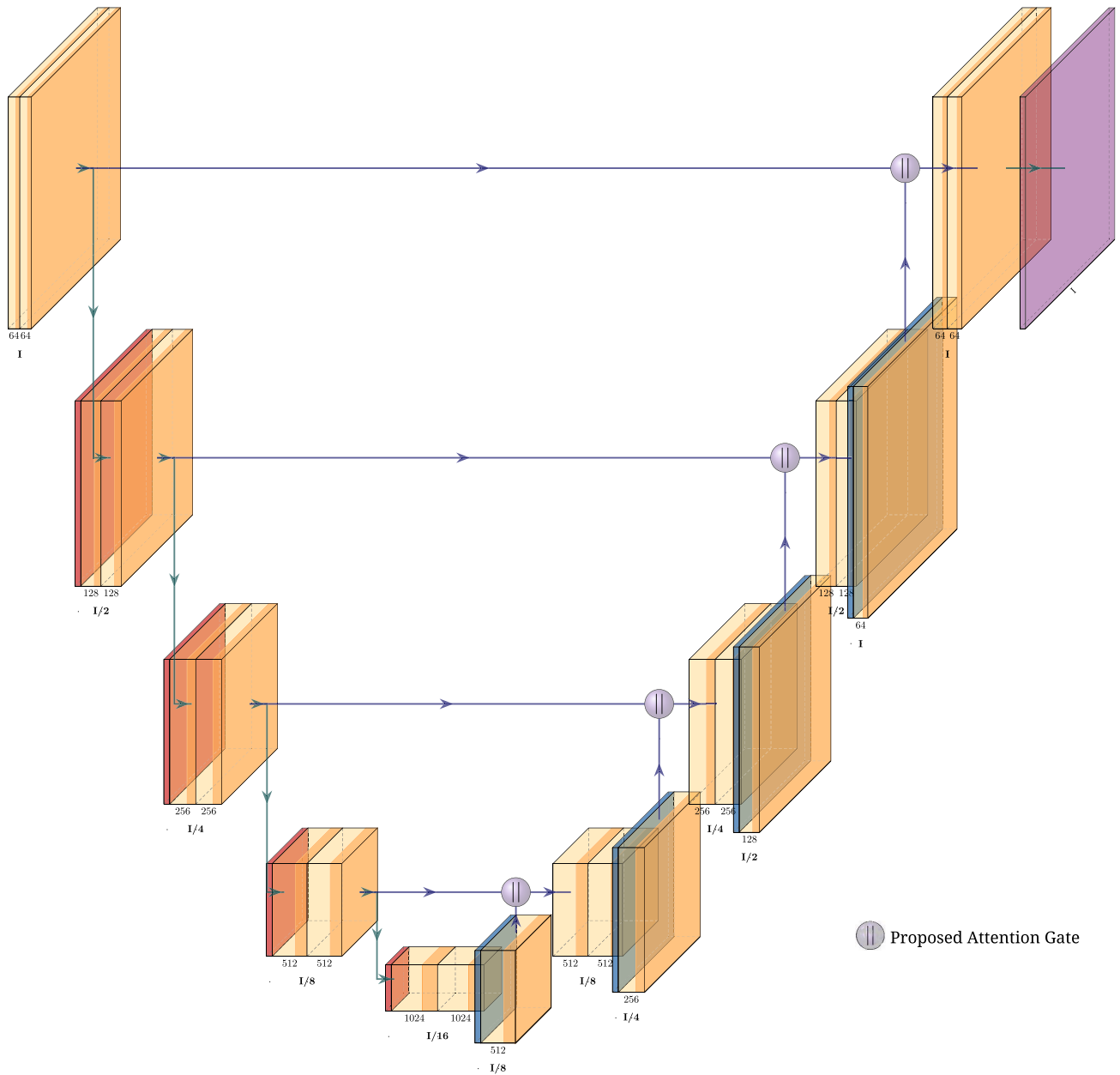


FIGURE 1. Modified U-Net architecture.

A. ARCHITECTURE

In the field of semantic segmentation in recent days, a popular class of deep learning architecture used is the encoder-decoder network. The encoder maps from the image space to a smaller latent space with convolutions, activation functions, and pooling layers. The decoder part of the network maps from this latent space to the label space with transpose convolutions, activation functions, and up-scaling layers. We are modifying the existing U-Net [4] and DeepLabv3+ [6] architectures for segmenting OD in fundus images. As an improvement to these architectures, we are using conditional random fields (CRFs) and a novel attention gating mechanism (AG) to

boost the segmentation accuracy. The following subsections give an insight into the custom modifications and enhancement mechanisms applied to these architectures.

1) MODIFIED U-NET

We use five 3×3 convolutional layers in each of the encoder and decoder modules. The encoder uses max pooling operation after each convolutional layer to reduce the size of the feature map. The decoder uses bi-linear interpolation to up-sample the feature maps after each convolutional layer. We also use batch normalization after each convolutional layer in both encoder and decoder modules. Between each

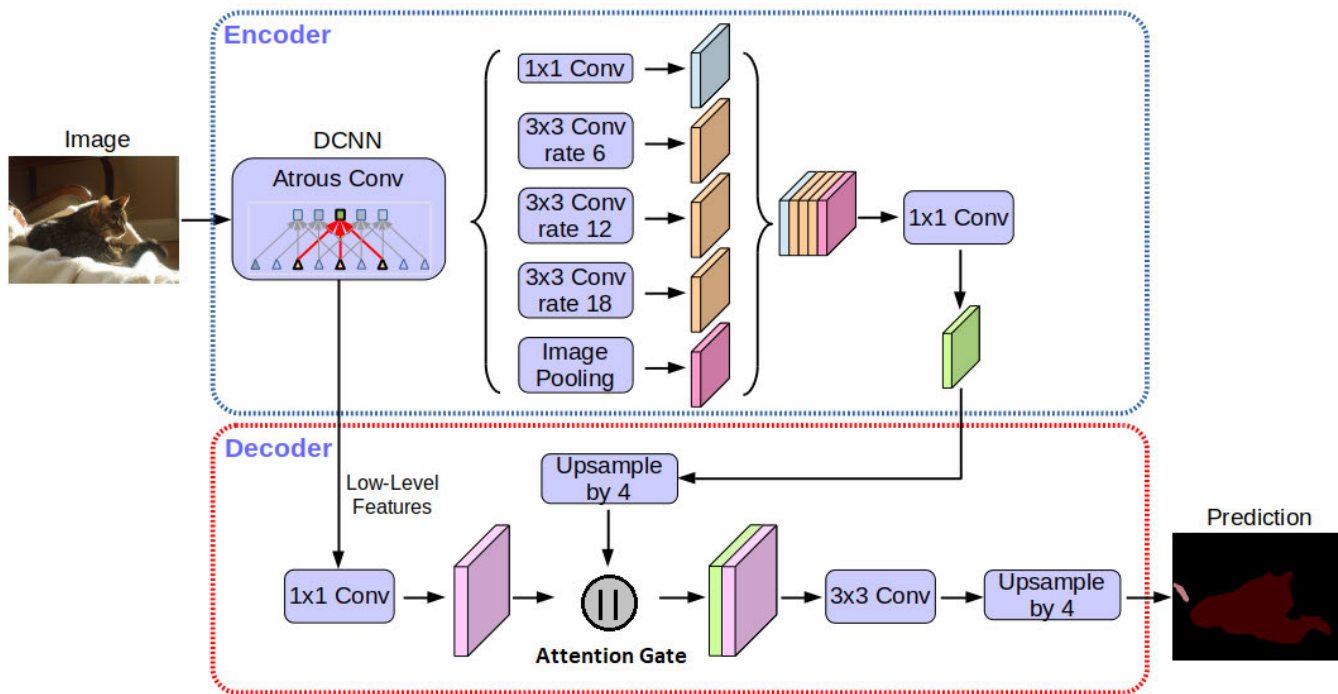


FIGURE 2. DeepLabv3+ architecture from [6].

skip connection from the encoder to the decoder, we replace the conventional concat operator with our proposed attention gate. Fig. 1 shows the proposed modified U-Net architecture.

2) MODIFIED DeepLabv3+

The DeepLabv3+ architecture works by using different backbone models depending on the computational constraints. We experiment with DeepLabv3+ as described in [6] with ResNet backbone and MobileNet backbone. ResNet uses the residual connections to allow the training of very deep networks without facing the problem of vanishing-gradients, and hence allowing for much more robust and expressive learning leading to many state-of-the-art results in computer vision benchmarks. MobileNet was designed for efficient computation, particularly, on mobile devices that use a low powered processor. It achieves this by replacing regular convolutions with depth separable convolution which requires far fewer FLOPS to compute. In our tests, the MobileNet based network achieved an inference speed that is 10x faster than the ResNet based model but with lower accuracy. We follow the same architecture as specified in [6] which is shown in Fig.2, but also add the proposed attention gate between the encoder and decoder modules similar to the previously described modified U-Net model.

For both modified U-Net and modified DeepLabv3+ architectures, we use Leaky ReLU as the activation function with α set to 0.1 to prevent the problem of dead neurons [33]. The final output of both these networks is passed through the sigmoid activation function $\sigma(x)$ to scale the network output

between (0, 1) to represent the probability of a pixel being part of the OD.

B. ATTENTION GATE

We use both spatial attention and channel-wise attention in our attention gating module but, instead of applying both one after the other, we compute each attention feature map separately and add them together at the end. We use additive attention because of its superior performance to multiplicative attention. The complete architecture of our proposed attention gate is shown in Fig. 3

The general formula of additive attention used in spatial attention and channel-wise attention is given by the equations (1), (2) and (3) respectively.

$$a = W_a^T (\sigma_1(W_x^T x + W_g^T g + b_g)) + b_a \tag{1}$$

$$\alpha = \sigma_2(a) \tag{2}$$

$$y = \alpha \times x \tag{3}$$

For spatial attention, W_a , W_x and W_g in the equation (1) are 1×1 2D conv layers that linearly map the feature vector x_i and gating vector g_i from (F, H, W) into matrices of (F', H, W) where F' is the number of intermediate feature channels. The functions σ_1 and σ_2 in equation(1) are Leaky ReLU and sigmoid activation functions respectively. The result is the spatial attention vector α_s which is multiplied element-wise across the channels of x resulting the output y_s . The spatial attention gate captures the important feature across the spatial dimensions of the feature vector x using g as the gating vector.

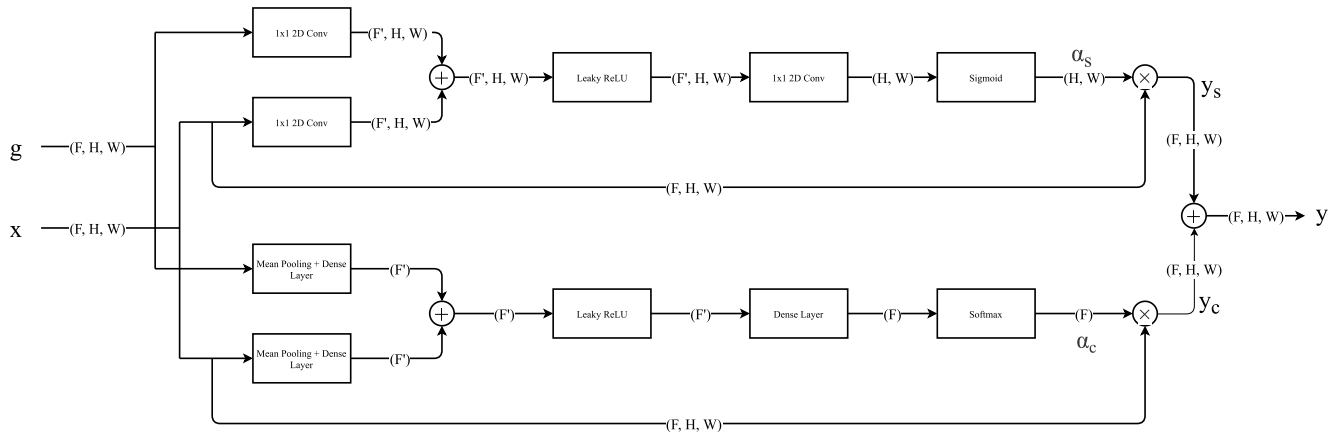


FIGURE 3. Proposed attention gate.

For channel-wise attention, W_a , W_x and W_g are fully connected dense layers. x and g are first mean pooled resulting in vectors of size (F) , and the dense layers convert the vectors into (F') where F' is size of the intermediate vector. The functions σ_1 and σ_2 in the equations are the Leaky ReLU and softmax activation functions respectively. The result is the channel attention vector α_c which is multiplied with x resulting in the channel attention output y_c . The channel attention gate assigns weights to each channel which capture different semantic features of the vector.

The final attention output y is a sum of both the spatial y_s and channel y_c attention gates.

In the U-Net architecture, x is the feature vector in the main branch from the decoder and g is the skip-connection from the encoder. Since the dimensions of x and g are the same, we need no interpolation for matching the size of α_s and x . In DeepLabv3+, we use the attention gate between the encoder and decoder, where x is the concated spatially pooled vector, and g is the intermediate vector. We use bilinear interpolation to make sure that the dimension of α_s and x match.

After trying out a range of value for F' , we used 16 channels for both spatial and channel attention. This was a good trade off between the GPU memory required for training the models and the gain in accuracy.

Attention gates also reduce the number of parameters in the network since the output vector of the attention gate has the same number of channels as the input, as opposed to the previous concat operator which doubles the number of layers. In our implementation of the base MobileNet model we used the summation operator instead of concat to reduce variables as much as possible.

C. CONDITIONAL RANDOM FIELDS

CRFs are mostly used on the final segmentation masks generated by the segmentation networks to refine the boundaries between different objects. The authors in [7] used a fully-connected CRF to overcome the limitations of short-range

CRFs to smooth the output masks further rather than recovering detailed local structures.

The CRF model employs the energy function (4) as given in [7]:

$$E(x) = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j) \tag{4}$$

where x_i is the label for i^{th} pixel. The unary potential is given by: $\theta_i(x_i) = -\log P(x_i)$, where $P(x_i)$ is the label assignment probability. The pairwise potential $\theta_{ij}(x_i, x_j)$ has a form that allows for efficient inference as in [7], [34], which take probability values (Grayscale values) of the pixels as well as their positions into the consideration. $E(x)$ forces pixels with similar position and color to have similar labels while ensuring that spatial proximity is considered for smoothness.

However, the need for CRFs in OD segmentation is not very necessary due to two reasons. Firstly, the ground truth masks are often smooth with little to no sharp deformation; secondly, our proposed networks capture the boundary between the OD and the retinal background quite distinctly, hence further minimizing the need for the CRFs. We nonetheless report the segmentation result using CRFs in the post-processing step.

To reduce computational load, we resize the prediction of the network from 720×576 to 180×144 and then feed it to the CRF during training and inference.

D. DATASET

We have trained both the proposed networks on our private, labeled dataset consisting of 300 fundus images with a resolution of 720×576 . The boundary of OD is labeled by an expert for 3 different times for each image and then averaged to get the ground truth.

Fig. 4 shows few samples from our private dataset with some illustrated labels for better visualization.

The dataset is split into 250 images for training, and 50 for testing. Since the dataset is so small, and convolutional networks require thousands of samples to converge

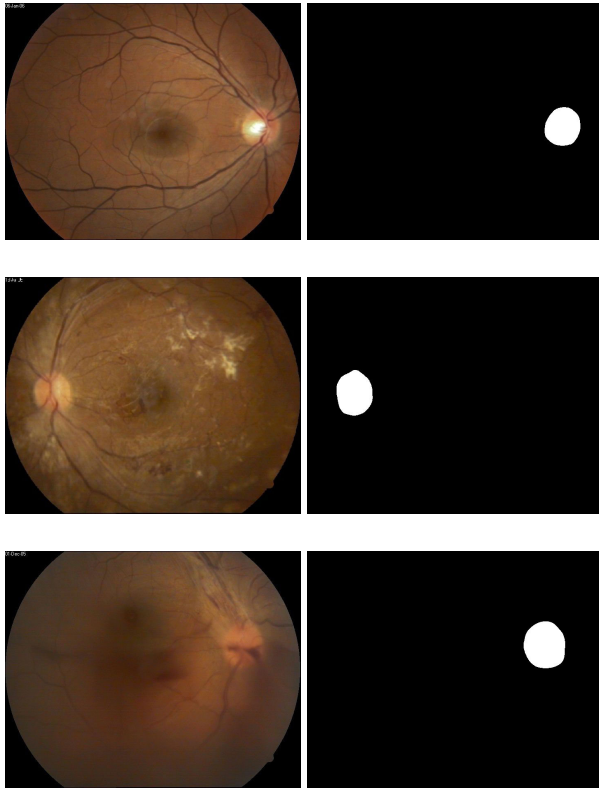


FIGURE 4. Sample images and labels from our private dataset.

successfully without overfitting, we make use of intense data-augmentation to increase the size of the training set artificially.

The following data augmentation techniques are used:

- Horizontal flips
- Random rotation between -10° to $+10^\circ$
- Random perturbation to the contrast of the image
- Random perturbation to the color of the image

All augmentations are applied during run-time for training of networks. The images are also divided by 255, to normalize the network's input.

Random elastic affine deformations of the images is a key concept used to train a segmentation network with very few labelled images. We generate displacement vectors on a 3×3 grid from a Gaussian distribution with a standard deviation of 10 pixels. The displacement vectors are used for smooth deformations with bi-cubic interpolation for calculating per-pixels displacements.

E. LOSS FUNCTION

The output mask generated by the network has a size of 720×576 , with each pixel value ranging between 0 to 1, which is the probability of that pixel lying in the OD. During training, the probabilistic output mask \hat{y} is compared with the ground label y for the loss calculation.

Since the output and the ground truth are probabilities of having or not having an OD, we may use binary cross-entropy

loss function during the network training given in (5).

$$L_{BCE} = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \quad (5)$$

However, L_{BCE} only acts as a proxy to the true objective that we're trying to optimize for, i.e., mean-IoU score. To overcome this problem with L_{BCE} , a soft-dice loss function [35] given by (6) can be used to train the network which is quite similar to the Sorensen-Dice coefficient [36] extended to work for non-binary vectors.

$$L_{DICE} = 1 - 2 \frac{|y\hat{y}|}{y^2 + \hat{y}^2} \quad (6)$$

Using L_{DICE} also takes care of the class-imbalance problem, but the gradients when using L_{DICE} are more complex compared to using L_{BCE} . This leads to unstable training, which may not converge smoothly.

To train the network, we use a combined loss function L_{comb} that is a weighted sum of L_{BCE} and L_{DICE} as given in (7).

$$L_{comb} = \alpha L_{DICE} + (1 - \alpha) L_{BCE} \quad (7)$$

where $\alpha = 0.3$ gives us the best results.

F. TRAINING

For optimization, we used Adam with an initial learning rate of 10^{-2} and a momentum of 0.9. The higher than usual learning rate is due to the high batch size afforded by using a Google Cloud TPU. We used a batch size of 256 for the U-Net based networks, and 128 for DeepLab-v3+ based networks.

Since the dataset is small, apart from batch normalization, we also use L2 regularization (8) to prevent overfitting:

$$L_{final} = L_{comb} + \lambda \|\theta\|^2 \quad (8)$$

where θ is the weight vector of the model, and λ is the regularization model, set to 10^{-6} .

All hyperparameters are tuned on the validation set using random grid searches. The model's parameters are initialized using the popular Xavier Initialization scheme.

We also tried to use the moving average of the weights during optimization as the final network weights, but this did not yield better results.

G. TESTING

The probability output mask \hat{y} of the network is converted to a binary mask using Otsu's method [37].

In traditional binary thresholding, a fixed value is used to differentiate between the two different class. The value is usually chosen to be 0.5, where all values less than 0.5 become 0, and all values greater than or equal to 0.5 become 1. This value is independent of the mask, and remains constant.

In Otsu's method we exhaustively search for the threshold value that minimizes the intra-class variance, defined as a weighted sum of variances of the two classes as given by (9).

$$\sigma_w^2(t) = w_0(t)\sigma_0^2(t) + w_1(t)\sigma_1^2(t) \quad (9)$$

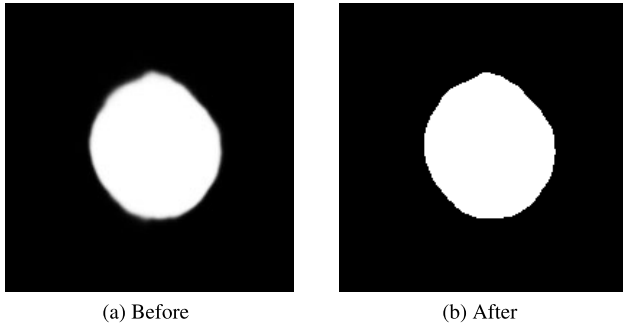


FIGURE 5. Otsu's method.

Weights w_0 and w_1 are the probabilities of the two classes separated by a threshold t , and σ_0^2 and σ_1^2 are the variances of these two classes.

In Fig. 5, we can see the predicted mask before and after applying Otsu's method.

The performance metrics used for the evaluation of segmentation accuracy are the Jaccard Index or mean-IoU (10), and the Dice coefficient (11).

$$JC(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (10)$$

$$DC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \quad (11)$$

where A and B are binary vectors representing the predicted and ground truth mask, respectively.

H. RADIUS OF OPTIC DISC

Once we obtain the binary segmentation mask of the optic disc, a beneficial metric to calculate is the diameter of the optic disc. However, since the optic disc is better represented as an ellipse instead of a circle, we use the method proposed in [32] to fit an ellipse to the mask, and report the measure of semi-major and semi-minor axis lengths in pixels.

IV. RESULTS

We are following a twofold experimentation method to evaluate the performance of proposed modified architectures. Firstly, we use our private dataset to check how these architectures perform on a wide range of fundus images with differing qualities. Secondly, we compare the architecture performing best on our private dataset with other benchmark methods on publicly available OD segmentation datasets.

We report both Dice Coefficient and Jacard Index metrics for parity between other works. All results are averaged over 5 runs to ensure reproducibility. We also report the time taken for inference on a GTX 1080Ti GPU system.

A. COMPARISON OF PROPOSED ARCHITECTURES ON THE PRIVATE DATASET

Table 1 shows the performance of our proposed architectures on our private dataset. As mentioned previously, the use of CRF added a very little boost in performance, and they are also significantly slower at inference due to the extra

TABLE 1. Proposed methods on our private dataset.

Architecture	DC(F)	Infer Time	Weights ($\times 10^6$)
U-Net	0.85	450ms	12.57
U-Net + CRF	0.86	2s	12.57
U-Net + AG	0.886	500ms	11.85
U-Net + AG + CRF	0.90	2s	11.85
DeepLabV3+ ResNet	0.963	650ms	30.49
DeepLabV3+ ResNet + AG	0.974	670ms	29.71
DeepLabV3+ MobileNet	0.92	80ms	3.50
DeepLabV3+ MobileNet + AG	0.933	110ms	3.57

TABLE 2. Drishti-GS.

Author	DC(F)	JC(O)	Acc
Sedai et al.	0.95	-	-
Nawalgi et al.	-	-	0.99
Zilly et al.	0.973	0.914	-
Oktoeberza et al.	-	-	0.9454
Al-Bander et al.	0.949	0.9042	0.9969
Best proposed architecture	0.96	0.92	0.996

TABLE 3. RIM-ONEv3.

Author	DC(F)	JC(O)
Sevastopolsky et al.	0.94	0.89
Shankaranarayana et al.	0.977	0.897
Al-Bander et al.	0.9036	0.8289
Best proposed architecture	0.97	0.94

TABLE 4. DRIONS-DB.

Author	DC(F)	JC(O)
Sevastopolsky et al.	0.94	0.89
Abdullah et al.	-	0.851
Zahoor et al.	-	0.886
Al-Bander et al.	0.9415	0.8912
Best proposed architecture	0.954	0.91

optimization step of minimizing the energy function of the fully-connected CRF. Table 5 shows the outputs of various models on selected images from our private test set. The segmentation of OD in Image 4 is a failure case due to no clear delineation of the optic disc boundary and its improper label.

The effect of attention gates (AG) is more evident in the case of U-Net as opposed to DeepLabv3+, and this is due to DeepLabv3+'s spatial pyramid pooling module that does a great job of capturing multi-scale information and thus making the attention gating mechanism redundant. The MobileNet backbone with AG has the fastest inference time of all the models without sacrificing accuracy too much. This model would be useful if the model is deployed for automatic segmentation in low powered devices such as mobile phones. The ResNet backbone with AG achieved the best OD segmentation accuracy on our private dataset. We select this modified architecture as our *best proposed architecture* (DeepLabV3+ ResNet + AG) for the comparison with other existing benchmark OD segmentation methods.

To compare the complexity of the various models, we have also listed the number of trainable weights in each network in Table 1. The more the number of weights, the longer it takes to train and for inference. However, a fully deep learning based model (the ones without CRFs) can be easily GPU

TABLE 5. Outputs from various models on selected images of private dataset.

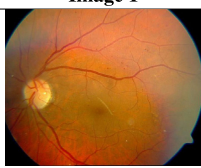

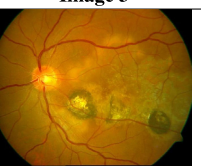

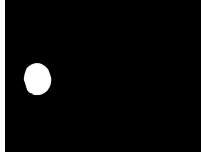
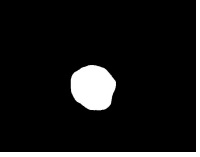
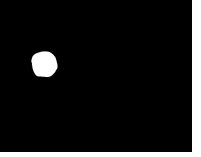
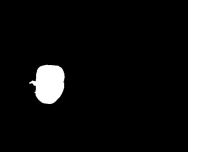
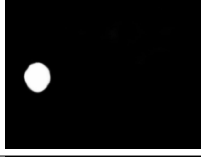
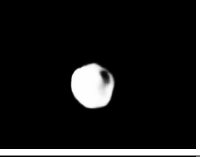
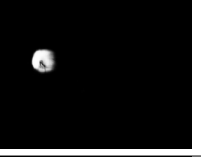
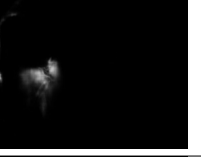
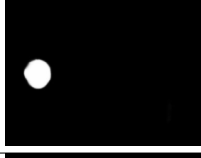
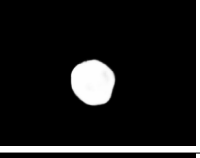
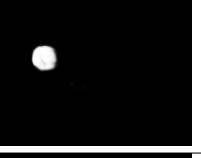

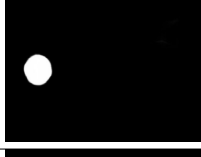
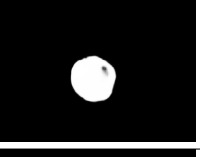
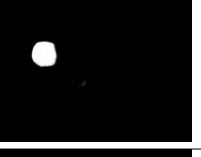
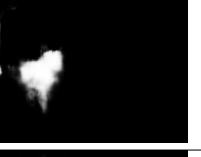
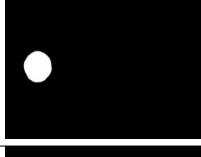
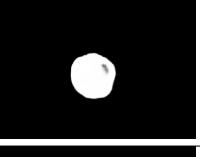
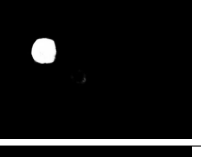

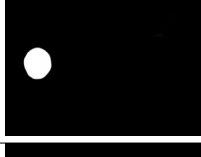
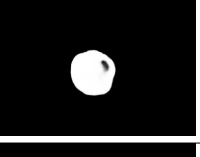


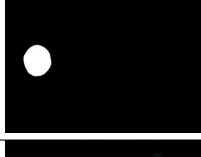
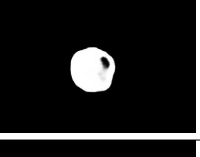
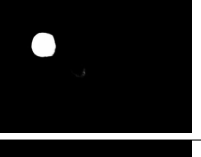

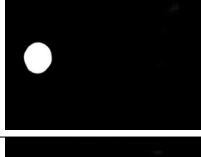
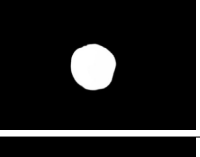
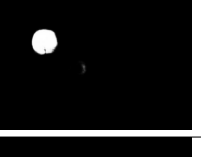


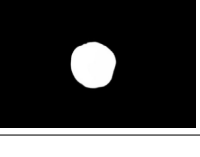
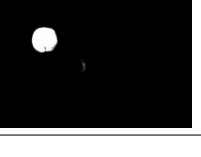




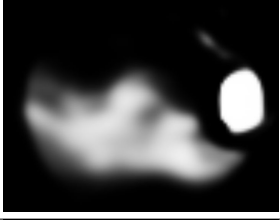


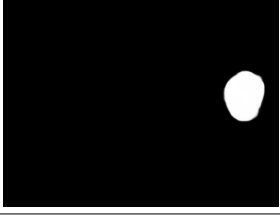
Method	Image 1	Image 2	Image 3	Image 4
Images				
Labels				
U-Net				
U-Net + CRF				
U-Net + AG				
U-Net + AG + CRF				
DeepLabv3+ (ResNet)				
DeepLabv3+ (ResNet) + AG				
DeepLabv3+ (MobileNet)				
DeepLabv3+ (MobileNet) + AG				

TABLE 6. Visualization of activation maps of intermediate attention gates for the Unet backbone.

Image Type and Original Resolution	Image (α_s)
Input image, (720 × 576)	
Level 1 α_s , (23 × 18)	
Level 2 α_s , (45 × 36)	
Level 3 α_s , (90 × 72)	
Level 4 α_s , (180 × 144)	
Level 5 α_s , (360 × 288)	
Predicted mask, (720 × 576)	

accelerated by modern deep learning libraries (TensorFlow, PyTorch, etc.) but CRFs are not easily GPU accelerated and

are trained using CPU only, and this leads to their slower performance.

B. COMPARISON OF BEST ARCHITECTURE WITH BENCHMARK METHODS ON THE PUBLIC DATASETS

Table 2, 3 and 4 shows the performance of our best proposed architecture against other benchmark methods for the OD segmentation on three popular datasets: Drishti-GS, RIM-ONEv3 and DRIONS-DB.

The proposed architecture performs as well or better than the existing benchmark methods across the datasets. Some of the listed works do both optic disc as well as optic cup segmentation, whereas we only measure the optic disc segmentation accuracy.

V. DISCUSSION AND LIMITATION

To ensure that the attention gates in our network are actually learning features of the optic disc, we have visualized the activation maps of the spatial attention branch (α_s) as shown in Table 6. We have used the attention gates of the U-Net backbone for the result demonstration because of the reason that U-Net uses the multiple hierarchical attention gates which better shows the effectiveness of the gating mechanism. The DeepLabv3+ also exhibits similar attention activation however, it only has a single attention gate. α_s is a vector of values from [0, 1] so we multiply by 255 to create a grayscale image. From the table, it is evident that each successive attention gate more finely centres on and picks up the bright spots of the image, until it narrows down on the optic disc region.

The methodology suffers the same drawbacks as most machine-learning based approaches [40]. If the distribution of the training data differs from the testing data, then we have degraded performance. A prominent case that we observed in testing was when the boundary between the optic disc and retinal background were not clearly delineated as in Image 4 of Table 5.

VI. CONCLUSION

In this paper, we show that attention mechanisms and CRFs can be used to boost the performance of deep convolutional neural network based models for OD segmentation. The proposed architecture is generic and modular; as such, it can be easily applicable to other biomedical segmentation tasks. Experimental results have demonstrated that the proposed network outperforms or matches the existing methods to achieve state-of-the-art results on publicly available datasets.

In the future, we would like to experiment using various gating structures and backbone architectures. We plan to exploit Deformable convolutions [38], which are now being more popularly used in encoder-decoder architectures to dynamically learn the structure of the primary convolution operator accounting for geometric distortions in the objects of the image.

After a few more rounds of validation, we would also like to publicly release our private training dataset to add another benchmarking dataset for optic disc segmentation.

Fully-connected CRFs result in slow training and inference, and it would be beneficial to replace it with convolutional CRFs [39] that reformulate the inference in terms of convolutions, which can be efficiently implemented on GPUs leading to much faster training and inference.

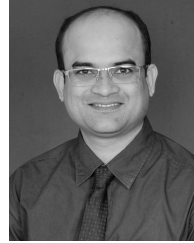
ACKNOWLEDGMENT

We thank Kasturba Medical College (KMC), Manipal, India, for providing the fundus image dataset, and TensorFlow Research Cloud (<https://www.tensorflow.org/tfrc>) for providing the computing resources. We sincerely thank the efforts of experts for labeling the private dataset required for the training. (Bhargav J. Bhatkalkar and Dheeraj R. Reddy contributed equally to this work.)

REFERENCES

- [1] L. Xiong and H. Li, "An approach to locate optic disc in retinal images with pathological changes," *Comput. Med. Imag. Graph.*, vol. 47, pp. 40–50, Jan. 2016.
- [2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
- [4] O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [6] 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)*, 2018, pp. 801–818.
- [7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
- [8] 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.
- [9] C. Farabet, C. Couprie, L. Najman, and Y. Lecun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.
- [10] S. Basu, "Image segmentation by semantic method," *Pattern Recognit.*, vol. 20, no. 5, pp. 497–511, Jan. 1987.
- [11] A. Garcia-Garcia, S. Orts-Escobedo, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*. [Online]. Available: <https://arxiv.org/abs/1704.06857>
- [12] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *Int. J. Multimed. Inf. Retr.*, vol. 7, no. 2, pp. 87–93, Jun. 2018.
- [13] J. G. Zilly, J. M. Buhmann, D. Mahapatra, "Boosting convolutional filters with entropy sampling for optic cup and disc image segmentation from fundus images," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, Oct. 2015, pp. 136–143.
- [14] M. B. Salah, A. Mitiche, and I. B. Ayed, "Multiregion image segmentation by parametric kernel graph cuts," *IEEE Trans. Image Process.*, vol. 20, no. 2, pp. 545–557, Feb. 2011.
- [15] J. Zilly, J. M. Buhmann, and D. Mahapatra, "Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation," *Comput. Med. Imag. Graph.*, vol. 55, pp. 28–41, Jan. 2017.
- [16] K. K. Maninis, J. Pont-Tuset, P. Arbeláez, L. Van Gool, "Deep retinal image understanding," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, pp. 140–148.

- [17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: <https://arxiv.org/abs/1409.1556>
- [18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [19] J. H. Tan, U. R. Acharya, S. V. Bhandary, K. C. Chua, and S. Sivaprasad, "Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network," *J. Comput. Sci.*, vol. 20, pp. 70–79, May 2017.
- [20] A. Sevastopolsky, "Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network," *Pattern Recognit. Image Anal.*, vol. 27, no. 3, pp. 618–624, Jul. 2017.
- [21] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," *J. VLSI Signal Process.-Syst. Signal, Image, Video Technol.*, vol. 38, no. 1, pp. 35–44, Aug. 2004.
- [22] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018.
- [23] J. Xu, O. Chutatape, E. Sung, C. Zheng, and P. C. T. Kuan, "Optic disk feature extraction via modified deformable model technique for glaucoma analysis," *Pattern Recognit.*, vol. 40, no. 7, pp. 2063–2076, Jul. 2007.
- [24] D. Mohan, J. R. H. Kumar, and C. S. Seelamantula, "High-performance optic disc segmentation using convolutional neural networks," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4038–4042.
- [25] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4151–4160.
- [26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: <https://arxiv.org/abs/1409.0473>
- [27] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2015, pp. 2048–2057.
- [28] A. Shapiro, "Monte Carlo sampling methods," in *Handbooks Operations Research And Management Science*, vol. 10. Amsterdam, The Netherlands: Elsevier, 2003, pp. 353–425.
- [29] 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.
- [30] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.
- [31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [32] A. Fitzgibbon, M. Pilu, and R. Fisher, "Direct least square fitting of ellipses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 5, pp. 476–480, May 1999.
- [33] L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis, "Dying ReLU and initialization: Theory and numerical examples," 2019, *arXiv:1903.06733*. [Online]. Available: <https://arxiv.org/abs/1903.06733>
- [34] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 109–117.
- [35] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.
- [36] T. A. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons," *Biol. Skar.*, vol. 5, pp. 1–34, May 1948.
- [37] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
- [38] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9308–9316.
- [39] M. T. Teichmann and R. Cipolla, "Convolutional CRFs for semantic segmentation," 2018, *arXiv:1805.04777*. [Online]. Available: <https://arxiv.org/abs/1805.04777>
- [40] S. Rabanser, S. Günnemann, and Z. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1394–1406.



BHARGAV J. BHATKALKAR received the B.E. and M.Tech. degrees in computer science and engineering from Vishweshwaraiah Technological University, India, in 2007 and 2012, respectively. He is currently pursuing the Ph.D. degree in medical image analysis and automated disease diagnostics using machine learning techniques from the Manipal Academy of Higher Education, Manipal. His areas of research interest are image processing, pattern recognition, medical image analysis,

machine learning, and data security.

He started his career as a Software Developer, and later, he switched to the teaching profession. He has 12 years of experience in teaching undergraduate students at the university level. He is currently working as an Assistant Professor-Senior Scale with the Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal, India.

Mr. Bhatkalkar is a Lifetime Associate Member of the Institution of Engineers, India.



DHEERAJ R. REDDY is currently pursuing the B.Tech. degree in computer science and engineering with a minor specialization in computational mathematics with the Manipal Institute of Technology, Manipal, India.

He was the Team Leader of Project MANAS at the institute level which is a funded project for the design and development of an autonomous driving car. His areas of interest are deep learning, machine learning, and robotics.



SRIKANTH PRABHU received the Ph.D. degree in biometric systems design from IIT Kharagpur. He is currently working as an Associate Professor with the Department of Computer Science and Engineering, MIT, Manipal, India.

He has more than 20 years of experience in teaching undergraduate and postgraduate university students. He is a member of various the Doctoral Advisory Committees in the institute and the Coordinator for Computer Applications in Modern

Algebra. His areas of research interest are pattern recognition, pattern classification, fuzzy logic, image processing, and parallel processing.



SULATHA V. BHANDARY received the M.B.B.S. and M.S. degrees in ophthalmology. She is currently a Professor and the Head of the Ophthalmology Department, KMC, Manipal, India.

She has more than 17 years of working experience as a Surgeon. She teaches undergraduates, postgraduates, and medical students. She is also a Ph.D. thesis Guide. Her areas of interest are cataract, glaucoma, retina, trauma, keratoplasty, management of lacrimal diseases, and community ophthalmology. Her areas of expertise include phacoemulsification (cataract surgery), retinal vitreo-retinal diseases including diabetic retinopathy, venous occlusions, retinal detachment, retinopathy of prematurity, ocular trauma, glaucoma management, and lacrimal surgeries. Her research interests are cataract diagnosis and treatment, and retinal imaging.

Dr. Bhandary has affiliations to various professional bodies such as the member of Karnataka Ophthalmic Society, in 2004; the member of the All India Ophthalmological Society, in 2007; the Member of the Karnataka Ophthalmic Society, in 2008; the member of the All India Ophthalmic Society, in 2009; and the member of the Glaucoma Society of India, in 2013.

• • •