Scientific Research Publishing

Open Journal of Medical Imaging, 2020, 10, 165-185
https://www.scirp.org/journal/ojmi
ISSN Online: 2164-2796
ISSN Print: 2164-2788

Segmentation of Diabetic Retinopathy Lesions: The Common Fallacy and Evaluation of Real Segmenters

Pedro Furtado

University of Coimbra, DEI/CISUC Polo II, Coimbra, Portugal
Email: pnf@dei.uc.pt

How to cite this paper: Furtado, P. (2020) Segmentation of Diabetic Retinopathy Lesions: The Common Fallacy and Evaluation of Real Segmenters. *Open Journal of Medical Imaging*, **10**, 165-185.
https://doi.org/10.4236/ojmi.2020.104016

Received: September 1, 2020
Accepted: October 25, 2020
Published: October 28, 2020

Abstract

In the context of automated analysis of eye fundus images, it is an important common fallacy that prior works achieve very high scores in segmentation of lesions, and that fallacy is fueled by some reviews reporting very high scores, and perhaps some confusion with terms. A simple analysis of the detail of the few prior works that really do segmentation reveals scores between 7% and 70% in sensitivity for 1 FPI. That is clearly sub-par with medical doctors trained to detect signs of Diabetic Retinopathy, since they can distinguish well the contours of lesions in Eye Fundus Images (EFI). Still, a full segmentation of lesions could be an important step for both visualization and further automated analysis using rigorous quantification or areas and numbers of lesions to better diagnose. I discuss what prior work really does, using evidence-based analysis, and confront with segmentation networks, comparing on the terms used by prior work to show that the best performing segmentation network outperforms those prior works. I also compare architectures to understand how the network architecture influences the results. I conclude that, with the correct architecture and tuning, the semantic segmentation network improves up to 20 percentage points over prior work in the real task of segmentation of lesions. I also conclude that the network architecture and optimizations are important factors and that there are still important limitations in current work.

Keywords

Semantic Segmentation, Diabetic Retinopathy, EFI, Deep Convolution Neural Networks

1. Introduction

When a trained medical doctor looks at an Eye Fundus Image (EFI) he is able to

distinguish sufficiently well not only coarse regions around lesions but also the contours and areas of many individual lesion instances. Likewise, the better the automated segmentation procedure is at classifying each pixel, the more precise automated analysis and quantification will be (e.g. areas, contours, number of instances). Difficulties segmenting lesions include the size of those lesions, frequent lack of contrast and varying shapes. Most prior works frequently said to be related to this task do not segment at all. The subtle difference is that to segment an approach needs to receive as input the whole image and needs to trace the contours of each lesion. Most "lesion detection" approaches in related work do not even find locations of lesions in the image. Instead, they receive as input small squares and classify those squares as a certain lesion or as background. In other words, they do the easy work and leave out the difficult work. Worse still, surveys such as [1] [2] [3] can report scores of 90% to 100% in tasks that are segmentations of lesions but are not in reality, and I look at the details of the prior work to show that. There are three main reasons for the confusion: first of all, most works do not segment at all, as just explained before and review in related work; secondly, for those that do segment, when looking into the details of their experiments one discovers that those works only report very high scores for identification of lesions at image level (whether there is a lesion or not in an image), reporting much lower scores in sensitivity versus FPI in true segmentation of lesions. The other problem that was also mentioned for instance in [4] is that many scores are artificially high, because in the context of eye-fundus images, the background is huge and hence the term TN (true negatives) is also huge, making specificity, ROC and AUC inviable as scores. Sensitivity is also inviable if reported alone because it lacks recognition of FP (false positives), and in the context of segmentation of lesions the default classification threshold of 0.5 is associated with too many FP. That is why authors reporting segmentation scores correctly use the sensitivity when FP varies between 1 to 10 false positives per image instead of just sensitivity. The last problem also frequent is that the approaches are evaluated based on the degree of overlap between segments, which means that, for instance, a large region matching only 20% with the true segment can be considered a match if the overlap threshold is 20%. An overlap of 20% or even 50% between regions, no matter the shape of size, is a very bad tracing of contours that is accounted for as 100% correct in those works.

The focus of this work is on explaining what prior works propose, expose the common fallacy and then build and evaluate semantic segmentation networks that really segment the lesions, comparing to prior work, to show that a good-performing segmentation network is better than prior work. This destroys a common fallacy and shows the way for future improvements in segmentation of lesions in eye-fundus-images.

This article is very relevant because it identifies an important common fallacy concerning the perception of what prior works achieve in segmentation of diabetic retinopathy lesions, and compares true segmenters with what they do in

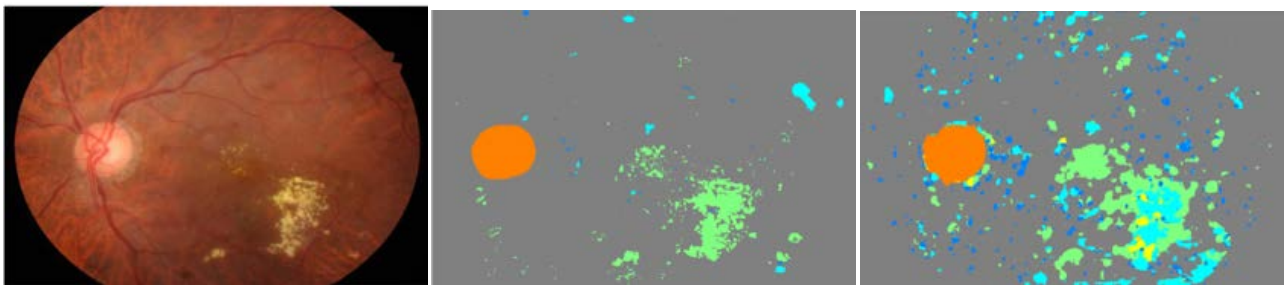
terms of segmentation.

1) *The Lesions segmentation problem*

Lesions that are characteristic of Diabetic Retinopathy (DR) in different stages include micro-aneurisms (MA), which are small red and rounded regions resulting from augmented capillaries, exudates (hard and soft HE and SE), which are yellowish deposits of lipids and proteins, and hemorrhages (HA), larger blood stains that are a serious signal of advancing conditions. Proliferative Diabetic Retinopathy also exhibits neo-vascularization and other affections [5]. **Figure 1(a)** shows the original Eye Fundus Image (EFI) and **Figure 1(b)** shows the corresponding groundtruth pixelmap with hard and soft exudates (EX, HE, SE), microaneurysms (MA) and hemorrhages (HA). It also shows the optic disc (OD). The segmentation Deep Convolution Neural Network (DCNN) is given a large dataset with images similar to the one shown in **Figure 1(a)** and pixelmaps similar to **Figure 1(b)** and learns how to classify each pixel of the image to obtain a pixelmap as close as possible to the pixelmap shown in **Figure 1(b)**. **Figure 1(c)** is an example of a segmentation result, showing the lesions and optic disk that were detected by the deep learning network. In summary, semantic segmentation is the classification of each pixel as one of the possible classes that includes each type of lesion plus the background (BK, the background is everything that is not a lesion and covers more than 93% of all the EFI area). Evaluation is based on assessing the quality of segmentation of all pixels or pixels of a certain class (e.g. lesion), using a typical set of metrics.

Figure 2 shows a slightly different definition of groundtruths and evaluation most used in related work. The images are taken from DIARET-DB1 dataset [7]. In this case the groundtruth segments are shapes with varied sizes that enclose lesions or sets of lesions. The quality of segmentation is assessed by analyzing the amount of overlap between the segments each pixel belongs to and the groundtruth segments. A match is found if the overlap is above a certain threshold σ between 0 and 1, in which case all pixels in the segment are considered a hit (e.g. TP or TN depending on the case).

Semantic segmentation DCNNs can be applied and evaluated with both types of groundtruth and evaluation approach. Since prior work evaluates using the second approach and DIART-DB1 is one of the most frequent datasets used in



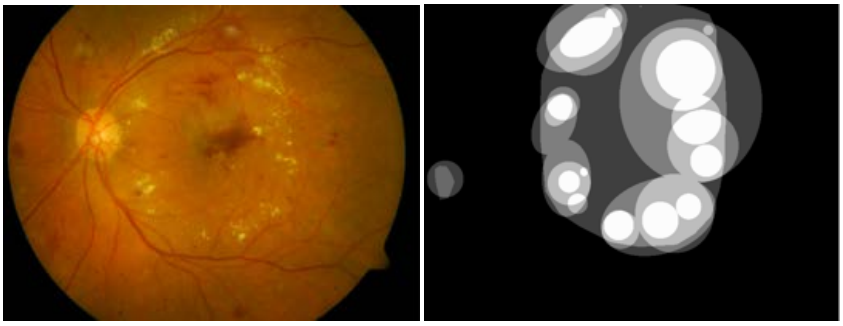
(a) (b) (c)

Figure 1. Eye Fundus Image (EFI), groundtruth and output pixelmaps (IDRID dataset [6]). (a) Lesions on EFI Image; (b) Lesions Map; (c) segments.

those evaluations, it is used here as well to compare with prior works, while the first dataset and approach is used to evaluate and compare the quality among segmentation networks because it is more pixel-wise.

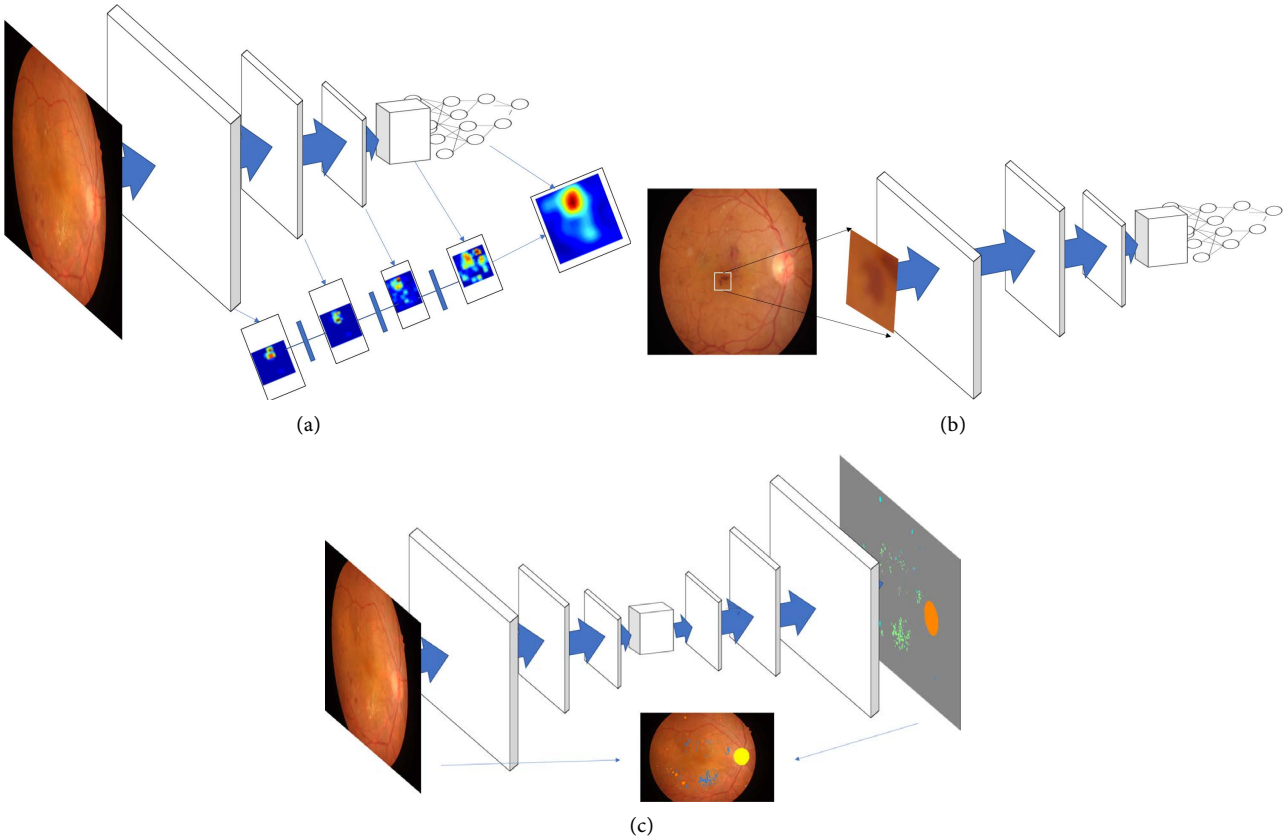
2) *Deep Learning Approaches to Segment Lesions*

Most proposals described as segmentation and/or identification of lesions in related work follow two main approaches shown in **Figure 3(a)** and **Figure 3(b)**. **Figure 3(a)** shows the generation of lesions CAM/heatmaps from the inner coefficients of a Diabetic Retinopathy Classification network (e.g. [8] [9]).



(a) (b)

Figure 2. Example of EFI and groundtruth (DIARET-DB1 dataset [7]). (a) Lesions in EFI; (b) groundtruth.



(a) (b)

(c)

Figure 3. Types of segmentation or lesion identification networks. (a) CAM/Heatmap generation; (b) Small windows lesion classifier; (c) Segmentation network.

Figure 3(b) is a classifier of small windows (e.g. works [10] [11] [12] [13], with some typical window sizes being 25 × 25, 45 × 45, 65 × 65, 129 × 129). The first is trained with EFI images classified as DR or not DR, the second is trained with windows containing lesions of a single type and negative windows containing background and no lesions. Both approaches are based on a classification DCNNs, which consist of a set of convolution layers with pooling and ReLU that encode the image or window into features, followed by a fully connected neural network that produces the classifications. However, the classifier of small windows is not by itself a real segmentation approach, because all it does is classify windows given to it directly, it does not process the image to segment contours. The most difficult part, actually localizing the lesions and the contours, is completely avoided, and up to hundreds of thousands of small windows (as many as the pixels) would have to be classified to segment the image using such an approach. Processing hundreds of thousands of windows would take a huge amount of time. It will be shown that the only small window classifier that processes the whole image uses trivial thresholding for hand-crafted segmentation of MA lesions, which is sub-par compared with end-to-end learning and produces lower scores. On the contrary, the segmentation network shown in Figure 3(c) trains and outputs a segmentation pixelmap with the classification of each pixel. It includes encoder and decoder stages. The encoder is in essence a DCNN without the last fully-connected stage that extracts features and compresses, the decoder is made of a set of de-convolution layers that together reinstate the image size and output a segmentation pixelmap. These alternatives have to be compared. In related work limitations of the proposed approaches will be discussed, and they will be compared in the experimental section.

3) ***Contributions and organization of the work***

This work answers the following questions: 1) what are the referenced prior works really doing, are they segmenting at all, and those that segment, how good are they really in segmentation? 2) How do they compare with a real segmentation network? 3) How do different segmentation architectures compare? 4) How do the results vary depending on the lesion being detected? 5) What are the limitations in terms of segmentation quality that need to be handled in future work?

The remainder of this work is structured as follows: Section 2 reviews the contributions of others and how their approaches relate to ours. Section 3 presents alternative segmentation network designs. Section 4 is experimental setup. Section 5 presents experimental results and section 6 discusses the results and concludes regarding experimental work. Experimental work includes the comparison of architectures, optimizations of training options, comparison of the top-performing segmentation network to the solutions proposed in prior work, showing the superiority of segmentation network, and investigation of limitations of current state-of-the-art. Finally, a few visualizations are shown and conclusions are provided in section 7.

2. Related Work

Lesions detection and localization from eye fundus images using either classical Machine Learning (ML) or Deep Learning (DL) was surveyed most recently in works such as [1] [2] [3], where most tasks report very high performance scores (90% and 100%) measures in metrics such as sensitivity, ROC/AUC and others. However, if we look closer only at approaches that are reviewed and claim to "segment" lesions, the list is already filtered to a much smaller number, including Prentasic *et al.* [10], Gondal *et al.* [8], Quellec *et al.* [9] (exudates, hemorrhages and microaneurisms), Haloi *et al.* [11], van Grinsven *et al.* [12], Orlando *et al.* [14] and Shan *et al.* [13] (microaneurisms, hemorrhages or both), and if we look at their performance results in detail and using the adequate metrics, it is possible to see that scores are not as high. Also importantly, there are other details that should be taken into account when analyzing related work, which we discuss next.

For a work to propose a way to segment, it should input an EFI image and output all segments of its lesions or at least of a specific type of lesion that exists in the EFI image, and we should expect experiments in those papers to include a setup that does and evaluates that. Prentasic *et al.* [10], Haloi *et al.* [11], Van Grinsven [12] and Shan *et al.* [13] do not fit this definition of segmentation. Those works are classification DCNNs that classify small square windows as being of a specific type of lesion or not. They can also be seen as classifiers of a single pixel, the center pixel of the squared window since the classification they output does not intend to guarantee that all the other pixels in the window have that same type. Although an image typically has hundreds of thousands of pixels, those works do not clarify how the classifier would be applied to the whole image or how segmentation of the whole image would work based on the classifier. Not only those works in general involve no segmentation, as the test windows for evaluating the approaches use windows picked based on the groundtruths with the lesions perfectly centered. Such experiments are totally unrealistic from the perspective of segmentation, where the input should be the EFI image and there is no prior knowledge of the location of lesions. The authors make no attempt in the papers at applying this as a building block to segment lesions, and there is also no good solution provided or experimented to scale this to real-time operation classifying all pixels of an EFI in a way that could enable segmentation. For instance, [10] is a pixel classifier that takes a window around a pixel to be classified, achieving 77% sensitivity classifying exudates, but the small test squares used for evaluation, as well as the train ones, were picked based on the groundtruth to get squares centered on a lesion (and also negative squares with background only). All that is done is the classification of those squares, the authors do not propose or experiment with scaling this to classify or segment all pixels in an EFI image. [12] and [13] have exactly the same limitations. [12] reports, at 1 false positive per image (FPI), a maximum sensitivity of 0.786 detecting Hemorrhages.

Haloi [15] is another pixel/window classification approach, but there the authors propose to preprocess the image with a color threshold to leave out trivial non-MA pixels to reduce computation time. To detect MA in an unseen image, they first apply a mask to get all pixels of interest, removing the usual black region, and apply a color threshold to leave out trivial non-MA pixels to reduce computation time. Then a window of size 129 × 129 centered at each image pixel is extracted. The classifier is used to classify every of the extracted pixels, resulting in a probability map that is then post-processed to remove false detections by analyzing connected regions using the concept of convexity and area. The authors report sensitivity for 1 false positive per image (FPI) around 50% (70% for 10 FPI). This is an interesting work that, contrary to the previous ones, deals with the whole EFI image, but it also has some limitations. First of all, the colour threshold is a very basic hand-crafted solution that easily returns too many false positives and also easily misses many MA, and the time cost of processing all remaining image pixels one-by-one, each one centered on a 129 × 129 square window around it can be large if the threshold is set to keep more pixels. A tradeoff would need to be studied between the colour threshold, precision of the approach and runtime. The main advantage of the segmentation network compared to this threshold is that it is optimized end-to-end. The backpropagation learning learns to extract MA and other lesions based not on a fixed manually-defined colour threshold, but rather on learning adjustments to convolution filter coefficients. Another limitation of [15] is that it is only dealing with MA, the procedure would have to be defined for other lesions as well, and it is not clear how the colour threshold and post-processing details would be defined for those.

The remaining related works Gondal *et al.* [8], Quellec *et al.* [9] and Orlando [14] all segment the EFI. [8] and [9] are based in DR classifier networks from which CAM/heatmaps are extracted to get the positions of lesions. Orlando [14] uses a different approach that combines DL with image processing to find candidate regions. These three works evaluate the quality of segmentation of lesions using the amount of overlap of connected componentes as criteria. In those works sensitivities reported for 1 false positive per image (1 FPI) are: (HA = hemorrhages, MA = micro-aneurisms, HE = hard exudates, SE = Soft Exudates): Quellec [9] (HA = 47%; HE = 57%; SE = 70% and MA = 38%), Gondal [8] (HA = 50%; HE = 40%; SE = 64% and MA = 7%) and Orlando [14] (HA: 50%, MA: 30%). These relatively low scores mean that improvements are welcome.

This related work section is ended by briefly reviewing deep segmentation networks. The segmentation network has two well-distinguished parts, the encoder, most frequently an existing Convolution Neural Network (CNN) without the final fully connected layers, a decoder that reinstates the full image size, and the pixel classifier layer that assigns a score for each class to each pixel. The Fully Convolutional Network (FCN) [16] uses a CNN for encoding (e.g. VGG16 [17]), replacing all the fully connected layers by convolutional layers with large recep-

tive fields and adds up-sampling layers based on simple interpolation filters. The U-Net [18] is another segmentation network especially designed for segmentation of biomedical images (around 75 layers). The architecture consists of "a contracting path to capture context" and a "symmetric expanding path that enables precise localization". Finally, the DeepLabV3 network [19] is used in experiments, using Resnet-18 encoder and applies some new techniques to improve the quality of segmentation, including Atrous Spatial Pyramid Pooling (ASPP) [20] that is implemented to better capture objects at multiple scales, and Conditional Random Fields (CRF) for improved localization of object boundaries using probabilistic graphical models. From the mentioned networks, best results are obtained with DeepLabV3 [21], which I hypothesize to be due both to the ASPP and CRF and to the use of Resnet-18 as the encoder CNN.

3. DCNN Segmentation Architectures

The DCNN (deep convolution neural network) networks used in segmentation are distinguished by different architectural choices and innovations. Our purpose is to choose the best possible architecture and to understand the influence of the architecture in the quality of the results. Next, the main characteristics and layers of the four architectures tested are summarized, and **Figure 4** shows a rough sketch of those architectures.

Simple: "Simple" is a basic encoder-decoder CNN architecture that was built from scratch to compare with deeper, more complex architectures. With only four convolution layers accompanied by max pooling and ReLU operations on each, and another four deconvolution layers accompanied by ReLU plus a softmax and pixel classification layer for output. The convolution layers apply 64 3 × 3 filters with stride [11], and the deconvolution layers apply 64 4 × 4 filters. **Table 1** summarizes the layers of "Simple".

Table 1. Layers of "Simple".

1 Image Input 2848 × 4288 × 3 images	13 Max Pooling 2 × 2 max pooling
2 Convolution 64 3 × 3 convolutions	14 Transposed Convolution 64 4 × 4
3 ReLU	15 ReLU
4 Max Pooling 2 × 2 max pooling	16 Transposed Convolution 64 4 × 4
5 Convolution 64 3 × 3 convolutions	17 ReLU
6 ReLU	18 Transposed Convolution 64 4 × 4
7 Max Pooling 2 × 2 max pooling	19 ReLU
8 Convolution 64 3 × 3 convolutions	20 Transposed Convolution 64 4 × 4
9 ReLU	21 ReLU
10 Max Pooling 2 × 2 max pooling	22 Convolution 5 1 × 1 convolutions
11 Convolution 64 3 × 3 convolutions	23 Softmax
12 ReLU	24 Pixel Classification (loss function)

Fully Convolution Network (FCN): the FCN, sketched in Figure 4(b) and whose layers are summarized in Table 2, uses a DCNN classification network for feature extraction or encoding (VGG-16 with 7 stages corresponding to 41 layers), plus a much smaller sequence of up-sampling layers (decoding stages) for a total network size of 51 layers. FCN also forwards feature maps (the pooled output of coding stage 4 is fused with an output of the first up-sampling layer, and the pooled output of coding stage 3 is fused with the output of the second up-sampling layer. Finally, the image input is also fused with the output of the third up-sampling layer, all this followed by the final pixel classification layer.

Segnet (and U-Net): Segnet shares a similar architecture with U-Net, with encoder stages and symmetric decoding stages. The Segnet uses VGG-16 for feature extraction (encoding), while the decoder up-samples using pooling indices computed in the max-pooling step of the corresponding encoder, to perform non-linear up-sampling (while the U-Net forwards cropped feature maps directly after ReLu regularization at each stage, which are concatenated with the corresponding stage outputs at the destination, Segnet forwards max-pooled outputs and unpools at the destination). Segnet can be configured with any number of stages, defined for this work 5 encoding layers and the symmetric decoding layers, for a total of 73 layers. Segnet (and U-Net) architecture is sketched in Figure 4(c).

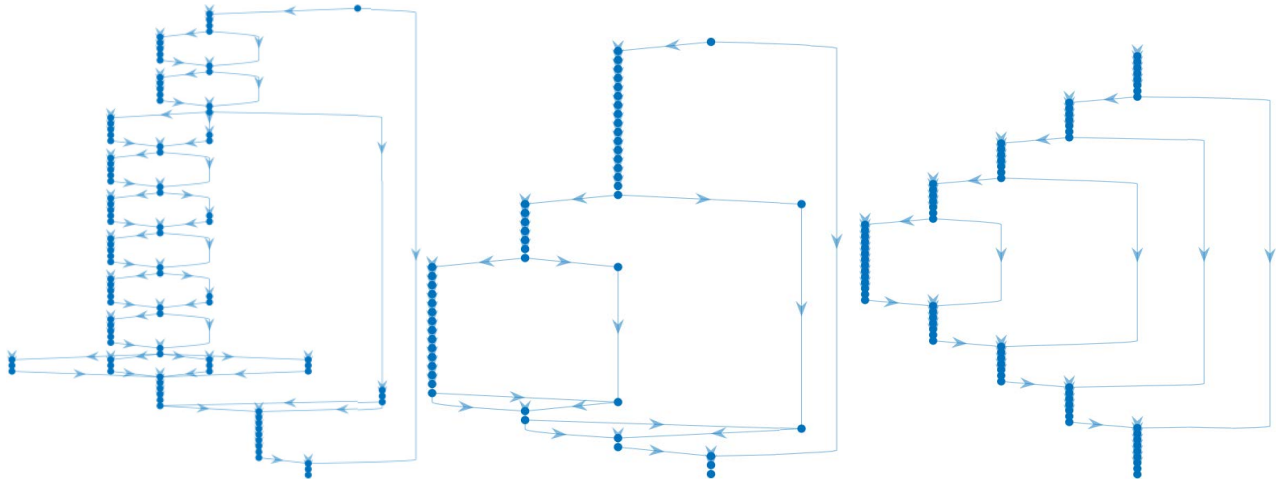
DeepLabV3: it is the deepest network tested in this work, with 100 layers listed in Table 3. The general layout of layers in DeepLabV3 is shown in Figure 4(a). DeepLabV3 uses Resnet-18 as feature extractor, with 8 stages totaling 71 layers, the remaining stages being ASPP plus the final stages. Forwarding connections are also added from encoding stages to the Atrous Spatial Pyramid

Table 2. Layers of "FCN".

1	Image Input	19	Conv 512 3 × 3 × 256 filters	37	ReLU + 50% dropout
2	Conv 64 3 × 3 × 3 filters	20	ReLU	39	Conv 5 1 × 1 filters
3	ReLU	21	Conv 512 3 × 3 × 512 filters	40	Transposed Conv 5 4 × 4 × 5
4	Conv 64 3 × 3 × 64 filters	22	ReLU	41	Addition
5	ReLU + max pool 2 × 2	23	Conv 512 3 × 3 × 512 filters	42	Transposed Conv 5 4 × 4 × 5
7	Conv 128 3 × 3 × 64 filters	24	ReLU + max pool 2 × 2	43	Addition
8	ReLU	26	Conv 512 3 × 3 × 512 filters	44	Transposed Conv 5 16 × 16 × 5
9	Conv 128 3 × 3 × 128 filters	27	ReLU	45	Crop 2D
10	ReLU + max pool 2 × 2	28	Conv 512 3 × 3 × 512 filters	46	Softmax
12	Conv 256 3 × 3 × 128 filters	29	ReLU	47	Pixel Classification Layer
13	ReLU	30	Conv 512 3 × 3 × 512 filters	48	Conv 5 1 × 1 filters
14	Conv 256 3 × 3 × 256 filters	31	ReLU + max pool 2 × 2	49	Crop 2D
15	ReLU	33	Conv 4096 7 × 7 × 512 filters	50	Conv 5 1 × 1 filters
16	Conv 256 3 × 3 × 256 filters	34	ReLU + 50% dropout	51	Crop 2D
17	ReLU + max pool 2 × 2	36	Conv 4096 1 × 1 × 4096 filters		

Table 3. Layers of "DeepLabV3".

1	Image Input	37	Batch Normalization + ReLU	72	Conv 256 3 × 3 filters
2	Conv 64 7 × 7 filters	39	Conv 256 3 × 3 filters	73	Batch Normalization + ReLU
3	Batch Norm + ReLU + maxPool [3 × 3]	40	Batch Normalization + Residue addition + ReLU	75	Conv 256 3 × 3 filters
6	Conv 64 3 × 3 filters	43	Conv 256 1 × 1 filters	76	Batch Normalization + ReLU
7	Batch Normalization + ReLU	44	Batch Normalization	78	Conv 256 3 × 3 filters
9	Conv 64 3 × 3 filters	45	Conv 256 3 × 3 filters	79	Batch Normalization + ReLU
10	Batch Norm + Res add + ReLU	46	Batch Normalization + ReLU	81	Conv 256 1 × 1 filters
13	Conv 64 3 × 3 filters	48	Conv 256 3 × 3 filters	82	Batch Normalization + ReLU
14	Batch Norm + ReLU	49	Batch Normalization + Residue addition + ReLU	84	Transposed Conv 256 8 × 8 filters
16	Conv 64 3 × 3 filters	52	Conv 512 3 × 3 filters	85	Crop 2D
17	Batch Norm + Res add + ReLU	53	Batch Normalization + ReLU	86	Conv 48 1 × 1 filters
20	Conv 128 3 × 3 filters	55	Conv 512 3 × 3 filters	87	Batch Normalization + ReLU
21	Batch Norm + ReLU	56	Batch Norm + addition + ReLU	89	Depth concatenation
23	Conv 128 3 × 3 filters	59	Conv 512 1 × 1 filters	90	Conv 256 3 × 3 filters
24	Batch Norm + Res add + ReLU	60	Batch Normalization	91	Batch Normalization + ReLU
27	Conv 128 1 × 1 filters	61	Conv 512 3 × 3 filters	93	Conv 256 3 × 3 filters
28	Batch Normalization	62	Batch Normalization + ReLU	94	Batch Normalization + ReLU
29	Conv 128 3 × 3 filters	64	Conv 512 3 × 3 filters	96	Conv 5 1 × 1 filters
30	Batch Norm + ReLU	65	Batch Norm. + addition + ReLU	97	Transposed Convolution 5 8 × 8 filters
32	Conv 128 3 × 3 filters	68	Depth concatenation	98	Crop 2D
33	Batch Norm + Res add + ReLU	69	Conv 256 1 × 1 filters	99	Softmax
36	Conv 256 3 × 3 filters	70	Batch Normalization + ReLU	100	Pixel Classification



(a) (b) (c)

Figure 4. Rough sketch of network architectures. (a) DeepLabV3; (b) FCN layers; (c) Segnet layers.

Pooling (ASPP) layers, for enhanced segmenting of objects at multiple scales. The outputs of the final DCNN layer are combined with a fully connected Conditional Random Field (CRF) for improved localization of object boundaries using mechanisms from probabilistic graphical models.

4. Experimental Setup

In experimental work two datasets are used, one to compare between segmentation networks that are very useful to evaluate semantic segmentation quality pixel-wise, the other one to compare with prior work because most prior works use it and are evaluated based on the connected components model (overlap between found segments and groundtruth segments). The first dataset is IDRID [6], a dataset that is publicly available for the study of automated detection of Diabetic Retinopathy and segmentation of characteristic lesions. It has groundtruth labelled data for each of 83 Eye Fundus Images (EFI), where most images have a large number of instances of each specific lesion, and the groundtruths represent the class that should be assigned to each individual pixel. IDRID contains the pixel groundtruths for micro-aneurisms, hemorrhages, exudates (hard and soft) and the optic disk. The equipment used to acquire the images was a Kowa VX-10 alpha digital fundus camera with 50-degree field of view (FOV), centered near the macula. Image resolution was 4288 × 2848, saved as jpg. Experts validated the quality of the images and their clinical relevance. For our work the dataset was divided randomly 80%/20% train/test images (and groundtruths), with cross-validation on five folds.

The second dataset, DIARET-DB1, consists of 89 color fundus photographs collected at the Kuopio University Hospital, in Finland [7]. Images were captured with the same fundus camera, a ZEISS FF450plus digital camera with a 50-degree field-of-view. Images all have a definition of 1500 × 1152 pixels. Independent markings were obtained for each image from four medical experts. The experts were asked to manually delineate the areas containing microaneurysms (or "small red dots"), hemorrhages, hard exudates and cotton wool spots (or "soft exudates") and to report their confidence (<50%, ≥50%, 100%) for each segmented lesion. Based on these annotations, only five images in the dataset are considered normal: none of the experts suspect these images to contain any lesions. The DIARET-DB1 dataset was randomly divided into 80% train and 20% test to ensure train/test independence, cross-validated with 5 folds as well.

The deep learning segmentation networks were all trained using the SGDM learning optimization function with learning rate 0.005. This was decided after preliminary tests to verify that the networks would converge to classify all lesions correctly. The loss function used was the default cross-entropy with class balancing added to the last pixel classification layer in order to counter the natural class imbalance that exists in this EFI context. The networks were trained for 300 epochs, after initial tests in which it was observed that they would stabilize much before that number of epochs. The minibatch size was configured to

32, and the momentum was 0.9. I also experimented with data augmentation to improve the robustness and quality of the network predictions. These consisted with the random transformations (rotations, small translations and scaling) and adding the two datasets. In terms of hardware, a machine running windows 10 was used. The hardware was an intel i5, 3.4 GHz, 16 GB of RAM 1TB SSD disk. A GPU was added to the PC, consisting of an NVIDEA GForce GTX 1070 GPU (the GTX 1070 has a Pascal architecture and 1920 cores, 8 GB GDDR5, with memory speed of 8 Gbps).

In terms of reporting the results of segmentation networks on IDRID, two scenarios are distinguished, one without data augmentation and one with data augmentation, defined in Table 4.

The same dataset, metrics and results that were used in prior works are also used in the comparison of the segmentation networks to those prior works, and the segmentation network with the best performance was chosen for those comparisons (deepLabV3). The main metric used for the comparison is sensitivity vs false positives per image (FPI), plus image-level sensitivities to compare lesion detection at image level (I also show the ROC curve for the segmentation network). Prior works compared used DIARET-DB1 [7].

For comparing the various segmentation networks, semantic segmentation evaluation metrics and the dataset IDRID [6] are used, which contains per-pixel classification groundtruths. In that case, the main metrics used are, in order of importance for the characterization of segmentation performance, are mean IoU, mean BFScore and then mean accuracy, weighted IoU and global accuracy. Metrics using the mean take the mean value over all classes (classes are each lesion plus the background), while weighted and global metrics take the average over all pixels directly. The values over all pixels are less informative than mean over classes because more than 90% of all pixels are background, therefore they characterize the quality of segmentation of the background mostly.

Training evolution curves

Figure 5 briefly shows plots of validation accuracy curves when training the IDRID images. It is apparent that Simple has more difficulties to converge to high accuracy, DeepLabV3 converged well and with fewer fluctuations than FCN.

5. Results

1) *Comparing segmentation network to prior work*

Table 5 compares lesion-level sensitivities between the segmentation network and the prior work that actually segments the lesions, *i.e.* [8] [9] and [14]. These

Table 4. Training options.

No DA	Fixed learn rate, no data augmentation, IDRID data
DA	Data augmentation, learn rate decay (85% every 25 epochs), IDRID plus DIARET-DB1 data

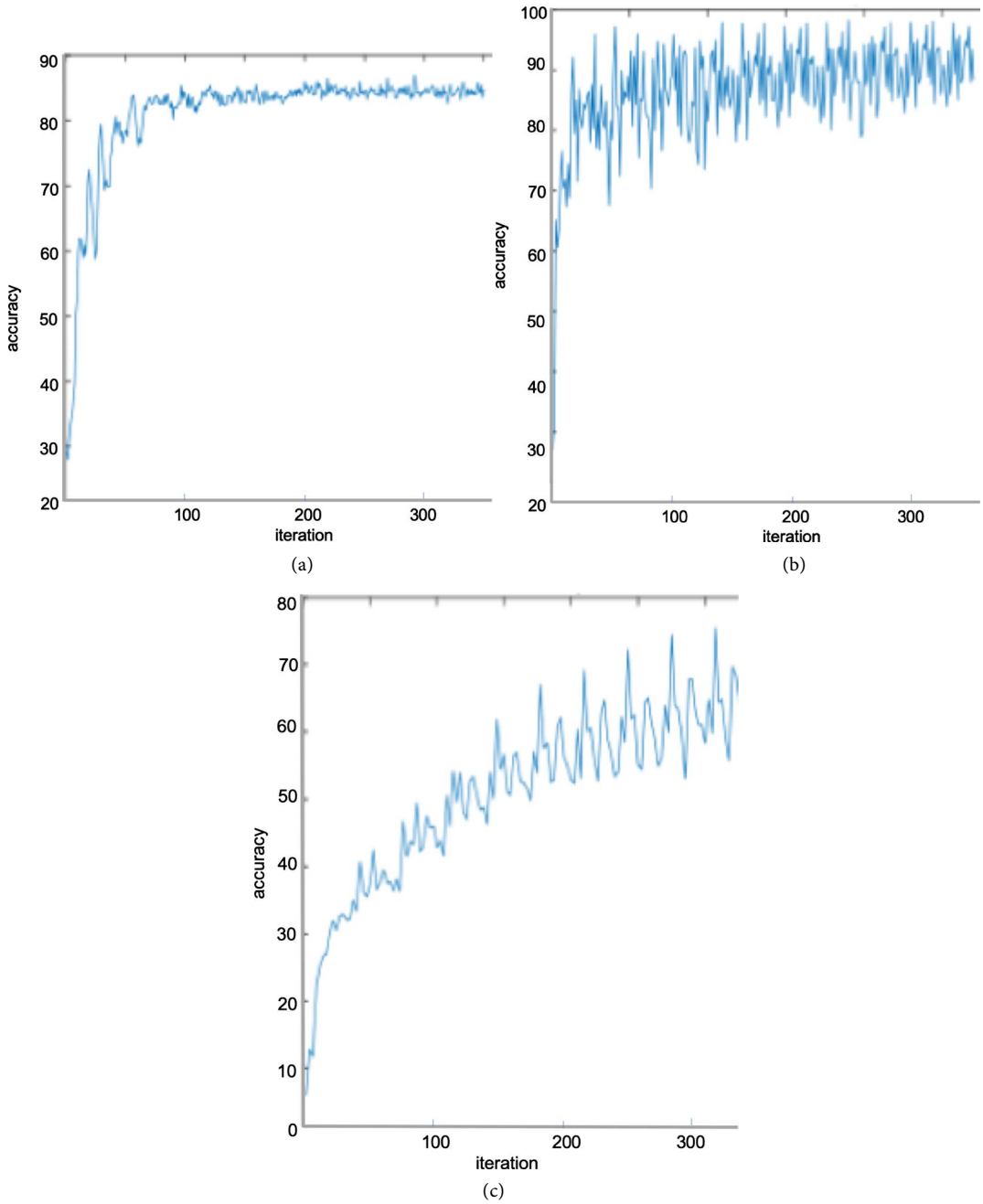


Figure 5. Training evolution. (a) Training DeepLabV3; (b) Training FCN; (c) Training Simple.

Table 5. Comparing lesion-level sensitivities.

Method	Hemorrhages		Hard Exudates		Soft Exudates		RSD (micro-aneurisms)	
	SE%	FPs/I	SE%	FPs/I	SE%	FPs/I	SE%	FPs/I
Quellec *et al.* [8]	71	10	80	10	90	10	61	10
Gondal *et al.* [9]	72	2.25	47	1.9	71	1.45	21	2
Orlando [14]	50	1					50	1
Segm. network DeepLabV3	**87**	10	**94**	2.76	**87.5**	3.92	48	6.4

results were obtained using the connected components model of evaluation (described for instance in [4]) with similar conditions as used in the compared works. The sensitivities are measured against the number of false positives per image (FPI), and both should be considered in the analysis of results. Note that we used the approach in [9], where FPIs can differ because they are obtained against the class classification threshold (0 to 1). Table 6 compares image-level detection of lesions for referral, where one can see again that DeepLabV3 ranks first in HA and SE and also ranks well in HE and MA when compared with the alternatives tested. Table 7 compares the segmentation network sensitivities to those of the small-window classifiers reviewed in the related work section. To finalize this first experiment Figure 6 shows the ROC curves of each lesion using DeepLabV3 semantic segmentation network.

2) Comparison between segmentation networks, improvements and limitations

Table 8 shows the comparison between different architectures of the segmentation DCNN networks using IDRiD dataset, 5-fold cross validation (80%/20%). These results were obtained without data augmentation. Table 9 shows the comparison of results with two scenarios: an initial scenario with no data augmentation and a second scenario with data augmentation (DA) that improves the results in general. Table 10 shows the quality of segmentation of each lesion measured as IoU and sensitivity.

3) ***Example visualizations of segmentations***

Figure 7 shows an example visualization of segmentation network results in IDRID, with the groundtruth on the left and the segmentation output on the right. Figure 8 shows an example for the dataset DIRET-DB1, with the labels

Table 6. Image-level sensitivities.

Method	HA	HE	SE	MA
Zhou *et al.* [22]	94.4	-	-	-
Liu *et al.* [23]	-	83	83	-
Haloi *et al.* [15]	-	**96.5**	-	-
Mane *et al.* [24]	-	-	-	**96.4**
Gondal [9]	97.2	93.3	81.8	50
Ours (DeepLabV3)	**100**	90	**87.5**	71

Table 7. Comparison with small-window lesion classifiers.

Work	Target lesion	sensitivity	Segmentation network sensitivity
Haloi [15]	Micro-aneurisms	50% (1 FPI), 70% (10 FPI)	48% (6.8 FPI)
Prentasic [10]	exudates	77% ($\sigma = 0.2$)	88 to 94% (3 FPI)
Van Grinsen [12]	hemorrhages	79% (1 FPI, 8 px center) 89% (10 FPI)	87% (10 FPI)

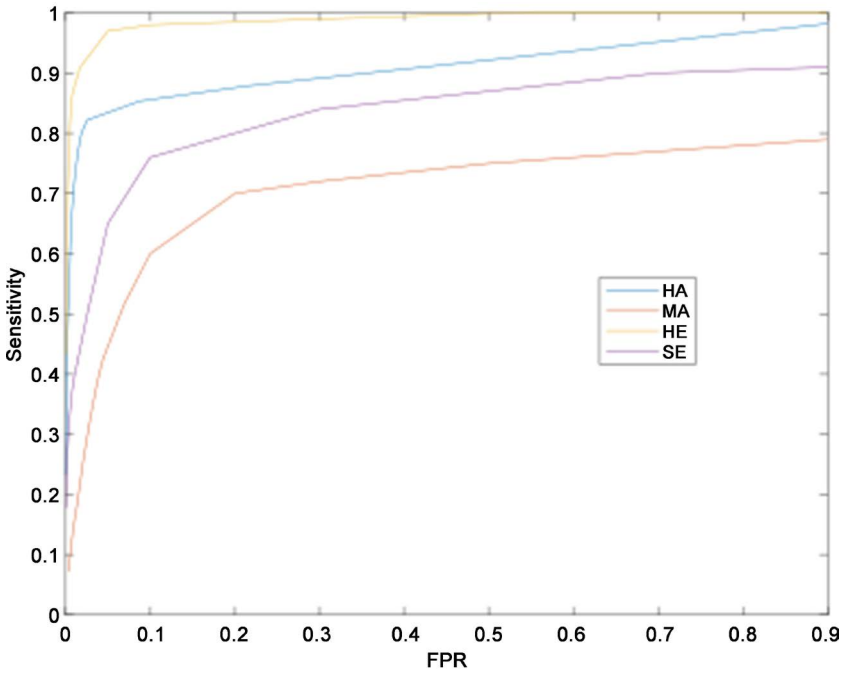


Figure 6. Lesions ROC (sens/FPR).

Table 8. Comparison between network architectures (basic scenario with no DA).

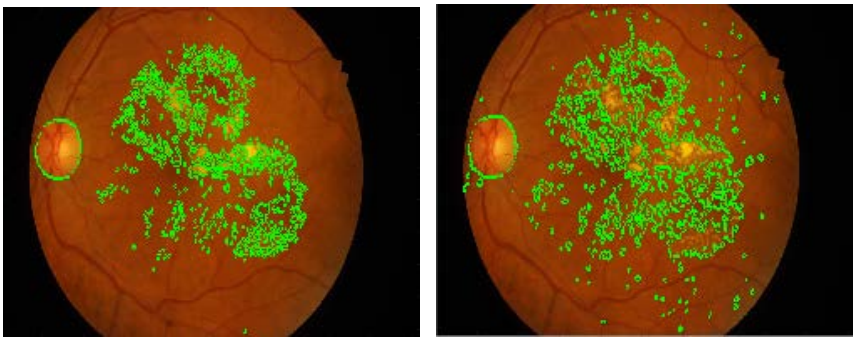
Method	Global Accuracy	Mean Accuracy	Weighted IoU	Mean IoU	Mean BF Score
FCN	89.5	74.6	87.5	37.9	48.5
DeepLabV3	81.2	84.1	78.5	32.8	33.6
U-Net	58.7	59.8	56.2	16.1	19.6
Segnet	52.7	45.4	50.2	14.2	17.5
Simple	49.0	54.6	46.4	11.6	19.1

Table 9. Improvement using data augmentation and tuning.

Network	Setup	Global Accuracy	Mean Accuracy	Weighted IoU	Mean IoU	Mean BF Score
deepLabV3	no DA	81	84	79	33	34
	DA	96	83	94	52	69
FCN	no DA	90	75	88	38	49
	DA	89	79	87	41	52

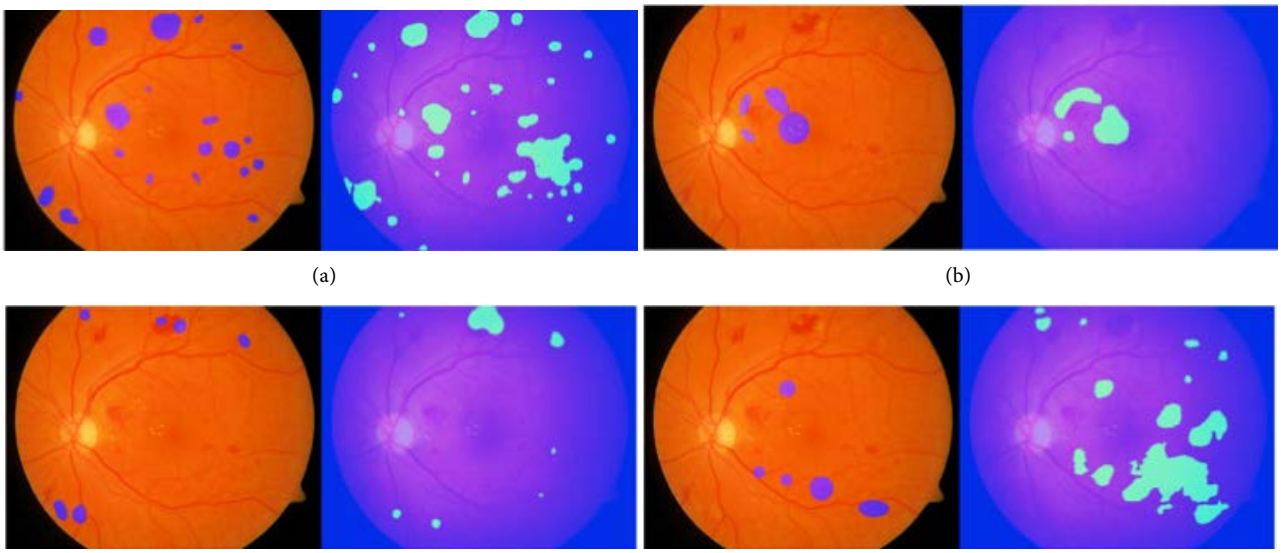
Table 10. Sensitivity and IoU in DIARET.

DIARET scores	sens = recall	IoU
BK	0.99	0.99
MA	**0.31**	**0.096**
HA	0.74	**0.33**
HE	0.78	**0.48**
SE	0.555	**0.287**



(a) (b)

Figure 7. Example visualization of test image on IDRID.



(a) (b)

(c) (d)

Figure 8. Example visualization of test image on DIARET-DB1. (a) HA labels and HA segmentation output; (b) HE and HE segmentation outputs; (c) SE and SE segmentation outputs; (d) MA and MA segmentation outputs.

and segmentation outputs for each class (MA, HA, HE, SE).

6. Discussion of Results

1) Comparing segmentation network to prior work

The comparison of lesion-level sensitivities between the segmentation network and prior work, in particular Quellec *et al.* [8], Gondal *et al* [9] and Orlando *et al.* [14] (Table 5) reveals that the segmentation network improves the sensitivity for most lesions, *i.e.* for HA, HE and SE, while for MA the results seem worse than [14] but in line with the results of the remaining prior works compared to. The prior works scored 50% to 72% in HA, 47% to 80% in HE, 71 to 90% in SE and 21% to 61% in MA, the segmentation network scored 87% in HA, 94% in HE, 87.5% in SE and 48% in MA. This means that the deep segmentation network with around 100 layers, Resnet-18 as encoder network and the innovations we discussed previously that include ASPP and CRF proved better than prior works for segmentation of lesions. Note also that segmentation net-

works are most often evaluated using Jaccard Index (JI), also known as intersect-over-the-union (IoU), or the Dice coefficient, and we include, at the end of the experimental section, an evaluation using IoU, but in this experiment we compared using exactly the same conditions and evaluation approaches that the prior works use.

The segmentation network was also better in terms of image-level detection of lesions for referral (**Table 6**), where we can see again that DeepLabV3 ranks first in HA and SE and also ranks well in HE and MA when compared with the alternatives compared (Zhou *et al.* [22], Liu *et al.* [23], Haloi *et al.* [15], Mane *et al.* [24], Gondal *et al.* [9]). The segmentation network achieved 100% HA, 90% HE, 87.5% SE and 71% MA, against 94% to 97% HA, 83% to 96.5% HE, 81 to 83% SE and 50% to 96.4% MA for the related work compared.

Table 7 compares the segmentation network sensitivities to some of the small-window classifiers reviewed in the related work section. The results can be interpreted as revealing that the segmentation network achieves worse sensitivity than Haloi [15] on MA lesions, but better sensitivity than Prentasic [10] on Exudates and a comparable sensitivity to van Grinsen [12] on HA. But, as we discussed in related work section, except for Haloi [15], but, as we discussed in related work section, except for Haloi [15], the remaining small window classifiers are only classifying lesions centered in small windows, and their evaluation is based on lesions centered and picked manually based on the groundtruth information.

The remaining small window classifiers are only classifying lesions centered in small windows, and their evaluation is based on lesions centered and picked manually based on the groundtruth information. In contrast, a segmentation algorithm inputs and processes the whole images to find and extract contours and, since the locations of the lesions are unknown.

The ROC curves in **Figure 6** complement the information for the segmentation network. It shows that MA is the hardest lesion to segment, with sensitivity around 58% for FPR 0.1, followed by SE (73% FPR 0.1), HA (85% FPR 0.1), and HE (97% FPR 0.1).

2) ***Comparison between segmentation networks, improvements and limitations***

For the remaining analysis we switched dataset (to IDRID) and evaluation approach, to consider semantic segmentation instead of overlap of coarse segments in the evaluation of the segmentation networks. Using this approach, **Table 8** shows that, without any training optimizations, DeepLabV3 [19] and FCN [16] compare favorably with the other segmentation network architectures that we described in the setup (UNET [18], SegNet and Simple) for the task of segmentation of eye fundus lesions. For instance, in terms of weighted IoU, FCN scored 87.5% and DeepLabv3 scored 78.5%, while the remaining ones scored between 46.5% and 56.2%. These results were obtained without data augmentation. **Table 9** shows the effects of data augmentation. Looking in particular at

weighted IoU, DeepLabv3 with DA improved from 78.5 to 94%, while FCN did not improve.

Even though the segmentation networks outperform prior work, Table 9 and Table 10 also show that there is still a long way to go in terms of optimizing semantic segmentation quality. DeepLabV3 achieves 94% weigthed IoU, but its mean IoU shown in Table 9 is 52%, and the quality of segmentation of each lesion (Table 10) is even lower for some lesions (MA and SE). We noticed that, while background, HA and HE have high scores (e.g. sensitivity of 74% to 99%), MA and SE have much lower scores (31% and 55% respectively). Additionally, significant difference between IoU and sensitivity signals a significant amount of false positives (e.g. background marked as lesions). These results show that segmentation networks could still benefit from further research in future to better deal with FP and also FN.

3) ***Visualizations and Conclusions from Experiments***

The visualizations shown in Figure 7 for IDRID and in Figure 8 for DIRET-DB1 help illustrate the capacity of the segmentation network, since we see that the lesions are reasonably well recognized in those figures, but they also show that there are still many false positives.

Our experimental results succeeded to show that the deep segmentation network improves the quality of lesions segmentation from Eye Fundus Image (EFI) when compared with prior work that also segments those lesions. We have also compared segmentation network architectures, showing that the simplest architectures (e.g. Simple, with 5 simple encoder and decoder layers) were unable to reach the segmentation performance of deeper and more complex networks. Our best results were achieved with DeepLabV3, a deep segmentation network with 100 layers using Resnet-18 as encoder CNN and integrating important innovations that include atrous spatial pooling (ASPP) and conditional random fields (CRF). FCN, another well-known deep segmentation network using VGG-16 also obtained quite reasonable results. Finally, we also revealed some limitations of current state-of-the-art in segmentation of lesions.

7. Conclusion

In this work, we have shown that most prior works on segmentation of diabetic retinopathy lesions only detect lesions approximately and do not segment them. We proposed the use of deep semantic segmentation networks to segment the lesions. Our main objective was twofold: to compare the segmentation networks solution with prior works, showing that they improve performance, and to compare between segmentation network architectures, to investigate whether the characteristics of those networks were important to improve the performance, and to see their limitations. We found out that the architecture is very important, achieving top performance the DeepLabV3 design with Resnet-18 as the encoder. Comparing with prior work segmenting EFI lesions, the advantage of the top-performing deep segmentation network reached 10 to 20 percentage

points in HA, HE and SE lesions, although it was slightly worse than some of the others segmenting micro-aneurisms. But we also revealed the limitations of current best performing networks. Future work needs to deal with the still existing imperfections, including many false positives.

Acknowledgements

We would like to acknowledge for the availability of the datasets [6] and [7] that were used in our experimental work. Other references to these datasets include [25] [26] [27] and [28].

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Qureshi, I., Ma, J. and Abbas, Q. (2019) Recent Development on Detection Methods for the Diagnosis of Diabetic Retinopathy. *Symmetry*, **11**, 749. https://doi.org/10.3390/sym11060749

[2] Asiri, N., Hussain, M., Al Adel, F. and Alzaidi, N. (2019) Deep Learning Based Computer-Aided Diagnosis Systems for Diabetic Retinopathy: A Survey. *Artificial Intelligence in Medicine*, **99**, Article ID: 101701. https://doi.org/10.1016/j.artmed.2019.07.009

[3] Raman, R., Srinivasan, S., Virmani, S., Sivaprasad, S., Rao, C. and Rajalakshmi, R. (2019) Fundus Photograph-Based Deep Learning Algorithms in Detecting Diabetic Retinopathy. *Eye*, **33**, 97-109. https://doi.org/10.1038/s41433-018-0269-y

[4] Zhang, X., Thibault, G., Decencière, E., Marcotegui, B., Laÿ, B., Danno, R., Chabouis, A., *et al.* (2014) Exudate Detection in Color Retinal Images for Mass Screening of Diabetic Retinopathy. *Medical Image Analysis*, **18**, 1026-1043. https://doi.org/10.1016/j.media.2014.05.004

[5] Wilkinson, C., Ferris III, F.L., Klein, R.E., *et al.* (2003) Proposed International Clinical Diabetic Retinopathy and Diabetic Macular Edema Disease Severity Scales. *Ophthalmology*, **110**, 1677-1682. https://doi.org/10.1016/S0161-6420(03)00475-5

[6] Porwal, P., Meriaudeau, F., *et al.* (2019) Indian Diabetic Retinopathy Image Dataset (IDRID). IEEE Dataport.

[7] Kauppi, T., Kalesnykiene, V., Kamarainen, J.-K., Lensu, L., Sorri, I., Raninen, A., Voutilainen, R., Pietila, J., Kalviainen, H. and Uusitalo, H. (2007) The DIARETDB1 Diabetic Retinopathy Database and Evaluation Protocol. *Proceedings of the British Machine Vision Conference* 2007, University of Warwick, UK, 10-13 September 2007, 15.1-15.10. https://doi.org/10.5244/C.21.15

[8] Gondal, W.M., Köhler, J.M., Grzeszick, R., Fink, G.A. and Hirsch, M. (2017) September) Weakly-Supervised Localization of Diabetic Retinopathy Lesions in Retinal Fundus Images. 2017 *IEEE International Conference on Image Processing* (*ICIP*), Beijing, 17-20 September 2017, 2069-2073. https://doi.org/10.1109/ICIP.2017.8296646

[9] Quellec, G., Charrière, K., Boudi, Y., Cochener, B. and Lamard, M. (2017) Deep Image Mining for Diabetic Retinopathy Screening. *Medical Image Analysis*, **39**, 178-193. https://doi.org/10.1016/j.media.2017.04.012

[10] Prentašić, P. and Lončarić, S. (2015) Detection of Exudates in Fundus Photographs Using Convolutional Neural Networks. 2015 9*th International Symposium on Image and Signal Processing and Analysis* (*ISPA*), Zagreb, 7-9 September 2015, 188-192. https://doi.org/10.1109/ISPA.2015.7306056

[11] Haloi, M. (2015) Improved Microaneurysm Detection Using Deep Neural Networks.

[12] Van Grinsven, M.J., van Ginneken, B., Hoyng, C.B., Theelen, T. and Sánchez, C.I. (2016) Fast Convolutional Neural Network Training Using Selective Data Sampling: Application to Hemorrhage Detection in Color Fundus Images. *IEEE Transactions on Medical Imaging*, **35**, 1273-1284. https://doi.org/10.1109/TMI.2016.2526689

[13] Shan, J. and Li, L. (2016) A Deep Learning Method for Microaneurysm Detection in Fundus Images. 2016 *IEEE First International Conference on Connected Health: Applications*, *Systems and Engineering Technologies* (*CHASE*), Washington DC, 27-29 June 2016, 357-358. https://doi.org/10.1109/CHASE.2016.12

[14] Orlando, J.I., Prokofyeva, E., del Fresno, M. and Blaschko, M.B. (2018) An Ensemble Deep Learning Based Approach for Red Lesion Detection in Fundus Images. *Computer Methods and Programs in Biomedicine*, **153**, 115-127. https://doi.org/10.1016/j.cmpb.2017.10.017

[15] Haloi, M., Dandapat, S. and Sinha, R. (2015) A Gaussian Scale Space Approach for Exudates Detection, Classification and Severity Prediction.

[16] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

[17] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition.

[18] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-net: Convolutional Networks for Biomedical Image Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[19] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L. (2017) Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

[20] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017) Feature Pyramid Networks for Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 2117-2125. https://doi.org/10.1109/CVPR.2017.106

[21] Furtado, P. (2020) Deep Semantic Segmentation of Diabetic Retinopathy Lesions: What Metrics Really Tell Us. In: *Medical Imaging* 2020*: Biomedical Applications in Molecular*, *Structural*, *and Functional Imaging*, Vol. 11317, International Society for Optics and Photonics, Bellingham, 113170. https://doi.org/10.1117/12.2549221

[22] Zhou, L., Li, P., Yu, Q., Qiao, Y. and Yang, J. (2016) Automatic Hemorrhage Detection in Color Fundus Images Based on Gradual Removal of Vascular Branches. *IEEE International Conference on Image Processing* (*ICIP*), Phoenix, 25-28 September 2016, 399-403. https://doi.org/10.1109/ICIP.2016.7532387

[23] Liu, Q., Zou, B., Chen, J., Ke, W., Yue, K., Chen, Z. and Zhao, G. (2017) A Location-to-Segmentation Strategy for Automatic Exudate Segmentation in Colour Retinal Fundus Images. *Computerized Medical Imaging and Graphics*, **55**, 78-86.

https://doi.org/10.1016/j.compmedimag.2016.09.001

[24] Mane, V.M., Kawadiwale, R.B. and Jadhav, D. (2015) Detection of Red Lesions in Diabetic Retinopathy Affected Fundus Images. *IEEE International Advance Computing Conference* (*IACC*), Banglore, 12-13 June 2015, 56-60. https://doi.org/10.1109/IADCC.2015.7154668

[25] Porwal, P., *et al*. (2018) Indian Diabetic Retinopathy Image Dataset (IDRiD). IEEE Dataport.

[26] Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V. and Meriaudeau, F. (2018) Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research. *Data*, **3**, 25. http://www.mdpi.com/2306-5729/3/3/25 https://doi.org/10.3390/data3030025

[27] Porwal, P., Pachade, S., Kokare, M., *et al*. (2020) IDRiD: Diabetic Retinopathy-Segmentation and Grading Challenge. *Medical Image Analysis*, **59**, Article ID: 101561.

[28] DIARETDB1—Standard Diabetic Retinopathy Database. http://www2.it.lut.fi/project/imageret/diaretdb1