Computerized Medical Imaging and Graphics 76 (2019) 101636

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Artery–vein segmentation in fundus images using a fully convolutional network

Ruben Hemelings[a,b,d], Bart Elen[d], Ingeborg Stalmans[a], Karel Van Keer[a], Patrick De Boever[c,d,*], Matthew B. Blaschko[b]

[a] Research Group Ophthalmology, KU Leuven, Kapucijnenvoer 33, 3000 Leuven, Belgium
[b] ESAT-PSI, KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
[c] Hasselt University, Agoralaan building D, 3590 Diepenbeek, Belgium
[d] VITO NV, Boeretang 200, 2400 Mol, Belgium

ARTICLE INFO

Article history:
Received 29 March 2018
Received in revised form 18 May 2019
Accepted 24 May 2019

Keywords:
Fundus image
Fully convolutional network
Artery–vein segmentation

ABSTRACT

Epidemiological studies demonstrate that dimensions of retinal vessels change with ocular diseases, coronary heart disease and stroke. Different metrics have been described to quantify these changes in fundus images, with arteriolar and venular calibers among the most widely used. The analysis often includes a manual procedure during which a trained grader differentiates between arterioles and venules. This step can be time-consuming and can introduce variability, especially when large volumes of images need to be analyzed. In light of the recent successes of fully convolutional networks (FCNs) applied to biomedical image segmentation, we assess its potential in the context of retinal artery–vein (A/V) discrimination. To the best of our knowledge, a deep learning (DL) architecture for simultaneous vessel extraction and A/V discrimination has not been previously employed. With the aim of improving the automation of vessel analysis, a novel application of the U-Net semantic segmentation architecture (based on FCNs) on the discrimination of arteries and veins in fundus images is presented. By utilizing DL, results are obtained that exceed accuracies reported in the literature. Our model was trained and tested on the public DRIVE and HRF datasets. For DRIVE, measuring performance on vessels wider than two pixels, the FCN achieved accuracies of 94.42% and 94.11% on arteries and veins, respectively. This represents a decrease in error of 25% over the previous state of the art reported by Xu et al. (2017). Additionally, we introduce the HRF A/V ground truth, on which our model achieves 96.98% accuracy on all discovered centerline pixels. HRF A/V ground truth validated by an ophthalmologist, predicted A/V annotations and evaluation code are available at https://github.com/rubenhx/av-segmentation.

1. Introduction

The retina is used to diagnose ocular diseases such as glaucoma and diabetic retinopathy, two of the leading causes of preventable, but incurable blindness (Prokofyeva and Zrenner, 2012). Glaucoma is characterized by a progressive cupping of the optic disc that can be subtle in early stages of the disease. Early stages of diabetic retinopathy are mainly associated with the appearance of exudates, retinal hemorrhages and cotton wool spots. During their development, these diseases are also associated with morphological and functional changes in the small blood vessels that can be observed in retinal fundus images. In the case of diabetic retinopathy, high blood sugar levels have weakened retinal blood vessel walls inducing changes in vessel dimensions such as caliber and tortuosity (Sasongko et al., 2011). In addition, changes have been recorded in the retinal blood flow (Bursell et al., 1996). The vascular theory of glaucoma suggests that retinal ganglion cells of the optical nerve are lost as a consequence of insufficient blood supply. To support this hypothesis, a higher venous oxygen saturation (Vandewalle et al., 2014) and a lower arteriovenous oxygen difference were observed by Ramm et al. (2014), resulting in a widened venous diameter. Other works revealed that the average retinal arteriole diameter might decrease in a case of glaucoma (Pekel and Pekel, 2016).

The retinal vasculature shares physiological, morphological and embryological characteristics with the cerebrovascular and coronary bed. Epidemiological research has pointed out that retinal blood vessel diameters are associated with cardiovascular risk fac-

[*] Corresponding author.
E-mail address: patrick.deboever@vito.be (P. De Boever).

https://doi.org/10.1016/j.compmedimag.2019.05.004

tors and cardiovascular morbidity (Tedeschi-Reiner et al., 2005; Wong et al., 2006, 2002). The structure and function of the retinal vessels is affected by the cardiovascular disease process, making them valuable candidate biomarkers to track the progression of cardiovascular diseases. Vessel changes in retinal fundus images are associated with increased blood pressure and hypertension. A meta-analysis found that retinal arteriolar narrowing and venular widening were independently associated with an increased risk of hypertension. Each 20 μm narrower arterioles at baseline was associated with a 1.12 mmHg greater increase in systolic blood pressure over 5 years (Ding et al., 2014). Wong and colleagues showed in the Atherosclerosis Risk in Communities (ARIC) prospective study that 3-year cumulative incidence of stroke was in association with retinal microvascular abnormalities, including exudates, retinal hemorrhages and vessel caliber changes (Wong et al., 2001). When performing a pooled analysis on prospective stroke studies, it was found that mainly the venular microvasculature was the dynamic compound responsive to changes in circulatory flow. A hazard ratio of 1.15 stroke events per 20 μm change in retinal vessel caliber was observed when adjusting for major confounding factors (McGeechan et al., 2009). A follow-up study in ARIC confirmed that narrower arterioles and wider venules conferred long-term risk of ischemic stroke. The inclusion of retinal vessel caliber in the risk model reclassified 21% of low-risk women as intermediate-risk (higher risk) (Seidelmann et al., 2016). More recently, the inclusion of information about retinopathy and retinal vessel calibers in a prediction model, already containing established risk factors, improved discrimination and overall reclassification of cardiovascular disease risk for diabetes patients. In this study, a higher cardiovascular disease risk was observed with narrower arteriolar calibers and wider venular calibers (Ho et al., 2017). Microvascular reactivity, measured as percentage of arteriolar dilatation after flicker light stimulus, was inversely associated with fasting blood glucose in diabetes patients (Sörensen et al., 2017). An association of vessel diameters with cause-specific mortality was found during a follow-up period of 25 years in the Rotterdam study (Mutlu et al., 2016). The latter authors concluded that there exists an opportunity to use retinal vascular imaging in clinical settings to help to identify high-risk patients of future cardiovascular events. However, until now, this type of information is not routinely used in clinical decision making.

Several metrics have been introduced to quantify retinal vessel abnormalities for application in risk prediction models. One of the most commonly employed is the arteriovenous ratio (AVR). The AVR is a ratio of the average width of the arterioles with respect to the venules, consisting of the central retinal artery equivalent (CRAE) and the central retinal vein equivalent (CRVE) (Xu, 2012). The consensus is to calculate CRAE and CRVE based on the revised Parr-Hubbard formulas in a region between two concentric circles (0.5 and 1 optic disc diameter) around the optic disc, also known as Zone B (Fig. 6) (Knudtson et al., 2003). In order to obtain CRAE, CRVE and derivative AVR, a trained grader needs to differentiate between arterioles and venules.

This work can be time-consuming; especially when high volumes of retinal images need to be analyzed in large population studies. An automated solution could speed up this analysis and remove variability introduced by the image interpretation done by different graders.

Since 2012, successful applications of convolutional neural networks (CNNs) to image classification problems have seen a significant growth, with popular architectures including AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015) and VGGNet (Simonyan and Zisserman, 2014) being used to obtain one global class probability per image. The related problem of semantic image segmentation is somewhat more complex, as the number of output probabilities is directly proportional to the number of pixels in an image. Upsampling the final downsampled image representation of a CNN to its original dimensions leads to a coarse map, hampering segmentation results. Fully convolutional networks (FCN) introduced in Long et al. (2014) are able to train end-to-end, with image information shared between the downsampling and upsampling path to improve segmentation output. U-Net (Ronneberger et al., 2015) was developed around the same time, and differs from an FCN in the decoder path. Whereas the FCN approach uses the skip connections to improve the segmentation output, U-Net uses them to improve the upsampled features in the decoder part of the network. Additionally, U-Net tackles the common data scarcity in biomedical imaging through the application of a random elastic deformation field prior to each training epoch. In this manner, the network is fed a unique training set at each epoch, and overfitting is alleviated. A 2D weight map is also implemented to counter class imbalance. The authors allocate the highest weight to the pixels bordering a minority class, to boost learning. Recently, such techniques based on CNNs have been successfully applied to the computer vision task of identifying and extracting the microvasculature from fundus images in an automated setting (Fraz et al., 2012b; González et al., 2010; Ishikawa et al., 2005; Türetken et al., 2011; Yedidya and Hartley, 2008). One step further is the discrimination of vessels into arteries and veins. Up until now, several works running on different architectures, notably graph theory and intensity-based features, have been successfully applied, achieving high classification accuracies (Xu et al., 2017; Estrada et al., 2015). A deep learning classification network, as a separate step to segmentation, was applied to this subproblem (Welikala et al., 2017), indicating the potential use of these novel architectures. The aim of the underlying work is to perform a segmentation between vessels and background, while simultaneously discriminating between arteries and veins in a unified framework, reporting significantly improved results over the previous state-of-the-art.

The main contributions of this work are:

- A novel application of fundus image segmentation based on deep learning that achieves A/V discrimination in an automated setting;
- Ablation study leading to insights for semantic segmentation applied to fundus images;
- Detailed benchmarking with previous work;
- Development of A/V ground truth for High Resolution Fundus Image Database. A/V annotations and evaluation code are available at https://github.com/rubenhx/av-segmentation.

An extended abstract of this work appeared in Hemelings et al. (2018).

The remainder of this paper is organized as follows: Section 2 reviews previous work on A/V discrimination in fundus images. Section 3 describes the data used, with Section 4 introducing our main methodology and optimal segmentation model. In Section 5, our experiments and results are given. Section 6 concludes.

2. Prior work

A large amount of research has been carried out in the fields of vessel segmentation (Fraz et al., 2012b; González et al., 2010; Ishikawa et al., 2005; Orlando et al., 2017; Türetken et al., 2011; Yedidya and Hartley, 2008) and vessel centerline extraction (Fraz et al., 2012a; Huang et al., 2012; Malek and Tourki, 2013; Sofka and Stewart, 2006). The classification of vessels into the two group groups of arteries and veins goes one step further. Performance benchmarking between different approaches for A/V segmentation is not evident, given that the majority of publications use proprietary data sets, or use incompatible evaluation metrics. Comparison using different data sets is not recommended, as there are many factors that influence label predictions, including image luminosity, resolution, and differences in the selection procedure of subjects to be imaged.

Several approaches have been introduced to discriminate between arterioles and venules in fundus images. One of the first approaches to the automated classification of vessels into arteries and veins was presented by Grisan and Ruggeri (2003). In that work, a particular region around the optic disc was divided into four quadrants, followed by color feature extraction using only the major vessels. After classification of vessels within these quadrants, a vessel tracking technique was employed to propagate the labels throughout the complete retinal vasculature. This work combines two of the main approaches to A/V classification investigated in subsequent research: (i) graph theory and (ii) intensity based feature extraction from color images.

Several approaches are primarily based on the use of vessel tracking techniques (Rothaus et al., 2009; Dashtbozorg et al., 2014). These approaches, stemming from graph theory, first establish a graph that represents the projected vasculature and then classify different parts of this graph as belonging to arterioles or venules using label propagation. Often, domain knowledge is taken into account in order to improve classification accuracy. An example is that it is highly unlikely that an arteriole will cross another arteriole, or that a venule would touch another venule. The advantages of applying graph theory to A/V discrimination are twofold. First, the integration of domain knowledge is beneficial, because it incorporates non-local physical constraints. Second, graph theory typically scales well both in computation and memory, with low order polynomial complexity for standard graph operations. The main drawback of strictly using graph theory is that it usually requires a partial manual labeling prior to training (see however Estrada et al. (2015), who partially overcome this issue via topology estimation).

In addition to vessel tracking techniques, researchers have focused on color intensity as main classification feature. Given the bright pixels located close to the vessel centerline, and darker pixels near the vessel boundaries, the color profile of vessels can be characterized by means of a Gaussian (Li et al., 2003; Relan et al., 2013). Vazquez et al. (2013) combined color information with a graph approach by use of a minimal path algorithm after the segmentation of a small ROI using color features. Kondermann et al. (2007) employed both a neural network (NN) and support vector machine (SVM) to classify vessels after principal component analysis (PCA). Niemeijer et al. (2011), Muramatsu et al. (2011), Mirsharif et al. (2013) and Xu et al. (2017) each applied an LDA classifier to various sets of color features. Zamperini et al. (2012) analyzed the various effective features for A/V discrimination, comparing color, spatial and size features, and concluded that color and position provided the best results.

The main benefit of using a fully-automated intensity-based classifier is the alleviated need for partial manual segmentation. Most of the research works discussed above achieve high results on primary, relatively wide vessels, but seem to encounter difficulties classifying their smaller variants. One possible cause could be the lack of central reflex in small vessels, often considered an important distinguishing feature. As a result, the evaluation of one's technique is often limited to an ROI or primary vessels.

Welikala et al. (2017) appear to report the first use of convolutional neural networks to the problem of A/V discrimination. However, this approach has a number of limitations. For instance, this network is based on 25 × 25 pixel patches, and therefore has limited capacity to learn larger scale features. Furthermore, their approach is based on a two-step segmentation and classification procedure, which may lead to suboptimal performance. Although the use of CNNs on this problem represents a methodological advance, the performance reported by Welikala et al. (2017) is surpassed by a more traditional approach (Xu et al., 2017). We show in the sequel that we have been able to surpass traditional approaches with a modified fully convolutional architecture, a joint segmentation procedure, and careful data augmentation.

3. Data

3.1. DRIVE

The DRIVE dataset (Staal et al., 2004) is publicly available, and has become a frequently used benchmark for research on retinal vessel segmentation. The set consists of both 20 training and 20 test fundus images, all having the same resolution of 584 × 565. These images were randomly selected from a diabetic retinopathy screening set of 400 Dutch participants suffering from diabetes. In this subset of 40 images, 33 of them are considered healthy, while the remaining 7 show early signs of diabetic retinopathy. Next to these images, a ground truth labeling for vessel extraction and binary masks for field of view (FOV) selection are also provided. The DRIVE set has also been used to evaluate A/V discrimination techniques mentioned in the previous section (Muramatsu et al., 2011; Mirsharif et al., 2013; Dashtbozorg et al., 2014; Estrada et al., 2015; Xu et al., 2017; Welikala et al., 2017). However, there is no ground truth labeling for arterioles and venules included in the official package, which hampers adequate benchmarking.

Qureshi et al. (2013) published a manually-labeled AV classified benchmark for DRIVE in 2013. Spurred by the aforementioned absence of a generally approved ground truth for A/V discrimination on DRIVE, the authors saw the need to share a gold standard, to allow for objective comparison between methods. The labeling itself was done by two computer vision experts and one ophthalmologist. Vessel class was determined through a majority vote among the three labelers. Two examples of the A/V discrimination ground truth are presented in the last column of Fig. 1. In addition, the ground truth described in Hu et al. (2013) was generated from the binary vessel segmentation ground truth of the second manual observer. In this paper, this ground truth serves as a proxy to evaluate our A/V predictions by an independent expert.

3.2. HRF

The High-Resolution Fundus (HRF) Image Database (Budai et al., 2013) contains 45 fundus images, equally distributed into three classes (healthy, glaucoma, diabetic retinopathy. As the name suggests, these images sport a high resolution of 3504 × 2336, equivalent to six and four times the width and height of DRIVE, respectively. For each image there exists a binary vessel segmentation ground truth (similar to the official package of DRIVE), generated through the collaboration of several experts and clinicians active in the field of retinal image analysis.

Unlike DRIVE, there exists no publicly available A/V ground truth for HRF. As DRIVE is often criticized for its unrealistic images (small resolution, outdated, limited symptoms of pathologies), this paper introduces a novel A/V ground truth for HRF. Similar to the A/V ground truth for DRIVE, the annotation process started with the official binary vessel segmentation ground truth. The initial labeling was carried out by an expert in retinal image analysis, and was subsequently carefully corrected by an ophthalmologist. The final result is characterized by three colors: red for arterioles, blue for venules, and green for uncertain pixels (unidentifiable vessels near the optic disc, crossovers, secondary vessels that have no clear origin and neovascularization). The first five images of each subcategory (healthy, diabetic retinopathy and glaucoma) were used as test images, which is equivalent to a third of the data. We encour-

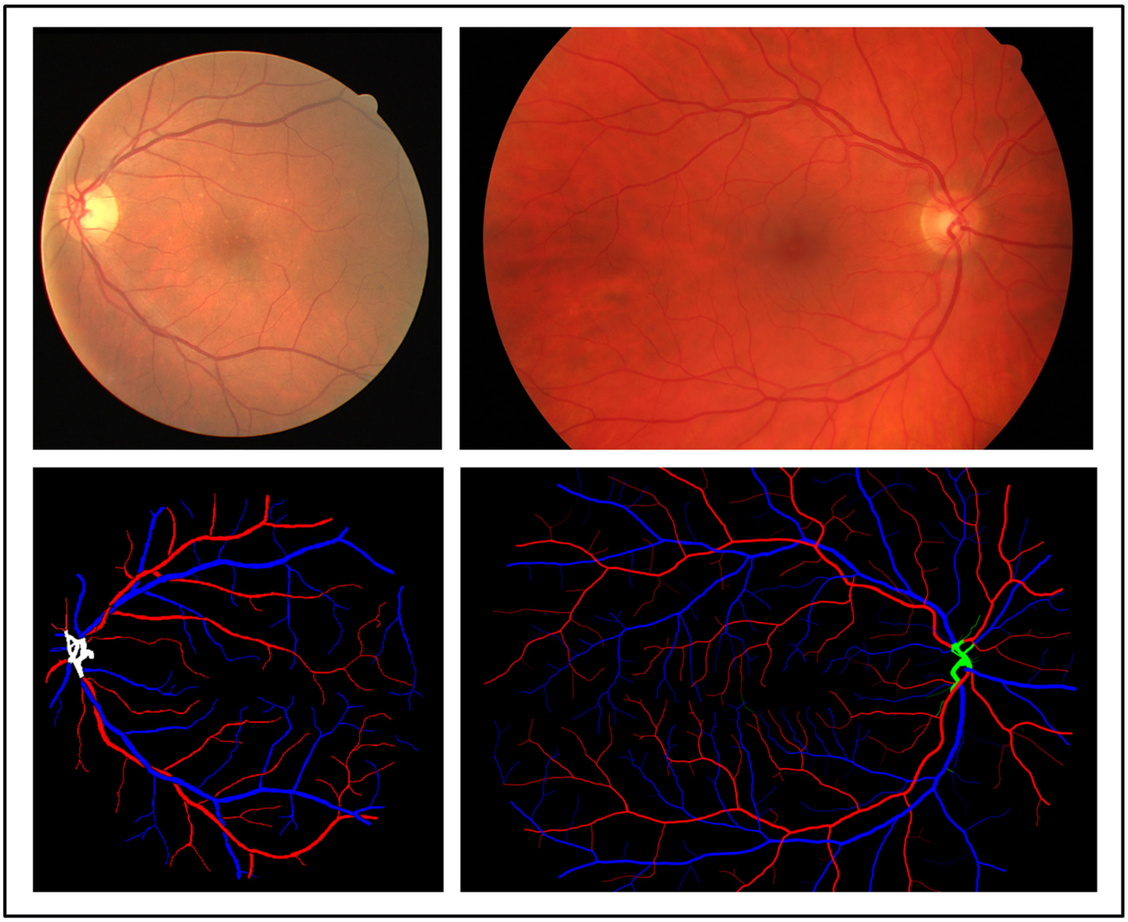

Fig. 1. Examples of annotated images from the DRIVE dataset (Staal et al., 2004) and HRF dataset (Budai et al., 2013). Arteries are colored red while veins are colored blue. From left to right: (i) original image DRIVE, (ii) original image HRF, (iii) DRIVE A/V ground truth from Qureshi et al. (2013), (iv) HRF A/V ground truth introduced in this work.

age researchers in the field of A/V discrimination to use the same split, in order to achieve proper benchmarking.

4. Methodology

In this section, we present an overview of the created architecture, based on the well-known U-Net (Ronneberger et al., 2015), for the ternary image segmentation task of discriminating arteries and veins in fundus images. The first motivation for the choice of U-Net stems from a successful application of this network on the binary segmentation task of vessel extraction applied to the DRIVE data set (Antiga, 2016). As mentioned in the introduction, the U-Net achieved groundbreaking results in the context of biomedical image segmentation (Ronneberger et al., 2015). One important consideration to be made is the motivation for ternary segmentation: given that there exists a near-perfect vessel extraction model, one could argue that a ternary segmentation for A/V discrimination is partly solving a problem that is already dealt with. Starting from a near-perfect ground truth with extracted vessels, the remaining task would become binary. However, the data processing inequality (Cover and Thomas, 2006) implies that a jointly trained system has the ability to learn at least as well as a system that is trained in two phases.

We base our implementation on a publicly available U-Net implementation for vessel extraction (Antiga, 2016), with several modifications. These include support for RGB input, proper ground truth conversion, and ternary output labels. The motivation for RGB input is due to the significant intensity similarities between retinal artery and vein: by increasing the information per pixel, we aim to get better results when compared with a grayscale setting.

4.1. Preprocessing

Before training, the DRIVE and HRF training images are preprocessed with several transformations in order to eliminate undesired variance and facilitate training.

Fundus images show significant variability in lighting within the image due to the curvature of the retina. To counter this unfavorable characteristic, a local contrast enhancement is implemented. Prior to this, the ROI is padded using the technique described in Soares et al. (2006) to prevent the excessive contrast enhancement at the border. In the case of DRIVE, a Gaussian filter with kernel of 65 × 65 pixels, zero mean and standard deviation σ of size 10 is applied to the fundus image. For the HRF data set, σ is increased to 60, corresponding to the larger image size when compared to DRIVE. Finally, the output from the convolution is subtracted from the original image (Fig. 2):

$$N(w,h) = I(w,h) - G(w,h), \tag{1}$$

where N is the normalized image, I is the original image, and G is the Gaussian blurred image.

In a second step, the images are normalized by subtracting the mean image computed over the training set, and dividing each pixel by the average standard deviation. That way, the variance among images in the training set (see Fig. 1) is lowered. Finally, other preprocessing techniques were considered (e.g. gamma correction and CLAHE), but did not result in improved performance.

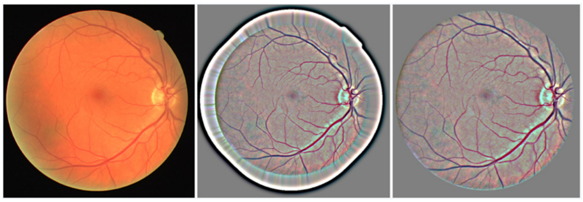

Fig. 2. Preprocessing steps applied in this work (cf. Equation (1)). The original image (left), application of local contrast enhancement (center) after padding the ROI, crop based on the mask corresponding to the FOV (right).

4.2. *Data augmentation - manipulation*

The U-Net architecture requires a large amount of training data. The 20 and 30 images of the DRIVE and HRF training set in their original form do not fulfill this requirement. Furthermore, a validation set is needed for model selection. A validation split of 15% is selected, resulting in a data split of 17 training, 3 validation and 20 test for DRIVE and 25 training, 5 validation and 15 test images for HRF. Due to data scarcity, powerful data augmentation techniques need to be applied. One of the most important data augmentation techniques used in training the model was random cropping. Random crops of size 512 × 512 (DRIVE) and 2048 × 2048 (HRF) are extracted randomly out of the original images with resolution 584 × 565 (DRIVE) and 3504 × 2336 (HRF). That way, the number of training instances can be greatly increased. Evidently, partially overlapping patches are unavoidable, and could still imply an unsolved data scarcity issue.

Differentiating arteries from veins is achieved in part by detecting small relative differences in color and brightness. We deemed it important to ensure the presence of both artery and vein pixels in each patch to allow the FCN to detect those relative differences. The use of a large patch size that is close to the original image size allows the network to learn global features. We opt for a dynamic augmented data set, where training samples are generated randomly at the start of each mini-batch.

The use of an increased patch size is strongly associated with the likelihood of overlap, hence the need for additional data augmentation. We artificially grow our data set by a factor of 8 through rotation at 90, 180 and 270 degrees and horizontal flips. In addition, data augmentation through elastic deformation was one of the key success factors behind the original U-Net paper and resulted in a substantial decrease in overfitting in our work as well. We have implemented elastic deformation by sampling control points on a regularly spaced 100 × 100 grid. Each control point has isotropic Gaussian noise added with $\sigma = 20$. This greatly increases the number of synthetic training images. However, the use of elastic deformation seems to come at the cost of harming the continuous structure of secondary vessels with a thickness of 1-2 pixels. Finally, a number of experiments employed random rotation at any degree between 0 and 360. As rotation resulted in poor performance on the validation set, we have not included this step in all experiments.

4.3. *Network architecture*

A key novelty in our work is the combination of vessel extraction and artery-vein discrimination; two processes that have always been dealt with separately in the literature. As previously discussed, CNNs have outperformed classic computer vision techniques in vessel extraction (Fraz et al., 2012b; González et al., 2010; Ishikawa et al., 2005; Türetken et al., 2011; Yedidya and Hartley, 2008), and showed promising results in the A/V discrimination method introduced by Welikala et al. (2017). We adapt the original U-Net approach to allow for three class segmentation in the context of A/V discrimination.

The model that achieved the best performance on the two data sets differs from the original U-Net in several ways. The input is set to 3 × 512 × 512 (equivalently 3 × 2048 × 2048 for HRF), which corresponds to the 3-channel RGB input and large patch size. At the end of the expanding path, the output segmentation map is modified to allow for four classes: ‘background’, ‘artery’, ‘vein’ and ‘unknown’. This is obtained through a final convolutional layer with four convolution filters of size 1 × 1. That way, four output maps are generated, to which softmax activation is applied. The unknown class, corresponding to the green pixels in the ground truth, is ignored in the computation of the loss function, and will not be predicted.

The number of feature maps generated at each layer is a quarter of the number described by Ronneberger et al. (2015) due to memory limitations (preference for larger mini-batch size) and data scarcity. We opt for a mini-batch size of four to reduce variance in the gradient updates, resulting in fast convergence. The authors of the original U-Net preferred a mini-batch with a single input image combined with an optimizer employing a high momentum over a larger mini-batch size. Our best model comprises a little over 5 million trainable weights, which is significantly smaller than the original network. Given that there are few (original) training data, we expect an increased risk of overfitting when moving to a larger network. An experiment that added an additional layer block to both contracting and expanding paths scored worse on the three holdout validation images. To compare with Welikala et al. (2017), additional experiments were carried out with a different data split (25 - 5 - 10). The best model on this new split successfully goes one level deeper, without overfitting, increasing the reported F1 score over the standard split slightly. Batch normalization layers with momentum = 0.99 and epsilon = 0.001 were added after each convolution to standardize the outputs of the latter. The kernel of each convolution was increased to 5 × 5, leading to an increase in the receptive field. The model for DRIVE did not benefit from the use of dilation, an operation that similarly expands the receptive field but at reduced memory consumption. However, the overall performance of the FCN applied to the HRF data increased significantly when dilation was employed. This method allows the same number of trainable parameters, while the receptive field grows in size. A detailed overview of the final network architecture is presented in Table 1.

4.4. *Training details*

Dropout layers of 20% at end of contracting path are added for explicit regularization, as done in the original U-Net paper. The use of dropout following each convolution had a significant negative impact on segmentation performance, explained by the loss of image information at the start of the network. A custom weight map is introduced that emphasizes vessel centerline pixels. The motivation behind this weightmap is threefold. Within a fundus image of the DRIVE data set, close to 90% of all pixels are non-vessel, resulting in a huge class imbalance. A custom weightmap does not only counter class imbalance; it leads to equal importance among vessels. Previous work reported high accuracies on primary vessels (e.g. vessel width at least three pixels), whereas the secondary vessels were often omitted in the evaluation. Finally, the use of a weight map also allowed us to ignore the pixels for which the labeling is unknown in the ground truth described by Qureshi et al. (2013), and the HRF variant introduced in this paper. The vessel centerline was obtained by skeletonizing the official vessel ground truth with the Image Processing SciKit Toolbox in Python (van der Walt et al., 2014). Categorical cross-entropy was maintained as loss metric, but in combination with the Adam optimizer (Kingma and Ba, 2015), which updates weights with a maximum step size of 0.001 (after convergence the learning rate is decreased to a tenth of its original value). The proposed U-Net segmentation network is implemented

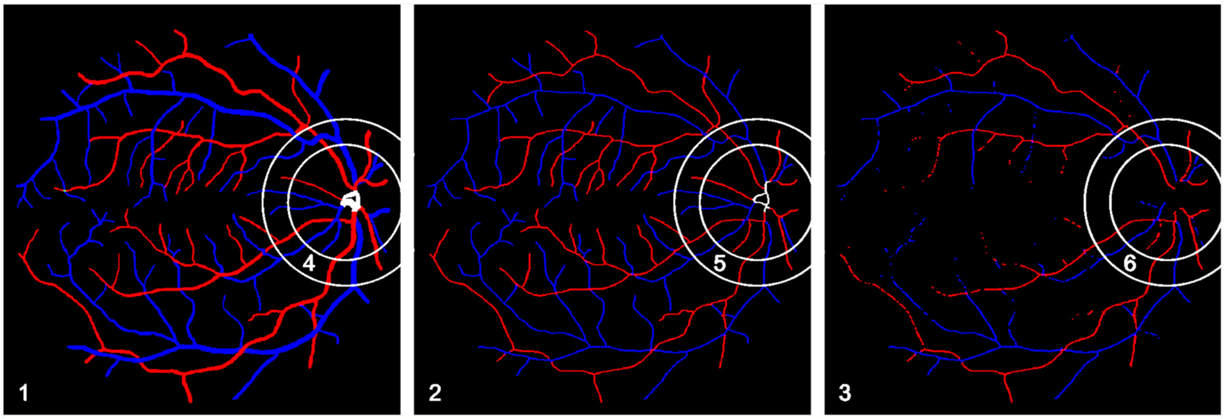

Fig. 3. Illustration of the different zones used in the evaluation of A/V segmentation algorithms (cf. Section 4.5 and Table 3). From left to right: (1) All vessel pixels, (2) Vessel centerline pixels, (3) Vessel centerline pixels, limited to vessels wider than two pixels. (4)–(6) limit evaluation to the ROI known as Zone B, obtained as described by Knudtson et al. (2003).

Table 1
FCN architecture for A/V discrimination on DRIVE. Each convolution was followed by a ReLU activation and Batch Normalization layer with default settings. In the decoder path, features from the encoder path at the same resolution are fused through concatenation. The FCN for HRF follows the same architecture, with the main difference the input size (and subsequent output sizes).

Layer	Output size	Filter size	Stride	Dropout
Inputs	3 × 512 × 512	-	-	-
Convolution 1	16 × 512 × 512	5 × 5	1	-
Convolution 2	16 × 512 × 512	5 × 5	1	-
Max pool 1	16 × 256 × 256	2 × 2	2	-
Convolution 3	32 × 256 × 256	5 × 5	1	-
Convolution 4	32 × 256 × 256	5 × 5	1	-
Max pool 2	32 × 128 × 128	2 × 2	2	-
Convolution 5	64 × 128 × 128	5 × 5	1	-
Convolution 6	64 × 128 × 128	5 × 5	1	-
Max pool 3	64 × 64 × 64	2 × 2	2	-
Convolution 7	128 × 64 × 64	5 × 5	1	-
Convolution 8	128 × 64 × 64	5 × 5	1	-
Max pool 4	128 × 32 × 32	2 × 2	2	-
Convolution 9	256 × 32 × 32	5 × 5	1	20%
Convolution 10	256 × 32 × 32	5 × 5	1	20%
Upsampling 1	256 × 64 × 64	2 × 2	2	-
Convolution 11	128 × 64 × 64	5 × 5	1	-
Convolution 12	128 × 64 × 64	5 × 5	1	-
Upsampling 2	128 × 128 × 128	2 × 2	2	-
Convolution 11	64 × 128 × 128	5 × 5	1	-
Convolution 12	64 × 128 × 128	5 × 5	1	-
Upsampling 3	64 × 256 × 256	2 × 2	2	-
Convolution 11	32 × 256 × 256	5 × 5	1	-
Convolution 12	32 × 256 × 256	5 × 5	1	-
Upsampling 4	32 × 512 × 512	2 × 2	2	-
Convolution 11	16 × 512 × 512	5 × 5	1	-
Convolution 12	16 × 512 × 512	5 × 5	1	-
Output	4 × 512 × 512	1 × 1	1	-

in Python 3 using the open-source neural network library Keras 2. Training and development was done on a single GPU (NVIDIA GeForce GTX 1080 Ti), offering a total of 11GB VRAM. We trained each model for 150 epochs (9.75 hours), with each epoch sampling 500 randomly augmented patches from the training images.

4.5. Evaluation protocol

Evaluation of our custom U-Net for A/V discrimination is carried out on all test images. From these images, crops of the same size that was used in training are extracted, but this time in an ordered way to facilitate recomposition after prediction is completed. Additionally, we employ common test-time augmentation techniques (overlapping crops, rotations and horizontal flips) to reduce prediction errors. The reported performance was computed using a single model, no ensemble techniques were used. After obtaining the averaged class probabilities, each pixel is assigned to the class corresponding to the largest probability. Finally the segmentation maps are compared to the masked ground truth of the test images, depending on the active evaluation zone.

Previous work on artery-vein discrimination often report their best results in terms of accuracy, but with differences in the details of the data set split, evaluation zone, and pixel mask (e.g. restricting to centerline pixels of vessels of varying minimal pixel width, usually two or three pixels) (Table 6). This inconsistency in evaluation hampers the objective comparison of methods. We have therefore opted for a comprehensive evaluation for DRIVE, in which we use one of three subsets of vessels (all, centerline, or centerline limited to vessels wider than two pixels), and either the whole image or only vessel pixels present in Zone B (Fig. 3). Given that we are the first to propose a method that combines both vessel extraction and artery-vein discrimination, we want to set an objective standard to facilitate future benchmarking. The binary masks that were used in the evaluation can be found in the project repository at https://github.com/rubenhx/av-segmentation.

5. Results

5.1. DRIVE

Our best model yields a vessel segmentation accuracy of 96.75%, which is in line with related work in that domain (Orlando et al., 2017). For A/V discrimination, performance metrics were computed over six different evaluation masks (Fig. 3), as introduced in Section 4.5. The average performance statistics are given, but also split into artery and vein, to highlight the model's performance on the two classes individually. Results on two-class segmentation are reported in Table 3, while three-class segmentation is given in Table 4.

The best model achieves an A/V discrimination accuracy of 94.25% in evaluation zone 3. Since these accuracies are comparable to most A/V discrimination results on the DRIVE data set reported in literature, they are highlighted in Table 3. As previously discussed, significant discrepancies in results can be observed between evaluation zones. This particular model obtained better accuracies for retinal arteries than retinal veins, especially in evaluation zone 2. At the same time, the F1 score for the retinal vein class is often higher than its artery equivalent, illustrating the limitations of the accuracy metric in this context.

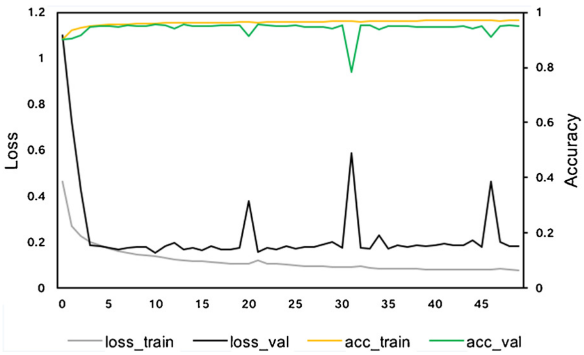

Fig. 4. DRIVE - Training evolution of accuracy and loss as a function of the number of training epochs. Weights at epoch 11 were selected, corresponding to the lowest loss on the three holdout validation images.

Table 2
DRIVE - Ablation study with a fixed learning rate at 0.001, assessed on F1 score in evaluation zone (1) (see Fig. 3). The largest factor in the performance of the final system is the background subtraction that counters lighting effects. The performance is comparatively stable with respect to other factors related to image normalization, dropout, kernel size, and dilation.

Network change	Effect on F1 score	
	without	with
Background subtraction (preprocessing)	0.8299	0.9450
Data set normalization (preprocessing)	0.9435	0.9450
20% dropout after each convolution	0.9513	0.9450
Kernel size 3 × 3	0.9575	0.9490
Kernel size 7 × 7	0.9575	0.9388
Centerline weights, value 2	0.9484	0.9535
Dilation rate 2 at each convolution	0.9595	0.9344
Elastic deformation	0.9475	0.9575
Standard rotation	0.9434	0.9575
Standard rotation + elastic deformation	0.9337	0.9575

Table 3
Quantitative results of the proposed FCN for A/V discrimination on all 20 DRIVE test images, given for each of the evaluation zones (in parentheses) introduced in Section 4.5 and illustrated in Fig. 3. This table restricts results to the two vessel classes (background is not included).

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Average						
Accuracy	0.9386	0.9190	**0.9425**	0.9442	0.9159	0.9466
F1	0.9383	0.9188	0.9425	0.9437	0.9161	0.9464
Artery						
Accuracy	0.9487	0.9466	**0.9442**	0.9617	0.9581	0.9617
F1	0.9317	0.9161	0.9370	0.9400	0.9184	0.9468
Vein						
Accuracy	0.9302	0.8942	**0.9411**	0.9284	0.8757	0.9312
F1	0.9439	0.9212	0.9472	0.9470	0.9140	0.9459

Table 4
Quantitative results of the proposed FCN for A/V discrimination on all 20 DRIVE test images, given for the evaluation zones (in parentheses, see Fig. 3) where the background is taken into consideration as well.

Average	**(1)**	**(4)**
Accuracy	0.9675	0.9479
F1	0.9663	0.9519

Fig. 4 visualizes the loss and accuracy evolution of our best model, for which we selected the weights at Epoch 11 (early stopping, corresponds to lowest validation loss). Qualitative results of our work are displayed in Fig. 6.

Table 2 highlights the most important ablation effects. Starting from the best model with learning rate fixed at 0.001, several changes were made to assess its impact on performance on three-class segmentation. Background subtraction did lead to an significant surge in performance, whereas the second preprocessing step had negligible impact. Both excessive dropout and dilation resulted in subpar performance.

Finally, Table 6 compares our work with other works that evaluate their method on the public DRIVE data set (although not necessarily the same ground truth, as described in the final column of the table). The proposed method makes 25% fewer mistakes on vessel centerline pixels (limited to vessels wider than two pixels) compared to the current state-of-the-art.

5.2. HRF

For the High Resolution Fundus data set, the performance of several experiments with varying patch size, centerline weight, and dilation rate were evaluated on five holdout validation images. Detailed quantitative results on both validation and test set are presented in Table 5.

Det% vessel (column 4) translates the number of vessel pixels in the ground truth that were found by the FCN (ignoring the true background pixels completely). *Acc - full image* and *F1 - full image* correspond to the performance in terms of three-class segmentation. Although these results are comparable across rows, they are hard to interpret due to the influence of the background pixels. Next, *Acc - detected centerl* and *F1 - detected centerl* yield the two-class metrics on the centerlines that were found by the model. Given that the detected centerline pixels vary across experiments, these metrics are not suited for objective comparison. Hence, the final two metrics *Acc - full centerline* and *F1 - full centerline* are introduced to provide a proper way to assess our system that achieves both vessel extraction and vessel discrimination. These metrics are computed by considering each vessel centerline pixel in the ground truth, and compare with the prediction, even if the model predicted a background pixel (i.e. it was not able to segment correctly). The last row of the table corresponds to the results of the best model on all 15 test images, yielding an accuracy of 96.98% on all discovered centerline pixels.

Fig. 5 provides a visualization of both the ground truth and predictions by all eight experiments on the region surrounding the fovea. For a macula-centered fundus image, this is typically the area with the smallest vessels. Hence, this visualization aims at facilitating a qualitative benchmarking between experiments.

Finally, ground truth and predicted segmentation on a complete test image are visualized in Fig. 7.

6. Discussion and conclusion

In this work, a successful application of the U-Net semantic segmentation architecture to the discrimination of arteries and veins in fundus images is presented. The first goal, to check whether a segmentation network based on the U-Net would be able to perform state-of-the-art A/V discrimination, is attained. Following novel modifications to the U-Net architecture based on performance on the training and validation data, results that surpassed those reported in previous work have been obtained. On DRIVE, the proposed model achieves accuracies of 94.42% and 94.11% on arteries and veins, respectively. This represents a decrease in error of 25% over the previous state of the art reported by Xu et al. (2017). On HRF, trained with the A/V ground truth presented in this work, an accuracy of 96.98% was measured on all discovered centerline pixels. Moving forward, we invite fellow researchers active in this field to report results on all centerline pixels present in the ground truth when developing a system that combines vessel extraction and A/V discrimination.

The proposed method gives predictions at every pixel, while related work do not. This is particularly useful in the domain of

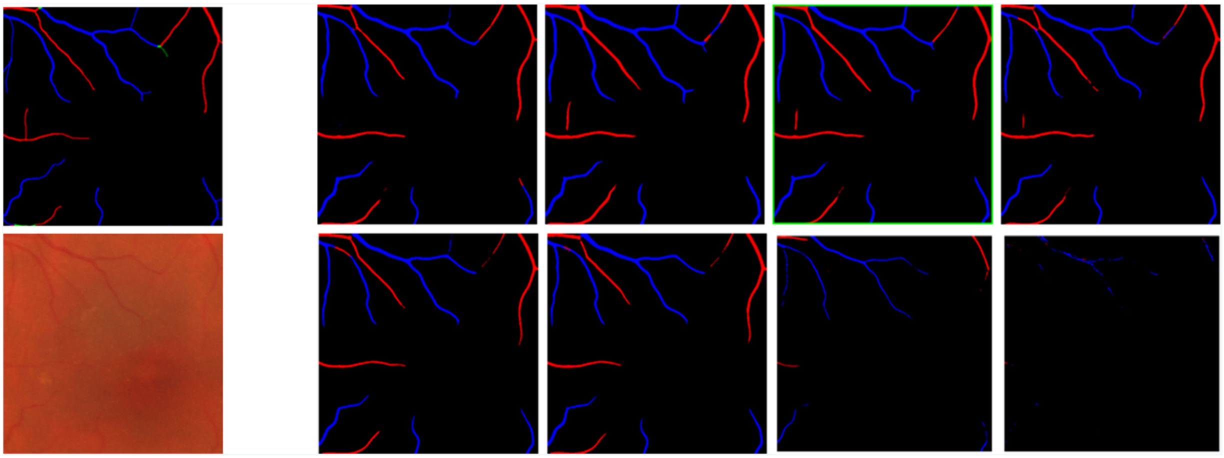

Fig. 5. Visualization of the region surrounding the fovea, with exclusively secondary vessels. The ground truth and actual image are visualized on the outer left. The rest of the images follow the order of the experiments in the table. The prediction of the best model is the third prediction on the first row (also highlighted in green).

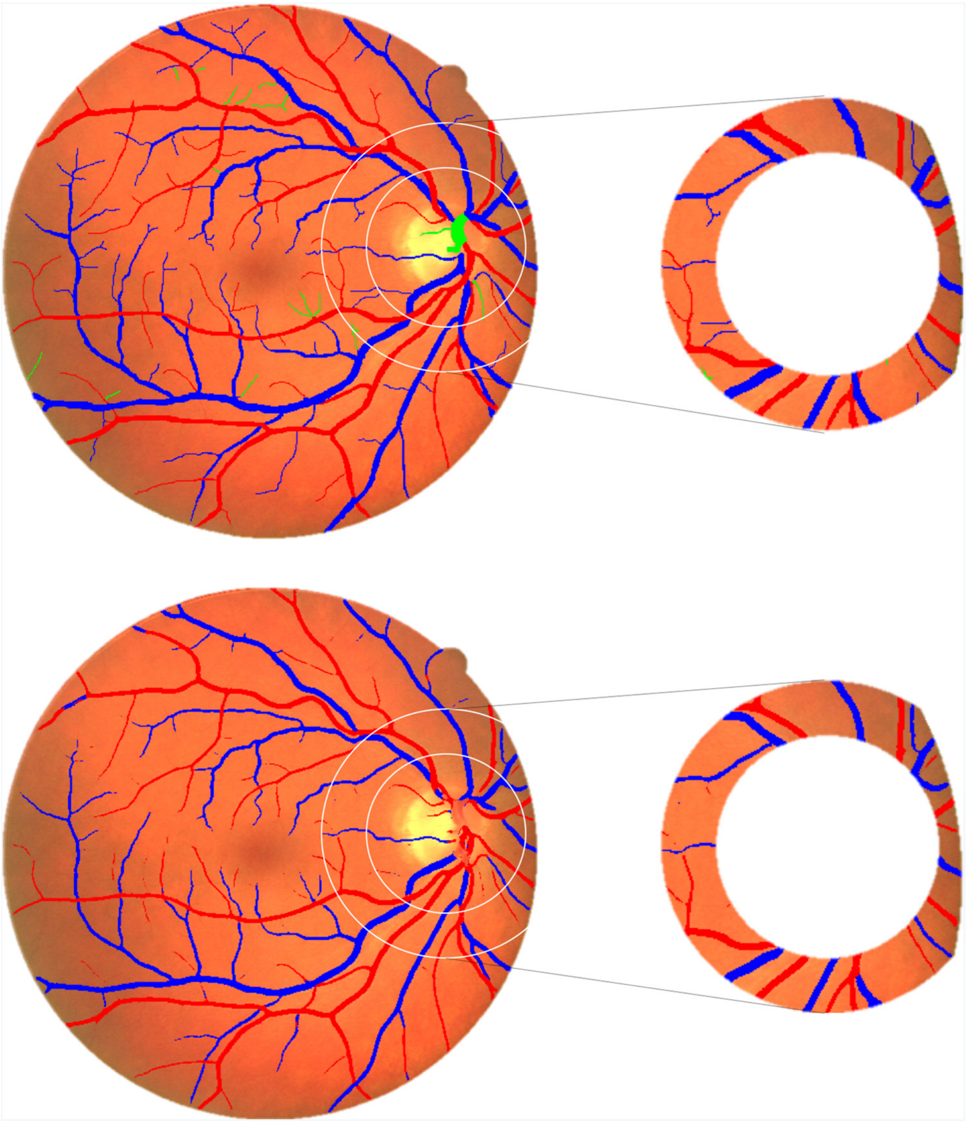

Fig. 6. Ground truth (top) and prediction of best model (bottom) of test image 2, along with extracted Zone B (referred to as evaluation zone (4)).

Table 5
Quantitative results of the proposed FCN for A/V discrimination on five HRF validation images. The 15 test images are evaluated using the best model, selected on highest F1 score computed on all centerline pixels (indicated with a * in the table).

Patch size	Centerline weight	Dilation rate	Det% vessel	Acc - full image	F1 - full image	Acc - detected centerl	F1 - detected centerl	Acc - full centerline	F1 - full centerline
512 × 512	10	3 × 3	0.7786	0.9688	0.9695	0.9092	0.9091	0.7109	0.7979
1024 × 1024	10	3 × 3	0.8577	0.9656	0.9679	0.9477	0.9477	0.8167	0.8773
2048 × 2048	10	3 × 3	**0.8617**	0.9694	0.9709	0.9642	0.9642	**0.8189**	**0.8856***
2048 × 2048	5	3 × 3	0.8016	0.9707	0.9714	0.9582	0.9583	0.7572	0.8456
2048 × 2048	2	3 × 3	0.8161	**0.9728**	**0.9733**	0.9703	0.9703	0.7192	0.8261
2048 × 2048	1	3 × 3	0.8231	0.9703	0.9713	0.9609	0.9609	0.7051	0.8133
2048 × 2048	1	1 × 1	0.6356	0.9716	0.9693	0.9743	0.9743	0.5419	0.6963
2048 × 2048	1	5 × 5	0.2904	0.9510	0.9380	0.5958	0.5051	0.1598	0.2213
Evaluation of best model on 15 test images									
2048 × 2048	10	3 × 3	0.8074	0.9681	0.9688	0.9698	0.9698	0.7992	0.8753

Table 6
Benchmark with related work (research that reported results on DRIVE data set).

Authors	Method	Dataset split	Performance (accuracy)	Description of evaluation
Muramatsu et al. (2011)	LDA classifier	Standard	93%	Limited to centerline pixels of 'major' vessels in Zone B
Mirsharif et al. (2012)	LDA classifier	Standard	84.05% (FOV) 90.16% (Zone B)	Vessel centerline, limited to vessels wider than three pixels
Dashtbozorg et al. (2013)	Graph-based	Standard	87.40%	Vessel centerline, limited to vessels wider than three pixels
Estrada et al. (2015)	Graph-based	Standard	91.70%	Vessel centerline, limited to vessels wider than two pixels
Xu et al. (2017)	LDA classifier	Standard	92.30%	Around 73,000 vessel centerline pixels segmented
Welikala et al. (2017)	CNN	Custom	91.97%	No information on amount of pixels
Our method	**FCN**	**Standard**	**94.25%**	Vessel centerline, limited to vessels wider than two pixels
			93.94%	Using labels from (Hu et al., 2013) as independent expert

automated AVR determination, which requires the vessel width in Zone B.

The insights of this work can be beneficial in the domain of automated analysis of fundus images with deep learning. We highlight two key concepts that led to results that surpass previous work.

First, this work had a strong focus on careful data augmentation. Deep learning architectures require a large number of unique training samples in order to generalize well and prevent overfitting. Hence the introduction of basic augmentation techniques including rotation, flips and random crops, but more importantly, elastic deformation. The advantage of elastic deformation is the sampling of a large amount of realistic artificial training data, allowing longer training time. For vessel segmentation, the sole drawback is the loss of connectedness observed in secondary vessels that are one pixel wide.

Next to that, a custom weightmap was introduced that emphasizes vessel centerline pixels. That way, equal importance is attributed towards all vessels, independent of their width. Previously, the evaluation of one's method for A/V discrimination was often limited to primary vessels that are easily distinguishable upon visual inspection, whereas performance on secondary vessels seemed to be much lower. The use of the custom weight map addresses this issue.

While our model outputs high accuracies on the DRIVE data set, there are some limitations to be discussed. Reporting on the DRIVE data set is recommended as it is a widely cited and evaluated data set in literature. At the same time, the data set itself is small, has low resolution and shows little variance. To provide a more realistic performance, we evaluated our model on the High Resolution Fundus Image Database, for which the ground truth is made publicly available through this work. The performance on the latter exceeds the one measured on DRIVE, most likely due to the availability of an increased training set that allows for better generalization. Still, even at accuracies nearing 97% on HRF, the model makes subtle mistakes that a human expert would never make (e.g. a sudden transition from one vessel type to another on a straight segment). Hence, there is room for improvement left, possibly by incorporating a traditional vessel tracking technique. A more novel way would be to segment the bifurcation and crossover areas as an additional class, after which segments could be classified based on the majority vote.

The prediction of reliable vessel diameters is of utmost importance in width-related retinal biomarkers such as AVR. Qualitative results for DRIVE seemed to indicate a consistently smaller diameter in segmentations produced by the FCN when compared to the ground truth. On average, predicted Zone B vessels are 0.34 pixels thinner than the reference standard of the DRIVE test set. As long as the error is consistent in both arterioles and venules, the effect on AVR value should be minimal. A quantitative comparison for images of HRF confirmed a consistent deviance of 1.89 pixels surplus in predicted segments. This is due to the different sample weight on the centerline (2 for DRIVE, 10 for HRF) which led to the best results for overall segmentation, but not necessarily vessel thickness. One future avenue of investigation is to formulate a loss function that takes vessel thickness into account, in order to optimize for AVR directly.

Moving forward to a general model that can achieve A/V discrimination on any fundus image, additional ground truth are needed. Data sets where a different camera (equivalently, resolution) was used, or containing pathologies, could only lead to a more robust model that could be employed in a real life scenario. The introduction of the HRF A/V ground truth represents a step in this direction.

Recently, it was demonstrated that deep learning applied to retinal fundus images could predict to some extent cardiovascular risk factors such as age, smoking status, and systolic blood pressure (Poplin et al., 2018). However, to fully extract the cardiovascular information hidden within the retinal microvasculature, it is believed that metrics describing the status of the arterial and venular tree is important. Many prospective population studies have fundus images included in their data acquisition protocol. The exploitation of these images can be time-consuming as one needs graders for manual or semi-automated analysis. Therefore, a framework for simultaneous segmentation and classification of vessels, such as the one presented in this paper, could be valuable in a workflow for the automated calculation of vessel widths, CRAE,

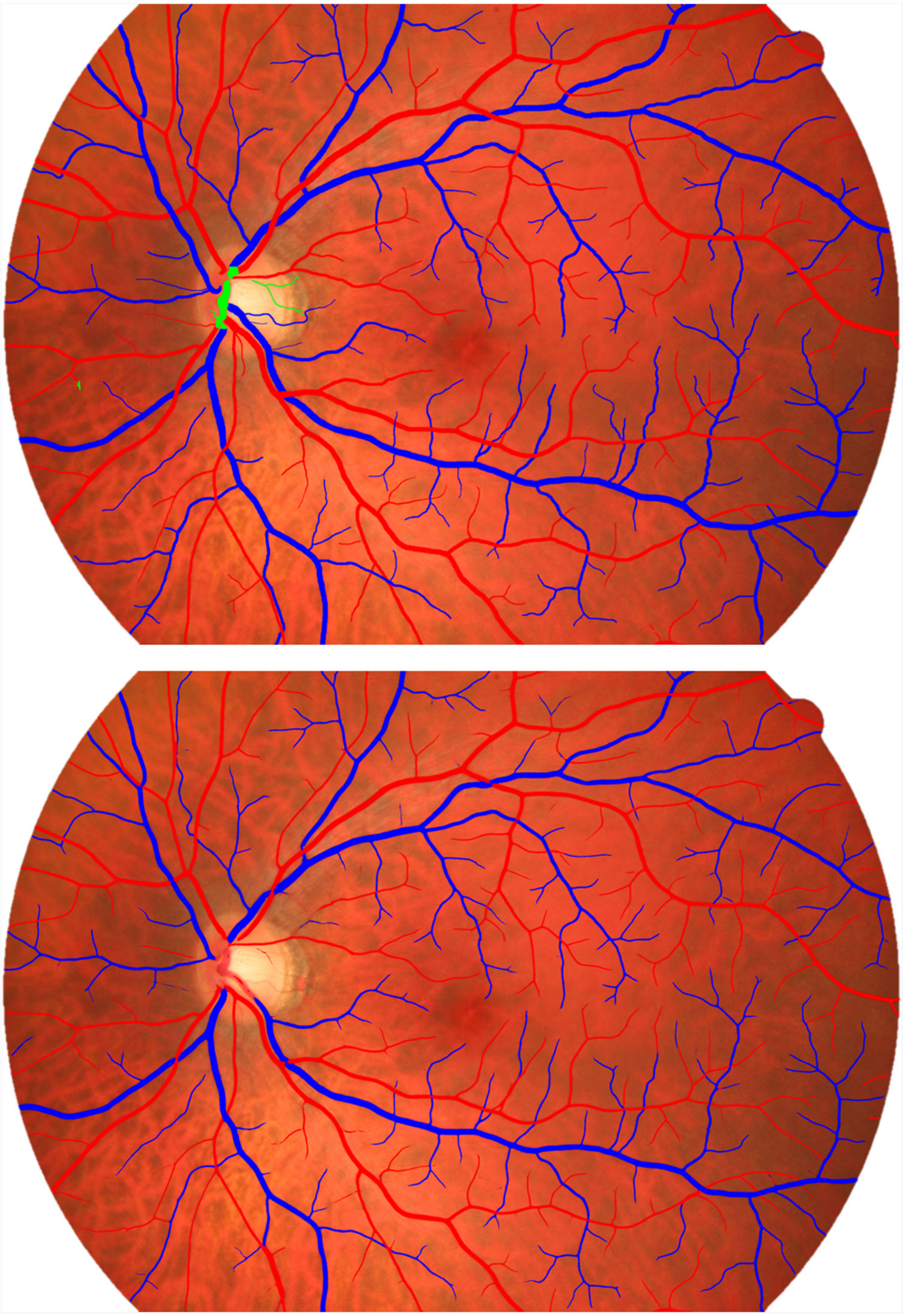

Fig. 7. Ground truth (top) and prediction of best model (bottom) of HRF healthy image number two.

CRVE and other metrics such as arteriolar and venular tree fractal dimensions.

Conflict of interest

The first author is jointly supported by the Research Group Ophthalmology, KU Leuven and VITO NV.

No outside entities have been involved in the study design, in the collection, analysis and interpretation of data, in the writing of the manuscript, nor in the decision to submit the manuscript for publication. Thus, the authors declare that there are no conflicts of interest in this work.

Acknowledgements

The first author is jointly supported by the Research Group Ophthalmology, KU Leuven and VITO NV. The authors would like to thank Bashir Al-Diri for providing artery-vein ground truth for the DRIVE data set.

No outside entities have been involved in the study design, in the collection, analysis and interpretation of data, in the writing of

the manuscript, nor in the decision to submit the manuscript for publication. Thus, the authors declare that there are no conflicts of interest in this work.

References

Antiga, L., 2016. Retina-unet. https://github.com/orobix/retina-unet.

Budai, A., Bock, R., Maier, A., Hornegger, J., Michelson, G., 2013. Robust vessel segmentation in fundus images. Int. J. Biomed. Imaging, 2013.

Bursell, S.E., Clermont, A.C., Kinsley, B.T., Simonson, D.C., Aiello, L.M., Wolpert, H.A., 1996. Retinal blood flow changes in patients with insulin-dependent diabetes mellitus and no diabetic retinopathy. Invest. Ophthalmol. Vis. Sci. 37, 886.

Cover, T.M., Thomas, J.A., 2006. Elements of Information Theory. Wiley.

Dashtbozorg, B., Mendonça, A.M., Campilho, A., 2014. An automatic graph-based approach for artery/vein classification in retinal images. IEEE Trans. Image Process. 23, 1073–1083.

Ding, J., Wai, K., McGeechan, K., Ikram, M., Kawasaki, R., Xie, J., Klein, R., Klein, B., Cotch, M., Wang, J., Mitchell, P., Shaw, J., Takamasa, K., Sharrett, A., Wong, T., 2014. Review: Retinal vascular caliber and the development of hypertension: A meta-analysis of individual participant data. J. Hypertens. 32, 207–215.

Estrada, R., Allingham, M.J., Mettu, P.S., Cousins, S.W., Tomasi, C., Farsiu, S., 2015. Retinal artery-vein classification via topology estimation. IEEE Trans. Med. Imaging 34, 2518–2534.

Fraz, M., Barman, S., Remagnino, P., Hoppe, A., Basit, A., Uyyanonvara, B., Rudnicka, A., Owen, C., 2012a. An approach to localize the retinal blood vessels using bit planes and centerline detection. Comput. Methods Programs Biomed. 108, 600–616.

Fraz, M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A., Owen, C., Barman, S., 2012b. Blood vessel segmentation methodologies in retinal images: A survey. Comput. Methods Programs Biomed. 108, 407–433.

González, G., T&ldquo;uretken, E., Fleuret, F., Fua, P., 2010. Delineating trees in noisy 2d images and 3d image-stacks. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2799–2806.

Grisan, E., Ruggeri, A., 2003. A divide et impera strategy for automatic classification of retinal vessels into arteries and veins. Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439), 890–893, Vol.1.

Hemelings, R., Elen, B., Stalmans, I., Van Keer, K., De Boever, P., Blaschko, M.B., 2018. https://openreview. net/pdf?id=SJ7ma1nsG.

Ho, H., Cheung, C.Y., Sabanayagam, C., Yip, W., Ikram, M.K., Ong, P.G., Mitchell, P., Chow, K.Y., Cheng, C.Y., Tai, E.S., Wong, T.Y., 2017. Retinopathy signs improved prediction and reclassification of cardiovascular disease risk in diabetes: A prospective cohort study. Sci. Rep. 7, 41492 EP -.

Hu, Q., Abràmoff, M.D., Garvin, M.K., 2013. Automated separation of binary overlapping trees in low-contrast color retinal images. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (Eds.), Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 436–443.

Huang, Y., Zhang, J., Huang, Y., 2012. An automated computational framework for retinal vascular network labeling and branching order analysis. Microvasc. Res. 84, 169–177.

Ishikawa, H., Geiger, D., Cole, R., 2005. Finding tree structures by grouping symmetries. Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 1132–1139, Vol. 2.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. International Conference on Learning Representations.

Knudtson, M.D., Lee, K.E., Hubbard, L.D., Wong, T.Y., Klein, R., Klein, B.E., 2003. Revised formulas for summarizing retinal vessel diameters. Curr. Eye Res. 27, 143–149, PMID: 14562179.

Kondermann, C., Kondermann, D., Yan, M., 2007. Blood vessel classification into arteries and veins in retinal images. Proc. SPIE 6512, 651247-651247-9.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 25. Curran Associates, Inc, pp. 1097–1105.

Li, H., Hsu, W., Lee, M.L., Wang, H., 2003. A piecewise Gaussian model for profiling and differentiating retinal vessels. Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on, pp. I-1069-72 vol.1.

Long, J., Shelhamer, E., Darrell, T., 2014. Fully convolutional networks for semantic segmentation, CoRR abs/1411.4038.

Malek, J., Tourki, R., 2013. Inertia-based vessel centerline extraction in retinal image. 2013 International Conference on Control, Decision and Information Technologies (CoDIT), 378–381.

McGeechan, K., Liew, G., Macaskill, P., Irwig, L., Klein, R., Klein, B.E.K., Wang, J.J., Mitchell, P., Vingerling, J.R., de Jong, P.T.V.M., Witteman, J.C.M., Breteler, M.M.B., Shaw, J., Zimmet, P., Wong, T.Y., 2009. Prediction of incident stroke events based on retinal vessel caliber: A systematic review and individual-participant meta-analysis. Am. J. Epidemiol. 170, 1323–1332.

Mirsharif, Q., Tajeripour, F., Pourreza, H., 2013. Automated characterization of blood vessels as arteries and veins in retinal images. Comput. Med. Imaging Graph. 37, 607–617.

Muramatsu, C., Hatanaka, Y., Iwase, T., Hara, T., Fujita, H., 2011. Automated selection of major arteries and veins for measurement of arteriolar-to-venular diameter ratio on retinal fundus images. Comput. Med. Imaging Graph. 35, 472–480.

Mutlu, U., Ikram, M.K., Wolters, F.J., Hofman, A., Klaver, C.C., Ikram, M.A., 2016. Retinal microvasculature is associated with long-term survival in the general adult dutch populationnovelty and significance. Hypertension 67, 281–287.

Niemeijer, M., Xu, X., Dumitrescu, A.V., Gupta, P., van Ginneken, B., Folk, J.C., Abràmoff, M.D., 2011. Automated measurement of the arteriolar-to-venular width ratio in digital color fundus photographs. IEEE Trans. Med. Imaging 30, 1941–1950.

Orlando, J.I., Prokofyeva, E., Blaschko, M.B., 2017. A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images. IEEE Trans. Biomed. Eng. 64, 16–27.

Pekel, E., Pekel, G., 2016. Diagnostic ability of retinal arteriolar diameter measurements in glaucoma. Invest. Ophthalmol. Vis. Sci. 57, 2166.

Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M., Corrado, G., Peng, L., Webster, D., 2018. Predicting cardiovascular risk factors in retinal fundus photographs using deep learning. Nat. Biomed. Eng.

Prokofyeva, E., Zrenner, E., 2012. Epidemiology of major eye diseases leading to blindness in Europe: A literature review. Ophthalmic Res. 47, 171–188.

Qureshi, T.A., Habib, M., Hunter, A., Al-Diri, B., 2013. A manually-labeled, artery/vein classified benchmark for the DRIVE dataset. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 485–488.

Ramm, L., Jentsch, S., Peters, S., Augsten, R., Hammer, M., 2014. Investigation of blood flow regulation and oxygen saturation of the retinal vessels in primary open-angle glaucoma. Graefe's Arch. Clin. Exp. Ophthalmol. 252, 1803–1810.

Relan, D., MacGillivray, T., Ballerini, L., Trucco, E., 2013. Retinal vessel classification: Sorting arteries and veins. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 7396–7399.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation, CoRR abs/1505.04597.

Rothaus, K., Jiang, X., Rhiem, P., 2009. Separation of the retinal vascular graph in arteries and veins based upon structural knowledge. Image Vis. Comput. 27, 864–875, 7th IAPR-TC15 Workshop on Graph-based Representations (GbR 2007).

Sasongko, M.B., Wong, T.Y., Nguyen, T.T., Cheung, C.Y., Shaw, J.E., Wang, J.J., 2011. Retinal vascular tortuosity in persons with diabetes and diabetic retinopathy. Diabetologia 54, 2409–2416.

Seidelmann, S.B., Clagget, B., Bravo, P.E., Gupta, A., Farhad, H., Klein, B.E., Klein, R., Di Carli, M., Solomon, S.D., 2016. Retinal vessel calibers in predicting long-term cardiovascular outcomes. Circulation 134, 1328–1338.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-scale Image Recognition, CoRR abs/1409.1556.

Soares, J.V.B., Leandro, J.J.G., Cesar, R.M., Jelinek, H.F., Cree, M.J., 2006. Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. IEEE Trans. Med. Imaging 25, 1214–1222.

Sofka, M., Stewart, C.V., 2006. Retinal vessel centerline extraction using multiscale matched filters, confidence and edge measures. IEEE Trans. Med. Imaging 25, 1531–1546.

Sörensen, B.M., Houben, A.J.H.M., Berendschot, T.T.J.M., Schouten, J.S.A.G., Kroon, A.A., van der Kallen, C.J.H., Henry, R.M.A., Koster, A., Dagnelie, P.C., Schaper, N.C., Schram, M.T., Stehouwer, C.D.A., 2017. Cardiovascular risk factors as determinants of retinal and skin microvascular function: The maastricht study. PLOS ONE 12, 1–18.

Staal, J., Abràmoff, M., Niemeijer, M., Viergever, M., van Ginneken, B., 2004. Ridge based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging 23, 501–509.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. Computer Vision and Pattern Recognition (CVPR).

Tedeschi-Reiner, E., Strozzi, M., Skoric, B., Reiner, Z., 2005. Relation of atherosclerotic changes in retinal arteries to the extent of coronary artery disease. Am. J. Cardiol. 96, 1107–1109.

Türetken, E., González, G., Blum, C., Fua, P., 2011. Automated reconstruction of dendritic and axonal trees by global optimization with geometric priors. Neuroinformatics 9, 279–302.

Vandewalle, E., Pinto, L.A., Olafsdottir, O.B., Clerck, E.D., Stalmans, P., Calster, J.V., Zeyen, T., Stef'ansson, E., Stalmans, I., 2014. Oximetry in glaucoma: correlation of metabolic change with structural and functional damage. Acta Ophthalmol. (Copenh.) 92, 105–110.

Vazquez, S.G., Cancela, B., Barreira, N., Penedo, M.G., Rodriguez-Blanco, M., Pena Seijo, M., de Tuero, G.C., Barcelo, M.A., Saez, M., 2013. Improving retinal artery and vein classification by means of a minimal path approach. Mach. Vis. Appl. 24, 919–930.

van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T.a., 2014. scikit-image: Image processing in Python. PeerJ 2, e453.

Welikala, R., Foster, P., Whincup, P., Rudnicka, A., Owen, C., Strachan, D., Barman, S., 2017. Automated arteriole and venule classification using deep learning for retinal images from the UK Biobank cohort. Comput. Biol. Med. 90, 23–32.

Wong, T.Y., Islam, F.M.A., Klein, R., Klein, B.E.K., Cotch, M.F., Castro, C., Sharrett, A.R., Shahar, E., 2006. Retinal vascular caliber, cardiovascular risk factors, and inflammation: The multi-ethnic study of atherosclerosis (mesa). Invest. Ophthalmol. Vis. Sci. 47, 2341.

Wong, T.Y., Klein, R., Couper, D.J., Cooper, L.S., Shahar, E., Hubbard, L.D., Wofford, M.R., Sharrett, A.R., 2001. Retinal microvascular abnormalities and incident stroke: the atherosclerosis risk in communities study. The Lancet 358, 1134–1140.

Wong, T.Y., Klein, R., Sharrett, A.R., Duncan, B.B., Couper, D.J., M.,T.J, Klein, B.E., Hubbard, L.D., 2002. Retinal arteriolar narrowing and risk of coronary heart disease in men and women: The atherosclerosis risk in communities study. JAMA 287, 1153–1159.

Xu, X., 2012. Automated delineation and quantitative analysis of blood vessels in retinal fundus image. University of Iowa, Ph.D. thesis.

Xu, X., Ding, W., Abràmoff, M.D., Cao, R., 2017. An improved arteriovenous classification method for the early diagnostics of various diseases in retinal image. Comput. Methods Prog. Biomed. 141, 3–9.

Yedidya, T., Hartley, R., 2012. Tracking of blood vessels in retinal images using Kalman filter. 2008 Digital Image Computing: Techniques and Applications, 52–58.

Zamperini, A., Giachetti, A., Trucco, E., Chin, K.S., 2012. Effective features for artery-vein classification in digital fundus images. Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on, 1–6.