A Three-stage Deep Learning Model for Accurate Retinal Vessel Segmentation

Zengqiang Yan, Xin Yang, and Kwang-Ting (Tim) Cheng, *Fellow, IEEE*

***Abstract*—Automatic retinal vessel segmentation is a fundamental step in the diagnosis of eye-related diseases, in which both thick vessels and thin vessels are important features for symptom detection. All existing deep learning models attempt to segment both types of vessels simultaneously by using a unified pixel-wise loss which treats all vessel pixels with equal importance. Due to the highly imbalanced ratio between thick vessels and thin vessels (namely the majority of vessel pixels belong to thick vessels), the pixel-wise loss would be dominantly guided by thick vessels and relatively little influence comes from thin vessels, often leading to low segmentation accuracy for thin vessels. To address the imbalance problem, in this paper, we explore to segment thick vessels and thin vessels separately by proposing a three-stage deep learning model. The vessel segmentation task is divided into three stages, namely thick vessel segmentation, thin vessel segmentation and vessel fusion. As better discriminative features could be learned for separate segmentation of thick vessels and thin vessels, this process minimizes the negative influence caused by their highly imbalanced ratio. The final vessel fusion stage refines the results by further identifying non-vessel pixels and improving the overall vessel thickness consistency. The experiments on public datasets DRIVE, STARE and CHASE_DB1 clearly demonstrate that the proposed three-stage deep learning model outperforms the current state-of-the-art vessel segmentation methods.**

***Index Terms*—Deep learning, vessel segmentation, imbalance problem, retinal image analysis**

I. Introduction

RETINAL fundus images have been widely used for the diagnosis of eye-related diseases, such as macular degeneration, diabetic retinopathy, and glaucoma. Among various features in fundus images, retinal vessel features play a crucial role such as thin vessels for microaneurysm detection and thick vessels for vessel diameter measurement which are two important biomarkers for the diagnosis of diabetic retinopathy [1]. However, manual annotation of retinal vessels by a human observer is time-consuming, which makes the automated retinal vessel segmentation highly desirable. Generally, current retinal vessel segmentation methods can be classified into two main categories: unsupervised methods and supervised methods.

Unsupervised methods do not utilize any manual annotation for reference, and thus mainly rely on handcrafted features for vessel representation and segmentation. According to the types of features, existing unsupervised approaches can be further categorized into the filter-based [2], [3], [4], [5] and the model-based [6], [7], [8]. For the filter-based methods, Mendonca *et al.* [9] proposed four directional differential operators to detect vessel centerlines, and used an iterative region growing method combined with a morphological filter for vessel segmentation based on those vessel centerlines. Martinez-Perez *et al.* [10] applied a multi-pass region growing method to the first and second spatial derivatives of the corresponding intensity image. Similarly, Zhang *et al.* [11] segmented blood vessels by applying a matched filter to the first-order derivative of the Gaussian filtered image. Fraz *et al.* [12] used the first-order derivative of a Gaussian filter and a multi-directional morphological top-hat operator for feature extraction and vessel segmentation. Lam *et al.* [13] defined a multi-concavity model for vessel segmentation, including a differentiable concavity measure, a line-shape concavity measure, and a locally normalized measure. Azzopardi *et al.* [14] proposed a modified B-COSFIRE filter by combing the difference-of-Gaussian (DoG) filter with the mean shifting operation. Yin *et al.* [15] proposed an orientation-aware detector onto the energy distribution of the corresponding Fourier transformation. Zhang *et al.* [16] transformed fundus images into the lifted domain by wavelet transformation, and used a multi-scale second-order Gaussian filter for vessel segmentation. In terms of the model-based methods, Ali-Diri *et al.* [17] proposed an active contour model using two pairs of contours to locate each vessel edge, and Zhao *et al.* [18] proposed to solve an infinite active contour model by using hybrid region information.

In contrast to unsupervised methods, supervised approaches learn either vessel features or vessel pixel classifiers for segmentation from the annotated training images. Existing supervised methods can be further divided into traditional machine learning based methods and deep learning based methods. Traditional machine learning based methods mainly depend on handcrafted features and utilize typical classifiers for segmentation including the k-nearest neighbor classifier (KNN) [19] and the support vector machine (SVM) [20]. Ricci *et al.* [21] extracted features by a line detector and segmented vessels by using a support vector machine. Lupaşcu *et al.* [22] constructed the feature vector consisting of local intensity, spatial properties and geometry and adopted an AdaBoost classifier for segmentation. Soares *et al.* [23] used the two-dimensional Gabor wavelet transformation response and the pixel intensity as the feature vector, and conducted vessel segmentation based on a Bayesian classifier. Marin *et al.* [24] proposed a 7-D feature vector followed by a neural network for vessel segmentation. Fraz *et al.* [25] defined a feature

Z. Yan and K.-T Cheng are with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clearwater Bay, Kowloon, Hong Kong (e-mail: z.yan@connect.ust.hk, timcheng@ust.hk).

X. Yang is with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China (e-mail: xinyang2014@hust.edu.cn).

vector consisting of gradient, morphology, line strength and Gabor filter response, and used an ensemble system of bagged and boosted decision trees for vessel segmentation. Different from traditional machine learning based methods, deep learning based methods [26], [27], [28], [29] have shown strong ability to automatically learn features for accurate vessel segmentation. Orlando *et al.* [30] proposed a discriminatively trained fully connected conditional random field model for vessel segmentation. Fu *et al.* [31], [32] proposed to solve the vessel segmentation problem based on a fully convolutional neural network combined with a fully-connected Conditional Random Fields (CRFs). Li *et al.* [33] remolded the vessel segmentation task as a cross-modality data transformation problem, which was further modeled by training a deep learning model. Dasgupta [34] proposed a deep learning model to iteratively classify each pixel in fundus images. In [35], several deep learning architectures have been tested for vessel segmentation.

All existing deep learning models are trained by using a unified pixel-wise loss for the segmentation of both thick and thin vessels simultaneously. In the pixel-wise loss, all vessel pixels are treated with equal importance. However, due to the fact that the majority of vessel pixels in fundus images belong to thick vessels, the pixel-wise loss would penalize more on thick vessels than thin vessels, which in turn would guide the deep learning models for better segmentation of thick vessels to minimize the overall pixel-wise losses. Consequently, the trained deep learning models are able to learn robust features for the segmentation of thick vessels, while the ability for the segmentation of thin vessels is relatively limited. In this paper, we propose to further divide the vessel segmentation task into three sub-tasks: thick vessel segmentation, thin vessel segmentation, and vessel fusion. As sub-tasks of segmenting thick vessels and thin vessels are implemented by separate deep learning models, better discriminative features can be learned for thick vessels and thin vessels respectively. By adopting a deep learning model to automatically fuse the segmented thick vessels and thin vessels, the proposed three-stage deep learning model is able to achieve accurate vessel segmentation for both types of vessels. Experimental results on multiple public datasets demonstrate that the three-stage deep learning model considerably outperforms the current state-of-the-art methods.

The paper is organized as follows. A thorough analysis of the retinal vessel segmentation problem is presented in Section II. Section III presents details of the proposed three-stage deep learning framework. In Section IV, we evaluate the effectiveness of the proposed three-stage framework on public datasets. Section IV provides a discussion about the proposed model on dealing with challenging cases and Section V concludes the paper.

II. Problem Analysis

In retinal vessel segmentation, both thick vessels and thin vessels are important features for symptom detection in the diagnosis of eye-related diseases. However, accurately segmenting both thick and thin vessels simultaneously is challenging due to the following reasons:

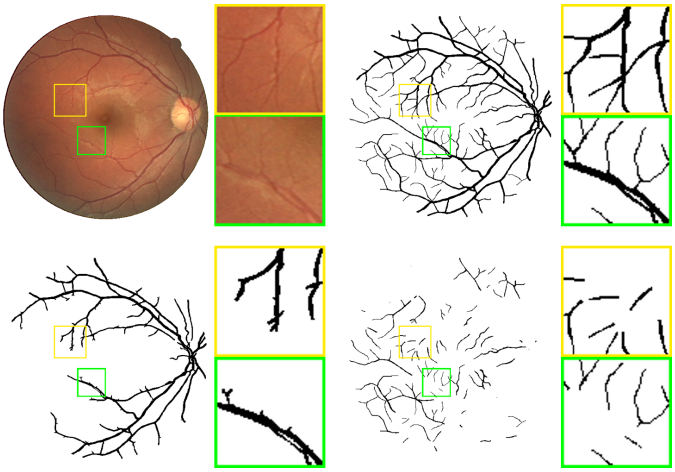

Fig. 1: Analysis of the retinal vessel segmentation problem. Row 1: from left to right, the fundus image and the enlarged patches, and the manual annotation and the annotations of the two fundus image patches. Row 2: from left to right, the manually annotated thick vessels and the annotated thick vessels in the two fundus image patches, and the manually annotated thin vessels and the annotated thin vessels in the two fundus image patches.

1) **Imbalance:** In fundus images, the majority of vessel pixels belong to thick vessels. Given the manual annotation in Fig. 1, if the vessels whose thickness is less than 3 pixels are denoted as "thin vessels" and the rest are denoted as "thick vessels", nearly 77% of vessel pixels belong to the "thick vessels" and the "thin vessels" only account for 23%. As all existing deep learning models are trained by pixel-wise losses which treat all vessel pixels with equal importance, the trained models tend to more accurately segment thick vessels when minimizing the overall losses while the segmentation of thin vessels would be less important as analyzed in [36]. As a result, using a pixel-wise loss to simultaneously segment both thick and thin vessels would put thin vessels at a disadvantage.
2) **Feature difference:** From the two enlarged patches and the corresponding annotated thick and thin vessels in Fig. 1, one basic observation is that thick vessels usually have much higher contrast and signal to noise ratio (SNR) than those of thin vessels. Therefore, the features for segmenting thick vessels might not be applicable to effectively segment thin vessels.

Based on the above two reasons, simultaneously segmenting both thick and thin vessels using pixel-wise losses would result in better segmentation performance for thick vessels than thin vessels.

III. Methodology

Figure 2 illustrates the overall framework of the proposed three-stage deep learning model, which consists of three separate models, namely *ThickSegmenter* for thick vessel segmentation, *ThinSegmenter* for thin vessel segmentation and *FusionSegmenter* for vessel fusion respectively. The three models

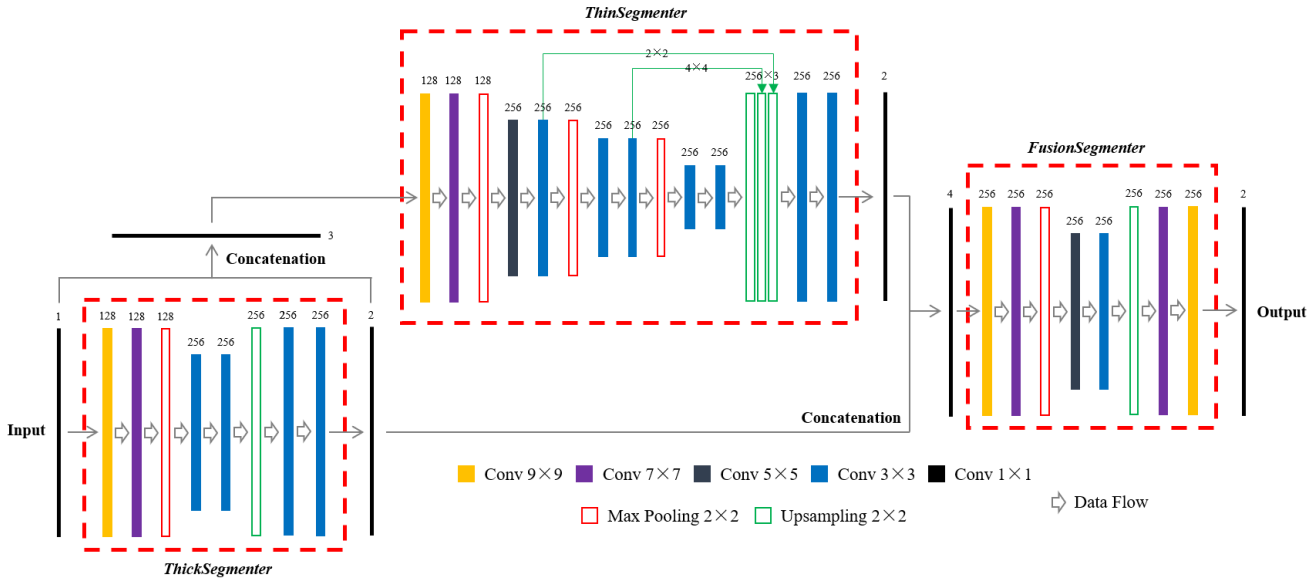

Fig. 2: The overview of the proposed three-stage deep learning framework. The framework consists of three separate models, namely *ThickSegmenter* for thick vessel segmentation, *ThinSegmenter* for thin vessel segmentation and *FusionSegmenter* for vessel fusion respectively.

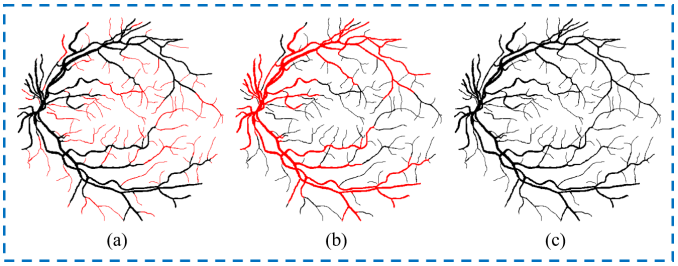

Fig. 3: Manual annotations, denoted in black, used for training different models. Red pixels represent those vessel pixels that are not counted for loss calculation and back propagation. From left to right: (a) the annotated thick vessels for training *ThickSegmenter*, (b) the annotated thin vessels for the training of *ThinSegmenter*, and (c) the annotated vessels for training *FusionSegmenter*.

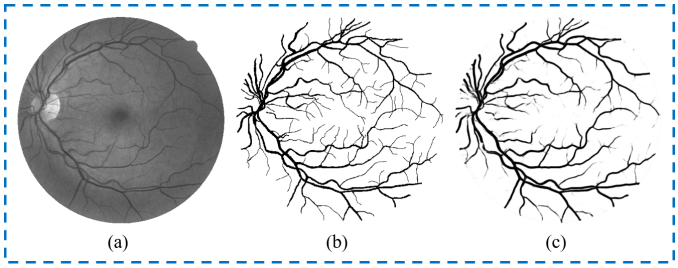

Fig. 4: Exemplar results of thick vessel segmentation obtained by *ThickSegmenter*. From left to right: (a) the input fundus image for training *ThickSegmenter*, (b) the complete manual annotation (with both thick and thin vessels), and (c) the predicted thick vessels by *ThickSegmenter*.

(i.e. *ThickSegmenter*, *ThinSegmenter* and *FusionSegmenter*) are trained separately and sequentially, and the corresponding annotations required for training different models are shown in Fig. 3. Details of the training strategy are as follows:

1) **Training *ThickSegmenter*:** For training the *ThickSegmenter* model, only those annotated thick vessels, as shown in black in Fig. 3(a), are used as ground truth for training. That is, the losses generated by the thin vessels are ignored and only the losses of thick vessels are used for back propagation. The input for training *ThickSegmenter* is the green channel of each fundus image (shown in Fig. 4(a)), and the output is the predicted thick vessels (shown in Fig. 4(c)).
2) **Training *ThinSegmenter*:** To train *ThinSegmenter*, only those annotated thin vessels, as shown in black in Fig. 3(b), are used as ground truth. That is, the losses generated by the thick vessels are ignored and only the losses of thin vessels are used for back propagation. The input for training *ThinSegmenter* is the concatenation of the green channel of each fundus image and the predicted thick vessels by *ThickSegmenter*. The output is the predicted thin vessels.
3) **Training *FusionSegmenter*:** For the *FusionSegmenter* model, all annotated vessels are used for training as shown in Fig. 3(c). The input for training *FusionSegmenter* is the concatenation of the predicted thick vessels by *ThickSegmenter* and the predicted thin vessels by *ThinSegmenter*. The output of *FusionSegmenter* is the final predicted probability map.

We employ the widely used pixel-wise cross-entropy loss function to train each model in the proposed three-stage framework.

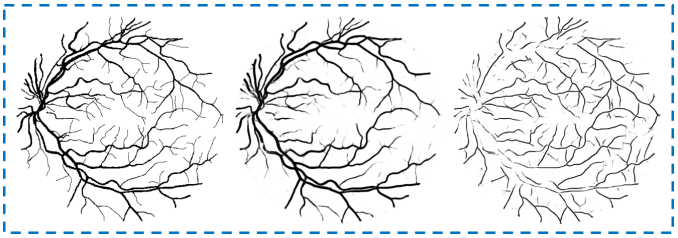

Fig. 5: Exemplar results of thin vessel segmentation by *ThinSegmenter*. From left to right: the complete manual annotation (with both thick and thin vessels), the predicted thick vessels by *ThickSegmenter* and the predicted thin vessels by *ThinSegmenter*.

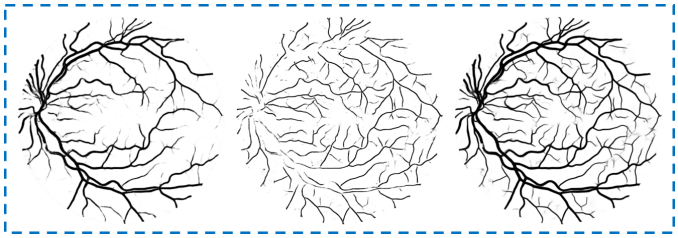

Fig. 6: Exemplar results of vessel fusion obtained by *FusionSegmenter*. From left to right: the predicted thick vessels by *ThickSegmenter*, the thin vessels generated by *ThinSegmenter* and the final vessel segmentation obtained by *FusionSegmenter*.

A. ThickSegmenter

Based on the observation that thick vessels generally have higher contrast and SNR than thin vessels as discussed in Section II, we directly adopt a model with only one pooling layer to extract robust local features as shown in Fig. 2. From the segmentation results generated by *ThickSegmenter* as shown in Fig. 4, we find that the simple model can successfully segment almost all thick vessels but few thin vessels, which demonstrates that the features for the segmentation of thick vessels might not be applicable for segmenting thin vessels.

B. ThinSegmenter

For thin vessel segmentation, we adopt a simplified *FCN* model [37] which contains multiple pooling layers for global feature extraction, based on the above observation that thin vessels usually have relatively lower local contrast and SNR. In addition to the fundus image, we further utilize the segmented thick vessels by *ThickSegmenter* as another input for training *ThinSegmenter*. This is because that most thin vessels are connected with thick vessels as shown in the manual annotation of Fig. 5, and using the segmented thick vessels as guidance would help differentiate thin vessels from those non-vessel pixels. Here, the fundus image and the predicted thick vessels by *ThickSegmenter* are concatenated together as input to train *ThinSegmenter*. As a result, the probability map generated by *ThinSegmenter* is quite "clean" as shown in Fig. 5, which means that vessel pixels and non-vessel pixels are effectively separated. Comparing the generated probability map to the corresponding manual annotation, we find that most thin vessels are successfully detected and the only problem is the thickness inconsistency between the segmented thin vessels and those annotated thin vessels. The thickness inconsistency problem mainly is due to the resolution limitation of fundus images. In addition, in the generated probability map in Fig. 5, only partial thick vessels are detected by *ThinSegmenter*, which further validates the fact that thick vessels and thin vessels have distinct feature characteristics and thus the features for segmenting thin vessels might not be applicable for the segmentation of thick vessels.

C. FusionSegmenter

FusionSegmenter refines the results to further improve the overall vessel thickness consistency, as the segmented thin vessels produced by *ThinSegmenter* usually are thicker than those annotated vessels. Comparing the segmented vessels in the final probability map by *FusionSegmenter* with those previously segmented vessels by *ThickSegmenter* and *ThinSegmenter* in Fig. 6, the fused vessels indeed have better thickness consistency with those manually annotated vessels.

IV. Evaluation

A. Datasets

Three public datasets DRIVE [19], STARE [38] and CHASE_DB1 [39] are used for evaluation.

DRIVE contains 40 equal-sized fundus images (565×584 pixels) with a 45° FOV. Among all images, 7 of them contain pathology. The dataset is officially and equally split into the training set and the test set. For the training set, only one manual annotation is provided for each image, while for the test set two manual annotations from two observers are provided. We utilize the same evaluation strategy as other methods and leverage the annotations by the first observer as the ground truth for performance measurement.

STARE consists of 20 equal-sized images (700×605 pixels). Among all images, 10 of them contain pathology. As training and test sets are not explicitly specified, the same leave-one-out cross validation is adopted for performance evaluation, where models are iteratively trained on 19 images and tested on the rest image. Same as other methods, manual annotations generated by the first observer are used for both training and test. Since FOV masks are not provided and there exists no uniform generation method, the masks provided in [24] and [30] are used for comparison.

CHASE_DB1 comprises 28 equal-sized fundus images (999×960 pixels) with a 30° FOV. We adopt the same split strategy as described in [33] to divide the entire dataset into training and test sets. That is the first 20 images are selected for training and the rest 8 images are used for testing.

B. Preprocessing

To reduce the training complexity, each fundus image in the training set is converted to gray scale by extracting the green channel. Then, each fundus image is cropped into 128×128 patches, and those patches in which the ratio of background pixels (pixels located outside the FOV mask) is greater than

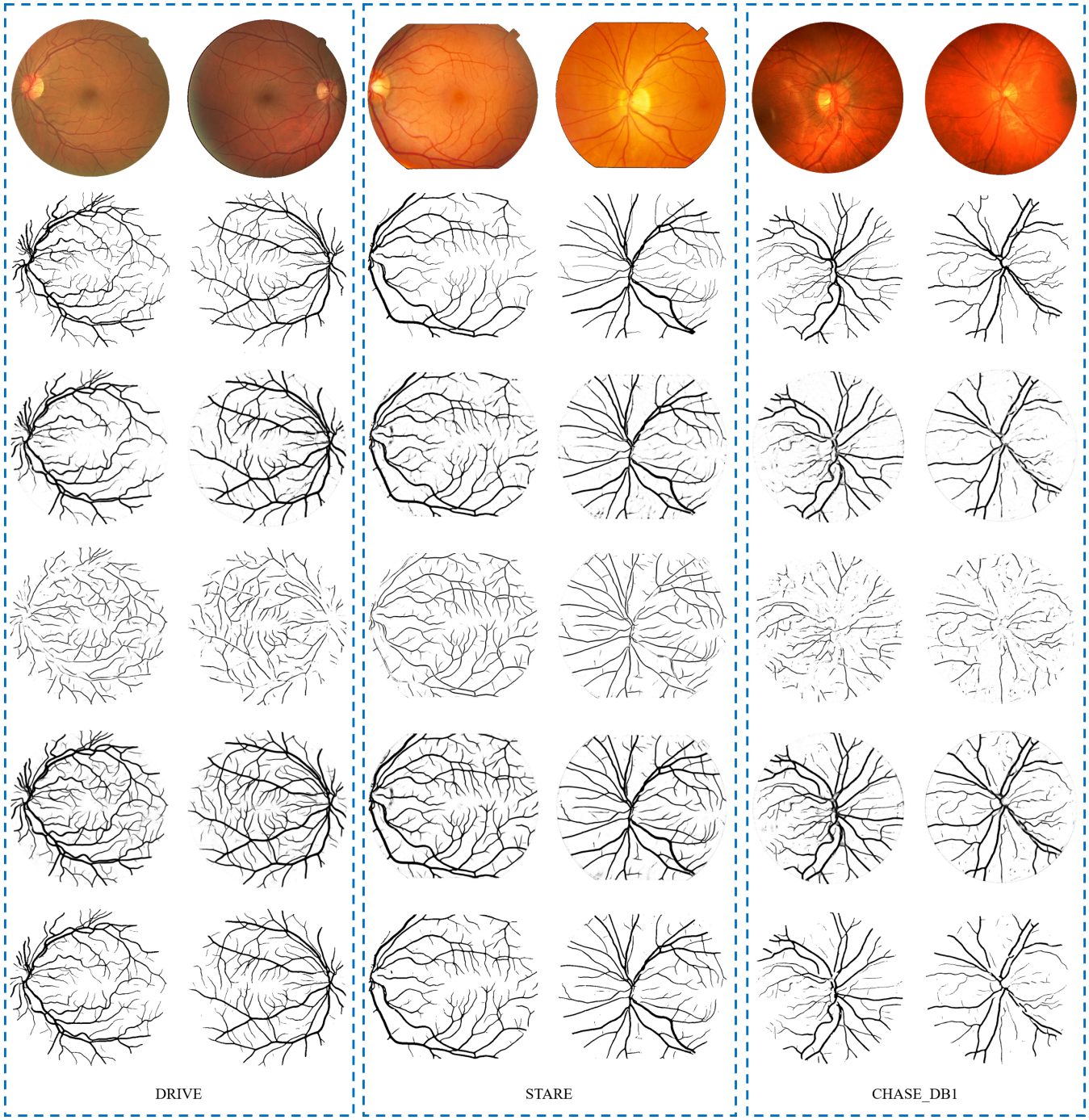

Fig. 7: Exemplar vessel segmentation results by the proposed three-stage deep learning model. From top to bottom: the fundus images, the manual annotations, the probability maps of thick vessels, the probability maps of thin vessels, the final probability maps and the corresponding hard segmentation maps.

50% are further discarded. To enlarge the training set, multiple data augmentation strategies have been utilized, including flipping, rotation, resizing and enhancing contrast.

To separate thick vessels and thin vessels for training different models, given a manual annotation, we first extract the corresponding skeletons by applying the skeletonization method [40]. Then, for each skeleton pixel, we calculate the minimum inscribed circle centered at the pixel which is completely covered by vessel pixels and use the diameter as its vessel thickness. For those skeleton pixels whose vessel thickness is below a fixed threshold, all pixels covered by the minimum inscribed circle are denoted as thin vessel pixels. The rest vessel pixels are classified as thick vessel pixels.

C. Implementation Details

The proposed three-stage deep learning model was implemented based on the open-source deep learning library Caffe [41]. The initial learning rate was set at 10^{-4} and decreased by

TABLE I: Comparison Results on the DRIVE, STARE and CHASE_DB1 Datasets

		DRIVE				STARE				CHASE_DB1			
Methods	Year	*Se*	*Sp*	*Acc*	*AUC*	*Se*	*Sp*	*Acc*	*AUC*	*Se*	*Sp*	*Acc*	*AUC*
2nd Human Observer	-	0.7760	0.9724	0.9472	-	0.8952	0.9384	0.9349	-	0.8105	0.9711	0.9545	-
Zhang [11]	2010	0.7120	0.9724	0.9382	-	0.7177	0.9753	0.9484	-	-	-	-	-
Fraz [12]	2012	0.7152	0.9759	0.9430	-	0.7311	0.9680	0.9442	-	-	-	-	-
Roychowdhury [4]	2015	0.7395	0.9782	0.9494	0.9672	0.7317	0.9842	0.9560	0.9673	0.7615	0.9575	0.9467	0.9623
Azzopardi [14]	2015	0.7655	0.9704	0.9442	0.9614	0.7716	0.9701	0.9497	0.9563	0.7585	0.9587	0.9387	0.9487
Yin [15]	2015	0.7246	0.9790	0.9403	-	**0.8541**	0.9419	0.9325	-	-	-	-	-
Zhang [16]	2016	0.7743	0.9725	0.9476	0.9636	0.7791	0.9758	0.9554	0.9748	0.7626	0.9661	0.9452	0.9606
You [20]	2011	0.7410	0.9751	0.9434	-	0.7260	0.9756	0.9497	-	-	-	-	-
Marin [24]	2011	0.7067	0.9801	0.9452	0.9588	0.6944	0.9819	0.9526	0.9769	-	-	-	-
Fraz [25]	2012	0.7406	0.9807	0.9480	0.9747	0.7548	0.9763	0.9534	0.9768	0.7224	0.9711	0.9469	0.9712
Li [33]	2016	0.7569	0.9816	0.9527	0.9738	0.7726	0.9844	0.9628	**0.9879**	0.7507	0.9793	0.9581	0.9716
Fu [31]	2016	0.7603	-	0.9523	-	0.7412	-	0.9585	-	0.7130	-	0.9489	-
Orlando [30]	2017	**0.7897**	0.9684	-	-	0.7680	0.9738	-	-	0.7277	0.9715	-	-
Dasgupta [34]	2017	0.7691	0.9801	0.9533	0.9744	-	-	-	-	-	-	-	-
Proposed	2018	0.7631	**0.9820**	**0.9538**	**0.9750**	0.7735	**0.9857**	**0.9638**	0.9833	**0.7641**	**0.9806**	**0.9607**	**0.9776**

a factor of 10 every 20000 iterations until it reached 10^{-6}. For quality evaluation, we adopted the same method to select the threshold to generate the corresponding binary segmentation map from a given probability map as described in [33], where the optimal threshold was set as the threshold maximizing the overall accuracy on the training set.

D. Evaluation Metrics

Given a binary segmentation map and the corresponding manual annotation, correctly detected vessel pixels are denoted as true positives (*TP*) and those wrongly classified as non-vessel pixels are counted as false negatives (*FN*). Similarly, correctly segmented non-vessel pixels are denoted as true negatives (*TN*) and those incorrectly detected as vessel pixels are counted as false positives (*FP*). Then, sensitivity (Se), specificity (Sp) and accuracy (Acc) are defined as

$$Se = \frac{TP}{TP+FN}, Sp = \frac{TN}{TN+FP}, Acc = \frac{TP+TN}{N}, \quad (1)$$

where $N = TN + TP + FN + FP$. The receiving operator characteristics (ROC) curve is computed with the true positive ratio (Se) versus the false positive ratio ($1-Sp$) with respect to a varying threshold, and the area under the ROC curve (AUC) is calculated for quality evaluation.

E. Experimental Results

Exemplar results generated by the proposed three-stage deep learning model are shown in Fig. 7. The probability maps in Fig. 7 show that the proposed model is able to effectively segment both thick and thin vessels. In addition, the probability maps demonstrate the effectiveness of the proposed model in classifying non-vessel pixels from vessel pixels. It should be pointed out that the produced probability maps contain some vessels which were not annotated by one human observer while labeled as vessels by the other observer, and thus these vessels should be regarded as true vessels.

Comparison results of both the proposed model and the current state-of-the-art methods are summarized in Table I. For the DRIVE dataset, the proposed three-stage model achieves 0.7631, 0.9820, 0.9538 and 0.9750 for Se, Sp, Acc and AUC respectively, among which the results for Sp, Acc and AUC are better than all the current state-of-the-art methods. In terms of Se, Orlando [30] achieves the best results but the scores for Sp, Acc and AUC are much lower. Considering the highly imbalanced ratio between vessel pixels and non-vessel pixels, the overall performance of the proposed model is much better.

For the STARE dataset, the proposed three-stage deep learning model achieves 0.7735, 0.9846, 0.9638 and 0.9833 for Se, Sp, Acc and AUC respectively, among which the results for Sp and Acc are better than other methods and the score for Se is the best among all supervised methods. Compared to the results by Yin [15], though the score for Se achieved by the proposed model is 0.0806 lower, the scores for Sp and Acc are 0.0427 and 0.0313 higher. Thus, the proposed model has better overall performance.

For the CHASE_DB1 dataset, according to the objective results in Table I, the proposed three-stage deep learning model achieves 0.7641, 0.9806, 0.9607 and 0.9776 for Se, Sp, Acc and AUC respectively, which consistently outperforms all the current state-of-the-art methods. The comparison results in Table I indicate that the proposed model consistently achieves better results on different datasets, demonstrating the robustness of the proposed model.

We further assess the extendibility of the proposed three-stage deep learning model by conducting the cross-training evaluation as in [33]. Different from the cross-training method in [33] which retrained the deep learning model, we directly apply the deep learning model trained on one dataset to other datasets without retraining. Exemplar results of the cross-training evaluation experiment are shown in Fig. 8. For the DRIVE dataset, when transferring the model trained on the STARE dataset, though most thick vessels are successfully detected, the majority of thin vessels are missing. It is because that the manual annotations for the STARE dataset mainly contain thick vessels. In the proposed three-stage deep learning model, as the thin vessel segmentation results would further affect the vessel fusion results, the influence of those missing thin vessels would be slightly magnified. As a result, when transferring the model trained on the STARE dataset onto the DRIVE dataset, the corresponding scores for Se, Sp, Acc

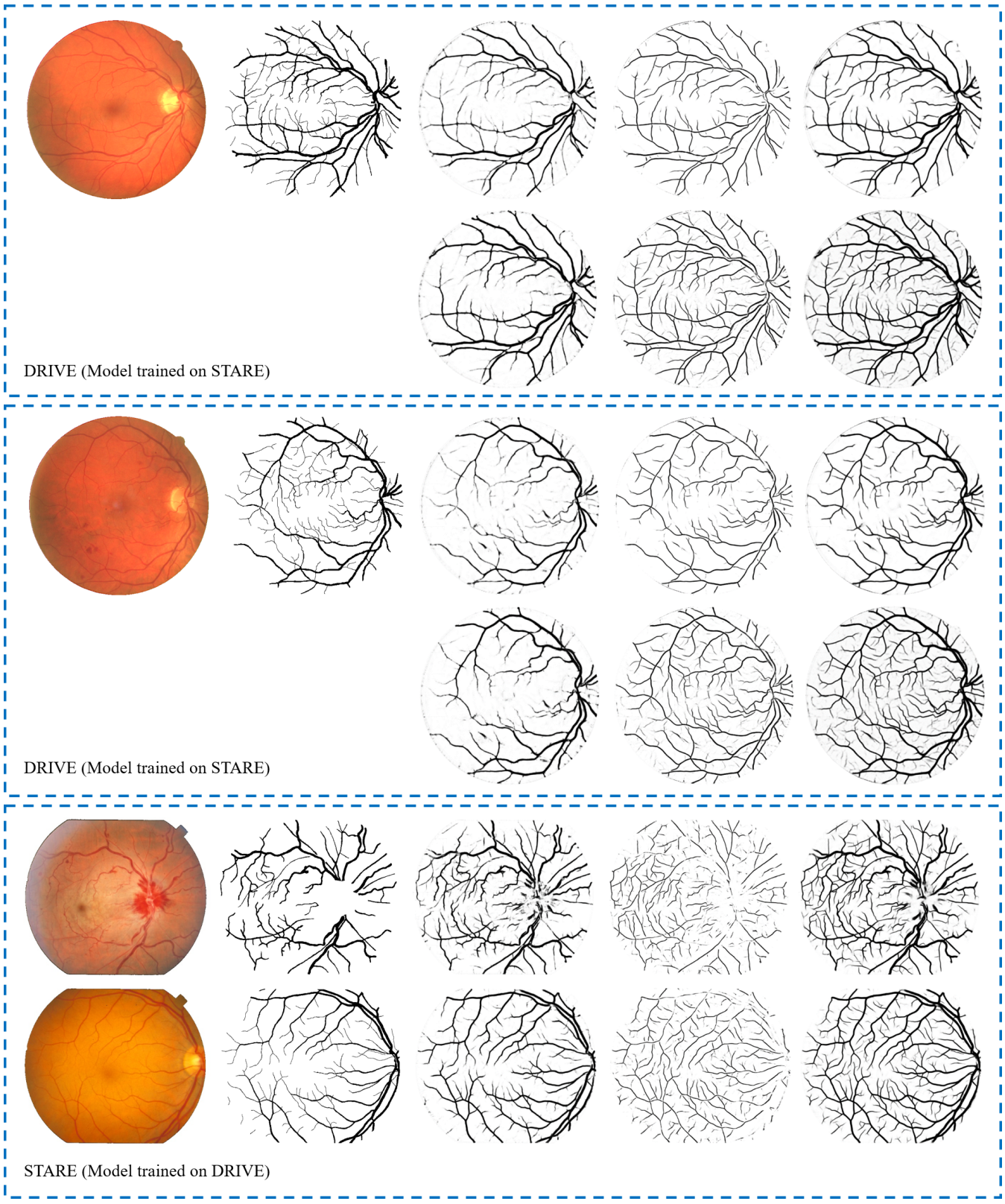

Fig. 8: Exemplar results of the cross-training evaluations on datasets DRIVE and STARE. For each example of the DRIVE dataset (from left to right), Row 1: the fundus image, the manual annotation, the probability map of thick vessels, the probability map of thin vessels and the final probability map generated by the originally trained model on the STARE dataset; Row 2: the corresponding probability maps generated by the retrained model using the other manual annotations. For the STARE dataset (from left to right), Rows 1&2: the fundus images, the manual annotations, the probability maps of thick vessels, the probability maps of thin vessels and the final probability maps generated by the originally trained model on the DRIVE dataset.

TABLE II: Results of the cross-training evaluation

Test Dataset	Methods	*Se*	*Sp*	*Acc*	*AUC*
DRIVE	Soares [23]	-	-	0.9397	-
	Ricci [21]	-	-	0.9266	-
	Marin [24]	-	-	0.9448	-
	Fraz [25]	0.7242	0.9792	0.9456	0.9697
	Li [33]	0.7273	0.9810	0.9486	0.9677
	Proposed	**0.7014**	**0.9802**	**0.9444**	**0.9568**
	Proposed	*0.7443*	*0.9814*	*0.9509*	*0.9720*
STARE	Soares [23]	-	-	0.9327	-
	Ricci [21]	-	-	0.9464	-
	Marin [24]	-	-	0.9528	-
	Fraz [25]	0.7010	0.9770	0.9495	0.9660
	Li [33]	0.7027	0.9828	0.9545	0.9671
	Proposed	**0.7319**	**0.9840**	**0.9580**	**0.9678**

Italic results represent the results achieved by the retrained model on the STARE dataset by using the manual annotations which contain both thick vessels and thin vessels.

and AUC are slightly lower than other methods. To better evaluate the extendibility of the proposed model, we retrain the model on the STARE dataset by using the annotations provided by the other human observer which contain both thick and thin vessels. As shown in Fig. 8, when transferring the retrained model onto the DRIVE dataset, more thin vessels are successfully detected while the segmentation results of thick vessels are quite similar. Therefore, the corresponding scores achieved by the retrained model are 0.7443, 0.9814, 0.9509 and 0.9720 for Se, Sp, Acc and AUC respectively, all better than the original results and those of other methods.

Conversely, as the manual annotations for the DRIVE dataset contain more thin vessels, transferring the model trained on the DRIVE dataset onto the STARE dataset is able to segment more thin vessels as shown in Fig. 8. Comparing the produced probability maps to the original fundus images, we find that most segmented thin vessels are indeed true vessels. As highlighted in the previous analysis, successfully segmenting thin vessels would further positively affect the vessel fusion results, especially improving the overall thickness consistency. Therefore, when transferring the model trained on the DRIVE dataset onto the STARE dataset, the corresponding scores for Se, Sp, Acc and AUC are 0.7319, 0.9840, 0.9580, 0.9678, all better than other methods.

F. Evaluation on Thin Vessel Segmentation

In this section, we evaluate the effectiveness of the proposed model on thin vessel segmentation by comparing it with the state-of-the-art model in [29]. Before focusing the evaluation on thin vessel segmentation, we first compare the overall performance based on the above-mentioned evaluation metrics and the evaluation metric used in [29], i.e. *Dice Coefficient*. According to the reported results in [29], the optimal score of *Dice Coefficient* on the DRIVE dataset is 0.8210 and the corresponding score of the proposed model is 0.8141, which is slightly lower. To better compare the performance, we conducted additional experiments based on the probability maps provided by [29]. For both our method and [29], the Otsu automatic thresholding algorithm is applied to the probability maps for generating the final segmentation results (as the optimal threshold to reproduce the results in [29] is unknown). Exemplar results of such adaptive-thresholding experiment are shown in Fig. 9. Visually, the probability maps generated by the proposed model are much "cleaner" than those probability maps in [29]. In other words, the proposed model can better distinguish vessel pixels from non-vessel pixels, which helps achieve better thickness consistency compared to the annotated vessels, especially for those thin vessels. When applying the above evaluation metrics to the segmentation maps generated by the proposed model and the model in [29] on the DRIVE dataset, the proposed model achieves better performance in terms of Sp and Acc as shown in Table III. The decrease in Se is mainly due to that the segmented vessels in the probability maps of [29] are much thicker than those segmented vessels by the proposed model, and thus result in a higher Se and a lower Sp. In comparison, the segmented vessels by the proposed model are thinner and have better thickness consistency. As a result, small location variations of those segmented thin vessels would lead to a non-trivial decrease in Se. In fact, small location variations of thin vessels among observers are quite common and acceptable, and this issue has been extensively discussed in [36].

According to the analysis in [36], the above evaluation metrics would emphasize more on the segmentation of thick vessels and less on that of thin vessels. Therefore, we fur-

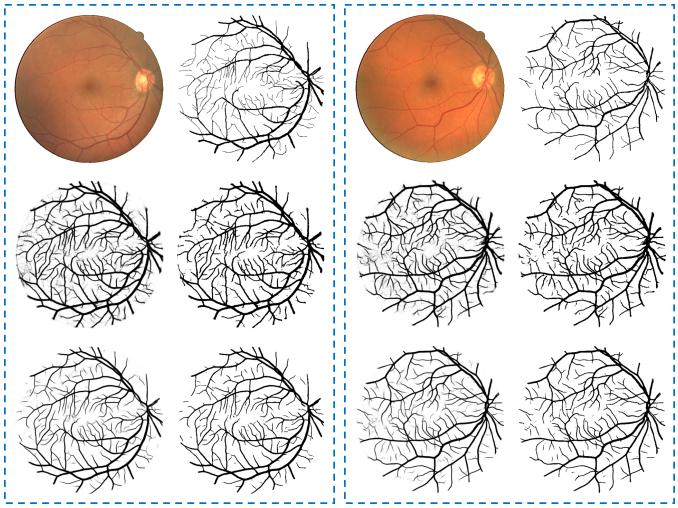

Fig. 9: Exemplar results of the adaptive-thresholding experiment on the DRIVE dataset. For each group (from top to bottom): Row 1: the fundus images and the corresponding manual annotations; Row 2: the probability maps generated by [29] and the corresponding hard segmentation maps by applying the Otsu automatic thresholding algorithm; Row 3: the probability maps obtained by the proposed model and the corresponding hard segmentation maps by applying the Otsu automatic thresholding algorithm.

TABLE III: Comparison results with Maninis [29] on the adaptive-thresholding experiment on the DRIVE Dataset

Vessels	Methods	*Se*	*Sp*	*Acc*	*CAL*
All	Maninis [29]	**0.9548**	0.8999	0.9069	0.7862
	Proposed	0.8939	**0.9466**	**0.9399**	**0.8224**

ther implement the $f(C, A, L)$ function proposed in [42] to emphasize thick and thin vessels more equally in the evaluation. In the $f(C, A, L)$ function, parameter C penalizes fragmented segmentations by comparing the number of connected segments in the manual annotation and in the generated segmentation map. Parameter A measures the degree of overlapping areas between the manual annotation and the generated segmentation map. Parameter L compares the lengths of vessels in the manual annotation and in the segmentation map. Obviously, the $f(C, A, L)$ function could better balance the relative importance between thick vessels and thin vessels for quality evaluation. According to the results in Table III, the proposed model can achieve 0.8224 for $f(C, A, L)$ while the corresponding score obtained by [29] is 0.7862, which offers another perspective for comparison. The above experimental results demonstrate that the proposed model achieves better overall performance compared to the model of [29].

Fig. 10: Vessel separation for quality evaluation on thin vessel segmentation. From left to right: the manual annotation, the identified thin vessels, and the defined vessel and non-vessel pixels (gray regions represent non-vessel pixels and black regions are vessel pixels) for quality evaluation of thin vessel segmentation.

TABLE IV: Comparison results with Maninis [29] on thin vessel segmentation on the DRIVE dataset

Vessels	Methods	*Se*	*Sp*	*Acc*
Thin	Maninis [29]	**0.9017**	0.6745	0.7255
	Proposed	0.8170	**0.8115**	**0.8127**

In addition to the implemented $f(C, A, L)$ function, we further evaluate the effectiveness of the proposed model on thin vessel segmentation by designing specific metrics. Given a manual annotation, we segment the entire vessel tree into small vessel segments and identify those vessel segments whose average thickness is less than 3 pixels. These identified vessels are considered as thin vessels as shown in Fig. 10. For each thin vessel segment, a 3-pixel range is assigned and the metrics Se, Sp and Acc are calculated based on pixels located within the range for evaluation (i.e., we only focus on the pixels in the 3-pixel range). Table IV shows the comparison results, which reaches similar conclusion to Table III. As noted before, those segmented thin vessels in the probability maps generated by [29] are much thicker which helps achieve a higher Se. However, the performance of Sp and Acc would be negatively impacted due to the false positives. From the comparison results on thin vessel segmentation, we conclude that the proposed model achieves better performance especially on thickness consistency and non-vessel pixels identification.

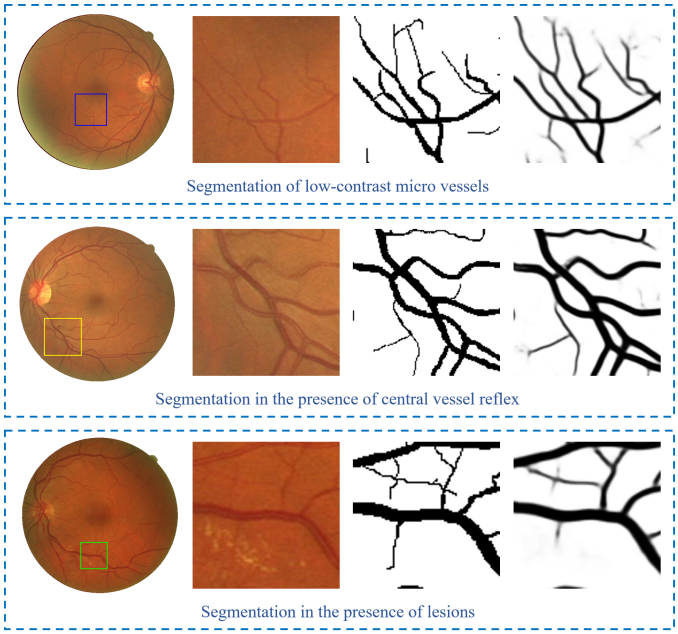

Fig. 11: Exemplar results in dealing with challenging cases. For each row, from left to right: the fundus images, the enlarged fundus image patches, the manual annotations and the probability maps generated by the proposed three-stage deep learning model.

V. Discussion

Though retinal vessel segmentation is an active research field, there still exist several key problems that need to be addressed: 1) segmentation of low-contrast micro vessels; 2) vessel segmentation in the presence of central vessel reflex; and 3) vessel segmentation in the presence of lesions. Exemplar results on these challenging cases are shown in Fig. 11. As the proposed three-stage deep learning framework contains a separate model for thin vessel segmentation, those low-contrast micro vessels can be effectively detected. In addition, the produced probability map also contains true vessels that are not annotated by the human observer. In the proposed three-stage deep learning model, thick vessels are first segmented by the *ThickSegmenter* model and then further refined by the *FusionSegmenter* model with guidance of the corresponding thin vessels. Therefore, the problem introduced by the presence of central vessel reflex can be largely solved as shown in Row 2 of Fig. 11. As the presence of lesions mainly affects the local features for thick vessel segmentation, the influence can be effectively removed by the *ThinSegmenter* model and the *FusionSegmenter* model. As a result, the presence of lesions would have little effect on the final vessel segmentation results. In summary, the proposed three-stage deep learning model can effectively address these challenging cases, as demonstrated by the experimental results.

VI. Conclusion

In this paper, we address the retinal vessel segmentation problem by considering thick and thin vessels separately.

Specifically, we propose to divide the retinal vessel segmentation task into three sub-tasks, each of which trains a deep learning model using a unique pixel-wise loss, so that the segmentation of thick and thin vessels can be separately implemented. Experimental results on multiple public datasets demonstrate that the proposed three-stage deep learning model outperforms the current state-of-the-art methods. Results on cross-training evaluation and challenging cases demonstrate excellent generalization ability and robustness of the proposed model.

REFERENCES

[1] M. K. Ikram et al., "Retinal vessel diameters and risk of impaired fasting glucose or diabetes," *Diabetes*, vol. 55, no. 2, pp. 506-510, 2006.

[2] G. B. Kande, P. V. Subbaiah, and T. S. Sacithri, "Unsupervised fuzzy based vessel segmentation in pathological digital fundus images," *J. Med. Syst.*, vol. 34, no. 5, pp. 849-858, 2010.

[3] T. Chakraborti, D. K. Jha, A. S. Chowdhury, and X. Jiang, "A self-adaptive matched filter for retinal blood vessel detection," *Mach. Vision Appl.*, vol. 26, no. 1, pp. 55-68, 2014.

[4] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, "Iterative vessel segmentation of fundus images," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 7, pp. 1738-1749, Jul. 2015.

[5] X. Yang, K. -T. Cheng, and A. Chien, "Accurate vessel segmentation with progressive contrast enhancement and canny refinement," in *Proc. ACCV*, 2014, pp. 1-16.

[6] W. Li, A. Bhalerao, and R. Wilson, "Analysis of retinal vasculature using a multiresolution hermite model," *IEEE Trans. Med. Imag*, vol. 26, no. 2, pp. 137-152, Feb. 2007.

[7] B. S. Y. Lam and Y. Hong, "A novel vessel segmentation algorithm for pathological retina images based on the divergence of vector fields," *IEEE Trans. Med. Imag*, vol. 27, no. 2, pp. 237-246, Feb. 2008.

[8] H. Narasimha-Iyer, J. M. Beach, B. Khoobehi, and B. Roysam, "Automatic identification of retinal arteries and veins from dual-wavelength images using structural and functional features," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 8, pp. 1427-1435, Aug. 2007.

[9] A. Mendonça and A. Campilho, "Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction," *IEEE Trans. Med. Imag.*, vol. 25, no. 9, pp. 1200-1213, Sep. 2006.

[10] M. Martinez-Perez, A. Hughes, S. Thom, A. Bharath, and K. Parker, "Segmentation of blood vessels from red-free and fluorescein retinal images," *Med. Image Anal.*, vol. 11, no. 1, pp. 47-61, 2007.

[11] B. Zhang, L. Zhang, L. Zhang, and F. Karray, "Retinal vessel extraction by matched filter with first-order derivative of Gaussian," *Comput. Biol. Med.*, vol. 40, no. 4, pp. 438-445, 2010

[12] M. M. Fraz, S. A. Barman, P. Remagnino, A. Hoppe, A. Basit, B. Uyyanonvara, A. R. Rudnicka, and C. G. Owen, "An approach to localize the retinal blood vessels using bit planes and centerline detection," *Comput. Methods Programs Biomed.*, vol. 108, no. 2, pp. 600-616, 2012.

[13] B. S. Y. Lam, Y. Gao, and A. W. -C. Liew, "General retinal vessel segmentation using regularization-based multiconcavity modeling," *IEEE Trans. Med. Imag.*, vol. 29, no. 7, pp. 1369-1381, Jul. 2010.

[14] G. Azzopardi, N. Strisciuglio, M. Vento, and N. Petkov, "Trainable COSFIRE filters for vessel delineation with application to retinal images," *Med. Image Anal.*, vol. 19, no. 1, pp. 46C57, 2015.

[15] B. Yin, H. Li, B. Sheng, X. Hou, Y. Chen, W. Wu, P. Li, R. Shen, Y. Bao, and W. Jia, "Vessel extraction from non-fluorescein fundus images using orientation-aware detector," *Med. Image Anal.*, vol. 26, no. 1, pp. 232-242, 2015.

[16] J. Zhang, B. Dashtbozorg, E. Bekkers, J. P. Pluim, R. Duits, and B. M. ter Haar Romeny, "Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores," *IEEE Trans. Med. Imag.*, vol. 35, no. 12, pp. 2631-2644, Aug. 2016.

[17] B. Al-Diri, A. Hunter, and D. Steel, "An active contour model for segmenting and measuring retinal vessels," *IEEE Trans. Med. Imag.*, vol. 28, no. 9, pp. 1488-1497, Sep. 2009.

[18] Y. Zhao, L. Rada, K. Chen, S. P. Harding, and Y. Zheng, "Automated vessel segmentation using infinite perimeter active contour model with hybrid region information with application to retinal images," *IEEE Trans. Med. Imag.*, vol. 34, no. 9, pp. 1797-1807, Mar. 2015.

[19] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501-509, Apr. 2004.

[20] X. You, Q. Peng, Y. Yuan, Y. M. Cheung, and J. Lei, "Segmentation of retinal blood vessels using the radial projection and semi-supervised approach," *Pattern Recog.*, vol. 44, no. 10, pp. 2314-2324, 2011.

[21] E. Ricci and R. Perfetti, "Retinal blood vessel segmentation using line operators and support vector classification," *IEEE Trans. Med. Imag.*, vol. 26, no. 10, pp. 1357-1365, Oct. 2007.

[22] C. A. Lupaşcu, D. Tegolo, and E. Trucco, "FABC: Retinal vessel segmentation using AdaBoost," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 5, pp. 1267-1274, Sep. 2010.

[23] J. V. B. Soares, J. J. G. Leandro, R. M. Cesar, H. F. Jelinek, and M. J. Cree, "Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification," *IEEE Trans. Med. Imag.*, vol. 25, no. 9, pp. 1214-1222, Sep. 2006.

[24] D. Marin, A. Aquino, M. E. Gegndez-Arias, and J. M. Bravo, "A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features," *IEEE Trans. Med. Imag.*, vol. 30, no. 1, pp. 146-158, Jan. 2011.

[25] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman, "An ensemble classification-based approach applied to retinal blood vessel segmentation," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 9, pp. 2538-2548, Sep. 2012.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proc. NIPS*, 2012, pp. 1097-1105.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, 2015, pp. 3431-3440.

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *Proc. ICLR*, 2015.

[29] K. K. Maninis, J. Pont-Tuset, P. Arbelaez and L. Van Gool, "Deep retinal image understanding," in *Proc. MICCAI*, 2016, pp. 140-148.

[30] J. Orlando, E. Prokofyeva, and M. Blaschko, "A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 1, pp. 16-27, Jan. 2017.

[31] H. Fu, Y. Xu, D. W. K. Wong, and J. Liu, "Retinal vessel segmentation via deep learning network and fully-connected conditional random fields," in *Proc. ISBI*, 2016, pp. 698-701.

[32] H. Fu, Y. Xu, S. Lin, D. W. K. Wong, and J. Liu, "DeepVessel: Retinal vessel segmentation via deep learning and conditional random field," in *Proc. MICCAI*, 2016, pp. 132-139.

[33] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, and T. Wang, "A cross-modality learning approach for vessel segmentation in retinal images," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 109-118, Jan. 2016.

[34] A. Dasgupta, and S. Singh, "A fully convolutional neural network based structured prediction approach towards the retinal vessel segmentation," in *Proc. ISBI*, 2017, pp. 18-21.

[35] P. Liskowski, and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 11, pp. 2369-2380, Mar. 2016.

[36] Z. Yan *et al.*, "A skeletal similarity metric for quality evaluation of retinal vessel segmentation," *IEEE Trans. Med. Imag.*, vol. 37, no. 4, pp. 1045-1057, Apr. 2018.

[37] J. Long *et al.*, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, pp. 3431-3440, 2015.

[38] A. D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Trans. Med. Imag.*, vol. 19, no. 3, pp. 203-210, Mar. 2000.

[39] C. G. Owen, A. R. Rudnicka, R. Mullen, S. A. Barman, D. Monekosso, P. H. Whincup, J. Ng, and C. Paterson, "Measuring retinal vessel tortuosity in 10-year-old children: Validation of the computer-assisted image analysis of the retina (CAIAR) program," *Invest. Ophthalmol. Vis. Sci.*, vol. 50, no. 5, pp. 2004-2010, 2009.

[40] L. Lam, S. -W. Lee, and C. Y. Suen, "Thinning methodologies-A comprehensive survey," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 14, no. 9, pp. 869-885, Sep. 1992.

[41] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[42] M. E. Gegundez-Arias et al., "A function for quality evaluation of retinal vessel segmentations," *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 231-239, Feb. 2012.