Supplementary Material for "Weakly-Supervised Vessel Detection in Ultra-Widefield Fundus Photography Via Iterative Multi-Modal Registration and Learning"

Li Ding, *Student Member, IEEE*, Ajay E. Kuriyan, Rajeev S. Ramchandran, Charles C. Wykoff, and Gaurav Sharma, *Fellow, IEEE*

S.I. Overview

This document provides Supplementary Material for the paper [1]. Section S.II provides implementation details for the proposed approach, including network architectures and the training protocol. Section S.III provides a summary of the evaluation metrics. Section S.IV provides additional validation and results and visualizations illustrating the working of the EM-based learning from noisy labels in the proposed approach. In Section S.V, we include additional visual results of vessel detection on the PRIME-FP20 dataset. In Section S.VI, we provide details on our evaluation of the prior alternative approach proposed in [2] for joint vessel detection and registration. Finally, we report the complete results of vessel detection evaluated on the narrow-field fundus photography in Section S.VII.

S.II. Implementation Details

A. Network Architectures

In the proposed framework, we adopt the U-Net [3] model that is an encoder-decoder architecture with skip connection. The encoder architecture is

$$C_1^e(64) \text{ - } C_2^e(128) \text{ - } C_3^e(256) \text{ - } C_4^e(512) \text{ - } C_5^e(512),$$

where $C_i^e(n)$ denotes the i-th layer in the encoder, which consists two consecutive convolutional layers followed by a max-pooling layer. The decoder architecture is

$$C_4^d(256) \text{ - } C_3^d(128) \text{ - } C_2^d(64) \text{ - } C_1^d(64) \text{ -} C_{out}(1),$$

where $C_i^d(n)$ denotes the i-th layer in the decoder that has the skip connection to the layer C_i^e in the encoder, and $C_{out}(1)$ is the output convolutional layer that returns the probabilistic vessel maps. The convolutional layers $C_i(n)$ have 3×3 kernel size, n output channels, and ReLU activation. The output layer $C_{out}(1)$ uses a 1×1 kernel and sigmoid activation.

B. Training Protocol

The input to the U-Net are 256×256 patches extracted from the training images with a stride of 128. Patches that are not completely in the FOV masks are not included. Data augmentation techniques are applied to enlarge the size of training data. To do so, we randomly apply a sequence of transformations to image patches, including (1) rotation with an angle randomly selected between -90° and 90°, (2) horizontal and vertical flip, (3) blurring with Gaussian filter, and (4) contrast and brightness adjustment. We use Adam optimizer [4] with a fixed learning rate of 0.0001. The parameters that are used for calculating the gradient averages and its square are set to 0.9 and 0.999, respectively. We shuffle the training dataset in each epoch and set the batch-size to 16. The network is trained on a Nvidia Tesla V100 GPU.

L. Ding and G. Sharma are with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627-0231, USA (e-mail: {l.ding, gaurav.sharma}@rochester.edu).

A. E. Kuriyan is with Retina Service, Wills Eye Hospital, Philadelphia, PA 19107-5109 & the University of Rochester Medical Center, University of Rochester, Rochester, NY 14642-0001, USA (e-mail: ajay.kuriyan@gmail.com).

R. S. Ramchandran is with the University of Rochester Medical Center, University of Rochester, Rochester, NY 14642-0001, USA (e-mail: rajeev_ramchandran@urmc.rochester.edu).

C. C. Wykoff is with Retina Consultants of Houston and Blanton Eye Institute, Houston Methodist Hospital & Weill Cornell Medical College, Houston, TX 77030-2700, USA (e-mail: ccwmd@houstonretina.com).

Models	# of Parameters ↓	BIC [7] ↓	mLLH ↑
Gaussian Mixture Model	5	-1.226×10^7	0.957
Beta Mixture Model [8]	5	-1.404×10^7	1.095
Prop. Exp+Gaussian Model	4	-1.420×10^7	1.107

TABLE S.I: Number of parameters, Bayesian information criterion (BIC) goodness-of-fit, and the mean log-likelihood (mLLH) for different mixture models used to fit the loss distribution for the proposed approach.

S.III. Description of Evaluation Metrics

We report three metrics to quantify the performance of vessel detection, i.e., the area under the Precision-Recall curve (AUC PR), the Dice coefficient (DC), and the CAL metric [5]. The PR curve is plotted as the precision versus the recall obtained by binarizing the predicted vessel map with thresholds τ ranging 0 to 1. Precision, recall, and DC are computed as

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN}, \quad DC = \frac{2TP}{2TP+FP+FN},$$

where TP, FP, and FN are true positive, false positive, and false negative, respectively.

We do not choose the Receiver Operating Characteristic (ROC) curve, which is a plot of the true positive rate against the false positive rate, as the evaluation metric. For assessing the performance of vessel detection, the ROC curve is not informative because the ground truth labels are highly skewed where the majority is the negative labels (background pixels in the UWF FP). In this setting, the false positive rate, computed as the ratio between the number of false positive detection to the total number of negative labels, is dominated by the negative labels. As noted in [6], the PR curve is more preferable than the ROC curve when the dataset contains highly imbalanced labels.

Although the AUC PR and the DC are commonly reported in prior works, these metrics are based on pixel-wise comparison of the labeled ground truth and the predicted vessel map. However, as shown in Fig. S.1(c), the pixel-wise comparison is sensitive to the label ambiguities, particularly for pixels on vessel peripheries that can be partially belong to the vessel. In addition, the pixel-wise comparison does not reflect the performance with regard to the higher level structure of the vasculature, which is also of clinical interest. To overcome these concerns, we use the CAL metric [5] that provides resilience to labeling of ambiguous pixels on vessel peripheries and better agreement with human assessment (of higher level structure). The CAL metric assesses the consistency of the binary ground truth and the binary predicted vessel map using three individual factors, the connectivity (C), the area (A), and the length of skeleton (L). The connectivity factor C compares the number of connected vessel segments between the ground truth and the predicted vessel maps, as shown in Figs. S.1(e) and (f). The area factor A assesses the relative overlapping area between the ground truth vessel map and the predicted vessel map while disregard the labeling uncertainty in pixels on vessel peripheries using morphological dilation on binary vessel maps. It can be seen in Fig. S.1(g) that the area factor is more robust against label uncertainties than pixel-wise comparison. The length factor L assesses the consistency of the vessel skeleton obtained from the ground truth and the predicted vessel map. Similar to the area factor, the morphological dilation operation is performed to overcome the issue that the vessel skeleton may be slightly displaced in one image relative to the other. Figures S.1(d) and (h) show the evaluation of vessel skeleton obtained from the pixel-wise comparison and the CAL metric, respectively. The overall CAL metric is defined as the product of individual C, A, and L factors. We refer the readers to the original paper [5] for detailed computation of the CAL metric.

S.IV. Additional Validation and Illustration of Working of the EM-based Learning from Noisy Labels

In this section, we provide additional validation and results and visualizations illustrating the working of the EM-based learning from noisy labels in the proposed approach. First, we evaluate alternative mixture models using the goodness-of-fit criteria, which further supports the use of the proposed mixture model for fitting the loss distribution. Table S.I lists the number of parameters in each mixture model, mean log-likelihood, and Bayesian information criterion (BIC) [7]. The model with less parameters, higher log-likelihood, and lower BIC value is preferred for fitting the data. The proposed mixture model that consists of the exponential and the Gaussian distribution achieves the best result for fitting the loss distribution.

To illustrate how the EM-based learning from noisy labels works in the proposed method, we show intermediate results of noisy label correction in Figure S.2. The noisy training labels in the warped FA vessel maps are generated from the last iteration in the proposed framework. The fourth column shows the posterior probabilities estimated by the mixture model. The "false positive" labels, highlighted by the yellow arrows in the warped FA vessel maps in the second column, are successfully identified (see red pixels in the posterior probabilities) and removed in the updated vessel maps.

S.V. Additional Visual Results on PRIME-FP20 Dataset

We provide additional visual comparison of vessel detection on PRIME-FP20 dataset in Fig. S.3. The results reinforce the findings in the main manuscript that the proposed iterative framework offers significant improvement over existing methods.

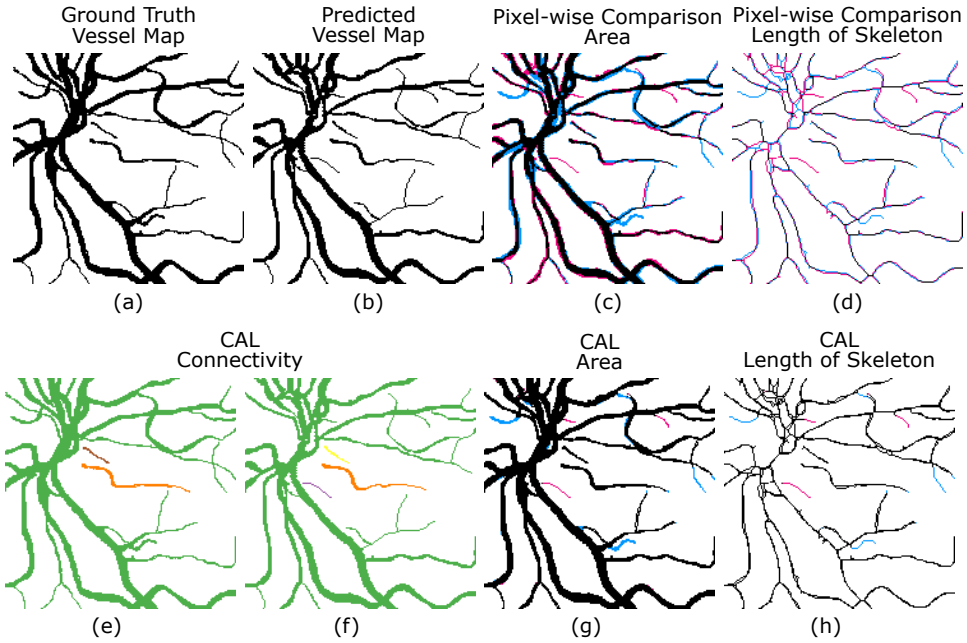

Fig. S.1: Schematic illustration of the pixel-wise based metrics and the CAL metric [5]. (a) and (b) show sample patches of the ground truth label and the binary predicted vessel map, respectively. (c) shows the pixel-wise comparison of the overlapping area, where true positive, false positive, and false negative are highlighted in black, red, and blue, respectively. (d) shows the pixel-wise comparison of the length of skeleton, which are obtained from the corresponding binary vessel maps. Figures in (e) - (h) illustrate the CAL metric that consists of three individual factors: the connectivity, the area, and the length of skeleton. (e) and (f) visualize the connected vessel segments in (a) and (b), respectively. (g) and (h) show the comparisons used in the CAL area factor and CAL length factor computations, respectively, demonstrating the resilience of these factors to differences in labeling of ambiguous pixels on vessel peripheries and slight displacements between vessel skeletons.

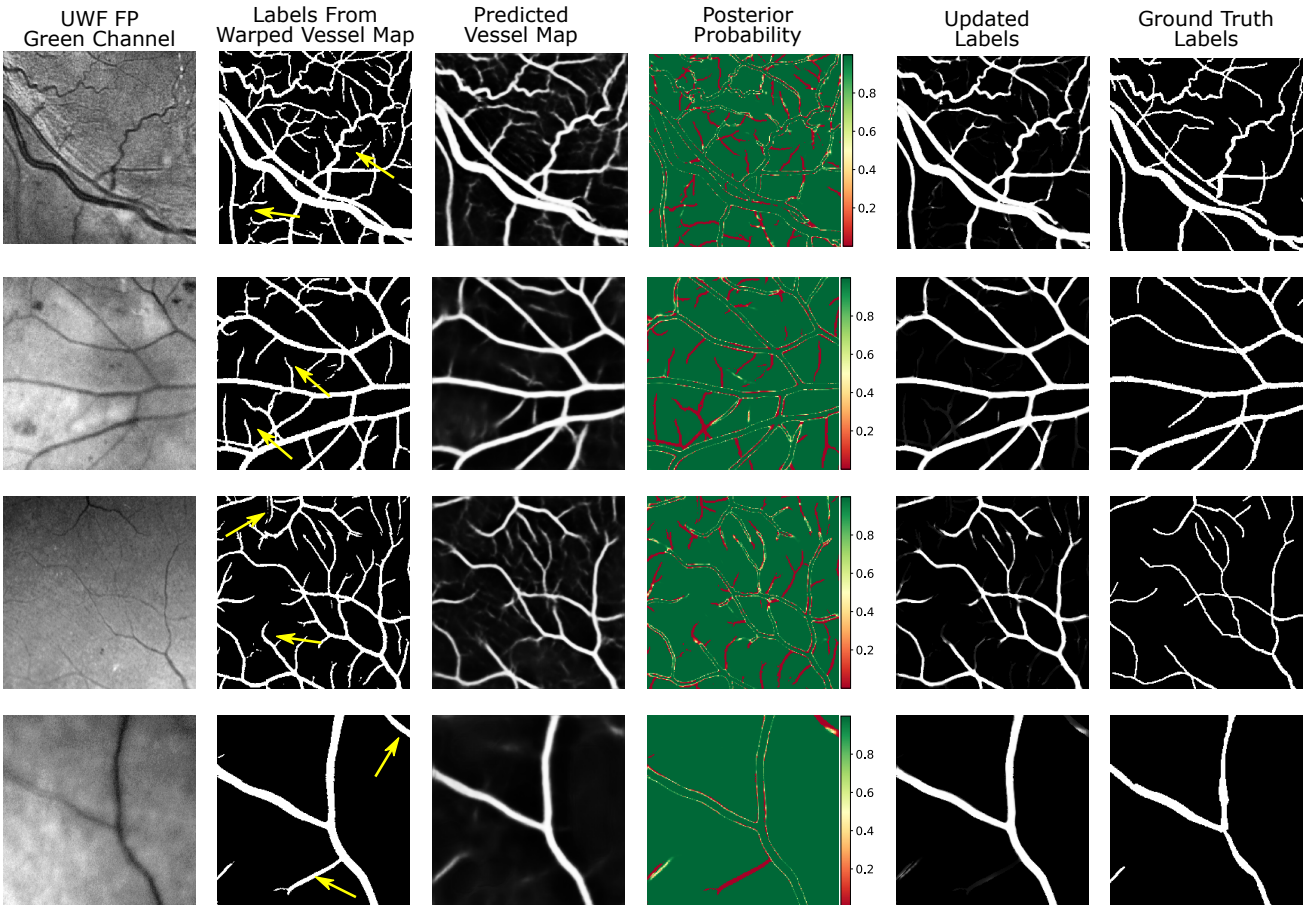

Fig. S.2: Intermediate results illustrating the working of the EM-based approach for learning from noisy labels in the proposed approach. The predicted vessel maps in the third column are obtained using the DNNs trained on the noisy dataset $\{(\boldsymbol{X}_c^i, \boldsymbol{Y}_{a\to c}^{i,t})\}_{i=1}^M$, where $t = 3$. The fourth column shows the posterior probabilities estimated by the mixture model obtained via EM (Eqns. (4) and (5) in the main manuscript), where the green pixels are predicted to have correct labels in the warped FA vessel map and the red pixels are predicted to have incorrect labels. The updated (probabilistic) labels are the linear combination of the the labels in the warped vessel map and the predicted vessel map where the coefficients are the posterior probabilities.

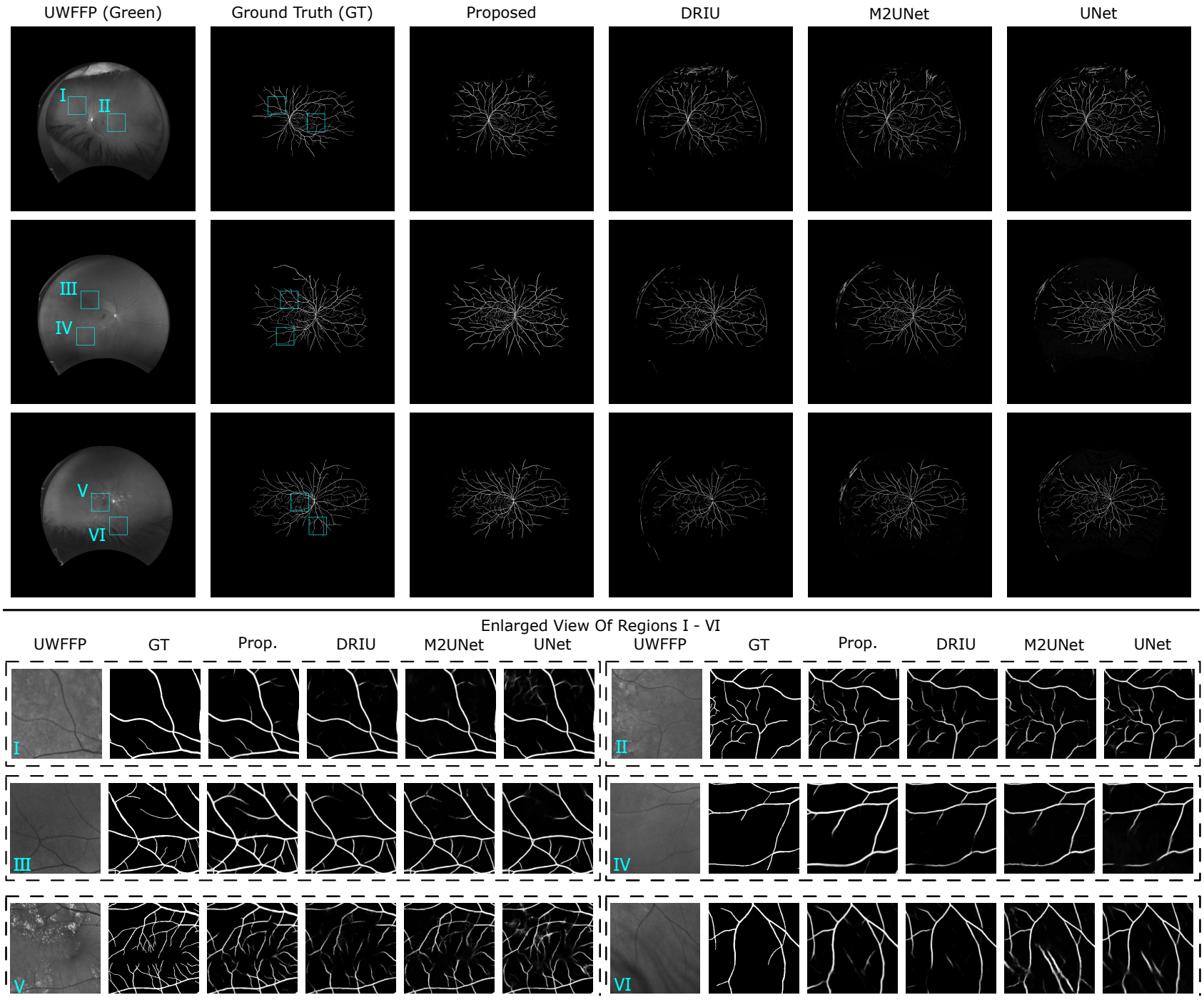

Fig. S.3: Additional sample images and detected vessel maps for the proposed approach and alternatives from the PRIME-FP20 dataset. Six contrast-enhanced enlarged views I-VI, marked by the cyan rectangles in the full image, are included.

Methods	Year	AUC PR	Max DC	CAL (C, A, L)
SegReg-NF [2]	2019	0.535	0.560	0.148 (0.953, 0.445, 0.340)
SegReg-UWF [2]	2019	0.344	0.376	0.078 (0.952, 0.312, 0.243)

TABLE S.II: Quantitative vessel detection results obtained with two versions of SegReg [2]. SegRef-NF and SegRef-UWF are trained on the narrow-field FP and FA dataset [9] and the PRIME-FP20 dataset, respectively.

S.VI. Evaluation of the Prior SegReg [2] Technique for Joint Segmentation and Registration

We evaluated SegReg [2], a method proposed for joint registration and vessel detection in FA and FP images. The model used in the SegReg publication[1] was trained on RGB narrow field images. The trained model provided with the code performs extremely poorly on the UWF FP images because of the differences in the capture modalities: UWF FP images captured by Optos system are pseudo-color consisting of red and green channels. We, therefore, modified the code to use only the green channel in the FP modality and retrained the network. We implement two versions of SegReg. The first one, SegReg-NF, is trained on narrow-field FP and FA dataset [9], which was also used in the original paper. The second version, SegReg-UWF, is directly trained on the PRIME-FP dataset. It is worth noting both models require a good initial alignment between FP and FA images. For SegReg-NF, we use the transformations provided in the original implementation, which were estimated by manually selecting corresponding points from two modalities. To ensure the same initialization between SegReg-UWF and the proposed framework, we use the transformations estimated by the chamfer alignment in the first iteration of the proposed method.

Quantitative results of SegReg-NF and SegReg-UWF are summarized in Table S.II. Both models perform rather poorly on the PRIME-FP20 dataset. For SegReg-NF, the model achieves an AUC PR of 0.535 and the maximum Dice coefficient of 0.560. Visual results of the detected vessel maps obtained with SegReg-NF are shown in Fig. S.4. The SegReg-UWF model fails to produce reliable vessel maps (AUC PR < 0.5) for UWF FP images in the PRIME-FP20 dataset.

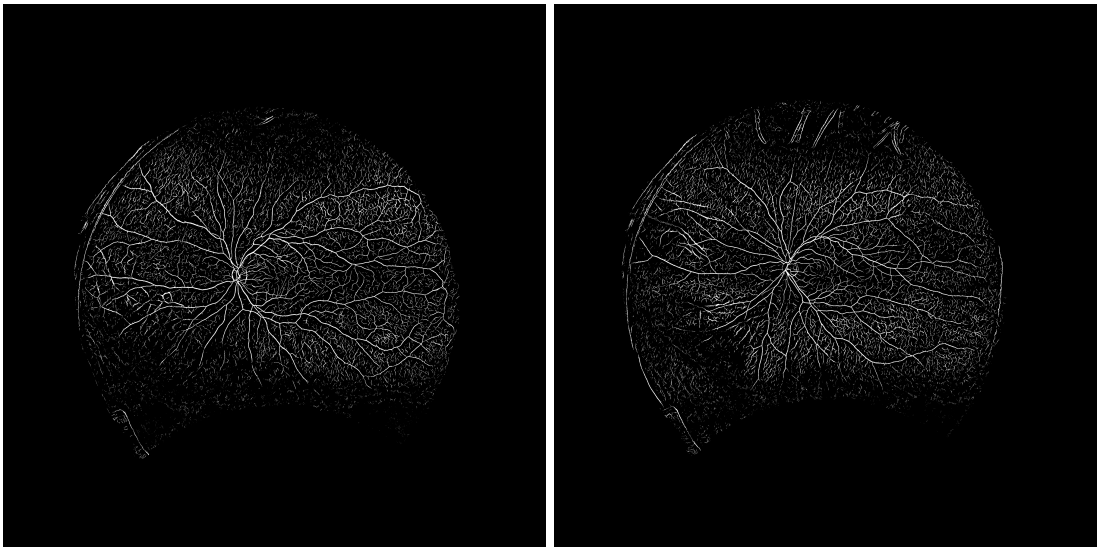

Fig. S.4: Visual results of the detected vessel maps for the SegReg method for joint registration and vessel detection. The corresponding UWF FP image is shown in Fig. 8 in the main manuscript.

The reasons for the poor performance are twofold. First, SegReg does not take into account the differences of the vessels in FP and FA modalities. Assuming the same vessels are presented in both modalities but not registered, SegReg proposes the mean square error between the FA vessel maps and the warped FP vessel maps as a "content loss" (Eq. (3) in [2]). This loss forces the registered vessel maps of FP and FA modalities to be close. However, this assumption does not hold, especially in the retinal periphery where FA imaging captures much more fine vessels than FP. Minimizing the content loss, therefore, yields "false positive" vessel detections in FP images, as shown in Fig. S.4. Second, the registration network in SegReg , which outputs a displacement field, can only handle small deformation [10] and therefore requires a good initial alignment. Compared to SegReg, the proposed framework is less sensitive to the initialization. In addition to the poor detection performance, SegReg has another disadvantage that the training process is much slower than the proposed framework. On an Nvidia V100 GPU, SegReg-UWF takes approximately 11.5 hours for training the network. For the proposed iterative framework, the registration

[1]Available at: https://github.com/JunkangZhang/RetinalSegReg

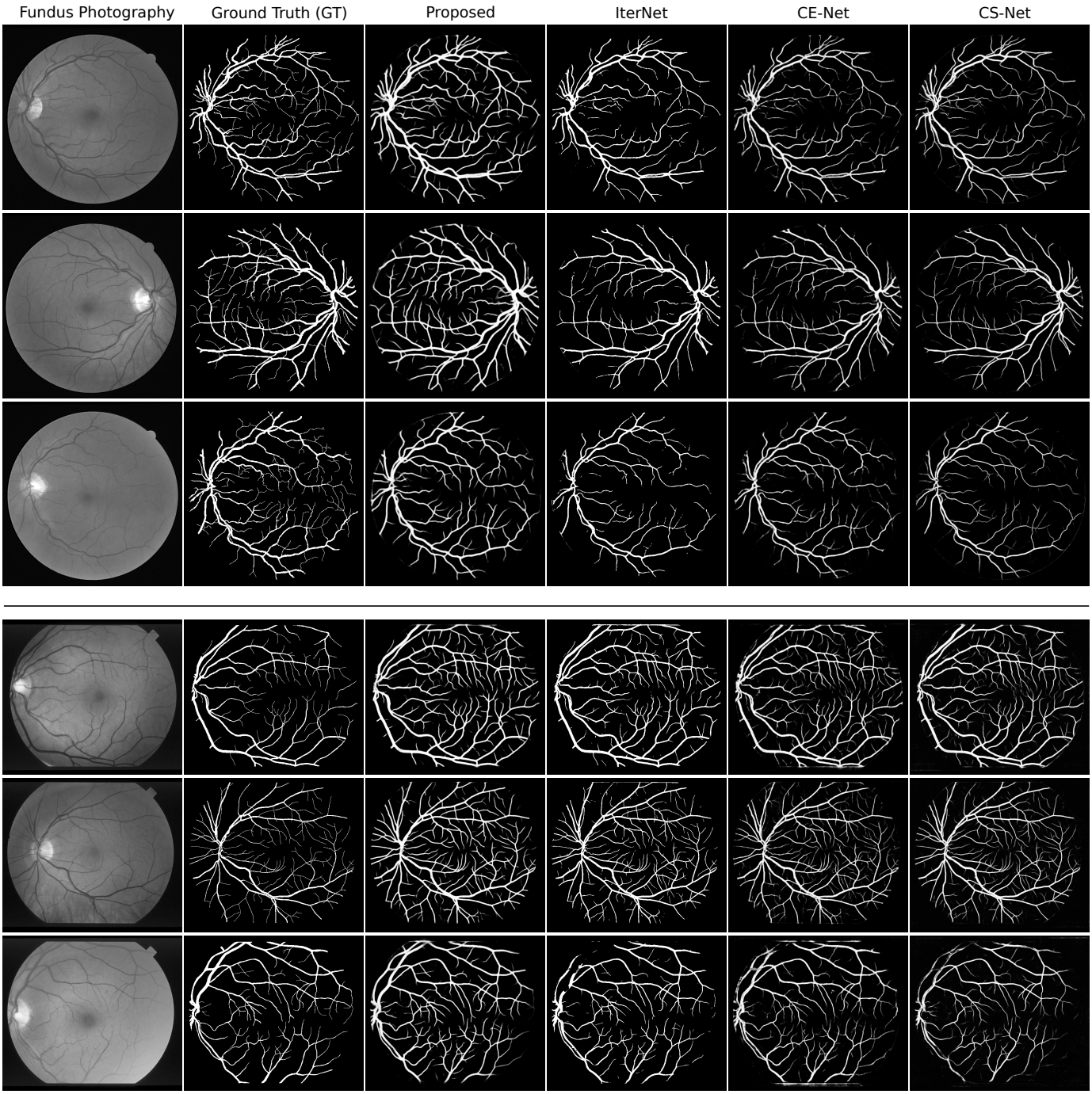

Fig. S.5: Sample images and detected vessel maps for the proposed approach and alternatives for cross-training evaluations on the DRIVE (Rows 1-3) and the STARE (Rows 4-6) datasets.

and the learning step only takes, respectively, 15 and 50 minutes in each iteration, which translates to a total of 3.25 hours for the entire training process.

S.VII. Results on Narrow-Field Fundus Photography

In this section, we provide complete results of cross-training evaluation where the training and the test data are from two independent sources. Table S.III lists the quantitative results. On the DRIVE dataset, the proposed framework has the best performance and significantly outperforms the existing alternatives, achieving an AUC PR of 0.886, the max DC of 0.803, and the overall CAL of 0.827. Although the proposed method is not the best performing method on the STARE dataset, the performance is only slightly worse than the best performing method (DRIU [11]). In Fig. S.5, we show sample results of the detected vessel maps on the DRIVE and the STARE datasets.

Method	Year	DRIVE (Trained On STARE)			STARE (Trained On DRIVE)		
		AUC PR	Max DC	CAL (C, A, L)	AUC PR	Max DC	CAL (C, A, L)
2nd Annotator	-	-	0.789	0.839 (1.000, 0.940, 0.892)	-	0.742	0.640 (1.000, 0.848, 0.753)
HED [12]	2015	0.879	0.797	0.743 (0.996, 0.900, 0.828)	0.838	0.748	0.574 (0.995, 0.773, 0.740)
U-Net [3]	2015	0.886	0.803	0.713 (0.997, 0.890, 0.803)	0.852	0.782	0.730 (0.996, 0.859, 0.842)
DRIU [11]	2016	0.877	0.793	0.629 (0.996, 0.847, 0.744)	**0.898**	**0.812**	**0.806 (0.996, 0.912, 0.886)**
NestUNet [13]	2018	0.877	0.795	0.688 (0.996, 0.876, 0.787)	0.892	0.805	0.786 (0.997, 0.895, 0.879)
M2U-Net [14]	2019	0.859	0.784	0.649 (0.995, 0.856, 0.760)	0.817	0.749	0.635 (0.995, 0.800, 0.785)
CE-Net [15]	2019	0.876	0.792	0.694 (0.997, 0.880, 0.790)	0.871	0.785	0.750 (0.997, 0.875, 0.855)
CS-Net [16]	2019	0.883	0.801	0.703 (0.996, 0.883, 0.798)	0.854	0.775	0.701 (0.996, 0.840, 0.821)
RU-Net [17]	2019	0.884	0.800	0.659 (0.996, 0.859, 0.769)	0.891	0.815	0.780 (0.996, 0.899, 0.869)
IterNet [18]	2020	0.845	0.795	0.698 (0.998, 0.882, 0.792)	0.815	0.794	0.727 (0.999, 0.861, 0.839)
Proposed	2020	**0.886**	**0.803**	**0.827 (0.998, 0.938, 0.883)**	0.884	0.795	0.756 (0.999, 0.880, 0.857)

TABLE S.III: Quantitative results of vessel detection obtained from different methods trained on the DRIVE and the STARE datasets. The best result is shown in bold.

REFERENCES

[1] L. Ding, A. E. Kuriyan, R. S. Ramchandran, C. C. Wykoff, and G. Sharma, "Weakly-supervised vessel detection in ultra-widefield fundus photography via iterative multi-modal registration and learning," *IEEE Trans. Med. Imaging*, accepted for publication, to appear. [Online]. Available: https://doi.org/10.1109/TMI.2020.3027665

[2] J. Zhang *et al.*, "Joint vessel segmentation and deformable registration on multi-modal retinal images based on style transfer," in *IEEE Intl. Conf. Image Proc.*, Sep. 2019, pp. 839–843.

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

[4] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Intl. Conf. Learning Representations*, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980

[5] M. E. Gegundez-Arias, A. Aquino, J. M. Bravo, and D. Marin, "A function for quality evaluation of retinal vessel segmentations," *IEEE Trans. Med. Imaging*, vol. 31, no. 2, pp. 231–239, Feb 2012.

[6] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Intl. Conf. on Mach. Learning*, 2006, pp. 233–240.

[7] G. Schwarz *et al.*, "Estimating the dimension of a model," *Annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[8] E. Arazo, D. Ortego, P. Albert, N. O'Connor, and K. McGuinness, "Unsupervised label noise modeling and loss correction," in *Intl. Conf. on Mach. Learning*, vol. 97, 2019, pp. 312–321.

[9] S. Hajeb-Mohammad-Alipour, H. Rabbani, and M. R. Akhlaghi, "Diabetic retinopathy grading by digital curvelet transform," *Computational and Mathematical Methods in Med.*, vol. 2012, 2012.

[10] J. Ashburner, "A fast diffeomorphic image registration algorithm," *NeuroImage*, vol. 38, no. 1, pp. 95 – 113, 2007.

[11] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, "Deep retinal image understanding," in *Intl. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2016, pp. 140–148.

[12] S. Xie and Z. Tu, "Holistically-nested edge detection," in *IEEE Intl. Conf. Comp. Vision.*, Dec. 2015, pp. 1395–1403.

[13] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Med. Image Analysis*, 2018, pp. 3–11.

[14] T. Laibacher, T. Weyde, and S. Jalali, "M2U-Net: Effective and efficient retinal vessel segmentation for real-world applications," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog. Wksp.*, June 2019, pp. 115–124.

[15] Z. Gu *et al.*, "CE-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imaging*, vol. 38, no. 10, pp. 2281–2292, Oct 2019.

[16] L. Mou *et al.*, "CS-Net: Channel and spatial attention network for curvilinear structure segmentation," in *Intl. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 721–730.

[17] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari, "Recurrent residual U-Net for medical image segmentation," *J. Med. Imaging*, vol. 6, no. 1, pp. 1 – 16, 2019.

[18] L. Li, M. Verma, Y. Nakashima, H. Nagahara, and R. Kawasaki, "IterNet: Retinal image segmentation utilizing structural redundancy in vessel networks," in *IEEE Winter Conf. Applications of Comp. Vision*, 2020, pp. 3645–3654.