Retinal Vessel Segmentation using Round-wise Features Aggregation on Bracket-shaped Convolutional Neural Networks

Cam-Hao Hua[1], Thien Huynh-The[2], and Sungyoung Lee[1]

***Abstract*— With the recent advent of deep learning in medical image processing, retinal blood vessel segmentation topic has been comprehensively handled by numerous research works. However, since the ratio between the number of vessel and background pixels is heavily imbalanced, many attempts utilized patches augmented from original fundus images along with fully convolutional networks for addressing such pixel-wise labeling problem, which significantly costs computational resources. In this paper, a method using Round-wise Features Aggregation on Bracket-shaped convolutional neural networks (RFA-BNet) is proposed to exclude the necessity of patches augmentation while efficiently handling the irregular and diverse representation of retinal vessels. Particularly, given raw fundus images, typical feature maps extracted from a pretrained backbone network are employed for a bracket-shaped decoder, wherein middle-scale features are continuously exploited round-by-round. Then, the decoded maps having highest resolution of each round are aggregated to enable the built model to flexibly learn various degrees of embedded semantic details while retaining proper annotations of thin and small vessels. Finally, the proposed approach showed its effectiveness in terms of sensitivity (0.7932), specificity (0.9741), accuracy (0.9511), and AUROC (0.9732) on DRIVE dataset.**

I. INTRODUCTION

In color fundus photography, abnormal changes in representation of retinal blood vessel may tell initial sign of common eye diseases comprising diabetic retinopathy (DR), glaucoma, ocular hypertension, cataracts, to name a few. For example, the phenomenon of capillary wall dilatation [1] at retinal vessels, namely microaneurysm, is the earliest indicator of suffering from DR. Therefore, efficiently extracting the vessel-based information can help ophthalmologists precisely diagnose and effectively deploy an optimal treatment plan for prevention and regulation of blindness and vision impairment for patients.

Recently, resulting from the expeditious growth of computational resources like Graphical Processing Units (GPU) as well as the quantity of image datasets, Convolutional Neural Networks (CNNs) has been widely employed in various domains of medical image processing with impressive performance thanks to the powerful feature representation. Specifically, in the field of retinal blood vessel segmentation in color fundus image, there are many attempts handling such kind of binary classification problem at pixel level, a.k.a. semantic segmentation given 2 classes (background and vessel), based on fully convolutional neural networks (FCN) architectures. For instance, CNN-RFs [2] utilized CNNs and Random Forest as feature extractor and corresponding classifier, respectively, for the vessel segmentation. Besides that, the authors of [3] introduced a model including a base CNN for extracting meaningful features along with additional layer blocks specialized for simultaneously segmenting optic disc and retinal blood vessel. On the other hand, since the ratio between the number of vessel and background pixels is massively imbalanced, many works split a given fundus image into multiple overlapping patches [4]–[8], which are considered as newly augmented images, in order to address the class-imbalancing issue as well as increase dataset size for combating overfitting matter. In specific, methods proposed in [4], [5] took into account RGB patches as inputs of CNNs formed by stacks of convolution, max-pooling, and fully connected layers. Meanwhile, Feng *et al.* [6] proposed a technique called local entropy sampling to generate grayscale patches from original fundus photography as inputs of a predefined FCN having skip-connection scheme. He *et al.* [7] implemented a similar approach but additionally took into consideration of differences between small and large vessel regions by a local de-regression along with regression based deep architecture. Furthermore, instead of converting RGB to grayscale as in [6], [7], Kassim *et al.* [8] only involved green channel of the raw fundus image to constitute patches for training a predefined 14-layer CNN. The readers may refer to [9] for an intensive review of existing literature of retinal vessel segmentation area.

As aforementioned, small patches generation can reduce the imbalance between the amount of retinal blood vessel and background pixels, which facilitates the semantic segmentation model to encode features more effectively. However, it is obvious that such kind of patch-based approaches brings in expensive computations in both data preprocessing and execution stage for trading-off better performance. Therefore, in this paper, we propose a methodology, namely RFA-BNet, be able to efficiently partition the blood vessels in color fundus photography without the necessity of costly processing small patches of raw images for training the deep learning network. Concretely, it can be realized that because of being pretrained with large-scale dataset, classification-based CNNs like VGG-Net [13], ResNet [10] can delineate the objects of interest at different levels of feature repre-

*This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-0-01629) supervised by the IITP (Institute for Information & communications Technology Promotion). This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2017-0-00655).

[1]The Department of Computer Science and Engineering, Kyung Hee University, Gyeonggi-do, 446-701, South Korea hao.hua@oslab.khu.ac.kr, sylee@oslab.khu.ac.kr

[2]ICT Convergence Research Center, Kumoh National Institute of Technology, Gumi, South Korea thienht@kumoh.ac.kr

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

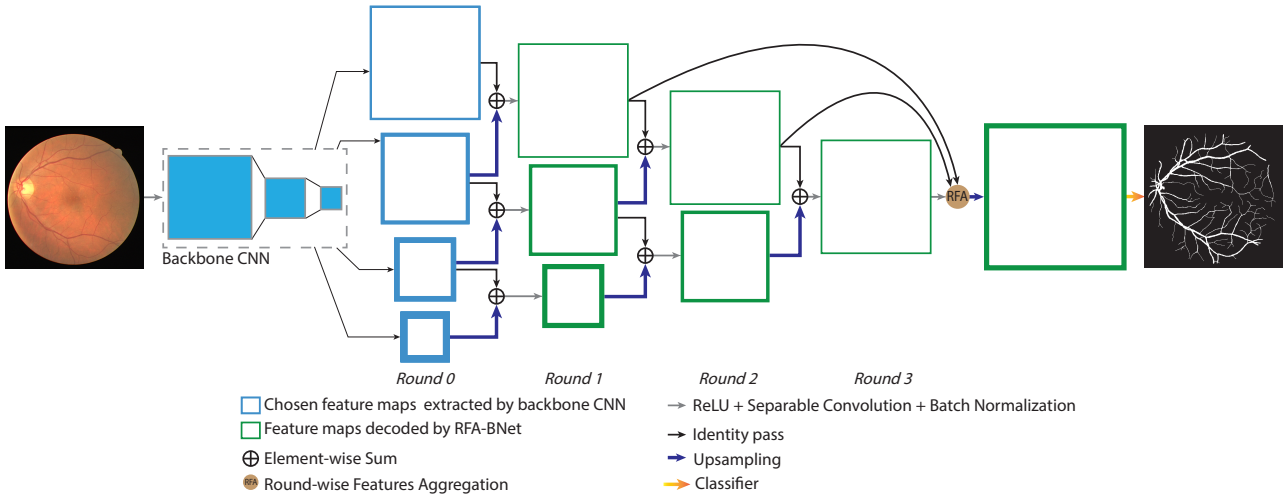

Fig. 1. Architecture of the proposed RFA-BNet. Let an input color fundus image be fed into the backbone ResNet-101 [10], final outputs of the initial convolution layer and three first residual blocks, i.e., with strides of 2, 4, 8, 16, respectively, are involved in the bracket-manner decoding process (in three rounds) for retinal blood vessel labeling. Briefly, these fine-to-coarse feature maps are densely combined via the element-wise summation along with non-linear learning (ReLU, separable convolution [11], and batch normalization [12]) to infer outputs which repeat the same operations round-by-round until only one decoded feature map is left. Then, the highest-resolution decoded feature maps at each round are aggregated via depth-wise concatenation procedure before its upsampled version goes through the predefined classifier for pixel-wise segmentation. Since the finest-resolution feature maps decoded at each round of the bracket-style CNN are aggregated to produce remarkable representation of retinal blood vessels, such process is called Round-wise Features Aggregation on Bracket-shaped Network (RFA-BNet). Note that area and thickness of rectangles demonstrate spatial and depth size of the corresponding feature maps, respectively.

sentation, i.e., from finely patterned to semantically rich features. This leads to the hypothesis that leveraging the utilization of those finely patterned features, which should be continuously enhanced the semantically rich information during the pixel-wise prediction map construction process, is capable of labeling small objects more precisely in case of heavily class-imbalancing issue. Hence, we propose Round-wise Features Aggregation (RFA), as the step of exhaustively utilizing finely patterned features, embedded into the B-Net architecture [14] with sorts of specialized manipulations. As a consequence, it is able to comprehensively exploit semantic context of middle-scale features onto the final per-pixel prediction map for the ultimate purpose of segmenting retinal blood vessels, which appear diversely and irregularly in terms of middle- to small-sized objects. In fact, experimental results on DRIVE dataset [15] show that the proposed RFA-BNet achieves competitive performance with state-of-the-art patch-based deep learning techniques in terms of sensitivity (recall), specificity, accuracy, precision, AUROC (Area Under the Receiver Operating Characteristics).

II. METHODOLOGY

A. Bracket-shaped Convolutional Neural Networks

Lately, Hua *et al.* [14] have proposed a novel deep learning based semantic segmentation model, namely B-Net, wherein a bracket-style decoding process is introduced to construct the final pixel-wise labeled map from typical feature maps of various scales learned at backbone VGG16-Net [13]. The idea is motivated from the observation that middle-scale features along a classification-based CNN's feed-forwarding path are not exploited intensively for the segmentation problem although they possess valuable balances between fine details and semantically contextual information, which is clearly profitable for the decoding (i.e., per-pixel prediction map inference) process. Accordingly, to leverage those features' usage, every pair of scale-adjacent feature maps chosen from the backbone network passes through predefined lateral connection modules to infer newly decoded outputs, which continuously repeat the same procedure round-by-round until one final prediction map of finest-resolution is obtained. Note that each round of such decoding approach is defined by the process in which n feature maps combining with neighboring versions to yield $n-1$ outputs possessing enhanced semantic information. In other words, the major contribution of their work is that feature maps at middle levels of spatial resolution are comprehensively utilized to simultaneously (i) contribute semantically richer contexts to the adjacent higher-resolution map and (ii) refine ambiguously coarse details in upsampled version of the adjacent lower-resolution one. Consequently, middle- to small-sized object representation is handled effectively in the final labeled map by the B-Net. Since the appearance of the retinal blood vessel is somewhat suitable to target function of the method proposed in [14], we apply the bracket-style CNN concept with several variations compared to the original work in order to maximize the retinal vessel segmentation performance.

In particular, as illustrated in Fig.1, we utilize pretrained ResNet-101 [10] as backbone network of the proposed approach. Subsequently, four feature maps of different scales utilized for the decoding procedure are outputs of the initial convolution layer and three first residual blocks, with strides of 2, 4, 8, and 16 with respect to the input images' spatial dimension, respectively. Let these feature maps of *Round 0* (blue-line rectangles in Fig.1) densely combine

Approach	Sensitivity	Specificity	Accuracy	AUROC
Liskowski *et al.* [4]	0.7763	0.9768	0.9495	0.9720
Jiang *et al.* [5]	0.7540	0.9825	**0.9624**	**0.9810**
Feng *et al.* [6]	0.7811	**0.9839**	0.9560	0.9792
He *et al.* [7]	0.7761	0.9792	0.9519	N/a
Baseline (w/o RFA)	0.7807	0.9667	0.9484	0.9659
RFA-BNet	**0.7932**	0.9741	0.9511	0.9732

TABLE I

QUANTITATIVE RESULTS ON DRIVE [15] DATASET. **BOLDFACE NUMBERS** INDICATE THE BEST PERFORMANCE OF EACH MEASURE.

with their adjacency as described previously, three newly decoded outputs (green-line rectangles of *Round 1* in Fig.1) are inferred. Next, the same procedure takes place two more rounds until one finest-resolution feature map (having stride of 2) is remained before the RFA module. The continuous combination between two certain scale-adjacent feature maps during the bracket-structured decoding process is defined as follows

$$f_i^r = Conv\left[f_i^{r-1} \oplus \mathcal{U}(f_{i+1}^{r-1})\right] \tag{1}$$

where f_i^r is i^{th} feature map at r^{th} round, wherein $r = 1, 2, 3$ and $i = 1, \ldots, 4 - r$ (the larger value of i, the lower spatial resolution (i.e., larger stride) the corresponding feature map has); $\oplus$ stands for element-wise sum; $\mathcal{U}(.)$ represents transposed convolution operator for 2x upsampling; $Conv[.]$ consists of following operations on the sum feature map: Rectified Linear Unit (ReLU) activation, separable convolution [11], and batch normalization [12] for diminishing adverse effects during the upsampling progress. It is worth noting that the number of both the transposed and separable convolution layers is specified to be identical to channel size of the corresponding higher-resolution input at each combination step.

B. Round-wise Features Aggregation

Apparently, the exhaustive utilization of middle-scale features by the bracket-shaped decoder can effectively represent medium- to small-sized objects at pixel level, which is suitable for segmenting blood vessel in fundus photography. However, naively applying the original structure of the B-Net is obviously not an optimal strategy since the representation of retinal blood vessels is diverse and irregular (if compared with usual contents in natural images), e.g. more and more sudden branches of thin vessels emerge when being away from the optic disc. Also, another noticeable factor is that the ratio between vessel and background pixels are heavily imbalanced (e.g. around 1.3:8.7 in training set of the DRIVE dataset [15]). Therefore, in this work, we additionally propose an approach of RFA on top of the B-Net manipulated by another backbone network with lower output stride as specified in previous sub-section. Since the finest-resolution feature map at each round possesses different degrees of semantically rich features which may get rid of representation of thin and ambiguous vessels, the RFA module aims to aggregate finest-resolution feature maps of all rounds to make the built model flexibly learn weakly-to-strongly embedded semantic contexts while retaining proper annotations of fine details like thin vessels' edges. In concrete, we concatenate the finest-resolution feature map of each round along the depth dimension and then apply transposed convolution followed by a final classifier as demonstrated in Fig.1. Accordingly, the final per-pixel prediction map Y is produced as below

$$Y = \mathcal{U}(\mathcal{A}[f_1^1, f_1^2, f_1^3]) \tag{2}$$

where $\mathcal{A}[.]$ means depth-wise aggregation procedure.

III. EXPERIMENTS

Remarkably, the experimental procedures involving human subjects described in this paper were approved by the Institutional Review Board.

A. DRIVE Dataset

DRIVE stands for Digital Retinal Images for Vessel Extraction [15], which is used to validate studies on retinal blood vessel segmentation in fundus photography. The dataset pool consists of totally 40 images, half of which is designated for training and the remaining for testing. It is worth noting that image crop size around field of view (FOV) is fixed at 584×565. Also, mask of the FOV inside each fundus image is provided to specify regions of interest for fair performance evaluation. Hence, we define ground-truth labels of the retinal background, vessel, and non-FOV pixels as 0, 1, and 255, respectively, during the training stage. For evaluation, well-known metrics such as sensitivity, specificity, accuracy, precision, and AUROC are involved to validate the effectiveness of the proposed method.

B. Implementation Details

In this paper, we utilize Tensorflow [16] and Scikit-learn [17] to train and evaluate the proposed deep network on one NVIDIA 1080TI GPU, respectively. Since the dataset pool is small, we exhaustively augment the training images by following manipulations: random scale of {0.5, 0.75, 1.0, 1.25, 1.5, 2.0}, random crop with centered FOV subject to predefined spatial dimension (i.e., 585×565), depth-wise mean intensity normalization, random horizontal and/or vertical flip. Then, each batch of five augmented images is continuously fed into the proposed architecture. Subsequently, weighted cross-entropy loss function is utilized to assess the compatibility between the resulting pixel-wise prediction maps Y and corresponding ground-truth label maps G in the scenario of class imbalance as follows

$$\mathcal{L}(Y, G) = \sum_{Y_p} \left(\sum_{j=0}^{1} \alpha_j i_{p,j} log(s_{p,j}) \right)$$
$$i_{p,j} = \begin{cases} 1, & Y_{p,j} = G_{p,j} \\ 0, & otherwise \end{cases} \tag{3}$$

where Y_p represents considered pixels of prediction map Y, α_j stands for balancing coefficient of class $j \in \{0, 1\}$, $i_{p,j}$ indicates the predicted class j of Y_p with respect to its actual class in ground-truth label map G, $s_{p,j}$ denotes softmax score

of Y_p corresponding to class j. In this work, α_0 and α_1 are set to be 1.0 and 6.975, respectively, which exhibit the ratio between total number of background (label value of 0) and vessel (label value of 1) pixels in the training dataset. As can be seen from (3), non-FOV pixels (label value of 255) are ignored during the loss computation procedure. From the measured loss, to optimize parameters initialized by He's approach [18] in the RFA-BNet, Adam optimizer [19] with learning rate of 0.001 is adopted. In addition, weight decay of 0.0001 is included to boost the generalization capability of the proposed architecture.

C. Experimental Results

As quantitatively shown in Table I, compared to the baseline concept, the involvement of RFA scheme outperforms $0.0027 - 0.0125$ for all the measures. Moreover, the proposed RFA-BNet achieves state-of-the-art sensitivity (0.7932) among the compared methods. Meanwhile, the performance in terms of specificity, accuracy, and AUROC is still comparable to that of the patch-based methods with 0.9741 (< 0.0098 compared to the best performance reported in [6]), 0.9511 (< 0.0113 [5]), and 0.9732 (< 0.0078 [5]), respectively. Additionally, it can be observed from several typically qualitative results displayed in Fig.2 that irregular and diverse appearance of retinal blood vessel is carried out remarkably under various illumination conditions of input images compared to corresponding ground truth. These outcomes imply that the proposed architecture is able to effectively label challenging retinal vessel at pixel level without expensively utilizing patches augmented from the raw fundus photography.

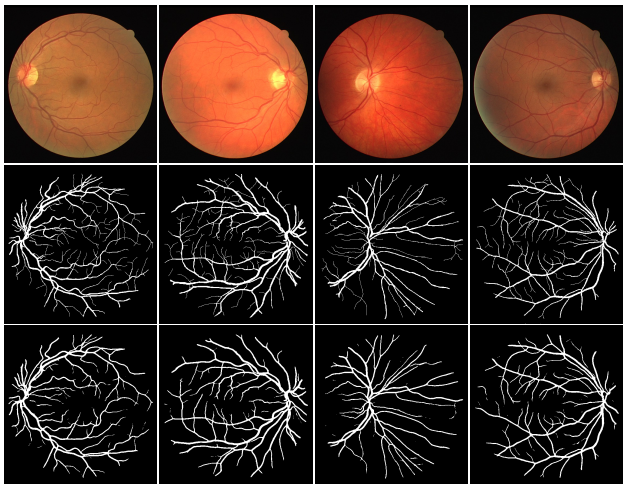

Fig. 2. Typically qualitative results of the proposed RFA-BNet on several testing fundus images of DRIVE [15] dataset. Top row: Raw fundus images; Middle row: Ground truth; Bottom row: Results of the proposed RFA-BNet.

IV. CONCLUSIONS

This paper introduced an approach using Round-wise Features Aggregation on Bracket-shaped convolutional neural networks for dealing with retinal blood vessel segmentation problem in color fundus image. The proposed method targets to efficiently infer pixel-wise labeled map without involving costly computation of generating patches from original color fundus image. For this objective, the bracket-style decoding manner combining with comprehensive aggregation between decoded feature maps of highest-resolution enables the proposed RFA-BNet to identify vessels' location flexibly and precisely at pixel level as shown by the experimental results. In the future, we continue to exploit further capability of the proposed network regarding multi-task segmentation of not only retinal vessel but also optic disc, microaneurysm, etc.

REFERENCES

[1] N. Cheung, P. Mitchell, and T. Y. Wong, "Diabetic retinopathy," *The Lancet*, vol. 376, no. 9735, pp. 124 – 136, 2010.

[2] S. Wang, Y. Yin, G. Cao, B. Wei, Y. Zheng, and G. Yang, "Hierarchical retinal blood vessel segmentation based on feature and ensemble learning," *Neurocomputing*, vol. 149, pp. 708 – 717, 2015.

[3] K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. V. Gool, "Deep retinal image understanding," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016.

[4] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 11, pp. 2369–2380, Nov 2016.

[5] Z. Jiang, H. Zhang, Y. Wang, and S.-B. Ko, "Retinal blood vessel segmentation using fully convolutional network with transfer learning," *Computerized Medical Imaging and Graphics*, vol. 68, pp. 1 – 15, 2018.

[6] Z. Feng, J. Yang, and L. Yao, "Patch-based fully convolutional neural network with skip connections for retinal blood vessel segmentation," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sep. 2017, pp. 1742–1746.

[7] Q. He, B. Zou, C. Zhu, X. Liu, H. Fu, and L. Wang, "Multi-label classification scheme based on local regression for retinal vessel segmentation," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 2765–2769.

[8] Y. M. Kassim, R. J. Maude, and K. Palaniappan, "Sensitivity of cross-trained deep cnns for retinal vessel extraction," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2018, pp. 2736–2739.

[9] S. Moccia, E. D. Momi, S. E. Hadji, and L. S. Mattos, "Blood vessel segmentation algorithms — review of methods, datasets and evaluation metrics," *Computer Methods and Programs in Biomedicine*, vol. 158, pp. 71 – 91, 2018.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[11] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1800–1807.

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[14] C.-H. Hua, T. Huynh-The, and S. Lee, "Convolutional networks with bracket-style decoder for semantic scene segmentation," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct 2018, pp. 2980–2985.

[15] 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 Transactions on Medical Imaging*, vol. 23, no. 4, pp. 501–509, April 2004.

[16] Martin Abadi. et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org.

[17] F. Pedregosa. et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.