Pattern Recognition 120 (2021) 107998

Contents lists available at ScienceDirect

Pattern Recognition

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

A nested U-shape network with multi-scale upsample attention for robust retinal vascular segmentation

Ruohan Zhao[a], Qin Li[b], Jianrong Wu[c], Jane You[a,*]

[a] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[b] Shenzhen Institute of Information Technology, Shenzhen, China
[c] Tencent healthcare, Shenzhen, China

ARTICLE INFO

Article history:
Received 6 December 2019
Revised 26 January 2021
Accepted 18 April 2021
Available online 24 April 2021

Keywords:
Vascular segmentation
Retinal imaging
Dense U-Net
Multi-scale attention
Deep learning

ABSTRACT

This paper presents a new nested U-shape attention network (NUA-Net) with improved robustness of lesions for effective vascular segmentation in retinal imaging. Unlike most of the current deep learning approaches which rely on vanilla upsample module to recover distinguishable features for segmentation, our attention-based multi-scale network extends the U-shape segmentation network by introducing a novel multi-scale upsample attention (MSUA) module to enhance vessel features in a hierarchical structure. The new approach connects encoder-decoder branches through a nested skip-connection pyramid architecture to extract discriminating retinal features from the rich local details. Experimental evaluations on five publicly available databases DRIVE, STARE, CHASE_DB, IOSTAR and HRF show the NUA-Net achieves 0.8043–0.8511 (Sensitivity), 0.9741–0.99 (Specificity) and 0.9646–0.9794 (Accuracy) respectively. The benchmark by cross-testing and separate-testing presents a state-of-the-art performance and better vessel preservation compared with other approaches.

1. Introduction

Vessel delineation of ocular images is widely utilized for the diagnosis of various ophthalmologic diseases [1]. Fundus disease may lead to irreversible visual impairment, and early diagnosis and treatment can help prevent the progression of disease. Subtle changes in retina can reflect early signs of disease, e.g., the hemorrhage occurs when a tiny blood vessel breaks just underneath the clear surface of eye, which is the symptom of diabetic retinopathy. The vessels on fundus images are usually manually annotated by ophthalmologists based on the condition of retinal imaging. However, the manual segmentation of vessels is a tedious and time-consuming task [2], and requires a proficient skill. Accordingly, fully-automated blood vascular segmentation is highly demanded in the diagnosis of eye diseases, to relieve the workload of manual segmentation and improve the processing speed.

The retinal vascular segmentation is a classic yet still challenging task due to the complexities of retinal vascular structures. Early works tackle the segmentation of vessels by formulating it into pixel-wise classification [3], exploiting intrinsic properties, e.g. vessels are elongated structures [2], employing matched filtering based methods [4] and etc. Free from hand-crarfted feature extractors, deep learning have been demonstrated to yield more discriminative representations of blood vessels [5]. Later, deep learning based methods achieve the state-of-art performances in vascular segmentation. Yan et al. [6] proposed a new pixel-wise loss for accurate segmentation of both thick vessels and thin vessels. Besides, Jin et al. [7] adopted deformation convolution module to capture the geometric transformations of diverse vessels. To segment vessels with smaller diameter and lower contrast, Wu et al. [8] applied multiscale network followed network. Despite above efforts on detecting variety of vasculature, it still remains a challenging to generate accurate segmentation of retinal vessels, particularly the capillaries with branches, crossing and reflex. Besides, few proposed approaches were designed in consideration of vascular segmentation with robustness of pathologies in abnormal retinal images. Practically, region in the form of pathological lesions is still easily mis-classified to be retinal vessels, leading to a high false positive rate.

In this paper, inspired in dense U-Net [9], we propose an end-to-end nested U-shape network with multi-scale attention mechanism to harness features from cross-layers for vascular segmentation. Specifically, to promote the information flow among feature maps, we replace skip connection [10] with nested connection module to exploit the potential of the network through feature reuse. By enhancing the mutual connection among multi-

* Corresponding author.
E-mail address: csyjia@comp.polyu.edu.hk (J. You).

https://doi.org/10.1016/j.patcog.2021.107998
0031-3203/

stages features via concatenation, our method can effectively sustain informative details of microvessels. To purify the feature channels toward the most discriminative information, a novel multi-scale upsample attention (MSUA) module is proposed and incorporated into the nested U-shape architecture. Owing to the designed attention module and nested U-shape architecture, our network achieves promising performance for vascular segmentation and shows robustness to lesions.

The main contributions of our work can be summarized as follows: (1) We proposed a novel encoder-decoder framework by integrating the new multi-scale upsample attention (MSUA) modules and dense blocks to harness features from multiple scales for effective vascular segmentation. (2) In comparison with other state-of-the-art methods, our experimental results present better performance on all publicly databases with different domains. (3) The NUA-Net works robustly under pathological conditions and is more capable of handling various challenging cases of vascular segmentation.

2. Related work

Blood vascular segmentation is a pixel-level prediction task. According to whether or not the annotated labels are used as supervision, the existing methods can be grouped into two categories: unsupervised methods and supervised ones.

2.1. Unsupervised methods

The unsupervised methods attempt to find inherent pattern of blood vessels without any prior labeling knowledge. Azzopardi et al. [11] proposed to use a combination of shifted filter to achieve orientation selectivity by computing the weighted geometric mean of filter responses. The vascular segmentation is achieved by summing up the responses of the two filters followed by thresholding. Annunziata et al. [12] proposed an inpainting filter, namely neighbourhood estimator before filling, to enhance vessel detection by inpainting nearby exudates to reduce false positives. Based on the mathematical morphology, Sazak et al. [13] introduced the bowler-hat transformation technique to detect innate features of vessel-like structures to first enhance the retinal images, followed by segmentation.

2.2. Supervised methods

The supervised learning methods generally learn vessel extraction from training dataset which contains manually annotated vascular segmentation. Soares et al. [14] utilized the multi-scale 2-D Gabor wavelet transform responses as feature vectors, followed by a Bayesian classifier for detecting vessels. Based on pixel-level classification, Marin et al. [3] computed a 7-D vector composed of gray-level and moment invariant features for pixel representation of vessels. By describing morphological attributes of vessels, Staal et al. [2] used ridge profiles as to construct features for vessels and applied a selection scheme for feature update followed by a KNN classifier. Later some approaches were proposed by considering heterogeneous feature for enhancing the representation of vessels. Fraz et al. [15] utilized a feature vector from heterogeneous source for constructing an ensemble system of mixed bagged and boosted decision trees to segment vessels. Zhang et al. [16] incorporated vessel filtering, wavelet transform features from orientation scores, and green intensity at multiple scales to represent vessels. Zhao et al. [17] proposed to combine local phase features and weighted geometric mean of the blurred and shifted responses to detect vessels from various modalities. Orlando et al. [18] used a fully connected conditional random field (CRF) model whose parameters were automatically learned in supervised manner and then adopted a structured output support vector machine for vascular segmentation. Wang et al. [19] enveloped a series of Mahalanobis distance classifiers to form a highly nonlinear decision for vessels by a one-pass feed forward process. However, these hand-crafted approaches show limited performance since they are highly dependent on artificial experience. They generally show limited robustness in representing vasculature and lack competing generalization compared to deep supervised learning methods, which achieved a huge success in vascular segmentation area.

2.3. Deep network

Liskowski et al. [5] started to use a Convolutional Neural Network (CNN) trained on large samples in the form of patches for vessel classification but the network was not designed in a pixel-level manner. DRIU [20] adopted a unified framework that constructed feature maps volumes for retinal vascular segmentation. The used networks, however, are easily impeded by multiple spatial pooling and stride convolution operations which result in coarse pixel-wise segmentation. To address this limitation, DeepVessel [21] integrated CNN with CRF as a boundary detection module for retinal vascular segmentation. To efficiently learn pyramid level features and ease the vanishing gradient problems, Ronneberger et al. [10] proposed an encoder-decoder network by adopting skip connection with concatenation operation. It made a great progress in vascular segmentation.

Made on the [10], several variations that were later raised and achieved a state-of-the-art performance. Hervella et al. [22] proposed to take of unlabeled multimodal data to learn the domain of vessels. Specially, a set of paired retinal images of different modalities were given for automatically constructing a network for segmenting vasculature in a self-supervised manner. In [23], to describe the vasculature, the proposed network was trained in the prediction of multi-instance heatmaps to model the likelihood of a pixel being a landmark location. Some new loss functions were also proposed for discriminating vessels. Hu et al. [24] designed an improved class-balanced entropy loss function and focused on the indistinguishable examples of blood vessels. Likewise, Yan et al. [6] proposed a novel joint pixel-wise and segment-level loss, aiming on handling the imbalance of various vessel types. In addition to modification of loss function, the architecture was optimized by introducing novel modules. A deformable convolution was introduced in Jin et al. [7] to capture vessels by adaptive adjusting to the vessels scales and shapes. Wu et al. [8] raised a cascaded followed network with applying multiscale models to segment multi-width retinal vessels. And, Alom et al. [25] came up with a novel recurrent residual network for better feature representation of vascular segmentation.

These implementations aimed at resolving the difficulty of discriminating various vessels but still remained a challenge to exactly extract vasculature. With respect to abnormal retinal images, few approaches demonstrated the robustness of lesions. To address the limitation, we demonstrated a state-of-the-art performance by utilizing our proposed multi-scale upsample attention module and dense module, which exploit the inter-channel relationship in a multi-scale manner. Several works of attention mechanism are briefly reviewed.

2.4. Attention mechanism

Attention mechanism was originally designed in the context of neural machine translation. Later it was proved to have a significant potential in various computer vision tasks, such as classification [26,27] and segmentation [28,29]. Among various attention mechanisms, channel attention focuses on exploiting the meaningful channels of the given feature and bias the allocation of fea-

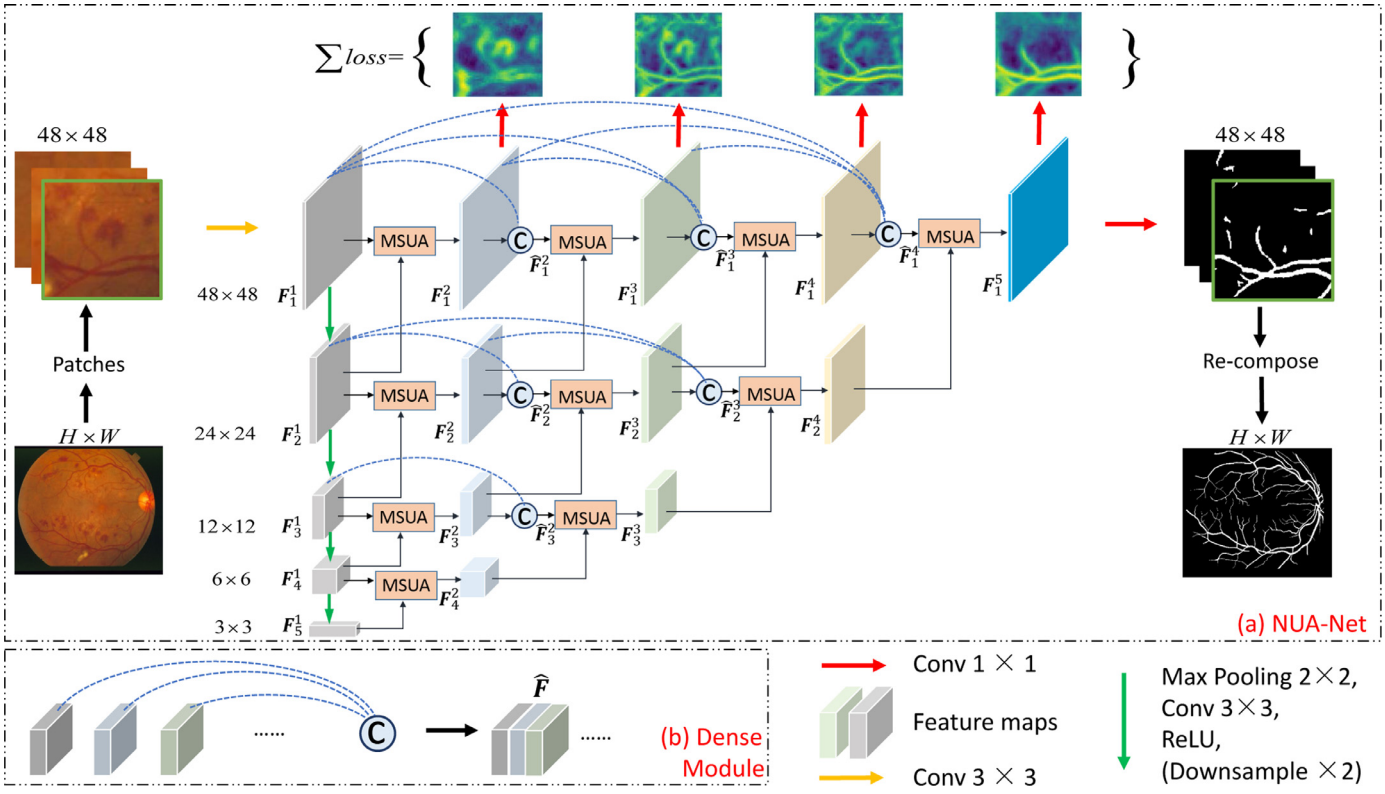

Fig. 1. (a) Framework of NUA-Net. (b) Pathway of dense module. Please note the figure is best viewed in color.

tures towards the most informative components. It generally first squeezes the spatial information of input features to derive the channel descriptor, and then applies the derived descriptor on the input features to re-allocate the channel components. Inspired by its effectiveness, we proposed to employ channel attention for retinal vascular segmentation. Specifically, different from most existing channel attention mechanisms which were performed on feature maps of the same scale, we proposed a multi-scale channel attention mechanism to adapt to the task of vascular segmentation, to additionally leverage the relationship among multi-scale features.

This paper is organized as follows: we introduce the detailed design of our proposed method in Section 3. In Section 4 we give the experimental results and analysis. We further discuss the importance of our proposed modules in Section 5, and also give cross-testing evaluations. The conclusions are finally given in Section 6.

3. Methodology

In this section, we describe in details our proposed nested U-shape attention network (NUA-Net) and also give metrics for measuring performance.

3.1. NUA-Net architecture

The overall structure of NUA-Net is depicted in Fig. 1. We extract patches from original image as inputs, and predict pixel-wise soft segmentation. We extend the U-net [10] with nested connections and our network consists of an encoding stage and 4 decoding stages. We refer to $\mathbf{F}_d^n$ as the feature map at the dth scale and nth stage, where $n = 1$ represents the encoding stage and $n \in [2, 5]$ is the decoding stage. The resolution of feature map is halved once the scale increases.

We firstly use a 3×3 conv layer to extract shallow feature, i.e., $\mathbf{F}_1^1$. In the encoding stage, to generate encoding units $\mathbf{F}_d^1$ at the dth scale (denoted as blue blocks in Fig. 1), we downsample the features at the $(d-1)$th scale by a factor of 2. The process can be formulated as: $\mathbf{F}_d^1 = \mathcal{H}(\mathbf{F}_{d-1}^1), d \in [2, 5]$ where $\mathcal{H}(\cdot)$ is a 2×2 max pooling followed by 3×3 convolution with batch normalization, ReLU and Dropout. The decoding stage adopts a bottom-up strategy to progressively derive the features at a larger scale. To harness the mutual relationship among multi-scale features for upsampling, we propose a Multi-Scale Upsampling Attention (MSUA) module. A new joint loss is further designed to put supervisions on each decoding stage.

3.1.1. Dense module

Vessels in fundus images are complex and diverse, making them difficult to detect. The widths of vessels vary largely, from one pixel to dozens of pixels. Besides, vessels are difficult to differentiate from the various lesions. To detect the diverse vessels, both local features and contextual features are of great importance. Local shallow features such as color, edges, local contrast are important signs for detecting tiny vessels which are however easily discarded in large-scale features due to the pooling layer. Comparatively, deeper features encode contextual information from a large receptive field, and are informative for differentiating irrelevant objects. To this end, our model employs multiple decoding stages where features are progressively derived by leveraging features from previous stages and larger scales, as can be seen from Fig. 1. Through implicit supervision on shallow feature, we utilize dense module to combine features from different stages with supervision to learn richer feature representations. In addition, the dense module also has the regularizing effect to allow flexible information flow among different layers, as illustrated in Huang et al. [30], which can ease the training of deep networks.

In each scale, we concatenate features from all preceding stages. The concatenation propagates the aggregation of all previous features for better context information preservation. We give formu-

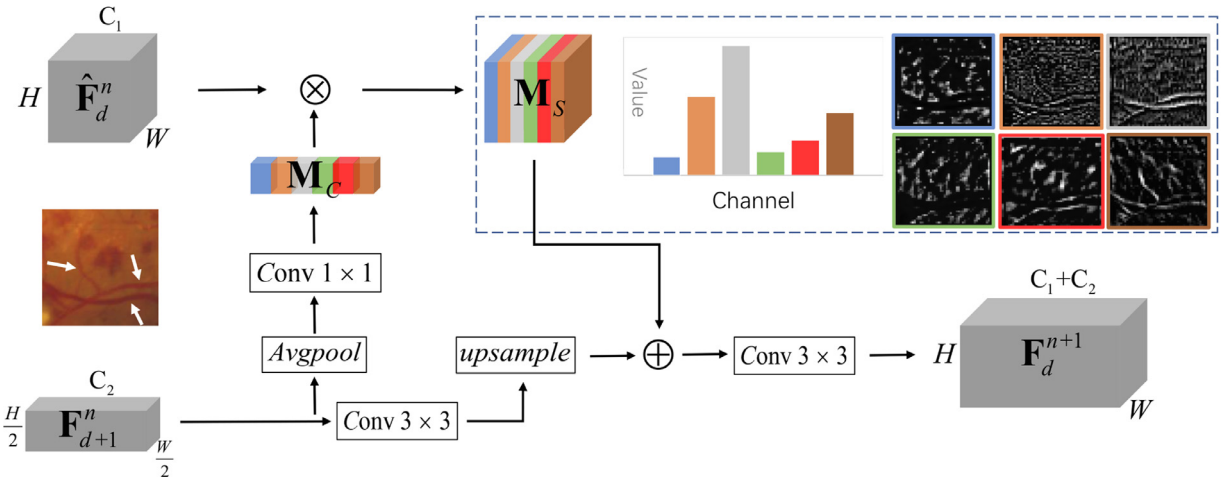

Fig. 2. An overview of MSUA module structure. The MSUA decodes the vessels' feature by weighting each channel from multi-scale architecture. The channels with higher activation are more likely to contain the feature of vasculature.

lation as follows:

$$\hat{\mathbf{F}}_d^n = [\mathbf{F}_d^1, \mathbf{F}_d^2, \ldots, \mathbf{F}_d^n] \tag{1}$$

where $[\cdot]$ denotes the concatenation operation of the feature maps. The dense future $\hat{\mathbf{F}}_d^n$ is then served as one of inputs for the MSUA module.

3.1.2. MSUA module

Channels in feature maps encode different information in fundus images. The channel attention can be regarded as a mechanism to automatically re-weight different attributes. The previous works [26,27] exploited channel dependencies by squeezing information through average pooling. And the generating channel descriptor is used to map a set of re-weighted channels. To obtain a channel descriptor for vasculature, we also adopt the average pooling. Our proposed attention mechanism is developed by attending information from larger scale to smaller scale, as discussed in Section 3.1.1. The main modification to existing channel attention mechanisms is that instead of applying self-attention on the input, we perform re-weighted channel descriptor on the shallower input which is concatenated through dense module to exploit multi-scale features. Accordingly, the multi-scale integration is performed in two folds: (1) The multi-scale attention and (2) concatenation of features. To fully utilize hierarchical information, our MSUA module separates out upsampling from convolution to compute features by using re-weighted channels of feature maps

The diagram of MSUA module is depicted in Fig. 2. The MSUA module takes as inputs the densely connected feature $\hat{\mathbf{F}}_d^n \in \mathbb{R}^{h\times w\times C_1}$ as well as larger scale $\mathbf{F}_{d+1}^n \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times C_2}$. In order to exploit channel dependencies, we squeeze features along the spatial dimension to encode all discriminative information and construct a channel descriptor. Note that the max-pooling only sustains the most discriminative regions since other values with lower scores are discarded. To this end, we aggregate spatial information by applying average pooling followed by 1×1 convolution layer to encode the channel-wise dependency. Hence, the channel attention descriptor $\mathbf{M}_C \in \mathbb{R}^{1\times 1\times C_1}$ is computed as:

$$\mathbf{M}_C = f^{1\times 1}\left(\frac{1}{\frac{h}{2}\times\frac{w}{2}}\sum_{i=1}^{\frac{h}{2}}\sum_{j=1}^{\frac{w}{2}}\mathbf{F}_{d+1}^n(i,j)\right) \tag{2}$$

where $f^{1\times 1}$ denotes 1×1 convolution operation mapping input to the same number of channels as $\hat{\mathbf{F}}_d^n$. We then apply the attention descriptor into $\hat{\mathbf{F}}_d^n$ for obtaining the re-weighted feature $\mathbf{M}_S$ by:

$$\mathbf{M}_S = \mathbf{M}_C \otimes \hat{\mathbf{F}}_d^n \tag{3}$$

where $\otimes$ denotes element-wise multiplication. Specifically, the values of channel-wise attention vector $\mathbf{M}_C$ are first broadcasted along the spatial dimension of $\hat{\mathbf{F}}_d^n$ before performing element-wise multiplication to obtain refined feature map $\mathbf{M}_S \in \mathbb{R}^{H\times W\times C_1}$. Larger scale $\mathbf{F}_{d+1}^n$ is then upsampled and combine with the re-weighted feature for multi-scale information, to obtain the decoding feature $\mathbf{F}_d^{n+1}$ as:

$$\mathbf{F}_d^{n+1} = f^{3\times 3}(\mathcal{U}(f^{3\times 3}(\mathbf{F}_{d+1}^n)) \oplus \mathbf{M}_S) \tag{4}$$

where $f^{3\times 3}$ denotes 3×3 convolution operation, $\oplus$ is a concatenation operation and $\mathcal{U}$ stands for bilinear interpolation. $\mathbf{F}_d^{n+1} \in \mathbb{R}^{H\times W\times(C_1+C_2)}$ is a hybrid feature map that incorporates cross-scale information in a hierarchy. Specially, the decoding feature $\mathbf{F}_d^{n+1}$ can be obtained through MSUA module $\mathcal{M}(\cdot)$ and formulated as follows:

$$\mathbf{F}_d^{n+1} = \mathcal{M}(\hat{\mathbf{F}}_d^n, \mathbf{F}_{d+1}^n) \tag{5}$$

We then apply a 1×1 conv layer to the feature of last decoding stage to obtain the final output.

3.1.3. Joint loss

We adopt cross-entropy loss function, which is commonly used in pixel-wise classification task, to train our network. Our NUA-Net contains four decoding stages. Each stage propagates the feature map to gradually increasing scale. Instead of using solely the final output as training loss, we propose a joint loss to put supervision on each decoding stage. With the usage of our joint loss, the gradient can directly flow through multiple stages at multi-scale. This implicit supervision helps the network learn shallow feature from earlier stages.

We use an additional conv layer following each largest scale feature map in decoding stage to decrease the channel dimension and predict the probability of the pixels belonging to the blood vessels. We denote the predicted probability of decoding stage n as $\hat{\mathbf{Y}}^n \in [0,1]$ and the ground-truth label as $\mathbf{Y} \in \{0,1\}$. We use $\{\mathbf{Y}^n\}_{n=2}^5$ to define our joint loss, as follows:

$$\mathcal{L} = -\frac{1}{4\times K}\sum_{n=2}^{5}\sum_{i=1}^{K}\left(\mathbf{Y}_i^n \log\left(\hat{\mathbf{Y}}_i^n\right) + \left(1-\mathbf{Y}_i^n\right)\log\left(1-\hat{\mathbf{Y}}_i^n\right)\right) \tag{6}$$

where K refers to the number of total pixels in the input feature map.

3.2. Performance measures

Retinal vascular segmentation can be regarded as a pixel-wise binary classification problem, where each pixel is classified as either vessel or non-vessel. We use four metrics: Sensitivity (Sen), Specificity (Spe), Accuracy (Acc) and Precision (Pr) for evaluation. The definition of these metrics is as follows:

$$\begin{aligned} Sen &= Recall = \frac{TP}{TP+FN}, \\ Spe &= \frac{TN}{TN+FP}, \\ Acc &= \frac{TP+TN}{TP+TN+FP+FN}, \\ Pr &= \frac{TP}{TP+FP}, \end{aligned} \tag{7}$$

where TP is the True Positive defined as the number of pixels that are correctly classified as vessels, TN is the True Negative defined as a pixel is correctly identified as non-vessel, FN is the False Negative where a pixel is mistakenly identified as non-vessel however in fact a vessel, and FP is the False Positive where a pixel is identified as a vessel but is in fact a non-vessel.

The Precision-Recall (PR) curve and Receiver Operating Characteristic (ROC) curve are computed by Area Under the Curve (AUC), which achieves 1 for a perfect system. PR curve is further plotted with Recall versus Pr. The closer a curve approaches the top right corner, the better the performance of the system.

4. Experiment

In this section, we conduct experiments to compare our NUA-Net with state-of-the-art methods on the widely-used benchmark datasets. We first introduce our preprocessing method, and then brief the used datasets. We finally give the experimental results and analysis.

4.1. Preprocessing

Several retinal image preprocessing methods are proposed to enhance feature extraction. To keep a wide color range in retinal images, we first transfer the raw image from RGB color space to Lab space and apply Contrast Limited Adaptive Histogram Equalization (CLAHE) [31] to the L color channel. The Lab image is then transferred back to RGB color space. The red channel is the brightest however with low contrast and the blue channel exhibits poor dynamic range. Comparatively, blood containing elements in the retinal layer are best represented and reach better contrast in the green channel [14]. To this end, we use the green channel images as network inputs, i.e., the input image is monochrome. We refer to $\mathbf{I}_j$ as the input image and $\{\mathbf{I}_j\}_{j=1}^{N}$ as the entire training dataset with N samples. We normalize $\mathbf{I}_j$ to the range of [0,1] as follows:

$$\mathbf{I}_j^{norm} = \frac{\mathbf{I}_j - \min\limits_{I \in \mathbf{I}_j} I}{\max\limits_{I \in \mathbf{I}_j} I - \min\limits_{I \in \mathbf{I}_j} I} \tag{8}$$

The normalized image $\mathbf{I}_j^{norm}$ is finally used as input to train our proposed network.

4.2. Materials

We evaluate our model on five public datasets: DRIVE, STARE, CHASE_DB1, HRF and IOSTAR. The STARE dataset contains two annotated segmentation results generated by two different experts, respectively. We follow previous works [3,5,11,18,21] to use the segmented results by the first human observer as our ground-truth labels. The binary masks of the field of view (FOV) for images in DRIVE database are available. Since the masks are not provided in STARE and CHASE_DB1 datasets, we follow [14] to manually create corresponding masks.

The DRIVE [2] stands for Digital Retinal Images for Vessel Extraction. The data was collected from Netherlands and the screening population consisted of 400 diabetic subjects between 25 and 90 years of age. The DRIVE dataset contains 40 images in total and 33 of them do not show any sign of diabetic retinopathy and the remaining 7 images show signs of mild, early diabetic retinopathy. The images were obtained using a Canon CR5 non-mydriatic 3CCD camera with a 45-degree FOV. The size of each image is 564×584. Images have been cropped around the FOV and masks are provided. There are 20 images in training and testing sets respectively without overlap. Manual segmentation from an experienced ophthalmologist is available as a gold standard. All images are restored in JPEG format.

The STARE [4] (Structured Analysis of the Retina) dataset contains 20 blood vessel images. The digitized slides were captured by a TopCon TRV-50 fundus camera at 35° FOV and digitized to 605×700 pixels. There are two observers manually segmenting all images. We use the segmentation labels by the first observer as ground-truth. Note that the diagnoses for each image are also provided. Since there is no standard division of training and testing set for STARE, we use the leave-one-out cross-validation. All images are represented by.ppm format.

The CHASE_DB1 [15] dataset is a new retinal reference dataset acquired from children. It is a part of the Child Heart and Health Study in England and is obtained from both left and right eyes of 14 children. The images denote connections between retinal vessel tortuosity and early risk factors for cardiovascular disease. They were collected by a hand-held Nidek NM-200-D fundus camera. Each image in this dataset is of 30° FOV with resolution 999×960. Since there is no standard split for training and testing subsets, we follow [32] to use the first 20 images for training, and the remaining 8 images for testing.

Fundus images in the HRF [33] dataset were captured with 60° FOV of resolution 3304×2336. It contains 15 images of healthy patients, 15 images of patients with diabetic retinopathy, and 15 images of glaucomatous patients. We followed setting [34] that a train/test split of 22/23 is adopted there. Unlike [18], our segmentation result is obtained and compared with original manual annotation without downsample.

The IOSTAR [35] dataset was collected with an EasyScan camera (i-Optics Inc., the Netherlands), which is based on a Scanning Laser Ophthalmoscopy (SLO) technique. It includes 30 images with a resolution of 1024×1024 pixels. The first 20 images were used for training and remaining ones for testing.

4.3. Implementation details

To prevent from overfitting, image random cropping was used as data augmentation during training. For each dataset, we cropped a total number of 22,000 patches (of size 48×48) to constitute our training samples. Patches were further randomly flipped horizontally or vertically and rotated by 30° as data augmentation. In inference stage, we divided each image into overlapped patches as inputs to feed into the network. The output segmented patches were then re-composed to construct an image. Detailedly we set $stride = 5$ when cropping patches. The predicted values at the overlapped regions are obtained by averaging multiple predictions.

The base channel number, i.e., the channel number of feature $\mathbf{F}_1^1$, was set as 16. In the encoding stage, we doubled the channel number as the feature resolution was halved. The channel number of the encoding features was $16 \times \{1, 2, 4, 8, 16\}$ respectively. As for the four decoding stages, the channel number was

Table 1
Quantitative comparison with other methods on DRIVE dataset.

Method	Year	Sen (%)	Spe (%)	Acc (%)	ROC AUC (%)
2nd Human Observer	–	77.60	97.24	94.72	–
Fraz et al. [15]	2012	74.06	98.07	94.80	97.47
Azzopardi et al. [11]	2015	76.55	97.04	94.42	96.14
Roychowdhury et al. [36]	2015	73.90	97.80	94.90	96.70
Li et al. [32]	2015	75.69	98.16	95.27	97.38
Orlando et al. [18]	2016	78.97	96.84	–	95.07
DeepVessel [21]	2016	76.03	–	95.23	–
Zhang et al. [16]	2017	78.61	97.12	94.66	97.03
Zhao et al. [17]	2017	77.40	97.90	95.80	97.50
Yan et al. [6]	2018	76.53	98.18	95.42	97.52
Hu et al. [24]	2018	77.72	97.93	95.33	97.59
R2U-Net [25]	2018	77.92	98.13	95.56	97.84
MS-NFN [8]	2018	78.44	98.19	95.67	98.07
Sazak et al. [13]	2019	71.80	98.10	95.90	94.60
Wang et al. [19]	2019	76.48	98.17	95.41	–
DUNet [7]	2019	78.94	**98.70**	96.97	98.56
NUA-Net	2019	**80.60**	98.55	**97.09**	**98.78**

Table 2
Quantitative comparison with other methods on STARE dataset.

Method	Year	Sen (%)	Spe (%)	Acc (%)	ROC AUC (%)
2nd Human Observer	–	89.52	93.84	93.49	–
Fraz et al. [15]	2012	75.48	97.63	95.34	97.68
Azzopardi et al. [11]	2015	77.16	97.01	94.97	94.97
Roychowdhury et al. [36]	2015	73.20	98.40	95.60	96.70
Li et al. [32]	2015	77.26	98.44	96.28	98.79
Orlando et al. [18]	2016	76.80	97.38	–	–
DeepVessel [21]	2016	74.12	–	95.85	–
Zhang et al. [16]	2017	78.82	97.29	95.47	97.40
Zhao et al. [17]	2017	78.80	97.60	95.70	95.90
Yan et al. [6]	2018	75.81	98.46	96.12	98.01
Hu et al. [24]	2018	75.43	98.14	96.32	97.51
R2U-Net [25]	2018	82.98	98.62	97.12	99.14
Sazak et al. [13]	2019	73.00	97.90	96.20	96.20
Wang et al. [19]	2019	75.23	98.85	96.40	–
DUNet [7]	2019	74.28	**99.20**	97.29	98.68
NUA-Net	2019	**85.11**	99.00	**97.94**	**99.44**

increased as concatenation layer was used in both dense module and MSUA module. Detailedly, the channel numbers of each decoding feature (from top to bottom) for the four decoding stages are $16 \times \{3, 6, 12, 24\}$, $16 \times \{10, 20, 40\}$, $16 \times \{34, 68\}$ and 16×114, respectively. The ReLU was used as activation function. We trained our network from scratch and all the weights were initialized using normal distribution. We used standard stochastic gradient descent (SGD) optimizer and set the learning rate as 0.0001 fixedly. The dropout rate was 0.2 and the batch size was set as 4. Code was written in Pytorch library. Our model was trained on NVIDIA Titan XP GPU, with 32GB of RAM. We trained our network for 200 epochs.

4.4. Results

The quantitative comparison on DRIVE dataset is shown in Table 1. One can see that our proposed method achieves the state-of-the-art performance on most metrics. The Sen, Acc and ROC AUC of our method are 0.8060, 0.9709 and 0.9878, respectively, which are superior to the second best method (DUNet [7]). As for the Spe metric, our method achieves nearly comparable performance compared with DUNet (0.9855 vs 0.9870). These obviously demonstrate the effectiveness of our NUA-Net, owing to the designed MSUA module and the integration of dense module.

In Table 2 we present the quantitative performance on STARE dataset. Note that half of the images in STARE dataset are abnormal where vessels are usually occluded by lesions and much more difficult to detect. From Table 2 one can see that our model achieves significantly better performance on Sen, Acc and AUC, compared with state-of-the-art methods. Specifically, the Sen of our NUA-Net is 0.8511, which is 2.5% higher than the second place, i.e., 0.8298 by R2U-Net [25]. This no-doubtfully demonstrates a remarkable segmenting capability of our model, especially in handling abnormal images. Besides, our method achieves 0.9900 for Spe which is slightly inferior to DUNet (0.9920), whereas the Sen of DUNet (0.7428) is largely inferior to our model (0.8511). This is probably because that DUNet tends to classify pixels as non-vessels with a high probability. This will lead to high Spe whereas a relatively low Sen simultaneously considering the imbalance distribution of vessels and non-vessels in abnormal images. Our NUA-Net, on the other hand, shows good performance for both Sen and Spe, demonstrating its robustness in segmenting vessels.

The quantitative comparison on CHASE_DB is shown in Table 3. Our proposed method obtains the best result among all competing methods in Spe, Acc and AUC, which are 0.9859, 0.9748 and 0.9884 respectively. The Sen of our model is 0.8106, which is slightly lower than 0.8229 obtained by DUNet.

To further validate the application scope of proposed method, we conducted the experiment on other up-to-date public datasets: HRF and IOSTAR. HRF contains high resolution retinal images which express more details of vessel contours. For IOSTAR dataset, the fundus images were acquired based on the new SLO technique. From Table 4 on all metrics: Sen, Spe, Acc and AUC, NUA-Net achieves the 0.8554/0.8392, 0.9741/0.9754, 0.9648/0.9646 and 0.9824/0.9823 on HRF/IOSTAR respectively. In comparison with other state-of-the-art methods reported in the literature,

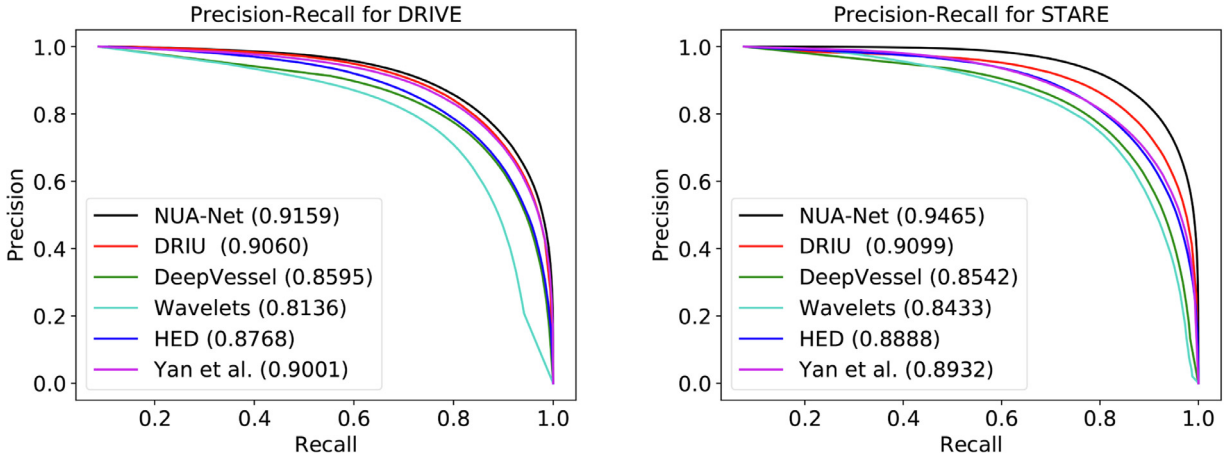

Fig. 3. Precision and Recall curves compared with different segmentation methods on DRIVE (left) and STARE (right) datasets respectively.

Table 3
Quantitative comparison with other methods on CHASE_DB dataset.

Method	Year	Sen (%)	Spe (%)	Acc (%)	ROC AUC (%)
2nd Human Observer	–	81.05	97.11	95.45	–
Fraz et al. [15]	2012	72.24	97.11	96.49	97.12
Li et al. [32]	2015	75.07	97.93	95.81	97.16
Azzopardi et al. [11]	2015	75.85	95.87	93.87	94.87
Zhang et al. [35]	2016	76.26	96.61	94.52	96.06
DeepVessel [21]	2016	71.30	–	94.89	–
Orlando et al. [18]	2017	72.77	97.12	–	95.24
Zhang et al. [16]	2017	76.44	97.16	95.02	97.06
Yan et al. [6]	2018	76.33	98.09	96.10	97.81
R2U-Net [25]	2018	77.56	98.20	96.34	98.15
MS-NFN [8]	2018	75.38	98.47	96.37	98.25
Wang et al. [19]	2019	77.30	97.92	96.03	–
DUNet [7]	2019	**82.29**	98.21	97.24	98.63
NUA-Net	2019	81.06	**98.59**	**97.48**	**98.84**

it demonstrates the best performance except that the result of Spe is slightly lower than the best Annunziata (0.9836) and Zhao (0.9772). According to experimental results on five publicly datasets, it is noted that our model obtains the value of Sen all higher than 0.80. In various scenarios, NUA-Net all achieves competing results with other algorithms on the benchmark.

Moreover, in the domain of fundus, the number of negative examples (non-vessels) greatly exceeds the number of positive examples (vessels). A huge shift in the value of False Positives (FP) can lead to a tiny change in the False Positive Rate ($\frac{FP}{FP+TN}$) used in ROC analysis. Thus, the Pr ($\frac{TP}{TP+FP}$) is more prevalent under the condition of class imbalance. We further present the PR curves with AUC score as quantitative metric, as shown in Fig. 3. We choose DRIU [20], DeepVessel [21], Wavelets [14], HED [37] and Yan et al [6] for comparison. From Fig. 3 one can see that our NUA-Net achieves the best performance on both DRIVE and STARE dataset. The improvement on STARE dataset is even more distinct, especially in the range [0.6, 1.0] of recall and precision axes, which further demonstrates the effectiveness of our NUA-Net.

To highlight the robustness of our model, we examine its performance on segmenting vessels under the condition of abnormality, as shown in Table 5. On HRF, the pathological images are classified as Diabetic Retinopathy and Glaucomatous. Comparatively, abnormal images on STARE are aggregated as one category even they have distinguishing lesions according to the diagnosis.[1] We evaluate our model specifically on abnormal images on STARE and HRF datasets respectively with other approaches whose results were reported in the literature. Note that NUA-Net outperforms others by a large margin in terms of Sen, Acc and AUC especially on STARE. For abnormal images on HRF and STARE, only the performance of Spe is inferior to the result from Annunziata. Undoubtedly, our model significantly promotes the robustness of detecting vessels against other approaches.

We present the segmentation results from each dataset to demonstrate qualitative analysis illustrated in Fig. 4 compared with manual annotations. In each pixel, red color in the segmentation maps indicates false negatives that predictions fail to get the vessels. And green color represents that non-vessels are wrongly classified as true blood vessels. As shown in the results, our method is capable of preserving vessels as possible from various fundus sources. It is worth mentioning that false negatives may stem from challenging cases which will be discussed next.

5. Discussion

5.1. Challenging cases

In this section we present the visualization results to evaluate the qualitative performance on dealing with some common segmentation challenges. The segmentation results are analyzed based on some challenging cases and difficult conditions of retinal images. In vascular segmentation these cases include tiny/wide vessels, parallel vessels, crossing branches of vessels and low intensity/contrast. It is noted that these difficulties could co-exist, making it harder to distinguish vessels from complicated condition. We choose the Orlando et al. [18], DeepVessel [21], Marin et al. [3] and Yan et al. [6] for comparison on STARE and DRIVE as shown in Fig. 5. In terms of FOV, illumination and vessel structure, fundus images in CHASE_DB dataset are quite different from others. The visual predictions on CHASE_DB are also given in Fig. 6 and we compare against DeepVessel and Yan reported in the literature. For HRF and IOSTAR dataset, segmentation results of other approaches are unavailable. From these comparisons, one can see that our proposed method can effectively preserve more imperceptible microvessels in a low-contrast or low-intensity condition. Besides, NUA-Net is able to accurately specify vascular width and distinguish vessels. While Yan's method is also based on the deep learning structure and achieves a comparable performance, our algorithm demonstrates more robust capacity of coping with vessel continuity preservation.

In another hard condition, different pathological manifestation in abnormal retinal images would significantly impact the perfor-

[1] https://cecas.clemson.edu/~ahoover/stare/diagnoses/all-mg-codes.txt.

Table 4
Quantitative comparison with other methods on HRF and IOSTAR dataset.

Method	Year	HRF				IOSTAR			
		Sen	Spe	Acc	AUC	Sen	Spe	Acc	AUC
Odstrcilik et al. [33]	2013	78.60	97.50	95.30	–	–	–	–	–
Annunziata et al. [12]	2015	71.28	**98.36**	95.81	–	–	–	–	–
Zhang et al. [35]	2016	79.78	97.17	95.56	96.08	75.45	97.40	95.14	96.15
Orlando et al. [18]	2017	78.74	95.84	–	–	–	–	–	–
Zhao et al. [17]	2017	74.90	94.20	94.10	97.10	77.20	96.70	94.80	96.00
Zhao et al. [34]	2018	76.08	98.13	–	–	79.15	**97.72**	–	–
Yan et al. [6]	2018	80.84	94.17	92.98	–	–	–	–	–
Sazak et al. [13]	2018	83.10	98.10	96.30	–	–	–	–	–
NUA-Net	2019	**85.54**	97.41	**96.48**	**98.24**	**83.92**	97.54	**96.46**	**98.23**

Table 5
Performance comparison using two sets of abnormal images on the STARE and HRF dataset. Pathological images on the HRF are denoted as Diabetic Retinopathy and Glaucomatous.

Method	STARE			HRF							
	Abnormal			Diabetic retinopathy				Glaucomatous			
	Sen	Spe	Acc	Sen	Spe	Acc	AUC	Sen	Spe	Acc	AUC
Soares et al. [14]	71.81	97.65	95.00	–	–	–	–	–	–	–	–
Fraz et al. [15]	72.62	97.64	95.11	–	–	–	–	–	–	–	–
Li et al. [32]	78.00	98.05	96.72	–	–	–	–	–	–	–	–
DeepVessel [21]	79.64	97.63	96.41	–	–	–	–	–	–	–	–
Orlando et al. [18]	–	–	93.93	–	–	–	–	–	–	–	–
Yan et al. [6]	80.87	97.48	96.35	–	–	–	–	–	–	–	–
Wang et al. [19]	–	–	96.24	–	–	–	–	–	–	–	–
Annunziata et al. [12]	–	–	95.65	69.97	**97.87**	95.54	–	75.66	**97.85**	96.03	–
Odstrcilik et al. [33]	–	–	–	74.63	96.19	94.45	95.89	79.00	96.38	94.97	97.04
Pandey et al. [38]	–	–	–	80.25	96.29	95.76	95.90	82.24	97.81	96.41	96.97
NUA-Net	**83.12**	**99.02**	**97.94**	**82.69**	97.01	**95.95**	**97.73**	**82.91**	97.62	**96.57**	**98.03**

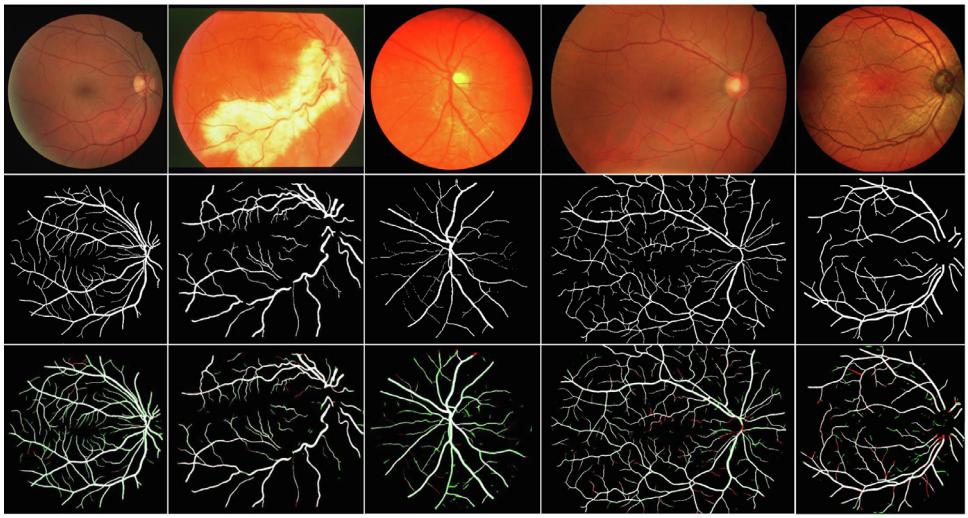

Fig. 4. Visual results of the segmentation. The first row to the third row: Fundus images, ground-truth and segmentation results. The first column to the last column: DRIVE, STARE, CHASED_DB1, HRF and IOSTAR datasets respectively.

mance of automatic vascular segmentation. In Fig. 7, our visual segmentation results of abnormal fundus are presented. The segmentation results reported in other approaches are listed for fair comparison. From top to bottom, fundus have hard exudates, central retinal artery/vein occlusion, drusen and hemorrhages respectively. These lesions are common in abnormal fundus images. And they are hard to discriminate due to analogous characteristics like the blood vessels. Accordingly, it can be seen that most of methods tend to give quantity of false positives in the region of lesions. NUA-Net, on the other hand, differentiates the pathologies and makes exact predictions for potential vessels. The robustness of NUA-Net owes to proposed attention mechanism MSUA module that has the strength on biasing vascular feature. For this reason, NUA-Net manages to preserve vascular structures in harsh condition, demonstrating noticeable promotion of specificity.

5.2. Ablation tests

In this section, we carry out ablation study on two major designs of our NUA-Net: the MSUA module and the dense module. In Fig. 8 we briefly illustrate the network architecture of ablation models. We first implement a baseline encoder-decoder network, which follows the architecture of U-net [10], denoted as *baseline_1*. The number of base filters and down/up-sampling levels are

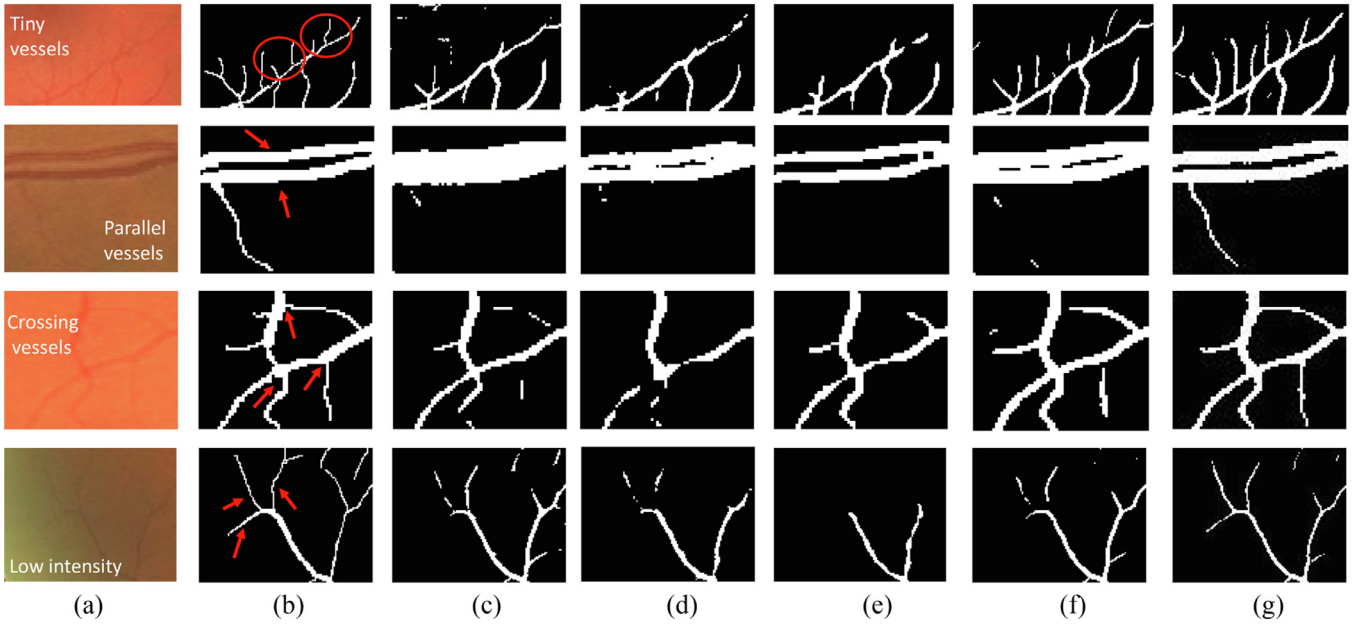

Fig. 5. Visual results compared with other methods on DRIVE and STARE. From left to right column: (a) Original fundus images (b) Ground-truth (c) Orlando et al. [18] (d) DeepVessel [21] (e) Marin et al. [3] (f) Yan et al. [6] (g) NUA-Net.

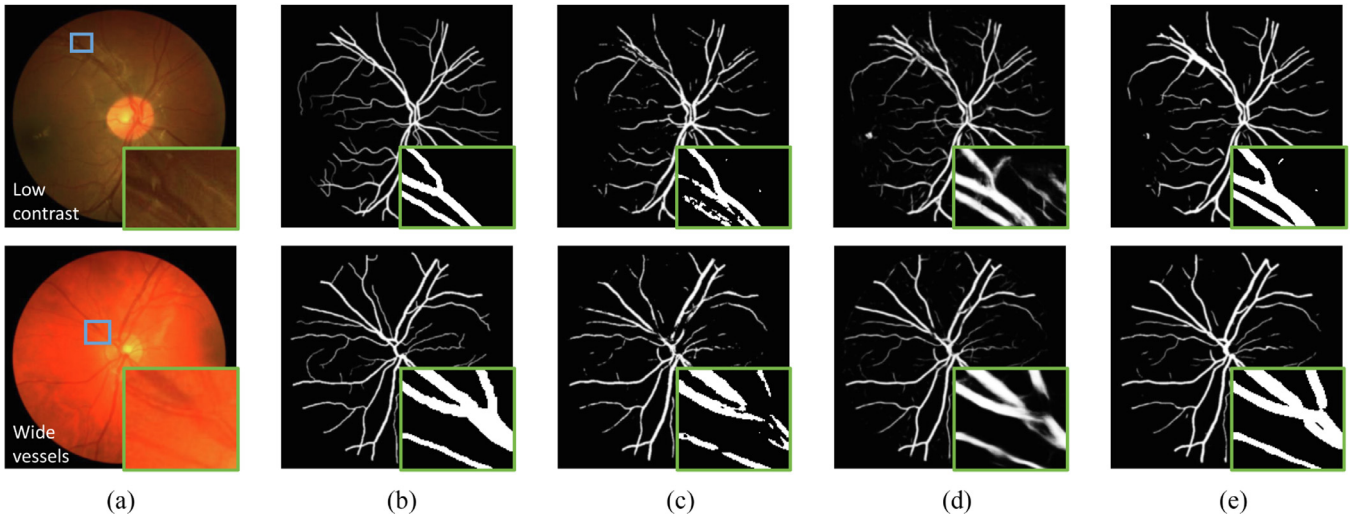

Fig. 6. Visual results compared with other methods on on CHASE_DB dataset. From left to right: (a) Original fundus images (b) Ground-truth (c) DeepVessel [21] (d) Yan et al. [6] (e) NUA-Net.

the same as our NUA-Net, i.e., 16 filters and 5 levels respectively. We use the cross entropy loss between the network output and groundtruth labels for training, considering that the proposed joint loss is not applicable in this model. To evaluate the effectiveness of dense model in our NUA-Net, we implement *model_1* by removing the dense connection from our NUA-Net. The architecture of *model_1* is illustrated in Fig. 8(a). Besides, to validate the effectiveness of MSUA module, we implement the architecture of dense U-net [9], denoted as *baseline_2* with the vanilla upsampling layer. To further validate the role of MSUA module, we gradually (stage by stage) add MSUA modules to *baseline_2*. We refer to k as the number of stages that employing MSUA module. A brief illustration of NUA-Net ($k = 2$) is shown in Fig. 8(b). For fair comparison, all the hyper-parameters were kept the same for training these ablation models. The same loss function, i.e., joint loss, is applied on these ablation models except for *baseline_1*. And the same pre-processing is used for all these models. We use the DRIVE dataset for ablation study and the results are shown in Table 6.

The Baseline_2 [9] which employs solely dense module achieves 0.7900, 0.9860, 0.9698 and 0.9865 for Sen, Spe, Acc and AUC respectively, which all outperform the Baseline_1 [10]. This validates the effectiveness of dense module for vascular segmentation. By comparing model_1 with the baseline models, one can see that the incorporation of MSUA model achieves 0.7857, 0.9919, 0.9671 and 0.9834 for Sen, Sep, Acc and AUC respectively. All metrics are promoted by a large margin and this undoubtedly proves the important role of MSUA module. Note that model_1 obtains the highest Spe of 0.9919 among all methods, which means that the incorporation of MSUA promotes the robustness to non-vessels. Moreover, by combining MSUA module and dense module ($k = 1$), the performance can be further boosted, e.g., improving Sen from 0.7900 to 0.7989. We also notice that as the increase of k, the overall performance could be promoted a little more.

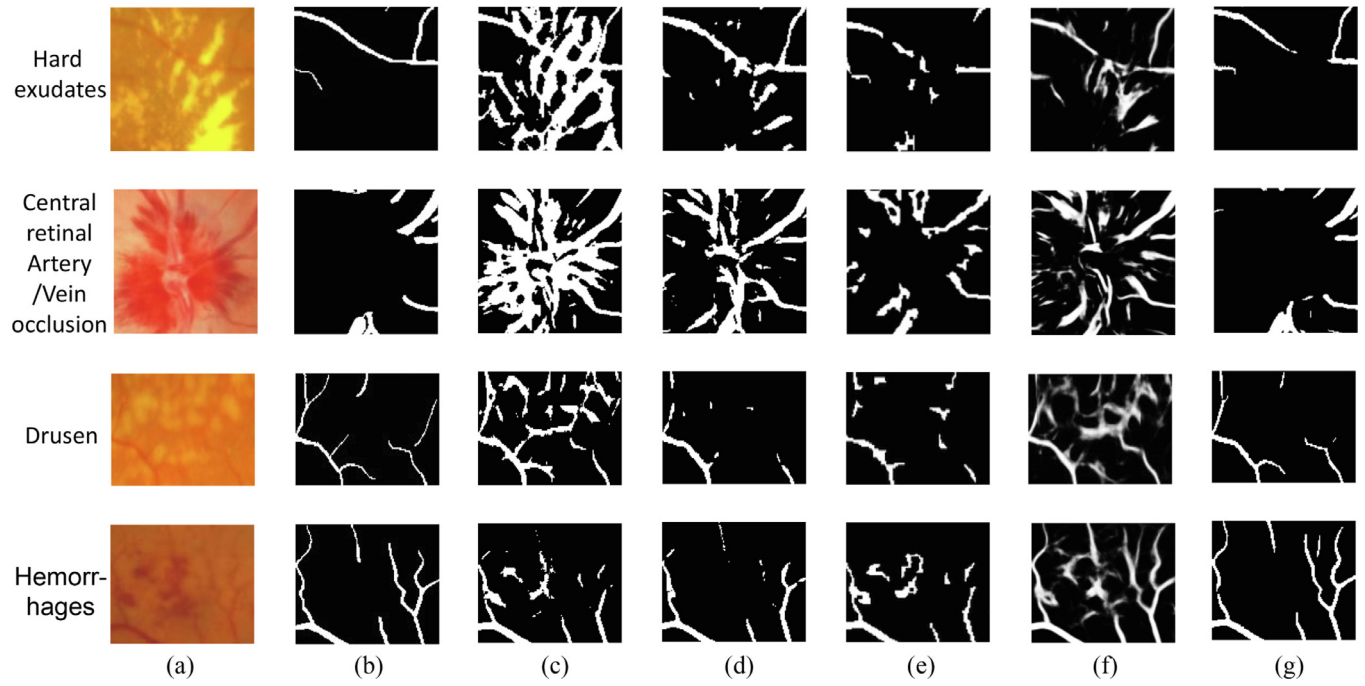

Fig. 7. Visual results of abnormal images compared with other methods. From left to right column: (a) Original fundus images (b) Ground-truth (c) Orlando et al. [18] (d) DeepVessel [21] (e) Marin et al. [3] (f) Yan et al. [6] (g) NUA-Net.

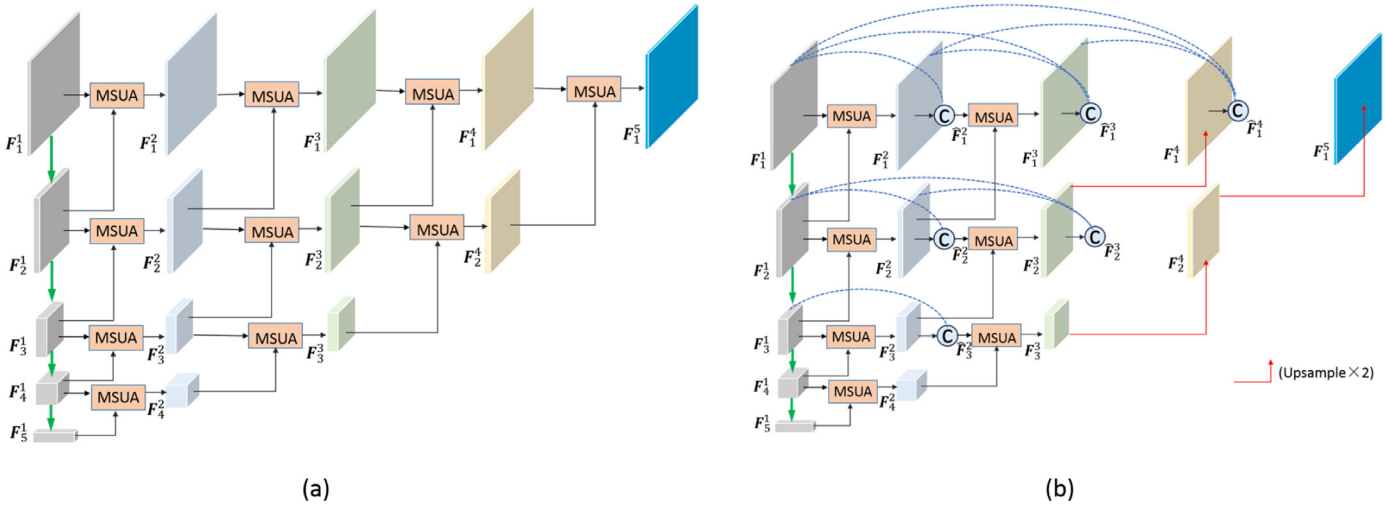

Fig. 8. Ablation models.(a) Model_1 (b) NUA-Net ($k = 2$).

Table 6
Performance comparison of different models on DRIVE dataset.

Method	Dense module	MSUA	Sen (%)	Spe (%)	Acc (%)	ROC AUC (%)
baseline_1 [10]			75.17	98.10	95.50	96.90
baseline_2 [9]	✓		79.00	98.60	96.98	98.65
model_1		✓	78.57	**99.19**	96.71	98.34
NUA-Net ($k = 1$)	✓	✓	79.89	98.62	96.98	98.67
NUA-Net ($k = 2$)	✓	✓	80.43	98.63	97.06	98.76
NUA-Net ($k = 3$)	✓	✓	80.58	98.61	97.08	98.77
NUA-Net ($k = 4$)	✓	✓	**80.60**	98.55	**97.09**	**98.78**

5.3. Cross-testing evaluations

Retinal images acquired by different devices have different FOV, illumination and structures. To verify the generalization capability of the proposed model, in this section we carry out cross-dataset evaluation. Specifically, we use network trained on one dataset to test on another dataset, and vice versa. Unlike [5,32], we do not collect samples from two or more datasets for training since it is more meaningful to measure generalization using a single dataset. We choose methods [5,6,15,32] which also conducted cross-testing for comparison. The cross-dataset results are shown in Table 7. Our NUA-Net achieves the best performance on all metrics when tested

Table 7
Performance of the cross-dataset evaluation.

Test set	Train set	Method	Sen (%)	Spe (%)	Acc (%)	ROC AUC (%)
DRIVE	STARE	Liskowski et al. [5]	–	–	94.16	96.05
		Fraz et al. [15]	72.42	97.92	94.56	96.97
		Li et al. [32]	72.73	98.10	94.86	96.77
		Zhang et al. [16]	–	–	94.47	95.93
		Yan et al. [6]	72.92	98.15	94.94	95.99
		Wang et al. [19]	–	–	94.95	–
		NUA-Net	**73.23**	**98.23**	**95.51**	**97.18**
STARE	DRIVE	Liskowski et al. [5]	–	–	95.05	95.95
		Fraz et al. [15]	70.10	97.70	94.95	96.60
		Li et al. [32]	70.27	98.28	95.45	96.71
		Zhang et al. [16]	–	–	94.88	96.76
		Yan et al. [6]	72.11	**98.40**	95.69	**97.08**
		Wang et al. [19]	–	–	95.73	–
		NUA-Net	**73.21**	98.37	**96.46**	96.69

on DRIVE dataset. This demonstrates the great generalization ability of our model, owing to the proposed MSUA model and dense block. As for test on STARE dataset, our model shows generalization ability on most of the metrics whereas the ROC AUC is sightly lower. This is probably because limited number of abnormal images in DRIVE will undermine MSUA's role. We can conclude that NUA-Net achieves more competitive performance and better generalization when trained on sufficient pathological patches.

6. Conclusion

In this paper, we propose a nested U-shape network with a new multi-scale upsample attention (MSUA) module for effective retinal vascular segmentation by leveraging multi-scale features. Our approach exploits the interdependence among hierarchical features in upsampling module owing to MSUA. The comparison studies and experimental results on all public retinal datasets demonstrate the state-of-the-art performance of our approach for vascular segmentation. We also evaluate the segmentation performance on pathological fundus images. Both quantitative and qualitative results demonstrate the robustness of our method in handling fundus image with lesions and microvessels. Still, due to GPU memory, the segmentation result is constructed by tiling sliding overlapped patches. Accordingly, it takes extra time to infer a complete high-resolution fundus image. For future work, we plan to investigate the optimization of current network to facilitate its processing time.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are most grateful for the constructive advice on the revision of the manuscript from the EiC and anonymous reviewers. The funding support from Hong Kong Government under its GRF scheme and the research grant from both Hong Kong Polytechnic University and Shenzhen Institute of Information Technology are greatly appreciated. The authors would like to express their appreciation for the support from NVIDIA Corporation with the donation of the Titan XP GPU for the research work.

References

[1] M.D. Abràmoff, M.K. Garvin, M. Sonka, Retinal imaging and image analysis, IEEE Rev. Biomed. Eng. 3 (2010) 169–208.
[2] J. Staal, M.D. Abràmoff, M. Niemeijer, M.A. Viergever, B. Van Ginneken, Ridge-based vessel segmentation in color images of the retina, IEEE Trans. Med. Imaging 23 (4) (2004) 501–509.
[3] D. Marín, A. Aquino, M.E. Gegúndez-Arias, 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. Imaging 30 (1) (2010) 146–158.
[4] A. Hoover, V. Kouznetsova, M. Goldbaum, Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response, IEEE Trans. Med. Imaging 19 (3) (2000) 203–210.
[5] P. Liskowski, K. Krawiec, Segmenting retinal blood vessels with deep neural networks, IEEE Trans. Med. Imaging 35 (11) (2016) 2369–2380.
[6] Z. Yan, X. Yang, K.-T. Cheng, Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation, IEEE Trans. Biomed. Eng. 65 (9) (2018) 1912–1923.
[7] Q. Jin, Z. Meng, T.D. Pham, Q. Chen, L. Wei, R. Su, Dunet: a deformable network for retinal vessel segmentation, Knowledge-Based Syst. 178 (2019) 149–162.
[8] Y. Wu, Y. Xia, Y. Song, Y. Zhang, W. Cai, Multiscale network followed network model for retinal vessel segmentation, in: A.F. Frangi, J.A. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, Cham, 2018, pp. 119–126.
[9] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: a nested U-Net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 3–11.
[10] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.
[11] G. Azzopardi, N. Strisciuglio, M. Vento, N. Petkov, Trainable cosfire filters for vessel delineation with application to retinal images, Med. Image Anal. 19 (1) (2015) 46–57.
[12] R. Annunziata, A. Garzelli, L. Ballerini, A. Mecocci, E. Trucco, Leveraging multiscale hessian-based enhancement with a novel exudate inpainting technique for retinal vessel segmentation, IEEE J. Biomed. Health Inform. 20 (4) (2015) 1129–1138.
[13] Ç. Sazak, C.J. Nelson, B. Obara, The multiscale bowler-hat transform for blood vessel enhancement in retinal images, Pattern Recognit. 88 (2019) 739–750.
[14] J.V. Soares, J.J. Leandro, R.M. Cesar, H.F. Jelinek, M.J. Cree, Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification, IEEE Trans. Med. Imaging 25 (9) (2006) 1214–1222.
[15] M.M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A.R. Rudnicka, C.G. Owen, S.A. Barman, An ensemble classification-based approach applied to retinal blood vessel segmentation, IEEE Trans. Biomed. Eng. 59 (9) (2012) 2538–2548.
[16] J. Zhang, Y. Chen, E. Bekkers, M. Wang, B. Dashtbozorg, B.M. ter Haar Romeny, Retinal vessel delineation using a brain-inspired wavelet transform and random forest, Pattern Recognit. 69 (2017) 107–123.
[17] Y. Zhao, Y. Zheng, Y. Liu, Y. Zhao, L. Luo, S. Yang, T. Na, Y. Wang, J. Liu, Automatic 2-D/3-D vessel enhancement in multiple modality images using a weighted symmetry filter, IEEE Trans. Med. Imaging 37 (2) (2017) 438–450.
[18] J.I. Orlando, E. Prokofyeva, M.B. Blaschko, A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images, IEEE Trans. Biomed. Eng. 64 (1) (2016) 16–27.
[19] X. Wang, X. Jiang, J. Ren, Blood vessel segmentation from fundus image by a cascade classification framework, Pattern Recognit. 88 (2019) 331–341.
[20] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, L. Van Gool, Deep retinal image understanding, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 140–148.
[21] H. Fu, Y. Xu, S. Lin, D.W.K. Wong, J. Liu, Deepvessel: retinal vessel segmentation via deep learning and conditional random field, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 132–139.

[22] Á.S. Hervella, J. Rouco, J. Novo, M. Ortega, Self-supervised multimodal reconstruction of retinal images over paired datasets, Expert Syst. Appl. 161 (2020) 113674.
[23] Á.S. Hervella, J. Rouco, J. Novo, M.G. Penedo, M. Ortega, Deep multi-instance heatmap regression for the detection of retinal vessel crossings and bifurcations in eye fundus images, Comput. Methods Prog. Biomed. 186 (2020) 105201.
[24] K. Hu, Z. Zhang, X. Niu, Y. Zhang, C. Cao, F. Xiao, X. Gao, Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function, Neurocomputing 309 (2018) 179–191.
[25] M.Z. Alom, M. Hasan, C. Yakopcic, T.M. Taha, V.K. Asari, Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation, arXiv preprint arXiv:1802.06955(2018).
[26] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, Cbam: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[27] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[28] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, Context encoding for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7151–7160.
[29] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, J. Jia, Psanet: point-wise spatial attention network for scene parsing, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 267–283.
[30] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[31] S.M. Pizer, E.P. Amburn, J.D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J.B. Zimmerman, K. Zuiderveld, Adaptive histogram equalization and its variations, Comput. Vis. Graph. Image Process. 39 (3) (1987) 355–368.
[32] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, T. Wang, A cross-modality learning approach for vessel segmentation in retinal images, IEEE Trans. Med. Imaging 35 (1) (2015) 109–118.
[33] J. Odstrcilik, R. Kolar, A. Budai, J. Hornegger, J. Jan, J. Gazarek, T. Kubena, P. Cernosek, O. Svoboda, E. Angelopoulou, Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database, IET Image Process. 7 (4) (2013) 373–383.
[34] H. Zhao, H. Li, S. Maurer-Stroh, Y. Guo, Q. Deng, L. Cheng, Supervised segmentation of un-annotated retinal fundus images by synthesis, IEEE Trans. Med. Imaging 38 (1) (2018) 46–56.
[35] J. Zhang, B. Dashtbozorg, E. Bekkers, J.P. Pluim, R. Duits, B.M. ter Haar Romeny, Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores, IEEE Trans. Med. Imaging 35 (12) (2016) 2631–2644.
[36] S. Roychowdhury, D.D. Koozekanani, K.K. Parhi, Iterative vessel segmentation of fundus images, IEEE Trans. Biomed. Eng. 62 (7) (2015) 1738–1749.
[37] S. Xie, Z. Tu, Holistically-nested edge detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1395–1403.
[38] D. Pandey, X. Yin, H. Wang, Y. Zhang, Accurate vessel segmentation using maximum entropy incorporating line detection and phase-preserving denoising, Comput. Vis. Image Underst. 155 (2017) 162–172.

Ruohan Zhao graduated from Beijing University of Chemical Technology with B.Sc. in Computer Science in 2014. Currently he is a Ph.D. candidate in Department of Computing, the Hong Kong Polytechnic University. His research interest includes pattern recognition, medical imaging and deep learning.

Qin Li received the Ph.D. degree from The Hong Kong Polytechnic University, China, in 2010. He is currently a Professional Teacher, a Senior Engineer, and a Shenzhen Peacock Scholar with the Shenzhen Institute of Information Technology College. His current research interests include image processing, pattern recognition, and biometrics based on mobile terminals.

Jianrong Wu received the Ph.D. degree from Graduate School of The Chinese Academy of Sciences, China, in 2012. He is currently a Senior Researcher in Tencent Healthcare, Designated representative of IEEE-SA's Artificial Intelligence Medical Device Working Group, Member of ITU/WHO's AI4H Focus Group. His current research interests include image/video processing, pattern recognition.

Jane You obtained her B.Eng. in Electronic Engineering from Xi'an Jiao tong University in 1986 and Ph.D. in Computer Science from La Trobe University, Australia in 1992. She was a lecturer at the University of South Australia and senior lecturer at Griffith University from 1993 till 2002. Currently she is a professor at the Hong Kong Polytechnic University. Her research interests include image processing, pattern recognition, medical imaging, biometrics computing, multimedia systems and data mining.