Medical Image Analysis 70 (2021) 102025

Contents lists available at ScienceDirect

Medical Image Analysis

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

SCS-Net: A Scale and Context Sensitive Network for Retinal Vessel Segmentation

Huisi Wu[a], Wei Wang[a], Jiafu Zhong[a], Baiying Lei[b,*], Zhenkun Wen[a], Jing Qin[c]

[a] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, 518060
[b] School of Biomedical Engineering, Health Science Centers, Shenzhen University, National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Marshall Laboratory of Biomedical Engineering, AI Research Center for Medical Image Analysis and Diagnosis, Shenzhen, China, 518060
[c] Center for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Article history:
Received 7 September 2020
Revised 24 February 2021
Accepted 25 February 2021
Available online 4 March 2021

Keywords:
Retinal vessel segmentation
Scale-aware feature aggregation
Adaptive feature fusion
Multi-level semantic supervision

ABSTRACT

Accurately segmenting retinal vessel from retinal images is essential for the detection and diagnosis of many eye diseases. However, it remains a challenging task due to (1) the large variations of *scale* in the retinal vessels and (2) the complicated anatomical *context* of retinal vessels, including complex vasculature and morphology, the low contrast between some vessels and the background, and the existence of exudates and hemorrhage. It is difficult for a model to capture representative and distinguishing features for retinal vessels under such large scale and semantics variations. Limited training data also make this task even harder. In order to comprehensively tackle these challenges, we propose a novel scale and context sensitive network (a.k.a., SCS^{-}Net) for retinal vessel segmentation. We first propose a scale-aware feature aggregation (SFA) module, aiming at dynamically adjusting the receptive fields to effectively extract multi-scale features. Then, an adaptive feature fusion (AFF) module is designed to guide efficient fusion between adjacent hierarchical features to capture more semantic information. Finally, a multi-level semantic supervision (MSS) module is employed to learn more distinctive semantic representation for refining the vessel maps. We conduct extensive experiments on the six mainstream retinal image databases (DRIVE, CHASEDB1, STARE, IOSTAR, HRF, and LES-AV). The experimental results demonstrate the effectiveness of the proposed SCS-Net, which is capable of achieving better segmentation performance than other state-of-the-art approaches, especially for the challenging cases with large scale variations and complex context environments.

1. Introduction

Retinal vessel segmentation plays a significant role in the diagnosis of many eye-related diseases such as hypertension, diabetic retinopathy, arteriosclerosis, and so on (Li et al., 2020; Zhang and Chung, 2018). The morphological information of retinal blood vessels, such as thickness, curvature, and density, can serve as important indicators for the detection and diagnosis of these diseases (Fan et al., 2020; Jin et al., 2019). In current clinical practice, however, manually visual inspection is usually employed to obtain this morphological information, which is laborious, time-consuming and subjective. In this regard, automatic and accurate retinal vessel segmentation from retinal fundus images is highly demanded and has attracted a lot of research interest (Li et al., 2018; Wang et al., 2020).

However, it remains a challenging task for the following reasons. First, retinal vessels vary greatly in scale and shape (see Fig. 1(a) and 1(b)). For example, the scale of retinal vessels usually varies between 1 and 20 pixels (Mo and Zhang, 2017). Second, the anatomical semantics around retinal vessels are quite complicated. There are many mimic and perplexing structures and areas in retinal fundus images, including optic disk regions, pathological areas, hemorrhage, and exudates (see Fig. 1(c) ~ 1(d)), which may easily incur false segmentation results of vessels. Third, in many areas, the low intensity contrast makes it difficult to separate vessels from background (see Fig. 1(e) ~ (f)).

A lot of effort has been dedicated to addressing these challenges. Early studies focused on harnessing various hand-crafted features to segment retinal blood vessels. Huang et al. (Huang and Yan, 2006) developed a vessel detection algorithm to quantitatively

* Corresponding author.
E-mail addresses: leiby@szu.edu.cn, leiby@szu.edu.cn (B. Lei).

https://doi.org/10.1016/j.media.2021.102025
1361-8415/

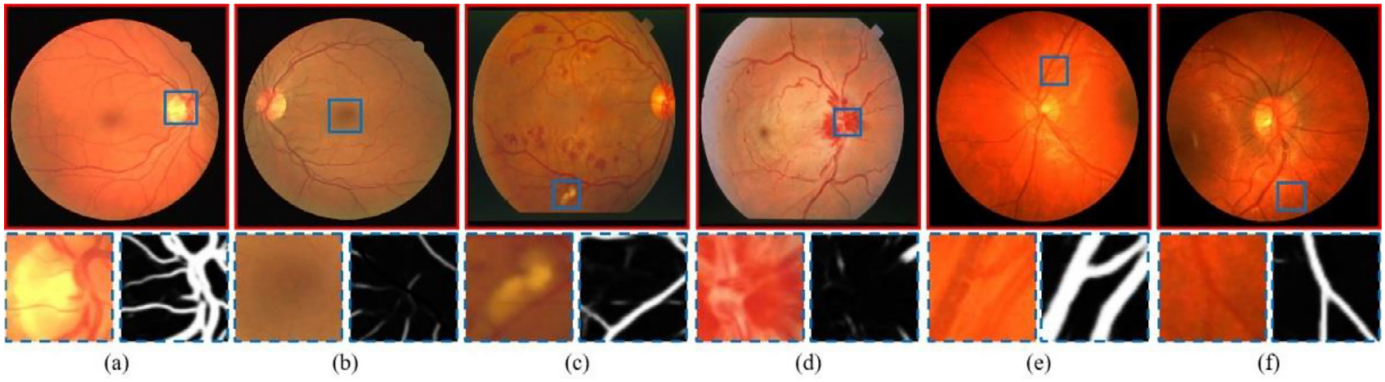

Fig. 1. The challenges in vessel segmentation from retinal images: (a) vessels with relatively large scale, (b) vessels with relatively small scale, (c) exudates, (d) hemorrhage, (e) and (f) low contrast with background. Each pair of small square boxes contains a partially enlarged view of the details and the corresponding ground truths.

measure the salient characteristics of retinal blood vessels, and combined these measurements through Bayesian decision to generate a confidence value for each detected vessel segment. Lam et al. (Lam et al., 2010) proposed a multi-concave modeling method to simultaneously deal with healthy and unhealthy retinas, which can deal with bright lesions in the perception space and remove dark lesions with different strength structures. After that, an automatic unsupervised vessel segmentation method has been proposed by (Zhang et al., 2015), which utilized self-organizing map (SOM) for pixel clustering and further adopt the Otsu method to classify each neuron in the output layer as retinal neuron or non-vessel neuron. Although these approaches achieved good results in some specific situations, the hand-crafted features are insufficient in representing the complicated semantics of retinal vessels and hence fail in the relatively large datasets with many complex cases.

Recently, many deep learning methods (López-Linares et al., 2018; Milletari et al., 2016; Zhang et al., 2020) based on fully convolutional networks (FCNs) (Long et al., 2015) have been proposed for medical image segmentation tasks and achieved remarkable performance. Among them, the U-Net (Ronneberger et al., 2015) and many of its variants have been proposed to improve the basic FCN by introducing various deep supervision mechanisms and further improve the segmentation performance (Çiçek et al., 2016; Khened et al., 2019; Zhou et al., 2018). These methods are mainly based on the symmetrical U-shape structure to gradually extract contextual features by continuously stacking the convolutional layers and the down-sampling layers.

Despite that these U-shaped structures achieve considerable performance; it is still insufficient for them to tackle the challenges of retinal vessel segmentation. In general, there are the following two limitations (Feng et al., 2020; Ibtehaz and Rahman, 2020). First, the networks usually lack the ability to effectively extract multi-scale contextual information. Many methods have been proposed to address this problem (Chen et al., 2018; Kamnitsas et al., 2017; Zhao et al., 2017). In these methods, however, the receptive fields of the feature maps are fixed which is incapable of thoroughly deal with the retinal vessels with such a large scale variation. Second, the vanilla skip connections in each stage usually directly combines the local information, which introduces too many irrelevant background noises and makes it difficult to distinguish retinal vessels from surrounding mimics and noise, particularly the tiny vessels. In principle, the high-level features obtained from the deep stages have abundant semantic information but lack sufficient resolution, while the low-level features from the shallow stages have rich spatial details but lack global semantic information (Zhang et al., 2018). Therefore, an intuitive idea is that the semantic information in the high-level features and the spatial information in the low-level features can be fully combined to boost efficient fusion between adjacent hierarchical features. In addition, it is essential to smartly extract semantic representations while suppressing mimics and noise.

Motivated by above thoughts, we propose a novel scale and context sensitive deep convolutional network to comprehensively tackle the challenges of retinal vessel segmentation; we call it SCS-Net. Our model is implemented based on classical encoder-decoder structure and consists of three core modules. First, in the top stage of the encoder, we propose a novel scale-aware feature aggregation (SFA) module to effectively extract multi-scale context information. Specifically, we first design a parallel dilated convolution structure with shared weight parameters. Then, the receptive fields are implicitly yet dynamically adjusted to adapt the variant scale of vessels by figuring out the importance of different scales in spatial space. Finally, a residual connection is developed to aggregate the multi-scale feature maps. Second, at each stage of the decoder, we replace the vanilla skip connection of the classic U-Net model with an adaptive feature fusion (AFF) module, which is able to adaptively combine semantic information and spatial information while suppressing the irrelevant background noise. By employing the AFF module, the network can effectively guide the fusion of adjacent hierarchical features to capture more distinguishable semantic information. In addition, a multi-level semantic supervision (MSS) module is also designed to drive the network to learn more semantic representations, thereby ultimately refine the vessel maps. Six mainstream fundus vessel datasets are utilized to validate the proposed SCS-Net, which consistently outperforms the state-of-the-art approaches. In summary, there are mainly three contributions in this paper:

(1) We propose a novel scale and context sensitive network (a.k.a., SCS-Net) to comprehensively tackle the challenges of retinal vessel segmentation. The proposed SCS-Net is capable of dealing with the large scale variation of retinal vessels and extracting representative features under complicated anatomical context.
(2) We propose a SFA module to effectively extract concealed multi-scale context information and aggregate the multi-scale features, which can improve the ability of the proposed SCS-Net in handling the complex cases where blood vessels vary heavily in sizes and shapes, and even many of them are intertwined.
(3) We propose an AFF module to adaptively combine semantic information and spatial information, which can not only suppress low-level irrelevant background noise but also retain more detailed local semantic information to disentangle blood vessels from hard mimics and noise. We further ap-

ply a MSS module to learn more global semantic representations from the side-output layers and improve the segmentation accuracy by assigning auxiliary supervisions to the early stages of the decoder network, which is able to refine the vessel maps.

The proposed techniques are general and can be employed in similar segmentation tasks where large scale variations and complex semantic environments are main challenges.

2. Related work

2.1. Retinal vessel segmentation

A lot of works have been proposed for retinal vessel segmentation in the last two decades. With the rise of the end-to-end fully convolutional network, Fu et al. (Fu et al., 2016) first applied a multi-scale convolutional neural network architecture to learn rich hierarchical representations, and combined conditional random fields to improve the performance of vessel segmentation. Then a method based on generative adversarial training was presented by (Son et al., 2017), which can generate a precise map of retinal vessels. To better address the challenges in retinal vessel segmentation, Mo and Zhang (Mo and Zhang, 2017) developed a fully convolutional network based on deep supervision, which improves the discriminative capability of features in lower layers of the deep network. Yan et al. (Yan et al., 2018) proposed a new segment-level loss framework. Compared with pixel-by-pixel loss, it can learn more distinguishable blood vessel segmentation features without significantly changing the network architecture to improve network performance. More recently, based on the U-Net architecture, Si et al. (Si et al., 2019) proposed to effectively establish the long-range dependence of different parts of the retinal image by combining an attention mechanism, which avoids the mistakes caused in some segmentation scenes. Shin et al. (Shin et al., 2019) embeds a graph neural network into a unified convolutional neural network architecture, which can effectively exploit the relationship that exists between vessel neighborhoods and help improve the vessel segmentation accuracy. To preserve the spatial information and extract the multi-scale context information, Wang et al. (Wang et al., 2019) proposed a novel U-Net structure with two encoders. However, simply stacking multi-scale modules with fixed receptive fields cannot effectively capture multi-scale features and thoroughly deal with the large scale variation of retinal vessels. Meanwhile, it introduces too many additional parameters. In addition, most of these existing methods did not fully consider the complicated anatomical context in retinal images and thereby design well-directed modules to disentangle blood vessels from hard mimics and background noise.

2.2. Multi-scale context extraction

Effective modeling of global contextual information is conducive to collecting information from a large receptive field (Ahn et al., 2019). ParseNet (Liu et al., 2015) first proposed to encode the global context by utilizing global pooling. Then DeepLab family (Chen et al., 2017a, 2017b, 2018) developed the atrous spatial pyramid pooling (ASPP) module to capture multi-scale features while PSPNet (Zhao et al., 2017) utilized the pyramid pooling module (PPM) to aggregate the context of different regions, which enhances the network's ability to exploit global context information. Instead of modeling global context information directly by parallel aggregating like ASPP and PPM module, Bilinski and Prisacariu et al. (Bilinski and Prisacariu, 2018) proposed to exploit the context information in a dense connection way. However, without effectively extracting concealed multi-scale context information and aggregating the multi-scale features, existing methods still cannot handle the complex cases where blood vessels vary extremely in sizes and shapes, and even many of them are intertwined. With the help of dynamic kernel selection mechanism, the SK-Net (Li et al., 2019) can effectively capture target objects with different scales. Inspired by this work, we design a scale-aware mechanism to extract multi-scale contextual information by implicitly and dynamically adjusting the receptive fields to refine the feature maps.

2.3. Feature fusion

Except for the semantic information, the precise spatial information is also essential for the retinal vessel segmentation task. One of the most prominent contributions of the U-Net (Ronneberger et al., 2015) network is the introduction of skip connection between the encoder and decoder to compensate the loss of spatial information caused by the down-sampling operation. Since the introduction of U-Net, many methods (Heinrich et al., 2019; Khened et al., 2019; Lei et al., 2020) have adopted this connection way to improve segmentation performance. To overcome the shortcomings of original skip connection, Chan et al. (Chen et al., 2018) introduced the low-level features into the decoder to improve segmentation performance. Although the spatial information is preserved through skip connection, there is a certain semantic gap between shallow level features and deep level features. Therefore, many methods have been proposed to improve the feature fusion process to bridge the gap. For instance, Zhang et al. (Zhang et al., 2018) improved this process by embedding more spatial resolution into high-level features and selectively fusing the features with different levels. Ibtehaz and Rahman et al. (Ibtehaz and Rahman, 2020) integrates several convolutional layers into the skip connection to better resolve the disparity between encoder and decoder features. A novel module (Hu et al., 2018) was also proposed to emphasize the help of attention mechanisms in the decoder of U-Net. To effectively handle the complicated anatomical semantics in retinal vessel segmentation task, we propose novel mechanisms to fuse adjacent-level features to learn more local and global semantic information for more accurate segmentation.

3. Methodology

The overall architecture of the proposed SCS-Net is shown in Fig. 2, which is implemented based on a U-shape structure and consists of three novel modules: SFA, AFF and MSS. The SFA module is proposed to dynamically adjust the receptive fields and efficiently aggregate multi-scale features so that the encoder in the U-shaped structure is capable of dealing with the large scale variations of retinal vessels. The AFF module is designed to (1) progressively guide the fusion between feature maps of adjacent levels to capture more semantic information, and (2) suppress the interference of the irrelevant background noise particularly introduced by low-level feature maps. Finally, the MSS module is developed to learn more semantic representation in the side-output layers; it is able to further learn more semantic details and refine the vessel maps for yielding more accurate segmentation results.

3.1. Scale-aware feature aggregation module

Given the large scale variations of retinal vessels, extracting scale context information and aggregating multi-scale features are essential to improve segmentation accuracy. It is, however, a challenging task to capture and fuse features in such a large range of scales. Despite its hierarchical feature representation capability, traditional U-Net, as well as its variants, still suffers limitations of fixed receptive fields, and hence the vessels substantially smaller

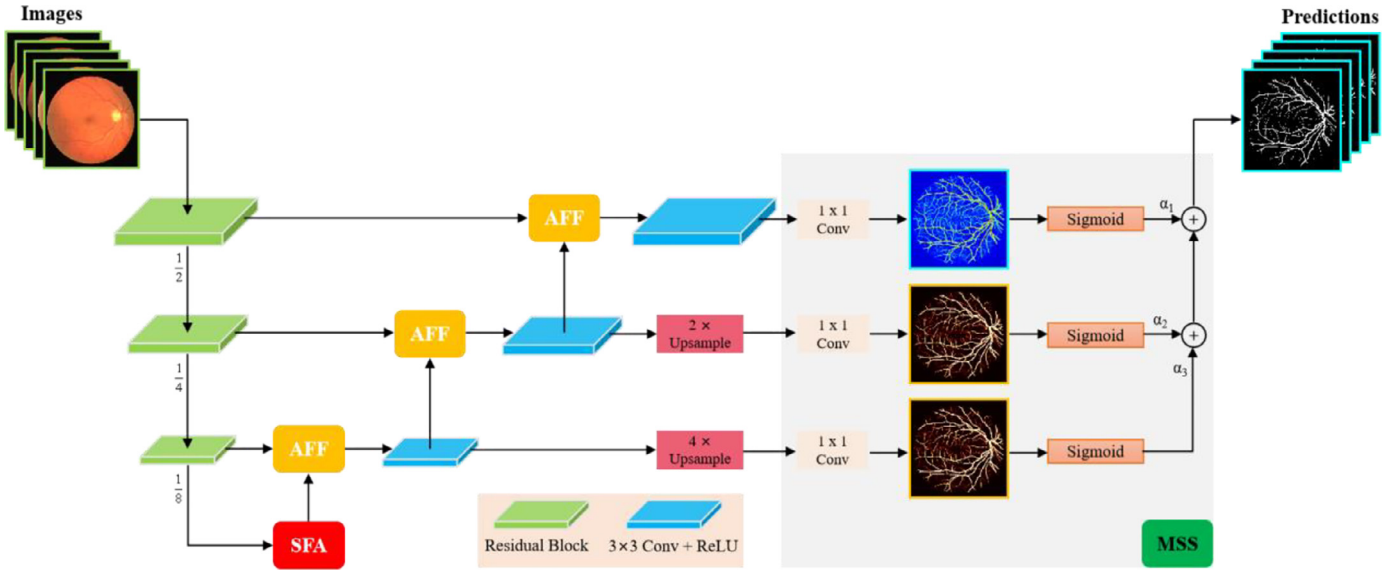

Fig. 2. An overview of our proposed SCS-Net, which is implemented based on a U-shape architecture and consists of three modules: SFA, AFF and MSS. The α_1, α_2, and α_3 indicate the weight coefficients of the side-output layers.

or larger than the fixed receptive field might incur false detection or produce fragmented segmentation. Several existing methods (Chen et al., 2017a; Zhao et al., 2017) have been dedicated to overcoming this challenge, but these methods still cannot effectively extract sufficient scale context information to yield satisfactory results, as discussed in Section 1.

In the proposed SCS-Net, we propose a SFA module to meet this challenge. Instead of capturing the multi-scale contexts by directly concatenating the feature maps with fixed receptive fields (Gu et al., 2019), our SFA dynamically adjusts the receptive fields by learning two correlation weights for every two adjacent scales, which reflect the importance of feature maps under different scales for retinal vessel segmentation. In other words, our SFA contains a dynamic feature selection mechanism to automatically select the appropriate receptive field for the feature map based on the learned correlation weights for every two adjacent scales. As illustrated in Fig. 3, the SFA module contains two components: Multi-scale feature extraction (Mfe) and Dynamic feature selection (Dfs).

3.1.1. Multi-scale feature extraction

Given an input feature map $\mathbf{F}_{in} \in \mathbb{R}^{C_{in} \times H \times W}$ extracted from the encoder, where C_{in}, H, and W are the number of channels, height and width of the feature map respectively, we first feed it into three dilated convolution layers with different atrous rates for multi-scale context modeling with multiple receptive fields. For efficiency, the weights of these three dilated convolutions are shared. To this end, we have obtained three feature maps ($\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$). We then concatenate the feature maps of two adjacent scales separately to preserve more scale information:

$$4 \rightarrow \mathbf{F}_{ij} = \mathbf{F}_i \copyright \mathbf{F}_j, \tag{1}$$

where i, j indicate the adjacent-scale feature maps and © denote the concatenation operation.

3.1.2. Dynamic feature selection

Considering the correlation of features between adjacent scales, we introduce a scale-aware mechanism to automatically select the appropriate receptive field for the feature map. In particular, taking the above branch $\mathbf{F}_{12}$ as an example, we further fuse the adjacent-scale features by employing a 3×3 convolution with C_{in} filters and then pass through a 1×1 convolution to unify the output channels to 2:

$$\mathbf{F}'_{12} = \widehat{\mathcal{F}}(\delta(\mathcal{F}(\mathbf{F}_{12}, \theta))), \tag{2}$$

where $\widehat{\mathcal{F}}$ and $\mathcal{F}$ function are the normal convolution operations with a kernel size of 3×3 and 1×1, respectively; θ is the related parameters; $\delta(\cdot)$ represent arbitrary nonlinear activation function; $\mathbf{F}'_{12} \in \mathbb{R}^{2 \times H \times W}$. In this work, we adopt the ReLU activation function (Nair and Hinton, 2010) since it is difficult to saturate and easy to optimize. After the fusion, a softmax function is utilized to generate two weight masks $\mathbf{W}_{\alpha_1}$ and $\mathbf{W}_{\beta_1}$, which reflect the importance of spatial information under different scales after considering the adjacent scale information. In this regard, we are able to send the importance information to the current feature map ($\mathbf{F}_1$, $\mathbf{F}_2$) by element-wise product operation; the whole process can be simply expressed as follows:

$$\mathbf{F}''_{12} = (\mathbf{F}_1 \otimes \mathbf{W}_{\alpha_1}) \oplus \left(\mathbf{F}_2 \otimes \mathbf{W}_{\beta_1}\right), \tag{3}$$

where $\otimes$ and $\oplus$ represents the element-wise product and addition, respectively; $\mathrm{F}''_{12} \in \mathbb{R}^{C_{in} \times H \times W}$. By doing so, we can efficiently aggregate multi-scale features and implicitly adjust the receptive fields for the feature maps with different scales to emphasize some areas with importance from the view of other scales. Note that the implementation of other branches (i.e., $\mathbf{F}''_{23}$) are similar with above-mentioned procedure. Finally, a residual connection is employed to aggregate the multi-scale feature maps:

$$\mathbf{F}_{out} = \delta(\hat{\mathcal{F}}\left(\mathbf{F}''_{12} \oplus \mathbf{F}''_{23} \oplus \mathbf{F}_{in}\right), \tag{4}$$

where $\delta(\cdot)$ and $\hat{\mathcal{F}}(\cdot)$ represents the ReLU activation functions and 1×1 convolution layer, respectively. In such a case, information from different scales is aggregated by the SFA module. Note that the three-scale branches here can be easily extended to multiple-scale branches.

3.2. Adaptive feature fusion module

Due to the proposed SFA module, multi-scale context information can be effectively extracted and fused in high-level feature maps. In order to harness multi-scale information to produce the segmentation results, however, we need to restore it to the original image spatial resolution. To do so, traditional U-Net and many

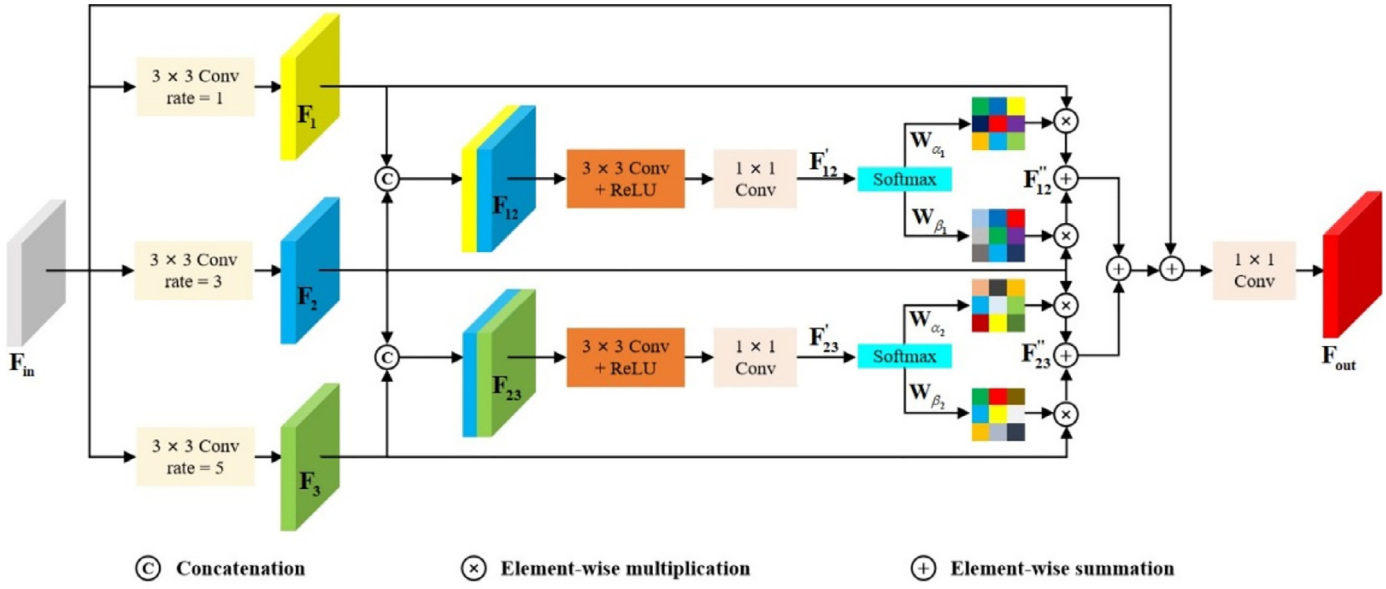

Fig. 3. The flowchart of the proposed SFA module.

of its variants directly concatenate the high-level features and the low-level features:

$$\gamma^{(t)} = \gamma_l^{(t)} \copyright Upsample(\gamma_h^{(t+1)}), \tag{5}$$

where t is the feature level; γ is the output of the corresponding stage; © is the concatenation operation; l and h indicate the low-level features and high-level features, respectively.

Unfortunately, such a naive connection is insufficient to consider the complementariness between high-level features and low-level features. High-level features contain rich semantic information, which can help low-level features to identify semantically important locations. However, it lacks of necessary spatial information due to relatively coarse resolution. By contrast, low-level features contain abundant spatial information, which is useful for high-level features to reconstruct precise details. However, it lacks necessary semantic information to identify the targeting objects from a global perspective. Therefore, the high-level features with rich semantic information and the low-level features with abundant spatial information are essentially complementary. Motivated by this, we propose an adaptive feature fusion (AFF) module to guide the fusion between adjacent layers based on a squeeze-and-excitation (SE) operation to model correlations among feature channels between two adjacent layers. As illustrated in Fig. 4, by calculating a weighting vector to re-weigh the low-level features and suppress the interference of the irrelevant background noise, the proposed network can preserve more important semantic context information for more precise localization. Specifically, we firstly concatenate the feature maps of adjacent levels and model the correlations among the combined feature channels:

$$\gamma_f^{(t)} = \Gamma\left(\gamma_l^{(t)} \copyright Upsample(\gamma_h^{(t+1)})\right), \tag{6}$$

where Γ denotes the squeeze and excitation operation (Hu et al., 2018), which is able to adaptively recalibrate channel-wise feature responses. After that, the output $\gamma_f^{(t)}$ is feed into a 1×1 convolution to reduce the dimension of the filter. The global average pooling (GAP) is then utilized to further extract the global context information. To suppress the interference of the irrelevant background noise, our weight vectors generated by the Sigmoid function are multiplied by the low-level features, and then we add the re-weighted low-level features to the high-level features to yield the final result:

$$\gamma^{(t)} = \gamma_h^{(t+1)} \oplus \left(\gamma_l^{(t)} \otimes \text{Sigmoid}\left(\text{GAP}\left(\hat{\mathcal{F}}\left(\gamma_f^{(t)}\right)\right)\right)\right), \tag{7}$$

where $\hat{\mathcal{F}}$ denotes the 1×1 convolution layer; $\oplus$ and $\otimes$ means the element-wise sum and multiplication. By employing AFF to progressively guide the fusion between high-level features and low-level features, the proposed SCS-Net can not only suppress irrelevant background noise, but also preserve more sematic information for more precise localization.

3.3. Multi-level semantic supervision

In order to obtain more semantic details to refine the segmentation results, we further devise a multi-level semantic supervision module, short for MSS in Fig. 2. To avoid gradient vanishing problem and achieve a higher performance in the decoder, deeply-supervised nets (Lee et al., 2015) proposed a deep supervision mechanism to generate auxiliary local output maps for early layers. Similarly, M-Net (Fu et al., 2018) also successfully utilized side-output layers to produce side-output prediction maps by directly averaging the multi-output loss. However, M-Net may ignore different characteristics of different output feature maps by setting the same weights for all side-output layers. For example, the shallow layers may have a smaller receptive field and can utilize more fine-grained feature information, ensuring that the network can capture more details. By contrast, deep-layer features may include more high-level semantic context after multiple convolutions and pooling operations, where the receptive fields also gradually increase.

Instead of directly averaging the multi-output loss in M-Net, we introduce the multi-level semantic supervision (MSS) module to improve the deep supervision mechanism by setting suitable weights for different side-output layers, which is able to learn better semantic representations from the side-output layers and improve the segmentation accuracy based on the auxiliary supervision in the early stages of the decoder. Specifically, we assign auxiliary supervisions to the early stages of the decoder network in order to learn more deep semantic representations from the side-output layers (as shown in Fig. 2). For each raw input image R and the corresponding ground truth binary vessel map G, G is a binary

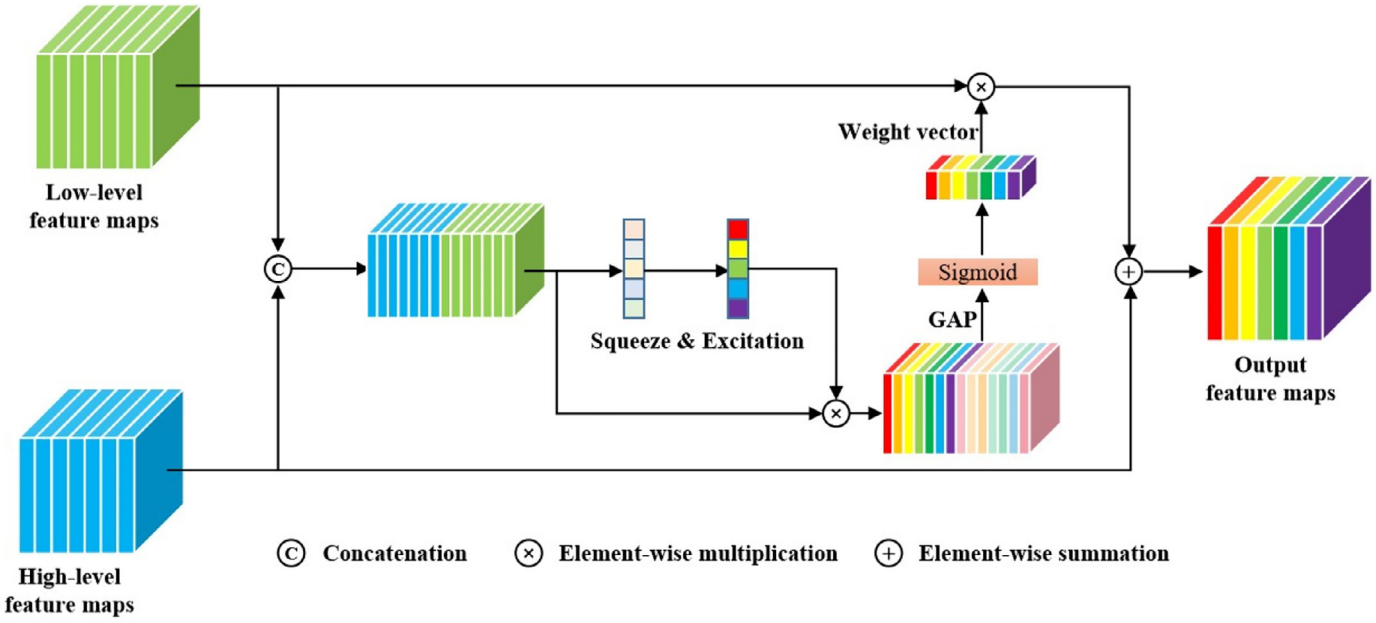

Fig. 4. Schematic diagram of AFF module. The green cube and blue cube represent the low-level and high-level feature maps, respectively. We firstly concatenate the feature maps of the adjacent level and perform a squeeze-and-excitation operation to model correlations among feature channels. Then we compute a weighting vector through GAP and Sigmoid operations to re-weigh the low-level features to suppress the interference of the irrelevant background noise. Finally, the re-weighted low-level feature maps are added to high-level feature maps as the AFF output. Note that GAP means global average pooling. . (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

image, $\mathbf{g}_j \in \{0, 1\}$. Let $Q = (R, G)$, where $R = \{\mathbf{r}_i,\ i = 1, 2, \ldots, |R|\}$ and $G = \{\mathbf{g}_j,\ j = 1, 2, \ldots, |G|\}$. Assuming that there are N side-output layers $S_n, n \in \{1, 2, \ldots, N\}$, we assign a weight coefficient α for each layer, then the side-output loss $\mathcal{L}_{side}$ can be defined as:

$$\mathcal{L}_{side} = \sum_{n=1}^{N} \alpha_n S_n(R, \theta_n), \tag{8}$$

where θ_n represents the relevant parameters of the n-th side-output layer in the network. In our implementation, we set $N = 3$ and employ a weighting strategy of $\alpha = (\alpha 1, \alpha 2, \alpha 3) = (0.6, 0.3, 0.1)$ for the three side-output layers. Unlike M-Net which directly averages the multi-output loss, we conduct extensive experiments to evaluate the impact of different weighting strategies in the MSS module for the retinal vessel segmentation and select the optimal weights for different side-output layers. The impact of the hyper-parameter $\boldsymbol{\alpha}$ will be discussed in Section 6.1.

3.4. Loss function

Considering the distributions of foreground and background in the vessel map might be heavily biased, here we use the binary Dice loss (Milletari et al., 2016) as the cost function:

$$DL = 1 - \frac{2 \cdot \left\langle v_{(h,w)}, \hat{v}_{(h,w)} \right\rangle + \varepsilon}{\left\| v_{(h,w)} \right\|_1 + \left\| \hat{v}_{(h,w)} \right\|_1 + \varepsilon} \tag{9}$$

where (h, w) indicate the pixel coordinate; $\hat{v}$ is the vessel mask ground truth; $\hat{v}_{(h,w)} = 1/0$ represent the pixel belonging to vessel/non-vessel; $0 \le v_{(h,w)} \le 1$ is the prediction probability for the pixel belonging to the vessel. In Eq. (9), we add the Laplace smoothing factor ε to the cost function, which is able to speed up the convergence. Here, ε is set to 10 in our experiments.

Finally, we design an overall loss function $\mathcal{L}_{\text{total}}$ with two elements, including the side output layer with $\mathcal{L}_{\text{side}}$, and a $\mathcal{L}_2$ regularization term, given by the following equations:

$$\mathcal{L}_{total} = \mathcal{L}_{side} + \frac{\lambda}{2}\|\omega\|_2^2, \tag{10}$$

$$\mathcal{L}_{\text{side}} = DL(side\_output,\ G), \tag{11}$$

where ω represents the network parameters and the weight decay rate λ is set to 0.001 in our experiment.

4. Experiments

4.1. Datasets

We conduct our experiments on six different retinal fundus datasets. As shown in Table 1, all images are RGB color images and they are in different formats. For convenience, we convert all pictures to PNG format. Note that, the FoV masks are not provided in CHASEDB1 and STARE, and we manually generate the corresponding masks for the sake of unity throughout the experiment (Soares et al., 2006; Wu et al., 2020).

DRIVE: The DRIVE (Digital Retinal Images for Vessel Extraction) dataset (Staal et al., 2004) contains 40 fundus images that were obtained from a diabetic retinopathy screening program in the Netherlands, in which 33 images do not show any signs of diabetic retinopathy and 7 images show signs of mild early diabetic retinopathy. For the test cases, two manual segmentations are available. One is used as a gold standard, and the other one is utilized to compare computer-generated segmentations with those of an independent human observer.

CHASEDB1: The CHASEDB1 (Child Heart and Health Study in England) dataset (Owen et al., 2009) contains 28 eye fundus images with a resolution of 990 × 960 taken from the eyes of 14 children. Each retinal fundus image was centered on the optic disk and captured at 30°FOV. There are two manual annotations available and the first manual annotation is adopted in this work.

STARE: The STARE (Structured Analysis of the Retina) dataset (Hoover et al., 2000) comprises 20 retinal fundus images, half of which contain signs of pathologies. Since there is no uniform division for the STARE dataset, we conduct training on this dataset with two types of techniques, which are reported in the published literature (Liskowski and Krawiec, 2016). The first technique is to

Table 1
An overview of the six publicly available databases. The total number of images, the training and test split, the image size (width × height), and the availabilities of Field-of-View (FoV) masks in each dataset are reported.

Dataset	Quantity	Train-test split	Resolution	Format	FoV mask
DRIVE	40	20–20	565 × 584	.tiff	√
CHASEDB1	28	20–8	999 × 960	.jpg	×
S TARE	20	18–2	700 × 605	.ppm	×
I OSTAR	30	25–5	1024 × 1024	.jpg	√
HRF	45	30–15	3504 × 2336	.jpg	√
LES-AV	22	20–2	1620 × 1444	.png	√

construct a set of training samples randomly extracted from all images, which is called the 'random samples' approach. Despite that this technique has been adopted by many published literature (Soares et al., 2006; Staal et al., 2004), it raises various concerns of excessively optimistic results due to the overlap samples exists in the training and test sets. The other technique is called 'leave-one-out' (Li et al., 2015; Liskowski and Krawiec, 2016; Mo and Zhang, 2017) and it leverages one image as testing image, while the remaining images are used for training. Note that, there is no overlap between the training and testing sample in this approach, because the test and training sample will be repeated 20 times, i.e., each iteration of one image forms the source of the testing set, and the remaining images are the source of the training set.

IOSTAR: The IOSTAR dataset (Zhang et al., 2016) includes 30 Scanning Laser Ophthalmoscopy (SLO) images with a resolution of 1024 × 1024 pixels. The images in IOSTAR database are acquired with an EasyScan camera (i-Optics Inc., the Netherlands), which is based on a SLO technique with a 45° Field of View (FOV). All the vessels in this dataset are annotated by a group of experts working in the field of retinal image analysis.

HRF: The HRF (High-Resolution Fundus) dataset (Köhler et al., 2013) includes three types of images: healthy patients, glaucomatous patients, and diabetic retinopathy, where 15 pictures of each type with a 3504 × 2336 resolution. Each image has a binary gold standard vessel segmentation image, which generated by a group of experts working in the field of retinal image analysis and clinicians from the cooperated ophthalmology clinics. Also the masks determining field of view (FOV) are provided for particular datasets.

LES-AV: The LES-AV dataset (Orlando et al., 2018) consists of 21 normal images with a resolution of 1620 × 1444 and 1 pathological image with a resolution of 1958 × 2196, respectively, with expert ground-truth for vessel segmentation.

4.2. Preprocessing

We follow the official division of DRIVE, using 20 images to train the model and the remaining 20 images to evaluate the model. Considering that CHASEDB1, STARE, IOSTAR and LES-AV do not have an official partition like DRIVE, we follow the ways reported in (Wang et al., 2019) to split the dataset. For CHASEDB1 and IOSTAR, we take the first 20 and 25 images as training data and the next 8 and 5 images as test data, respectively. Meanwhile, STARE is experimented with a 10-fold cross-validation method, i.e., taking 18 images as the training samples and remaining images as the test samples. Specifically, we repeat this process 10 times until the entire dataset is covered, which can reduce the deviation as much as possible and ensure the reliability of the experimental results. The LES-AV division is the same as STARE. Similar to most of existing splitting strategies, we utilize the first 5 samples from each type, i.e., glaucoma, diabetic, and healthy for training, and the remaining samples are for testing.

As many other medical image computing applications, the three fundus image datasets are quite small. In this case, we employ several sampling and data augmentation strategies before training in order to prevent overfitting and make the network more generalized. We first resize all input images as 512 × 512 or 1024 × 1024 (only for HRF dataset) and then employ eight data augmentation strategies to augment the datasets. We first conduct randomly horizontal and vertical flipping, and followed by the random rotation. The rotation is ranged from −15 to 15° for the input images. We also apply the random contrast, brightness, hue, and saturation enhancements to the image, which can reduce the interference caused by external environment factors. Random erase is also utilized to force the model to be more sensitive to boundary information during the training process. The above image enhancement method can increase the data capacity to a certain extent, and enhance the generalization ability of the network to prevent overfitting problem. Some images after data augmentation are shown in Fig. 5. Note that there are no major post-processing steps and pre-trained backbone feature extractors are needed in our implementation. Also, all methods are trained from scratch.

4.3. Evaluation metrics

In order to quantitatively analyze the experimental results, several important metrics are utilized, including sensitivity (SE), specificity (SP), accuracy (ACC), and F1-score (F1), which are calculated by the following equations:

$$\mathrm{SE} = \frac{|TP|}{|TP + FN|}, \tag{12}$$

$$\mathrm{SP} = \frac{|TN|}{|FP + TN|}, \tag{13}$$

$$\mathrm{ACC} = \frac{|TP + TN|}{|TP + TN + FN + FP|}, \tag{14}$$

$$\mathrm{F1} = \frac{2 \cdot |TP|}{2 \cdot |TP| + |FN| + |FP|}, \tag{15}$$

where *TP* and *FP* are the variables of true positive and false positive, which represent the number of blood vessel pixels correctly segmented and the number of background pixels that are incorrectly segmented by the model, respectively. Correspondingly, *TN* is the variable of true negative, which represents the number of background pixels that correctly segmented. *FN* is the variable of false negative, which represents the blood vessel pixel that is incorrectly marked as a background pixel. Additionally, the area under curve (AUC) of receiver operating characteristic curve (ROC) is also employed, which are based on the recall and precision to measure the segmentation performance. In order to evaluate the statistical significance of different methods, the p-value calculated by utilizing a paired *t*-test is also reported.

4.4. Implementation details

The implementation of our proposed SCS-Net is based on the PyTorch platform and an NVIDIA RTX 2080Ti graphics card with

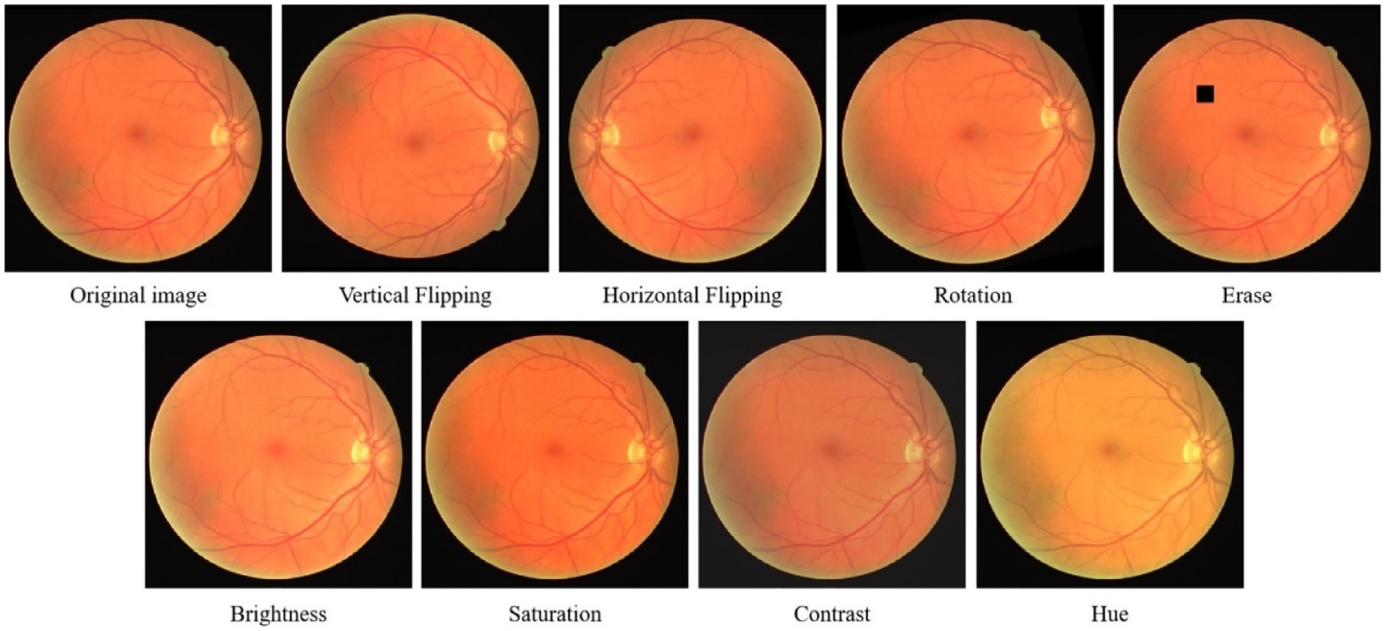

Fig. 5. Data augmentation. We apply eight strategies for data augmentation, including vertical flipping, horizontal flipping, random rotation, random erase, and random enhancements of brightness, saturation, contrast, and hue.

Table 2
Statistical comparison of ablation studies on DRIVE dataset.

Method	SE (%)	SP (%)	ACC (%)	AUC (%)	F1 (%)	p-value (AUC)
Baseline	76.39±0.25	**98.57±0.18**	96.60±0.07	97.72±0.01	79.57±0.09	4.38×10^{-3}
Baseline+SFA_w/o_Mfe	79.80±0.19	98.47±0.23	96.81±0.07	98.11±0.00	81.31±0.14	2.69×10^{-2}
Baseline+SFA_w/o_ Dfs	79.79±0.06	98.39±0.13	96.81±0.05	98.24±0.05	81.44±0.07	2.98×10^{-3}
Baseline+SFA	80.36±0.11	98.56±0.15	96.91±0.04	98.28±0.03	81.75±0.08	3.57×10^{-2}
Baseline+AFF	79.95±0.15	98.51±0.09	96.86±0.08	98.23±0.09	81.60±0.11	6.11×10^{-5}
Baseline+SFA+AFF	81.36±0.08	98.42±0.30	96.94±0.06	98.32±0.04	81.79±0.06	1.86×10^{-4}
Baseline+SFA+AFF+MSS (Ours)	**82.89±0.11**	98.38±0.29	**96.97±0.05**	**98.37±0.06**	**81.89±0.13**	–

11GB memory. During the training, we use the Adam algorithm with an initialization learning rate of 1e-3 as an optimization method, in which the weight decay is set to 0.001. In our experiments, the batch size is set to 2. In addition, we also adopt the weight decay strategy to halve the current learning rate when the loss on the validation set has not dropped for 10 consecutive epochs. In order to prevent the output loss gradient of the layer activation function from exploding or disappearing during the forward propagation of the deep neural network, the initialization method introduced in (He et al., 2015) is employed to initialize all encoders and decoder layers. Besides, dropout and batch normalization are also leveraged to reduce overfitting and gradient vanishing respectively. In our experiments, our model can be converged after no more than 150 epochs.

5. Results

5.1. Ablation studies

To demonstrate the effectiveness of our proposed SCS-Net, we first perform an ablation study to validate the effect of each component. The visual results for different components and statistical comparisons are shown in Fig. 6 and Table 2, respectively. As mentioned in Section 3, the SFA module contains two components: multi-scale feature extraction (Mfe) and dynamic feature selection (Dfs). Therefore, we further conduct experiments by replacing the atrous convolution with the vanilla convolution (referred to as 'Baseline + SFA_w/o_Mfe') to validate the effectiveness of the multi-scale feature extraction. In our ablation study, we took a U-shaped network consisting of an encoder with three residual blocks and a feature decoder as our 'Baseline'.

5.1.1. Effectiveness of the SFA module

First, we only add the proposed SFA module into the Baseline (referred to as 'Baseline + SFA') and apply it on the DRIVE dataset. Fig. 6 shows three typical examples of retinal vessel segmentation results, which clearly show that our proposed SFA module can effectively segment vessels with various scales, particular some tiny vessels, which cannot be well handled by the baseline network without SFA. As shown in Table 2, compared with 'Baseline', 'Baseline + SFA' improves the performance from 76.39%/79.57% to 80.36%/81.75% in terms of SE/F1. As shown in Table 2, compared with 'Baseline + SFA', we can see that the performance of 'Baseline + SFA_w/o_Mfe' decreases 0.70%/0.54% in SE/F1, which demonstrates that the multi-scale feature extraction is necessary to improve segmentation accuracy. Finally, we also directly concatenate the extracted multi-scale features without using dynamic feature selection (referred to as 'Baseline + SFA_w/o_Dfs') to evaluate the influence of the Dfs. Our experiment results also clearly demonstrate the importance of Dfs in the proposed SFA module.

5.1.2. Effectiveness of the AFF module

Second, we investigate the effectiveness of the AFF module. Compared with the 'Baseline', the proposed AFF module (referred to as 'Baseline + AFF') increases the SE/F1 by 4.66%/2.55% (from 76.39%/79.57% to 79.95%/81.60%). As can be seen from Fig. 6, compared with the 'Baseline', the model with AFF module is capable

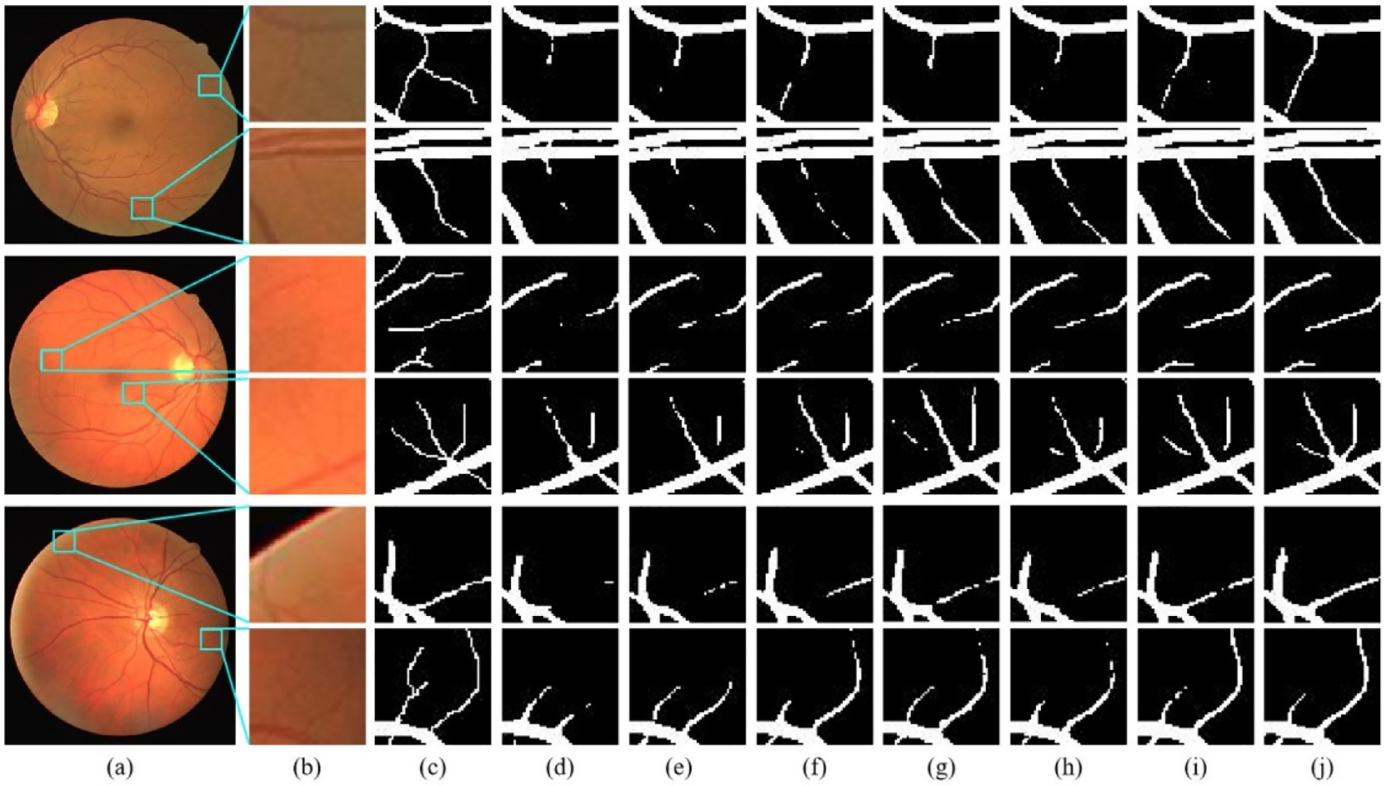

Fig. 6. Some typical visual results for different methods in our ablation study on DRIVE dataset. (a) Original image, (b) detailed view, (c) ground truth, (d) Baseline, (e) Baseline + SFA_w/o_Mfe, (f) Baseline + SFA_w/o_Dfs, (g) Baseline + SFA, (h) Baseline + AFF, (i) Baseline + SFA + AFF, (j) Baseline + SFA + AFF + MSS (Ours).

of obtaining a more consistent and complete vascular segmentation results, which indicates that the AFF can effectively guide the fusion of features in different levels to extract more semantics of targeting vessels, capturing relatively complete topology while surpassing background noise.

We further embed both SFA and AFF into the 'Baseline' (referred to as 'Baseline + SFA + AFF') to verify the complementarity between the two modules. As shown in Table 2, the segmentation accuracy is greatly improved, with an obvious improvement about 6.5%/2.79% in terms of SE/F1, which indicates that the combination of SFA and AFF in our SCS-Net is effective.

5.1.3. Effectiveness of the MSS module

Finally, to refine the hierarchical features and the final vessel map, the MSS module is also added to our network (Ours). As shown in Fig. 6, thanks to the MSS module, it is observed that our model can obtain refined segmentation results. It outperforms the 'Baseline' with an improvement of 0.65% in terms of AUC, as shown in Table 2. From both visual and statistical results of the ablation experiments, we can clearly see that our method achieves remarkable improvements by seamlessly combining SFA, AFF and MSS components, which demonstrates the effectiveness of the proposed SCS-Net in tackling the large scale variations and complicated semantics of retinal vessels in this challenging task.

5.2. Comparisons with the state-of-the-art methods

We further compare our proposed SCS-Net with six commonly used and state-of-the-art methods, including U-Net (Ronneberger et al., 2015), R2U-Net (Alom et al., 2018), AttU-Net (Schlemper et al., 2019), CE-Net (Gu et al., 2019), NFN+ (Wu et al., 2020), and IterNet (Li et al., 2020). We have implemented all competitors on six datasets (DRIVE, CHASEDB1, STARE, IOSTAR, HRF, and LES-AV) based on the same experimental settings and training strategies.

5.2.1. Visual comparison

The visual experimental results of our method and other competitors on typical images in six datasets are illustrated in Fig. 7 and Fig. 8. We can observe that as a general segmentation benchmark, the performance of U-Net is not satisfactory, which has more yellow pixels, indicating that it produces more false negatives. Based on the idea of feature reusing, R2U-Net applied the recurrent residual convolutional block and outperformed U-Net. By introducing a gated attention mechanism, AttU-Net can better focus on target structures and hence obtain a comparable performance against R2U-Net. As discussed in Section 1, accurate retinal vessel segmentation is a quite difficult task as blood vessels vary in sizes and shapes, and many of them are even intertwined, which results in an extremely complicated vascular structure.

It is clear to see that both U-Net and AttU-Net cannot segment small blood vessels very well because they are unable to extract effective multi-scale features. In contrast, the proposed SCS-Net method can accurately detect many tiny blood vessels, which are mainly attributed to the proposed SFA module. With the help of SFA, the network is able to dynamically adjust the receptive field according to the input target size and effectively extract the multi-scale features. By combing the residual multi-kernel pooling module and the dense atrous convolution module, CE-Net can enlarge the receptive field to segment more fine blood vessels. However, we can find that the segmentation map of CE-Net still has many discontinuous vessels due to the lack of sufficient semantic information, which makes it difficult to capture the complete vessel trees in the up-sampling process, as shown by the red pixels in Fig. 7. Thanks to the proposed AFF module, our SCS-Net is capable of utilizing the global context information to extract more seman-

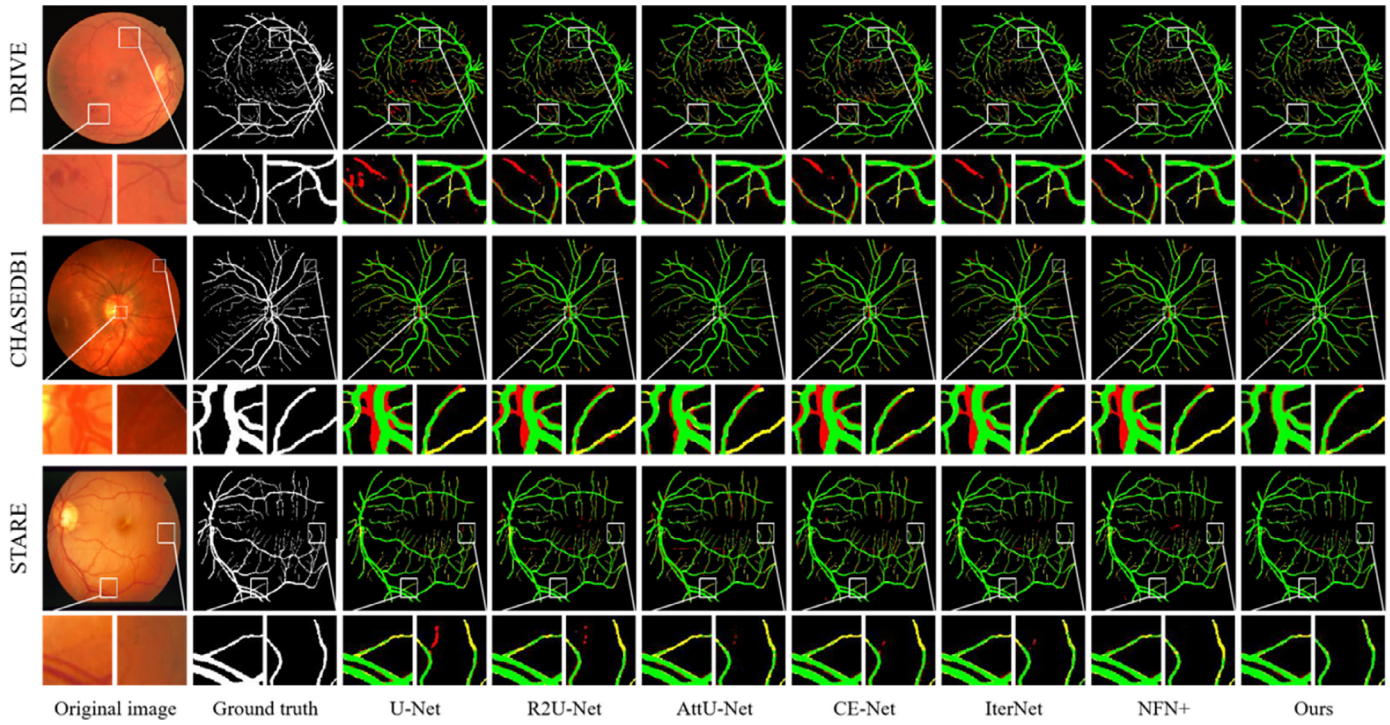

Fig. 7. Typical segmentation results of different methods on three classical datasets: DRIVE (top), CHASEDB1 (middle), STARE (bottom). The red pixel indicates false positives, the yellow pixel indicates false negatives, and the green pixels represent the true positives. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tic information and suppress many irrelevant background noises, which yields a more continuous segmentation map. On the other hand, despite IterNet and NFN+ can find obscured details of the vessel from the segmented vessel image itself by cascading multiple sub-networks, they cannot adapt well to the complexity vascular tree due to the limitations of the benchmark network itself. Overall, among the three datasets, our SCS-Net model generally outperforms the other competitors since the combination of SFA, AFF, and MSS modules can effectively deal with the large scale and anatomical semantics variations of retinal vessels.

5.2.2. Statistical evaluation

To quantitatively analyze the experimental results, we also perform a statistical comparison based on several important metrics including SE, SP, ACC, and AUC to evaluate our proposed SCS-Net and compare it with seven state-of-the-art methods on all six datasets. We first test the proposed SCS-Net on the DRIVE dataset. As we can see in the top of Table 3, the proposed SCS-Net outperforms the state-of-the-art methods in most metrics. Particularly, it achieves a SE of 82.89% in the DRIVE dataset. Moreover, it also achieves the highest accuracy of 96.97%, the highest AUC value of 98.37%, and a comparable SP of 98.38%, which demonstrates the superiority of the proposed SCS-Net.

We then conduct experiments on the CHASEDB1 dataset, and the comparison results of different methods are as shown in the middle of Table 3. We can observe that the SCS-Net achieves 83.65% and 97.44% in SE and ACC, respectively, better than other state-of-the-art methods. Compared with the classical U-Net, our SE increases from 76.17% to 83.65%. Meanwhile, ACC and AUC also increase from 97.16%/97.92% to 97.44%/98.67%, respectively, which demonstrates the effectiveness of the proposed SCS-Net. Note that the SP is slightly lower than the other methods due to a trade-off between SP and SE.

The performance comparisons on STARE dataset are summarized in the bottom of Table 3. The comparison results show that SCS-Net has a better performance in terms of AUC than other methods. Even compared with the latest NFN+ and IterNet, our SCS-Net still achieve improvements in terms of both SE and ACC. In addition, we also have reported the statistical comparison with state-of-the-art methods on the HRF, IOSTAR, and LES-AV datasets. As shown in Table 4, our method also generally obtains better segmentation accuracy than other state-of-the-art methods in terms of SE, SP, ACC and AUC indicators, even though the retinal vessels may have highly varying scale, illumination and complex structures, clearly demonstrating the advantages of our SCS-Net in retinal vessel segmentation. On the other hand, we also perform a commonly used non-parametric test method, Wilcoxon rank-sum test, to evaluate whether our performance improvement is statistically significant in terms of AUC metric compared to other current state-of-the-art methods, resulting in a p-value score. From the p-values shown in Tables 3-4, we can clearly see that our method has a statistically significant improvement in terms of AUC metric at the 5% level (all p-values are less than 0.05).

Moreover, we also incorporate ROC curves and AUC criteria to further assess the capability of different networks, as shown in Fig. 9. Based on ROC curves and AUC values calculated from the six datasets, we can clearly observe that our SCS-Net generally outperforms all competitors by achieving the highest ROC and AUC values.

6. Discussions

6.1. Key parameter analysis

As we mentioned above, the vessel map can be refined by the MSS module, which has a key parameter α will be to influence the performance of the segmentation. Therefore, we perform a set of experiments to evaluate the effect of the parameter on the performance of our model by setting different hyper-parameter α from 0 to 1 on the DRIVE database.

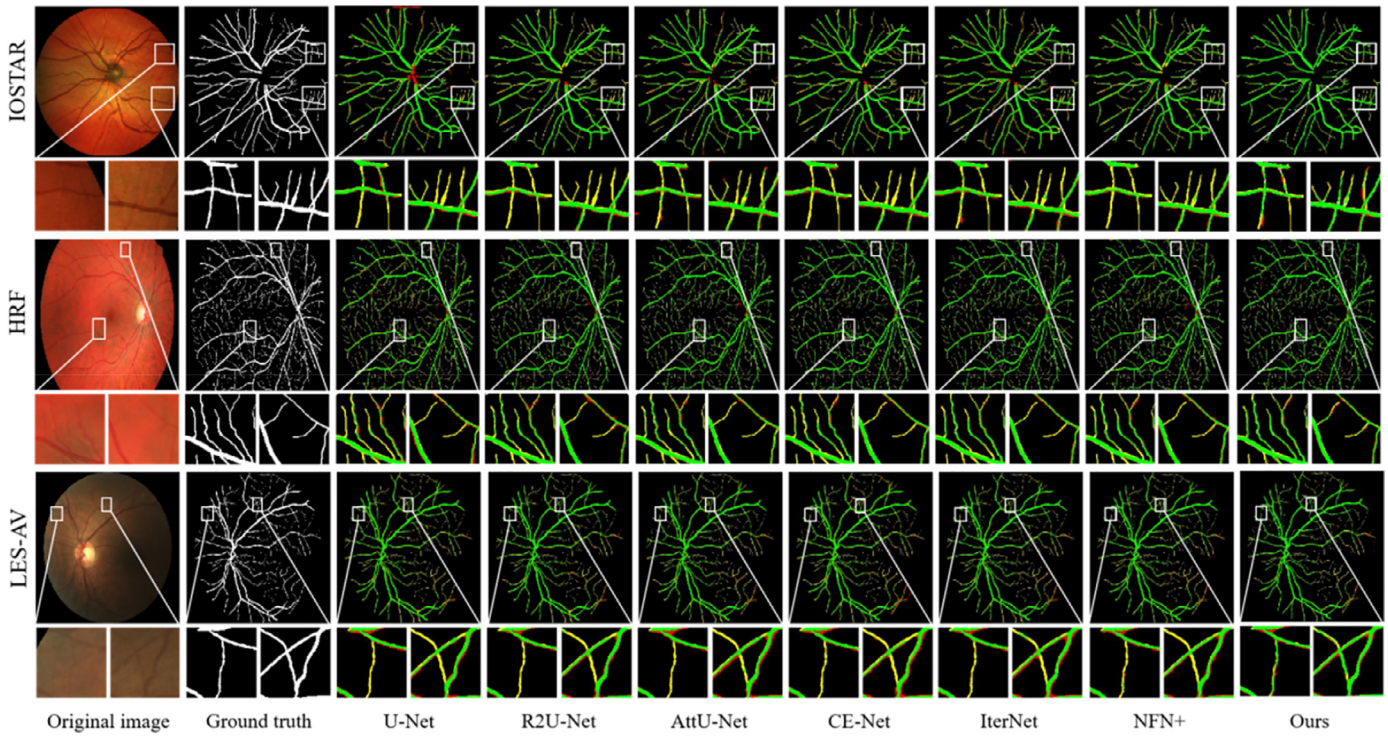

Fig. 8. Typical segmentation results of different methods on three newer datasets: IOSTAR (top), HRF (middle), LES-AV (bottom). The red pixel indicates false positives, the yellow pixel indicates false negatives, and the green pixels represent the true positives. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Statistical comparison with state-of-the-art methods on three classical dataset: DRIVE, CHASEDB1, and STARE.

DRIVE Dataset					
Method	SE (%)	SP (%)	ACC (%)	AUC (%)	p-value (AUC)
U-Net	79.15±0.23	98.08±0.31	96.40±0.13	97.64±0.08	1.89×10^{-6}
R2U-Net	79.23±0.56	98.03±0.48	96.54±0.07	98.02±0.08	3.73×10^{-5}
AttU-Net	78.82±0.17	**98.48±0.35**	96.49±0.19	98.03±0.07	1.10×10^{-5}
CE-Net	80.15±0.22	98.16±0.19	96.59±0.16	98.11±0.09	1.20×10^{-6}
IterNet	79.95±0.26	98.26±0.08	96.57±0.17	98.13±0.06	1.50×10^{-7}
NFN+	80.02±0.19	97.90±0.27	96.68±0.09	98.23±0.10	7.19×10^{-5}
SCS-Net (Ours)	**82.89±0.11**	98.38±0.29	**96.97±0.05**	**98.37±0.06**	–
CHASEDB1 Dataset					
Method	SE (%)	SP (%)	ACC (%)	AUC (%)	p-value (AUC)
U-Net	76.17±0.86	**98.61±0.69**	97.16±0.25	97.92±0.15	2.36×10^{-4}
R2U-Net	81.45±0.71	98.40±0.71	97.21±0.13	98.01±0.07	2.21×10^{-3}
AttU-Net	77.21±1.01	98.50±0.98	97.26±0.18	98.07±0.06	1.63×10^{-5}
CE-Net	80.42±0.39	98.39±0.33	97.23±0.36	98.06±0.09	1.90×10^{-6}
IterNet	79.97±1.55	98.47±1.05	97.31±0.24	98.26±0.12	3.45×10^{-6}
NFN+	79.33±0.95	98.55±0.88	97.35±0.18	98.32±0.04	8.30×10^{-4}
SCS-Net (Ours)	**83.65±0.69**	98.39±0.47	**97.44±0.10**	**98.67±0.05**	–
STARE Dataset					
Method	SE (%)	SP (%)	ACC (%)	AUC (%)	p-value (AUC)
U-Net	78.39±1.36	**98.71±0.96**	96.88±0.48	97.93±0.15	3.07×10^{-8}
R2U-Net	78.69±0.99	98.62±0.56	96.97±0.33	98.09±0.09	5.23×10^{-5}
AttU-Net	79.03±1.06	98.56±0.74	97.22±0.45	98.22±0.10	1.81×10^{-4}
CE-Net	79.16±0.86	98.53±1.11	97.15±0.25	98.17±0.07	7.39×10^{-6}
IterNet	80.86±0.53	98.46±0.68	97.23±0.36	98.29±0.07	1.83×10^{-4}
NFN+	80.96±0.78	98.43±0.93	97.27±0.41	98.44±0.05	9.38×10^{-3}
SCS-Net (Ours)	**82.07±0.66**	98.39±0.68	**97.36±0.28**	**98.77±0.06**	–

As we can see in Fig. 10, in general, with the help of the proposed MSS module, the network can detect more vessels and resulting in a more refined segmentation map. Meanwhile, we also see that the MSS module can improve the confidence of network prediction, which shows that our method can learn more discriminative and robust features in the early stage, and effectively address the information loss during forwarding propagation and improve the accuracy of details. As shown in Table 5, the network can

Table 4
Statistical comparison with state-of-the-art methods on three newer dataset: IOSTAR, HRF, and LES-AV.

HRF Dataset					
Method	SE (%)	SP (%)	ACC (%)	AUC (%)	p-value (AUC)
U-Net	80.44±0.79	97.76±1.35	96.38±0.06	97.75±0.02	4.42×10^{-7}
R2U-Net	80.02±0.16	97.89±0.31	96.46±0.13	97.86±0.22	4.66×10^{-3}
AttU-Net	78.55±0.83	98.14±1.08	96.58±0.32	97.93±0.08	1.96×10^{-4}
CE-Net	79.56±2.17	**98.23±4.16**	96.73±0.09	98.18±0.12	2.35×10^{-3}
IterNet	80.60±1.33	98.13±2.03	96.73±0.11	98.24±0.25	8.65×10^{-6}
NFN+	80.58±0.46	97.94±0.84	96.56±0.08	98.05±0.16	2.32×10^{-5}
SCS-Net (Ours)	**81.14±0.22**	**98.23±0.40**	**96.87±0.17**	**98.42±0.05**	–
IOSTAR Dataset					
Method	SE (%)	SP (%)	ACC (%)	AUC (%)	p-value (AUC)
U-Net	81.40±3.69	97.21±2.40	95.98±0.10	97.26±0.06	6.37×10^{-6}
R2U-Net	78.16±0.59	**98.57±1.64**	96.97±0.03	97.64±0.41	8.41×10^{-8}
AttU-Net	80.22±1.36	98.24±0.78	96.85±0.64	97.97±0.51	7.92×10^{-4}
CE-Net	80.04±0.99	98.00±0.55	96.60±1.65	98.15±0.35	6.99×10^{-3}
IterNet	80.13±0.30	97.98±1.74	96.45±0.98	98.01±0.15	3.57×10^{-5}
NFN+	79.21±2.17	98.12±3.48	96.83±0.61	98.03±0.67	4.18×10^{-3}
SCS-Net (Ours)	**82.55±1.78**	98.30±1.01	**97.06±0.32**	**98.65±0.20**	–
LES Dataset					
Method	SE (%)	SP (%)	ACC (%)	AUC (%)	p-value (AUC)
U-Net	80.99±1.33	98.57±1.37	97.15±0.65	97.53±0.03	6.84×10^{-5}
R2U-Net	80.04±4.98	98.72±2.56	97.29±0.97	97.71±0.07	9.33×10^{-6}
AttU-Net	81.33±0.98	98.74±0.78	97.33±0.88	97.88±0.05	8.18×10^{-5}
CE-Net	82.51±2.34	98.43±1.96	97.14±0.65	97.71±0.08	6.12×10^{-4}
IterNet	81.03±1.65	98.77±3.85	97.34±0.12	97.96±0.04	7.49×10^{-3}
NFN+	82.19±3.39	98.70±2.17	97.36±0.28	97.97±0.01	5.66×10^{-4}
SCS-Net (Ours)	**82.67±1.45**	**98.81±0.99**	**97.51±0.42**	**98.27±0.06**	–

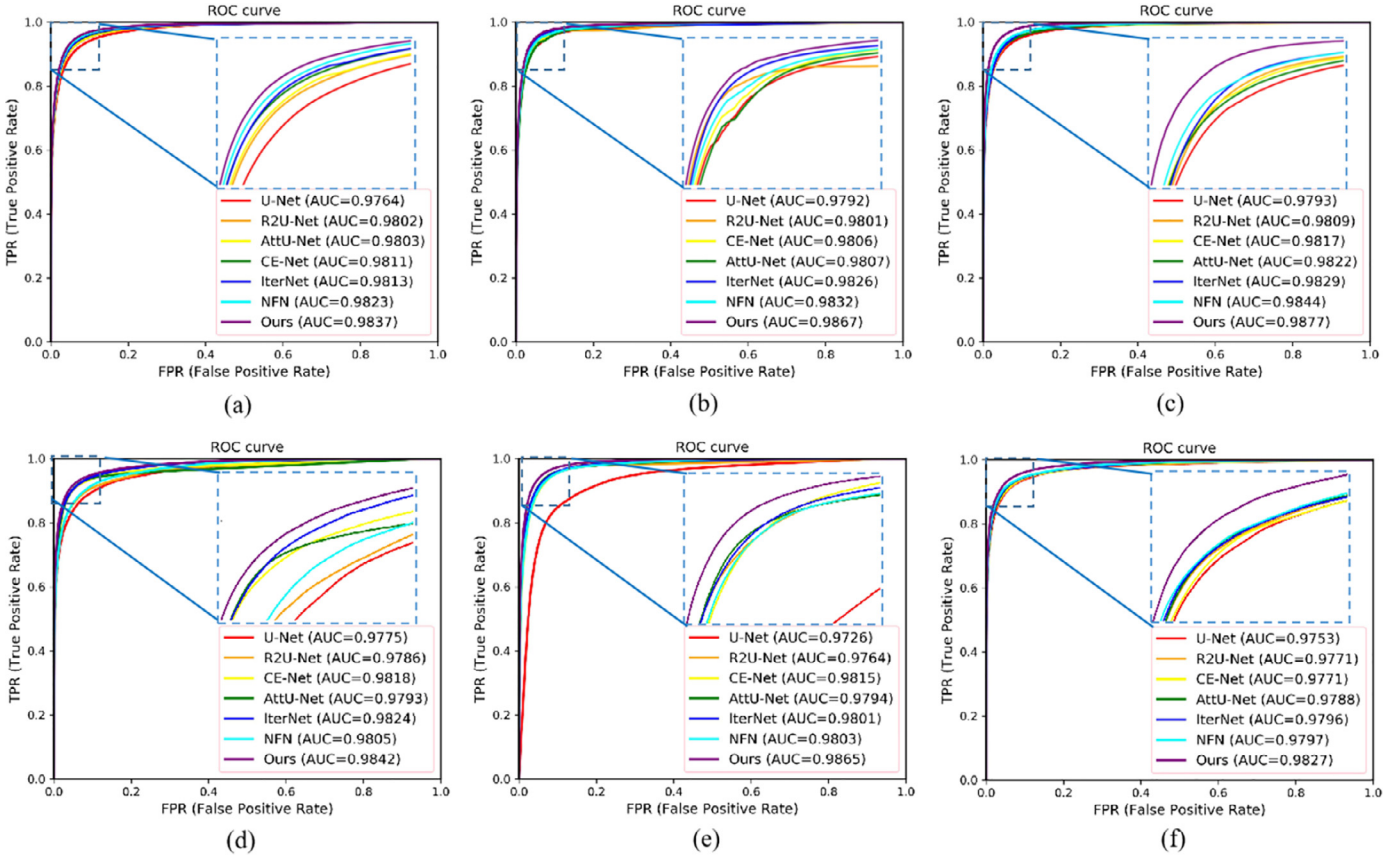

Fig. 9. ROC curves of different models for retinal vessel segmentation. (a) DRIVE, (b) CHASEDB1, (c) STARE. (d) IOSTAR, (e) HRF, and (f) LES-AV.

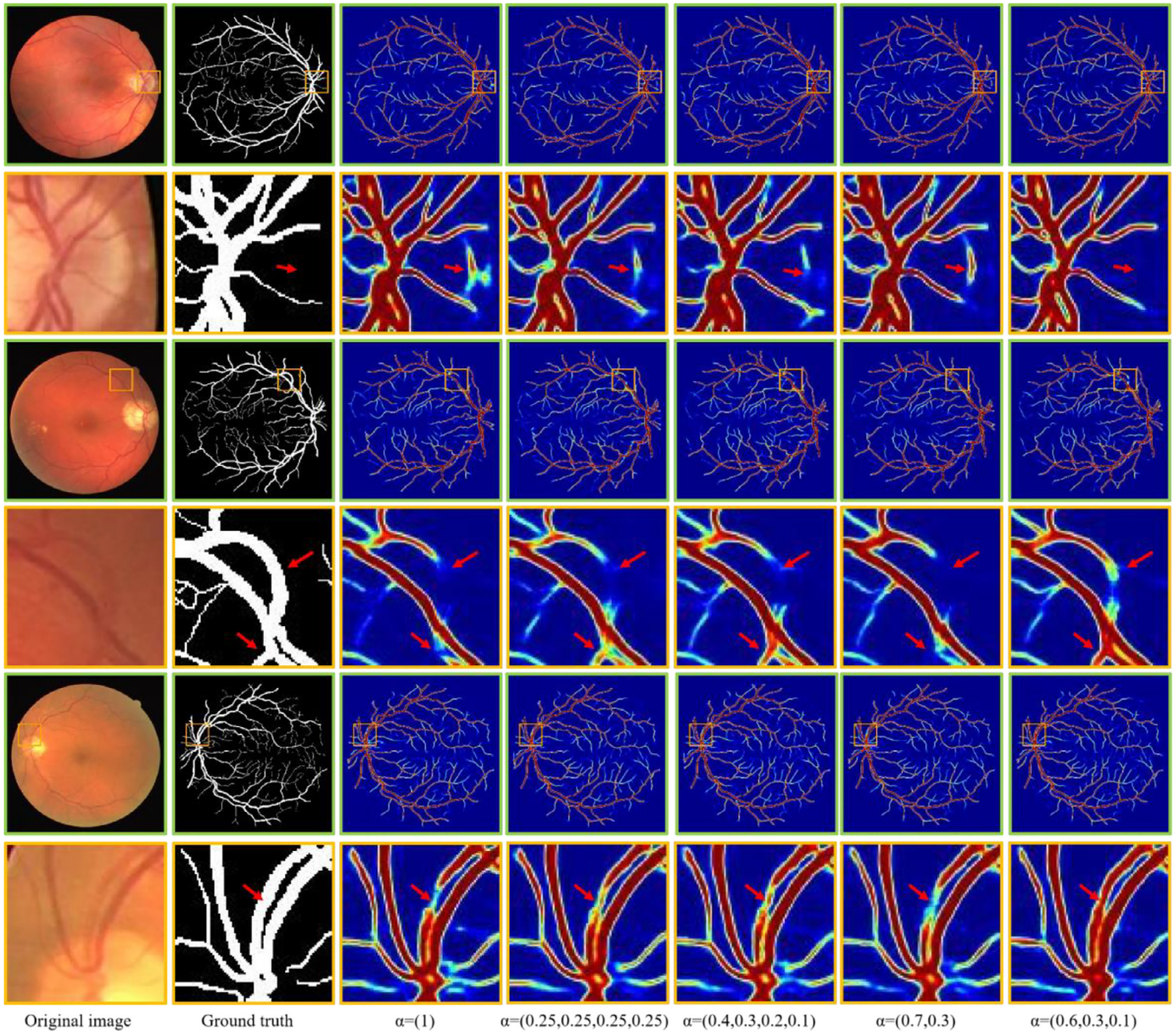

Fig. 10. Visual comparisons of three typical examples on DRIVE with different parameter α in the MSS module. The color indicates the confidence of each pixel, and the darker the color, the higher the probability of being predicted as a vessel.

Table 5
Accuracy comparison with different weighting strategies in the MSS module.

Weight $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$	ACC (%)	AUC (%)	F1 (%)	p-value (AUC)
α_1=1 (without MSS)	96.74±0.15	98.19±0.05	81.39±0.07	2.16×10^{-5}
α_1=0.25, α_2=0.25, α_3=0.25, α_4=0.25 (M-Net)	96.79±0.08	98.21±0.03	81.47±0.12	1.31×10^{-3}
α_1=0.4, α_2=0.3, α_3=0.2, α_4=0.1	96.81±0.06	98.25±0.05	81.52±0.09	7.25×10^{-4}
α_1=0.7, α_2=0.3	96.76±0.09	98.19±0.06	81.43±0.04	5.01×10^{-4}
α_1=0.6, α_2=0.3, α_3=0.1	**96.97±0.03**	**98.33±0.02**	**81.81±0.04**	–

obtain the highest performance when $\boldsymbol{\alpha} = (0.6, 0.3, 0.1)$, which increases by 0.52% in terms of F1 than the network without MSS. Therefore, we empirically set the key parameter $\boldsymbol{\alpha}$ to (0.6, 0.3, 0.1) for all our experiments in the comparison with state-of-the-art methods. The experimental results clearly demonstrate the advantages of our MSS. Compared with M-Net, we can generally obtain better retinal vessel segmentation performance, indicating that setting suitable weights for different side-output layers in a deep supervision mechanism is important.

6.2. Model complexity

This work proposes an innovative and efficient approach based on the U-shape structure for retinal vessel segmentation, which consists of three core modules: SFA, AFF, and MSS. The main motivation behind the model is to design an adaptive mechanism to address the major challenges of varying scale, shape and background imaging noises in vessel segmentation. Even though the vascular tree may appear with irregular shapes, different scales, and ir-

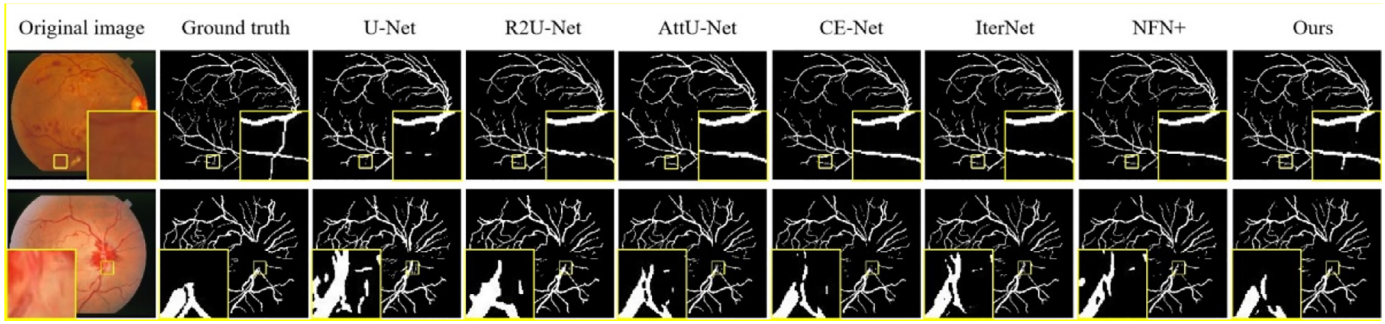

Fig. 11. Visual comparisons of failure cases with other state-of-the-art methods.

Table 6
Additional comparison with the state-of-the-art methods on CHASEDB1.

Method	MCC (%)	BM (%)	Parameter (M)
U-Net	78.35±0.88	76.74±0.38	**3.6**
R2U-Net	79.88±0.47	78.82±0.64	8.3
AttU-Net	80.04±0.15	77.63±0.22	7.2
CE-Net	80.26±0.27	80.34±0.31	14.4
IterNet	80.15±0.07	80.25±0.41	8.6
NFN+	80.23±0.34	80.37±0.27	4.5
Ours	**82.62**±0.05	**81.98**±0.48	3.7

relevant background noise, the above experiments show that our method still achieves satisfactory performance and improves the segmentation results well. The recent studies (Huang et al., 2017) have shown that increasing the complexity of the network usually enhances the ability of characterization and leads to a better performance. However, in many medical applications, it is not an optimal choice, as in many clinical settings, there is usually no sufficient computational resource to deploy and run models with high complexity. We compare the proposed SCS-Net with other state-of-the-art methods in terms of complexity by estimating their parameter amount. As shown in Table 6, compared with other competitors, our method achieves better performance with a highest SE/AUC score (83.65%/98.67%), without significantly increasing the number of parameters, which demonstrates that the improvement of our SCS-Net performance is not based on sacrificing the complexity of the network but achieved by proposing a set of novel and effective techniques to extract more scale and semantic information and generate more representative features. Compared with state-of-the-art methods, our SCS-Net is still a relative light-weight network, costing only 3.7 M parameters.

6.3. Limitations

Currently, all existing indicators for retinal vessel segmentation are pixel-based measurement, where thin vessels may not contribute much to the value of the indicators as they are only represented by a small number of pixels. However, above pixel-based metrics may suffer from the class imbalance problem, which commonly happens in medical imaging tasks, especially when the TP class (vessels) is significantly smaller than TN class (the rest of the image) in our retinal images. Obviously, the thin vessels only contain a small number of pixels in our retinal vessel segmentation task, where the detection of thin vessels is also much more challenging than the thick ones. Therefore, we should pay more attention on the thin vessels and attach higher weight to the smaller regions.

To demonstrate the advantage of our method in detecting very fine vessels, we further employed two additional metrics to evaluate our retinal vessel segmentation, including Matthews correlation coefficient (MCC) and bookmaker informedness (BM) (Luque et al., 2019), which are widely used for accuracy measurement and potentially more appropriate for the segmentation problem with category imbalance. The statistical results of comparisons among our proposed network and other six state-of-the-art methods on the CHASEDB1 dataset are as shown in Table 6. Compared with other pixel-based indicators (such as ACC and SP shown in Table 3 and Table 4), we can much more easily observe that our method significantly outperforms other competitors in terms of MCC and BM, which clearly verify the superiority of our proposed method in achieving better segmentation performance for the thin vessels.

We still find some failure cases in our experiments. As shown in Fig. 11, like most previous methods, when the contrast between the blood vessels and the background is extremely low, it is difficult for our model to accurately identify the vessels. On the other hand, for typical areas where the background noise is too severe, our method may also produce some false positives similar to other methods. However, even under such extreme conditions, our method still has considerable performance improvement compared with other state-of-the-art methods, and the segmentation results of our model are closer to the ground truth.

7. Conclusions

We present a novel retinal vessel segmentation network, namely SCS-Net, to comprehensively address the challenges of this task. The proposed SCS-Net is capable of effectively capturing multi-scale contextual information and promoting the fusion of the features at different levels in order to obtain more semantic representations. Three core modules, i.e. SFA, AFF, and MSS, are proposed. In the top stage of the encoder, we develop the SFA module to effectively extract the multi-scale contextual information by implicitly and dynamically adjust the receptive fields of the feature maps. In each stage of the decoder, the adjacent level features are adaptively fused by the AFF module, which fully combines the semantic information from the high-level features and the spatial information from the low-level features while simultaneously suppressing the background noise. In addition, the MSS module is harnessed to learn more semantic representations for refining the vessel segmentation maps. Extensive comparative evaluations on six publicly available retinal fundus databases (DRIVE, CHASEDB1, STARE, IOSTAR, HRF, and LES-AV) are implemented, which demonstrates the superiority of the proposed method over state-of-the-art approaches. The proposed techniques are general enough and we believe they can be easily extended to other medical image segmentation tasks where large scale variation and complicated anatomical semantics are main challenges.

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.

CRediT authorship contribution statement

Huisi Wu: Conceptualization, Funding acquisition, Project administration, Methodology, Writing - original draft, Writing - review & editing. **Wei Wang:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing. **Jiafu Zhong:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing. **Baiying Lei:** Conceptualization, Funding acquisition, Project administration, Methodology, Writing - original draft, Writing - review & editing. **Zhenkun Wen:** Supervision, Funding acquisition, Project administration. **Jing Qin:** Conceptualization, Funding acquisition, Project administration, Methodology, Writing - original draft, Writing - review & editing.

Acknowledgement

This work was supported partly by National Natural Science Foundation of China (nos. 61973221, 61871274, 61801305, 61872351, and 81571758), the Natural Science Foundation of Guangdong Province, China (nos. 2018A030313381 and 2019A1515011165), the Major Project or Key Lab of Shenzhen Research Foundation, China (nos. JCYJ20160608173051207, ZDSYS201707311550233, KJYY201807031540021294 and JSGG201805081520220065), the COVID-19 Prevention Project of Guangdong Province, China (no. 2020KZDZX1174), the Major Project of the New Generation of Artificial Intelligence (no. 2018AAA0102900), International Science and Technology Cooperation Projects of Guangdong (no. 2019A050510030), Key Laboratory of Medical Image Processing of Guangdong Province (no. K217300003). Guangdong Pearl River Talents Plan (2016ZT06S220), Shenzhen Peacock Plan (nos. KQTD2016053112051497 and KQTD2015033016104926), Shenzhen Key Basic Research Project (nos. JCYJ20180507184647636, JCYJ20170413161913429, JCYJ20180507184647636, and JCYJ20190808155618806), and the Hong Kong Research Grants Council of China under Grant PolyU 152035/17E and Grant 15205919.

References

Ahn, E., Kumar, A., Fulham, M., Feng, D., Kim, J., 2019. Convolutional sparse kernel network for unsupervised medical image analysis. Med. Image Anal. 56, 140–151.
Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K., 2018. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation arXiv preprint arXiv:1802.06955.
Bilinski, P., Prisacariu, V., 2018. Dense decoder shortcut connections for single-pass semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 6596–6605.
Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 834–848.
Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017b. Rethinking atrous convolution for semantic image segmentation arXiv preprint arXiv:1706.05587.
Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 801–818.
Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 424–432.
Fan, Z., Wei, J., Zhu, G., Mo, J., Li, W., 2020. ENAS U-Net: Evolutionary Neural Architecture Search for Retinal Vessel Segmentation arXiv preprint arXiv:2001.06678.
Feng, S., Zhao, H., Shi, F., Cheng, X., Wang, M., Ma, Y., Xiang, D., Zhu, W., Chen, X., 2020. CPFNet: Context Pyramid Fusion Network for Medical Image Segmentation. IEEE Trans. Med. Imaging 39 (10), 3008–3018.
Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X., 2018. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE Trans. Med. Imaging 37 (7), 1597–1605.
Fu, H., Xu, Y., Lin, S., Wong, D.W.K., Liu, J., 2016. Deepvessel: Retinal vessel segmentation via deep learning and conditional random field. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 132–139.
Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J., 2019. CE-Net: context encoder network for 2D medical image segmentation. IEEE Trans. Med. Imaging 38 (10), 2281–2292.
Heinrich, M.P., Oktay, O., Bouteldja, N., 2019. OBELISK-Net: Fewer layers to solve 3D multi-organ segmentation with sparse deformable convolutions. Med. Image Anal. 54, 1–9.
Hoover, A.D., Kouznetsova, V., Goldbaum, M., 2000. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Trans. Med. Imaging 19 (3), 203–210.
Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 7132–7141.
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 4700–4708.
Huang, K., Yan, M., 2006. A region based algorithm for vessel detection in retinal images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 645–653.
Ibtehaz, N., Rahman, M.S., 2020. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Netw 121, 74–87.
Jin, Q., Meng, Z., Pham, T.D., Chen, Q., Wei, L., Su, R., 2019. DUNet: A deformable network for retinal vessel segmentation. Knowl Based Syst 178, 149–162.
Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.
Khened, M., Kollerathu, V.A., Krishnamurthi, G., 2019. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. Med. Image Anal. 51, 21–45.
Köhler, T., Budai, A., Kraus, M.F., Odstrčilik, J., Michelson, G., Hornegger, J., 2013. Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. IEEE, pp. 95–100.
Lam, B.S., Gao, Y., Liew, A.W.C., 2010. General retinal vessel segmentation using regularization-based multiconcavity modeling. IEEE Trans. Med. Imaging 29 (7), 1369–1381.
Lei, B., Xia, Z., Jiang, F., Jiang, X., Ge, Z., Xu, Y., Qin, J., Chen, S., Wang, T., Wang, S., 2020. Skin Lesion Segmentation via Generative Adversarial Networks with Dual Discriminators. Med. Image Anal. 64, 101716.
Li, L., Verma, M., Nakashima, Y., Nagahara, H., Kawasaki, R., 2020. IterNet: Retinal Image Segmentation Utilizing Structural Redundancy in Vessel Networks. In: The IEEE Winter Conference on Applications of Computer Vision. IEEE, pp. 3656–3665.
Li, Q., Feng, B., Xie, L., Liang, P., Zhang, H., Wang, T., 2015. A cross-modality learning approach for vessel segmentation in retinal images. IEEE Trans. Med. Imaging 35 (1), 109–118.
Li, X., Wang, W., Hu, X., Yang, J., 2019. Selective kernel networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 510–519.
Li, Z., Zhang, X., Müller, H., Zhang, S., 2018. Large-scale retrieval for medical image analytics: A comprehensive review. Med. Image Anal. 43, 66–84.
Liskowski, P., Krawiec, K., 2016. Segmenting retinal blood vessels with deep neural networks. IEEE Trans. Med. Imaging 35 (11), 2369–2380.
Liu, W., Rabinovich, A., Berg, A.C., 2015. Parsenet: Looking wider to see better arXiv preprint arXiv:1506.04579.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3431–3440.
López-Linares, K., Aranjuelo, N., Kabongo, L., Maclair, G., Lete, N., Ceresa, M., García-Familiar, A., Macía, I., Ballester, M.A.G., 2018. Fully automatic detection and segmentation of abdominal aortic thrombus in post-operative CTA images using deep convolutional neural networks. Med. Image Anal. 46, 202–214.
Luque, A., Carrasco, A., Martín, A., de las Heras, A., 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognit 91, 216–231.
Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE, pp. 565–571.
Mo, J., Zhang, L., 2017. Multi-level deep supervised networks for retinal vessel segmentation. Int J Comput Assist Radiol Surg 12 (12), 2181–2193.
Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. In: Internation Conference on Machine Learning, pp. 807–814.
Orlando, J.I., Breda, J.B., Van Keer, K., Blaschko, M.B., Blanco, P.J., Bulant, C.A., 2018. Towards a glaucoma risk index based on simulated hemodynamics from fundus images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 65–73.
Owen, C.G., Rudnicka, A.R., Mullen, R., Barman, S.A., Monekosso, D., Whincup, P.H., Ng, J., Paterson, C., 2009. Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (CAIAR) program. Investig. Ophthalmol. Vis. Sci. 50 (5), 2004–2010.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: Learning to leverage salient regions in medical images. Med. Image Anal. 53, 197–207.
Shin, S.Y., Lee, S., Yun, I.D., Lee, K.M., 2019. Deep vessel segmentation by learning graphical connectivity. Med. Image Anal. 58, 101556.
Si, Z., Fu, D., Li, J., 2019. U-Net with Attention Mechanism for Retinal Vessel Segmentation. In: International Conference on Image and Graphics. Springer, pp. 668–677.
Soares, J.V., Leandro, J.J., Cesar, R.M., Jelinek, H.F., Cree, M.J., 2006. Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification. IEEE Trans. Med. Imaging 25 (9), 1214–1222.
Son, J., Park, S.J., Jung, K.H., 2017. Retinal vessel segmentation in fundoscopic images with generative adversarial networks arXiv preprint arXiv:1706.09318.
Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B., 2004. Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging 23 (4), 501–509.
Wang, B., Qiu, S., He, H., 2019. Dual Encoding U-Net for Retinal Vessel Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 84–92.
Wang, C., Oda, M., Hayashi, Y., Yoshino, Y., Yamamoto, T., Frangi, A.F., Mori, K., 2020. Tensor-cut: A tensor-based graph-cut blood vessel segmentation method and its application to renal artery segmentation. Med. Image Anal. 60, 101623.
Wu, Y., Xia, Y., Song, Y., Zhang, Y., Cai, W., 2020. NFN+: A novel network followed network for retinal vessel segmentation. Neural Netw 126, 153–162.
Yan, Z., Yang, X., Cheng, K.T., 2018. Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation. IEEE Trans Biomed Circuits Syst 65 (9), 1912–1923.
Zhang, J., Cui, Y., Jiang, W., Wang, L., 2015. Blood vessel segmentation of retinal images based on neural network. In: International Conference on Image and Graphics. Springer, pp. 11–17.
Zhang, J., Dashtbozorg, B., Bekkers, E., Pluim, J.P., Duits, R., ter Haar Romeny, B.M., 2016. Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores. IEEE Trans. Med. Imaging 35 (12), 2631–2644.
Zhang, J., Liu, M., Wang, L., Chen, S., Yuan, P., Li, J., Shen, S.G.-F., Tang, Z., Chen, K.-C., Xia, J.J., 2020. Context-guided fully convolutional networks for joint craniomaxillofacial bone segmentation and landmark digitization. Med. Image Anal. 60, 101621.
Zhang, Y., Chung, A.C., 2018. Deep supervision with additional labels for retinal vessel segmentation task. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 83–91.
Zhang, Z., Zhang, X., Peng, C., Xue, X., Sun, J., 2018. Exfuse: Enhancing feature fusion for semantic segmentation. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 269–284.
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2881–2890.
Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 3–11.