Global Transformer and Dual Local Attention Network via Deep-Shallow Hierarchical Feature Fusion for Retinal Vessel Segmentation

Yang Li, Yue Zhang, Jing-Yu Liu, Kang Wang, Kai Zhang,
Gen-Sheng Zhang, Xiao-Feng Liao, Guang Yang

***Abstract*—Clinically, retinal vessel segmentation is a significant step in the diagnosis of fundus diseases. However, recent methods generally neglect the difference of semantic information between deep and shallow features, which fail to capture the global and local characterizations in fundus images simultaneously, resulting in the limited segmentation performance for fine vessels. In this paper, a global transformer and dual local attention network via deep-shallow hierarchical feature fusion (GT-DLA-dsHFF) are investigated to solve the above limitations. First, the global transformer (GT) is developed to integrate the global information in the retinal image, which effectively captures the long-distance dependence between pixels, alleviating the discontinuity of blood vessels in the segmentation results. Second, the dual local attention (DLA), which is constructed using dilated convolutions with varied dilation rates, unsupervised edge detection, and squeeze-excitation block, is proposed to extract local vessel information, consolidating the edge details in the segmentation result. Finally, a novel deep-shallow hierarchical feature fusion (dsHFF) algorithm is studied to fuse the features in different scales in the deep learning framework respectively, which can mitigate the attenuation of valid information in the process of feature fusion. We verified the GT-DLA-dsHFF on four typical fundus image datasets. The experimental results demonstrate our GT-DLA-dsHFF achieves superior performance against the current methods and detailed discussions verify the efficacy of the proposed three modules. Segmentation results on diseased images show the robustness of our proposed GT-DLA-dsHFF. Our codes will be available on https://github.com/YangLibuaa/GT-DLA-dsHFF.**

***Index Terms*—Medical image analysis, global transformer, dual local attention, deep-shallow hierarchical feature fusion, retinal vessel segmentation.**

I. Introduction

STUDIES have shown that fundus diseases are one of the most important causes of blindness. Clinically, retinal image processing is adopted by doctors to screen for fundus diseases such as diabetic retinopathy, glaucoma, hypertension, and age-related macular degeneration [1, 2]. Ophthalmologists make the diagnosis by analyzing abnormal changes in fundus images [3]. However, annotating retina blood vessels is a time-costing and tricky task, which places extremely high requirements on the professional capabilities of physicians [4]. In addition, medical images have the characterizations of inferior contrast, noise, and the complex structure of vessels [5], which cause a series of difficulties in vessels segmentation. Thus, fast and accurate approaches for the segmentation of retinal vessels are one of the core technologies urgently needed in ophthalmology medicine.

Medical image segmentation approaches have been widely studied in the past decades [6]. Unfortunately, traditional approaches such as matched filtering [7], vessel tracking [8], and morphology processed vessels locally and failed to consider the global characteristics of vessels. For instance, Mendonca et al. identified blood vessels by combining centerline extraction and morphological reconstruction [9], where the centerline was captured in a single vessel from a local level, ignoring the global relationship among vessels. Moreover, local noises are normally present in fundus images due to interference during the acquisition process. These noises blur the edges of the vessels, resulting in the weak robustness of the segmentation methods which used local approaches [10]. Although the graph-based approaches propagated the global information through building the topology map [11], in terms of the complication, the traditional image processing-based methods had many hyper-parameters that need to be manually pruned [4], which might increase the computing complication.

To resolve the limitations, machine learning methods, which obtain global information under the condition of a few hyper-parameters, were applied in the image processing task [12, 13]. Moreover, convolutional neural networks (CNNs) used filters to extract semantic information in images, achieving more ideal results than previous methods [14-16]. The U-shaped network (U-Net) [17] utilized the skip connection within the encoder-

This work was supported in part by the National Natural Science Foundation of China [U1809209, 61671042, 61403016]; in part by the Beijing Natural Science Foundation [L182015, 4172037] and the Beijing United Imaging Research Institute of Intelligent Imaging Foundation [CRIBJQY202103]; in part by the BHF (TG/18/5/34111, PG/16/78/32402), the ERC IMI (101005122), the H2020 (952172), the MRC (MC/PC/21013), and the UKRI Future Leaders Fellowship (MR/V023799/1). (Corresponding authors: Yue Zhang, Gen-Sheng Zhang)

Yang Li, Yue Zhang and Jing-Yu Liu are with the School of Automation Sciences and Electrical Engineering, Beihang University, Beijing, China (E-mails: liyang@buaa.edu.cn; zhang_yue@buaa.edu.cn; liujingyu@buaa.edu.cn).

Kang Wang is with Department of Ophthalmology, Beijing Friendship Hospital, Beijing, China (Email: bnbn2000@163.com).

Kai Zhang and Gen-Sheng Zhang are with the Second Affiliated Hospital of Medical College, Zhejiang University, Zhejiang, China. (E-mails: zhangkai1993@zju.edu.cn; genshengzhang@zju.edu.cn).

Xiao-Feng Liao is with the College of Computer Science, Chongqing University, Chongqing, China (E-mail: xfliao@cqu.edu.cn).

Guang Yang is with the National Heart and Lung Institute, Imperial College London, London, SW7 2AZ, UK. (E-mail: g.yang@imperial.ac.uk).

decoder architecture, which connects the features generated by the encoder with the decoder features, refraining from the loss of multi-scale details due to the pooling operation. Additionally, numerous efforts have been dedicated to the improvement of the CNN model. For example, Wang et al. combined two identical multi-scale backbone networks with skip connections, effectively integrating the shallow and deep features of the network [18]. Wang et al. also proposed a feature pyramid cascade module to capture the scale changes of vessels and alleviate the problem of discontinuous segmentation results by aggregating local and global context information [19]. Wang et al. further investigated a hard attention network, which consists of an encoder and three decoders to solve the segmentation of vessel backbone, easy region, and difficult region respectively [20]. In order to refine segmentation results in different training stages, Lian et al. designed a global and local enhanced residual U-Net (GLUE). The GLUE cascaded two backbone networks and preprocesses retinal images with the contrast limited adaptive histogram equalization (CLAHE) to obtain global and local information of blood vessels [21]. Some studies were also aimed to improve the loss function in recent years. For instance, Yan et al. studied a loss function that combined the segment-level and pixel-level losses to balance the importance of vessels in different scales [22]. Guo et al. further adopted a deep supervision scheme to optimize the neural network and compared the vessel probability map obtained by each convolutional layer with the ground truth [23]. Some other methods were also improved on the connection mechanism of U-Net, such as increasing the number of skip connections and using multiple image coding paths to capture information [24-27]. Nevertheless, in spite of their exceptional representational power of the U-Net architecture, there are some limitations that need to be discussed [28]. The CNN-based approaches generally exhibit limitations for modeling explicit long-range pixel relation, due to the intrinsic locality of convolution operations [29, 30]. The Transformer was originally proposed for natural language processing [31], which utilized the self-attention mechanism to obtain long-distance dependencies of words and was applied to computer vision tasks in 2020, named Vision Transformer (ViT). In ViT, patches from different areas of the image were treated as different words, and the dependencies of different positions in the image were obtained through the self-attention mechanism [32]. Recently, the Transformer has been tried to apply in medical image processing, and the performance has approached or even surpassed traditional CNN methods [33]. For example, Huang et al. adopted a relational transformer network that utilizes self-attention and cross-attention mechanisms to integrate complex dependencies between different regions of fundus images [32]. For 3D medical images, Guo et al. proposed a transformer-based network to solve the anisotropy problem [34]. Liu et al. also studied a global pixel transformer based on U-Net and combined it with dense blocks to integrate global and local information using a multi-scale input strategy [29].

In contrast to the traditional image processing approaches, the U-Net-based methods obtained the global information more efficiently. However, the fixed receptive field of the convolution kernels in the U-Net results in inefficient information acquisition [35]. To tackle this issue, attention mechanism-based approaches strengthened local details by generating a weight map and are combined with the U-Net [36, 37]. For instance, Mou et al. studied a channel and spatial attention network (CS-Net), which utilizes convolution kernels in different directions to capture vessel structure and highlights the region of interest through weight maps generated by the channel and spatial attention mechanism [38]. In order to expand the receptive field, some recent methods used pooling operations or increased the size of the convolution kernel [39]. However, these methods may impose a burden on computational complexity. To adjust the receptive field more flexibly without increasing computational cost, the dilated convolution was applied to replace the original convolution for medical images segmentation tasks [40]. To address the edge attenuation in segmentation results, Zhang et al. also studied a boundary enhancement method with Sobel operators, alleviating the loss of vessel edge [41]. In addition, recent studies showed that the deep features and shallow features in a deep neural network can be complementary [42]. However, some approaches neglect the connection and difference between these two categories of features, leading to the loss of valid information [43].

Inspired by the discussions aforementioned, we studied a novel framework called global transformer and dual local attention network via deep-shallow hierarchical feature fusion (GT-DLA-dsHFF). First, we design the global transformer (GT) which implants the transformer into the U-Net architecture to capture the long-distance dependence among pixels, solving the problem of low efficiency in extracting global information. Second, the dual local attention (DLA) is constructed using the dilated convolutions with varied dilation rates, unsupervised edge detection method, and squeeze-excitation block, consolidating the edge details in the segmentation result. Finally, we propose a deep-shallow hierarchical feature fusion (dsHFF) algorithm which processes the deep and shallow features respectively We perform abundant experiments and discussions for our proposed model on the four public datasets, achieving more satisfactory segmentation performance against the current methods.

Our contributions are concluded below:

1) We design the global transformer to obtain the comprehensive characterization of the image within the encoder, solving the problem that low efficiency in extracting long-distance dependence between pixels and alleviating the discontinuity of blood vessels in the segmentation results.

2) A dual local attention that adopts the dilated convolution, unsupervised edge detection method, and squeeze-excitation block is investigated to expand the receptive fields and obtain local information of feature maps, consolidating the edge details of the retinal vessel.

3) We propose a deep-shallow hierarchical feature fusion to fuse the deep-shallow features which fully utilizes the complementary information between the deep and shallow features, avoiding the loss of information in the feature fusion.

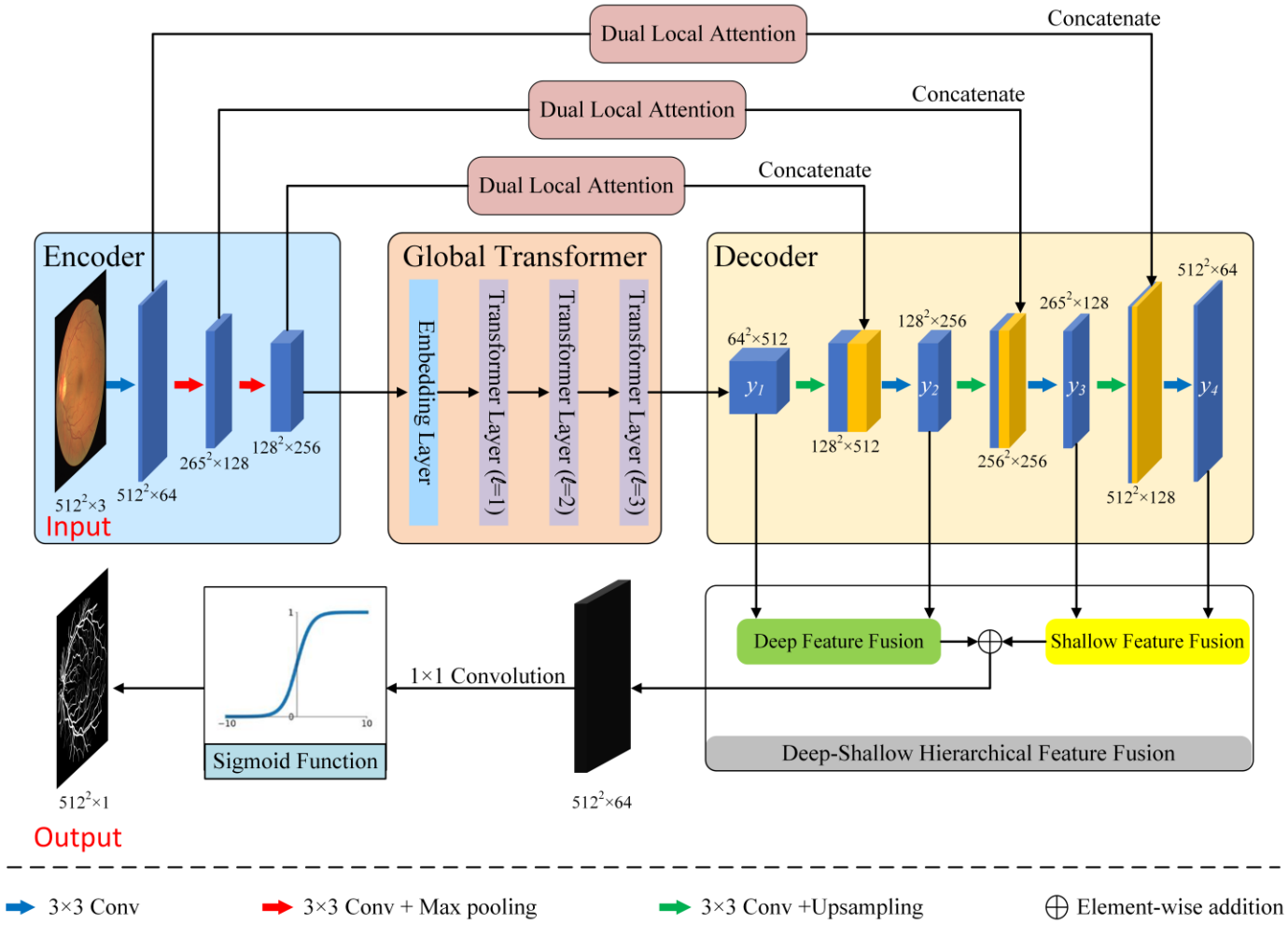


Fig. 1. The diagram of the proposed GT-DLA-dsHFF method.

II. Methodology

The proposed GT-DLA-dsHFF, which is designed on the U-Net framework with global transformer, dual local attention, and deep-shallow hierarchical feature fusion, is illustrated as Fig.1, and concluded as follows: 1) The GT is embedded between the encoder and decoder, which contains an embedding layer and three transformer layers. Each transformer layer has the multi-head self-attention (MHSA) and multi-layer perceptron (MLP), employing the residual connection to avoid gradient vanishing. We convert the image features into sequences via the embedding layer and feed them to the GT. Then the output of GT is reshaped to the size that fits the decoder so that obtain the global dependency of pixels. 2) The input of the DLA is the features generated by the encoder, and the local features extracted by the DLA are concatenated with the features generated in the decoder through skip connections to obtain the multi-scale supplementary information. 3) The dsHFF is applied to fuse the feature maps y_1, y_2, y_3, y_4 in deep and shallow levels hierarchically, avoiding the deficiency of the detailed vessel information and improving the segmentation performance of fine retinal vessels. 4) The segmentation results are generated from the features fused by dsHFF. The details of the GT, DLA, and dsHFF are given in the following subsections.

A. Global Transformer

As shown in Fig. 1, in the GT-DLA-dsHFF, the encoder generates high-level features of the image, which is regarded as the input feature x of GT here. The size of the input is $H \times W \times C$, where C denotes the number of image channels, and $H \times W$ is the size of the feature. The illustration of the GT is outlined in Fig.2. In order to transfer x to the sequence data so that the transformer layer can address the long-distance dependence between pixels, we first map x into a sequence k_0. Specifically, x is divided into N patches in dimension C at the size of $P \times P \times C$, where $N = \frac{H \times W}{P^2}$. These patches are reshaped to sequences $x_1, x_2, \ldots, x_N$ with the size of $1 \times C$ by the $P \times P$ convolution operator. It is worth pointing out that we employ a convolution operator to reshape all patches, which can reduce the number of parameters of the model and reduce inference time. In this paper, the output of the GT has the size of $\frac{H}{2} \times \frac{W}{2} \times 2C$ due to the parameter P being set to 2. The sequences $x_1, x_2, \ldots, x_N$ are mapped to the hidden space which has the dimension with $1 \times D$. In addition, the work in [31] indicates that unlike traditional recurrent neural networks, Transformer structures cannot obtain the position information of words in sequences (in this case image patches), so we need

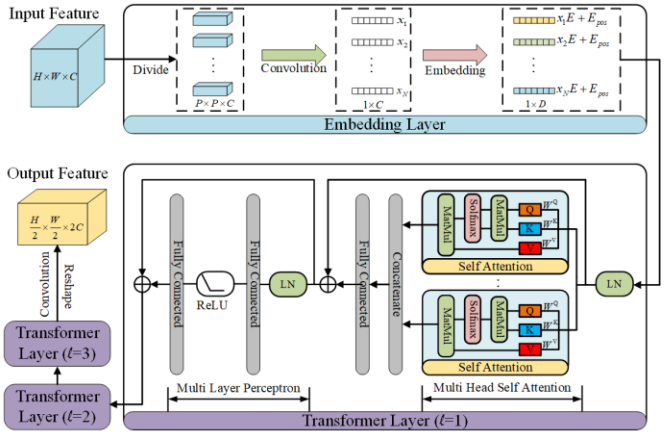


Fig. 2. The diagram of the global transformer.

to embed positions for the resulting sequences. We concatenate these vectors obtained above and then sum the concatenated vector with the position embedding amount to obtain the input k_0 of the first transformer layer. The $k_0 \in \mathbb{R}^{N\times D}$ can be calculated by:

$$k_0 = \boldsymbol{x}^T E + E_{pos}, \tag{1}$$

where $\boldsymbol{x} = [x_1{}^T, x_2{}^T, \ldots, x_N{}^T]$, $E \in \mathbb{R}^{C\times D}$ denotes projection embedding and $E_{pos} \in \mathbb{R}^{N\times D}$ denotes position embedding. Both of the parameter E and E_{pos} are learnable.

From Fig. 2, the transformer layer studied in this paper mainly consists of two parts: the multi-head self-attention mechanism MHSA(·) and the multi-layer perceptron MLP(·). The output of the ℓ-th ($\ell = 1,2,3$) transformer layer k_l can be calculated as:

$$k'_\ell = \mathrm{MHSA}\big(\mathrm{LN}(k_{\ell-1})\big) + k_{\ell-1} \tag{2}$$

$$k_\ell = \mathrm{MLP}\big(\mathrm{LN}(k'_\ell)\big) + k'_\ell, \tag{3}$$

where LN (·) denotes the layer normalization. For the MHSA, we denote the input of MHSA as x_{MHSA}, which has the size of $N \times D$. From Fig.2, the MHSA is composed of self-attention mechanisms and a fully connected layer. The MHSA fuses the outputs of the self-attention mechanisms by feature concatenation and a fully connected layer. The self-attention mechanism includes a query transform matrix $W^Q \in \mathbb{R}^{D\times d_k}$, a key transform matrix $W^K \in \mathbb{R}^{D\times d_k}$ and a value transform matrix $W^V \in \mathbb{R}^{D\times d_v}$, where d_k and d_v are integers. Therefore, the query Q, key K , and value V in Fig.2 can be defined by:

$$\begin{aligned} Q &= x_{\mathrm{MHSA}} W^Q \\ K &= x_{\mathrm{MHSA}} W^K \\ V &= x_{\mathrm{MHSA}} W^V. \end{aligned} \tag{4}$$

The output of the self-attention mechanism Z can be calculated as:

$$Z = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V. \tag{5}$$

Eq. (5) calculates the weight of the global spatial relationship between all positions in the feature map through the softmax function and fuses the features of all positions in the image space through matrix multiplication. Therefore, through the cascade of multiple transformer encoder layers, the global information of image features can be effectively extracted, which makes up for the drawback that the low efficiency of traditional CNN methods in capturing global information, thus improving the segmentation quality for fine vessels.

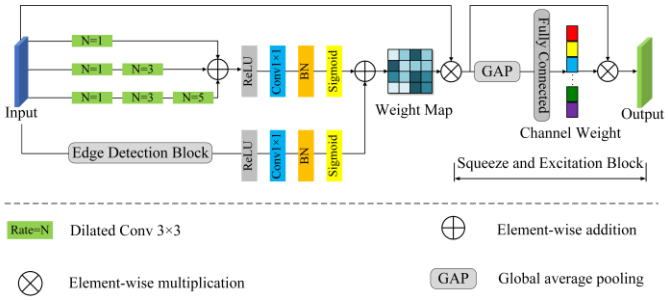


Fig. 3. The diagram of the dual local attention.

B. *Dual Local Attention*

It is well known that diversified convolutional receptive fields can extract multi-scale local information of images[35]. Most existing methods employed pooling operations or increase the size of the convolution kernel to expand receptive fields. Nevertheless, max-pooling inevitably causes the loss of the local spatial information and large convolution kernels cause the complexity burden of the model [4]. The dilated convolution, which expands the receptive area without parameters increasing, is widely adopted in vessel segmentation [39]. Additionally, to prevent the attenuation of the edge information of blood vessels, researchers employed a combination of unsupervised edge detection and deep networks [41]. Therefore, combining the dilated convolution and edge detection method is an effective method to capture the local characterizations of neighboring pixels while expanding the local receptive field of the network. Inspired by the above, we investigate dual local attention and embed it to the architecture in Fig.1.

Specifically, as outlined in Fig.1 and Fig.3, the encoder features x_{encoder} are regarded as the input of the DLA. To extract the multi-scale local information, we cascade dilated convolutions to form three convolution paths. The three paths embed dilated convolutions in different dilated rates. Each path possesses the unique receptive field to obtain multi-scale local information. We fuse the outputs of the three paths as D by element-wise addition, which can be defined by:

$$D = \sum_{p=0}^{2} \sum_{r\in\{1,3,\cdots,2p+1\}} \mathrm{dilated\ conv}_r(x_{\mathrm{encoder}}), \tag{6}$$

where dilated $\mathrm{conv}_r(\cdot)$ denotes the dilated convolution with a dilated rate of r. Additionally, we design an edge detection block to extract the blood vessel details in the encoder features. The multi-directional Sobel operator [44, 45] is adopted to detect gradient changes between retinal blood vessel pixels and background pixels. The output of the edge detection block can be calculated by:

$$G = \sqrt{{G_0}^2 + {G_{45}}^2 + {G_{90}}^2 + {G_{135}}^2} \tag{7}$$

where $G_0 = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * I$, $G_{45} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & -2 \end{bmatrix} * I$,

$G_{90^\circ} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * I$, $G_{135^\circ} = \begin{bmatrix} 0 & -1 & -2 \\ 1 & 0 & -1 \\ 2 & 1 & 0 \end{bmatrix} * I$, I denotes the image feature, $*$ is the convolutional operator, G is the feature generated by the edge detection block. Rectified linear unit function (ReLU), 1×1 convolution, batch normalization, and sigmoid function are used to generate a spatial weight map that highlights the crucial area and local details. Therefore, the feature weight map can be calculated as follows:

$$Weight\ map = \sigma\left(\mathrm{BN}\left(\mathrm{conv}_{1\times1}\left(\mathrm{ReLU}(D)\right)\right)\right) + \sigma\left(\mathrm{BN}\left(\mathrm{conv}_{1\times1}\left(\mathrm{ReLU}(G)\right)\right)\right) \quad (8)$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\mathrm{BN}(\cdot)$ is the batch normalization, $\mathrm{conv}_{1\times1}(\cdot)$ is the 1×1 convolution. In addition to the local information among pixels in the channel of features, we combine the feature x_{encoder} with the generated weight map before feeding the result into the squeeze-excitation block to acquire the output feature containing the importance of the channel. The squeeze-excitation block is mainly composed of global average pooling and fully connected layers [46]. The diagram of the squeeze-excitation block is also illustrated in Fig. 3. Therefore, the output of DLA can be calculated as:

$$\mathrm{LA}(x_{\mathrm{encoder}}) = \mathrm{SE}(x_{\mathrm{encoder}} \otimes Weight\ map), \quad (9)$$

where $\mathrm{SE}(\cdot)$ denotes the squeeze-excitation block, $\otimes$ denotes the element-wise multiplication.

C. Deep-Shallow Hierarchical Feature Fusion

In deep networks, shallow features contain the low semantic feature of the image, such as edges and textures. In contrast, deep features include the high-level semantic information of the image including shape and spatial connection. Therefore, the hierarchical fusion based on the semantic complementarity of deep and shallow features is conducive to retaining information in the network, preventing the fine structures in blood vessels from being ignored, and thus improving the segmentation performance of vessels.

Inspired by the above ideas, we propose the deep-shallow hierarchical feature fusion algorithm, which includes deep feature fusion (DFF) and shallow feature fusion (SFF). From Fig.1, we extract four feature maps y_1, y_2, y_3, y_4 in decoder from the deep to shallow layers. Features y_1, y_2 are regarded as deep features and y_3, y_4 are defined by shallow features. From Fig.4, we stack upsampling and 3×3 convolution to preserve the high semantic features in the DFF. The dimensional addition method is adopted to fuse the deep features, which have been processed by upsampling and 3×3 convolution. The output of DFF can be calculated as follows:

$$out_{DFF} = \mathrm{Conv}_{3\times3}\left(\mathrm{upconv}(y_1) + \mathrm{upconv}(y_2)\right), \quad (10)$$

where $\mathrm{upconv}(\cdot)$ represents upsampling and convolution operation, $\mathrm{Conv}_{3\times3}$ denotes convolution operation with a convolution kernel size of 3×3, and out_{DFF} is the output of the deep feature fusion module.

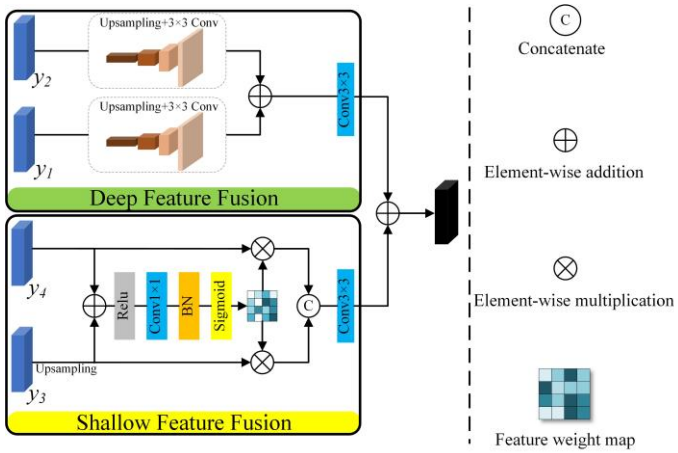


Fig. 4. The diagram of the deep-shallow hierarchical feature fusion.

Moreover, we propose the SFF to fuse shallow features. Similar to the DLA, the SFF generates a feature weight map to activate the detailed information in fundus images. From Fig.4, through upsampling and 1×1 convolution operation, the SFF adjusts the y_3, y_4 to the same size. Besides, the SFF gets the feature weight map by the operations outlined in Fig.4. Then the feature weight map multiply with y_3 , y_4 , respectively. Finally, the feature concatenation and 1×1 convolution is utilized to integrate shallow features. The output of SFF can be calculated as follows:

$$F = \sigma(\mathrm{BN}(\mathrm{conv}_{1\times1}(\mathrm{ReLU}(y_4 + \mathrm{upconv}(y_3))))), \quad (11)$$

$$out_{SFF} = \mathrm{Conv}_{3\times3}(\mathrm{Concatenate}(F \otimes y_4, F \otimes \mathrm{upconv}(y_3))), \quad (12)$$

where F represents feature weight map, $\mathrm{Concatenate}(\cdot)$ denotes the concatenate operation, and out_{SFF} is the output of the shallow feature fusion module.

Since the SFF avoids the use of continuous convolution and upsampling, the SFF can preserve structural details of vessels from losing during the feature fusion effectively. We fuse the output of DFF and SFF by element-wise addition to obtain the feature map, which aggregates the deep and shallow characterizations of the image. The output of dsHFF (out_{dsHFF}) can be calculated as follows:

$$out_{dsHFF} = out_{DFF} + out_{SFF}. \quad (13)$$

The segmentation result is obtained after processing the output of dsHFF by means of the convolution operation and sigmoid function.

D. Loss Function

The proposed GT-DLA-dsHFF has an end-to-end segmentation architecture. A loss function is supposed to design in the GT-DLA-dsHFF for measuring the distance between the label and predicted value. Since our target is a pixel-wise binary classification project, the cross-entropy loss function which calculates the difference between segmentation result $\hat{y}$ and the label y [15] is employed in this paper and is defined as:

$$\mathrm{CE}(\hat{y}, y) = -\frac{1}{N}\sum_{i=0}^{N}(y_i\ log\ \hat{y}_i + (1 - y_i)\ log(1 - \hat{y}_i)), \quad (14)$$

where N is the number of pixels in the retinal image.

III. Experiments and Results

A. Datasets

Our proposed GT-DLA-dsHFF is trained and tested on four retinal image datasets: DRIVE [47], STARE [48], CHASE_DB1 [49] and HRF datasets [50]. The key information of the four public datasets is summarized in Table I and four examples of four datasets are displayed in Fig. 5. The DRIVE dataset collects forty color retinal images with a 45° field of view (FOV), which has 584×565 pixels. Seven fundus images are collected from the diseased fundus, and the other images are from healthy subjects. In the DRIVE dataset, 20 images for training and the other 20 images are adopted in the test phase. Two labels completed by two experts are utilized for each image.

Additionally, to ensure our GT-DLA-dsHFF has satisfactory robustness on abnormal retinal images, the STARE dataset which contains some of the typical diseased retinal images is also adopted. The STARE dataset has 20 images with a size of 700×605, and each image is labeled by two observers. Since the STARE dataset lacks the training-test division, we use 4-fold cross-validation. Specifically, the 28 images in the STARE dataset are divided into four sets and each set contains seven fundus images. In the process of cross-validation, three sets were used as training sets to train a model, and the remaining one was regarded as a test set to validate the model performance. We repeated the above steps four times until all images were tested.

To further demonstrate that the GT-DLA-dsHFF is adapted for segmenting fundus images from low ages, CHASE_DB1 dataset which collects 28 color fundus images from 14 children is adopted. Each of the images has a size of 999×960. Each image has 30° FOV. We adopt the first 20 images of the CHASE_DB1 in the training stage while the other 8 images in the test stage. This method is also used in the dense dilated network [4]. Similar to the DRIVE dataset, each color fundus image is also manually annotated independently by two experts. We employ the first manual annotation as the label of vessels when we conduct experiments.

Retinal images acquired clinically have the characteristics of high resolution. To further demonstrate that our GT-DLA-dsHFF also has the ideal performance on clinical images, the HRF dataset is employed in our experiment. In the HRF, 15 images are from healthy subjects, 15 images are from patients with diabetic retinopathy, and 15 images are from patients with glaucoma. All images have the resolution of 3504×2336 pixels with a 60° FOV and each fundus image has a binary blood vessel annotation. In order to be consistent with other methods, we use the first five images of each set for training, and the remaining images are applied for evaluation of our GT-DLA-dsHFF in the experiment.

Since there are two manual annotations in the three datasets, in order to be consistent with other methods, we regard the work of the first expert as the ground truth and the second as the human observer. In recent studies on retinal blood vessel segmentation, due to the limitation of computing capacity, it is difficult for deep learning models to directly segment vessels in retinal images with high resolution. To address this issue, recent methods scaled images during the training stage and restored images to the original resolutions to compute performance metrics during the testing stage [4, 5, 19]. In this paper, we also adopt this strategy, which all the images in the four datasets are reshaped as 512×512, 512×512, 512×512, and 800×800 resolution in the training stage respectively. Similar to most recent methods, after obtaining preliminary segmentation results, we upsampled the obtained segmentation results to the same size as the original image in order to accurately calculate relevant performance indicators [4, 5, 19].

TABLE I
Summarization of the Four Public Datasets

Dataset	Subject	Training-test	Resolution	Reshape size
DRIVE	40	20-20	584×565	512×512
STARE	20	15-5	700×605	512×512
CHASE_DB1	28	20-8	999×960	512×512
HRF	45	15-30	3504×2336	800×800

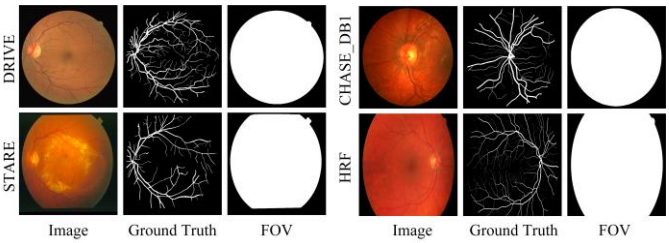


Fig. 5 Partial original images, labels and FOVs in the four datasets.

B. Implementation Details and Evaluation Metrics

The proposed GT-DLA-dsHFF is implemented on the Pytorch platform through NVIDIA Tesla V-100 GPU. When training our model, we employ the Adam optimizer for our GT-DLA-dsHFF. In order to update the learning rate adaptively, we utilize the cosine annealing strategy [51], which is summarized in **Algorithm 1**. The learning rate is 1×10^{-2} and the maximum epoch is initiated as 1000. The batch size of the training set is 2. The weight decay is 0.001. Common augmentation methods, such as random rotation, horizontal flip, and color enhancement, are utilized for image preprocessing.

To calculate the performance indicator of our GT-DLA-dsHFF, the pixels in the binary results obtained by a threshold can be divided into four categories: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). We evaluate the proposed GT-DLA-dsHFF by accuracy (Acc), sensitivity (Se), specificity (Sp), the area under the receiver operating characteristic (ROC) curve (AUC), and F1-Score [52, 53]. In addition, a large fraction of the image is the background pixels while the vessel pixels occupy a very small proportion. Thus, the vessel segmentation is a category imbalance problem. Matthews Correlation Coefficient (MCC) comprehensively

TABLE II
COMPARISON OF SEGMENTATION PERFORMANCE ON THE DRIVE AND STARE DATASETS

Method	Year	DRIVE					STARE				
		Acc	Se	Sp	AUC	*p*-value	Acc	Se	Sp	AUC	*p*-value
U-Net	2015	0.9634	0.7800	0.9810	0.9768	<0.05	0.9712	0.8167	0.9838	0.9857	<0.05
AttU-Net	2018	0.9680	0.8163	0.9826	0.9842	<0.05	0.9717	0.8153	0.9844	0.9862	<0.05
CE-Net	2018	0.9672	0.8090	0.9824	0.9835	<0.05	0.9727	0.8183	0.9861	0.9873	<0.05
CS-Net	2018	0.9692	0.8351	0.9820	0.9862	<0.05	0.9730	0.8325	0.9855	0.9877	<0.05
Iter-Net	2019	0.9687	0.8177	**0.9832**	0.9839	<0.05	0.9729	0.8287	0.9846	0.9874	<0.05
U-Net++	2020	0.9686	0.8256	0.9823	0.9854	<0.05	0.9733	0.8264	0.9851	0.9883	<0.05
Ours	**2021**	**0.9703**	**0.8355**	0.9827	**0.9863**	-	**0.9760**	**0.8480**	**0.9864**	**0.9905**	-

where the bold fonts denote the best results.

TABLE III
COMPARISON OF SEGMENTATION PERFORMANCE ON THE CHASE_DB1 AND HRF DATASETS

Method	Year	CHASE_DB1					HRF				
		Acc	Se	Sp	AUC	*p*-value	Acc	Se	Sp	AUC	*p*-value
U-Net	2015	0.9676	0.7713	0.9808	0.9783	<0.05	0.9663	0.7738	0.9827	0.9779	<0.05
AttU-Net	2018	0.9741	0.8225	0.9843	0.9847	<0.05	0.9656	0.7713	0.9822	0.9770	<0.05
CE-Net	2018	0.9743	0.8278	0.9841	0.9859	<0.05	0.9676	0.7933	0.9825	0.9808	<0.05
CS-Net	2018	0.9748	0.8196	0.9852	0.9852	<0.05	0.9667	0.7877	0.9819	0.9795	<0.05
Iter-Net	2019	0.9752	0.8303	0.9850	0.9861	<0.05	0.9690	0.8075	0.9828	0.9827	<0.05
U-Net++	2020	0.9753	0.8317	0.9850	0.9861	<0.05	0.9684	0.7950	**0.9832**	0.9811	<0.05
Ours	**2021**	**0.9760**	**0.8440**	**0.9858**	**0.9892**	-	**0.9698**	**0.8178**	0.9828	**0.9853**	-

where the bold fonts denote the best results.

Algorithm 1: The learning rate and parameter update in the proposed GT-DLA-dsHFF.

Input: Image data x, ground truth y, the maximum epoch τ, learning rate ξ, batch size k, the GT-DLA-dsHFF Net$(\cdot)$;
Output: The segmentation result O;

Initializing parameters in the GT-DLA-dsHFF as $\theta^{(0)}$;
Initializing data D_i in one batch with $i = 1,2\cdots,k$;
Initializing $\tau = 1000$, $\xi_{max} = 1\times10^{-2}$, $\xi_{min} = 1\times10^{-8}$, $\xi^{(0)} = \xi_{max}$, $q = 0, j = 0, \delta = 50$;
while $q \neq \tau$ **do**
  j++;
  for $i = 1$ to k **do**
    Generate the prediction result $\hat{y}_i = \text{Net}(\theta^{(q)}, D_i)$;
    Calculate the loss $J = \text{CE}(\hat{y}_i, y_i)$ by Eq. (10);
    Calculate the gradient $g = \nabla J$;
    Update the parameters: $\theta^{(q+1)} \leftarrow \theta^{(q)} - \xi^{(q)} \times g$;
  end for
  if $j \leq \delta$ **then**
    Update the learning rate:
    $\xi^{(q+1)} \leftarrow \xi_{min} - \frac{1}{2}(\xi_{max} - \xi_{min})\left(1 + \cos\frac{j\pi}{\delta}\right)$;
  else
    $\xi^{(q+1)} \leftarrow \xi_{max}$;
    $j = 0$;
  q++;
end
Get the segmentation result $O = \text{Net}(\theta^{(\tau)}, x)$.

considers the imbalance categories, which is commonly used to evaluate algorithms with unbalanced data. MCC essentially describes the correlation intensity between the segmentation results and golden standards.

$$\text{Acc} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}, \tag{15}$$

$$\text{Se} = \frac{\text{TP}}{\text{TP+FN}}, \tag{16}$$

$$\text{Sp} = \frac{\text{TN}}{\text{TN+FP}}, \tag{17}$$

$$\text{F1} = \frac{2\text{TP}}{2\text{TP+FP+FN}}, \tag{18}$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP+FP})\times(\text{TP+FN})\times(\text{TN+FP})\times(\text{TN+FN})}}. \tag{19}$$

C. Statistical Performance

To evaluate the efficacy of the proposed GT-DLA-dsHFF framework, we perform experiments on the four retinal image datasets. Table II and Table III list the performance indicators on the four datasets. We list the current methods in retinal vessel segmentation and compare our GT-DLA-dsHFF with them in Table II and Table III. It is worth pointing out that the experiment's environment of the U-Net [17], AttU-Net [36], CE-Net [39], CS-Net [38], Iter-Net [27], and U-Net++ [26] is the same as our proposed method. We faithfully reproduced these CNN-based deep learning methods. Specifically, for the

DRIVE dataset, our GT-DLA-dsHFF obtains 0.9703, 0.8355, 0.9827, and 0.9863 for the four performance indicators, respectively. Compared with other methods, our GT-DLA-dsHFF has the highest Acc, Se, and AUC. In terms of Sp, despite Iter-Net achieving the highest value (0.9832), the Se is lower (1.78%) than our GT-DLA-dsHFF, and the other metrics are inferior to the proposed GT-DLA-dsHFF. The obvious improvement on Se shows that our GT-DLA-dsHFF can identify more blood vessels compared with other methods.

In terms of the STARE and CHASE_DB1, our GT-DLA-dsHFF gets the highest indicator against the other methods. In addition, compared with the newly proposed U-Net++, Acc, Se, Sp, and AUC of GT-DLA-dsHFF increased by 0.27%, 2.16%, 0.13%, and 0.22% on STARE, respectively. As far as CHASE_DB1 dataset, the four metrics increased by 0.07%, 1.23%, 0.08%, and 0.31% compared with U-Net++.

For the HRF, our GT-DLA-dsHFF gets 0.9698, 0.8178, and 0.9853 on Acc, Se, and AUC, which are better than all of the methods in Table III. Although the Sp results on HRF are slightly inferior to others, as far as the comprehensive performance of all the evaluation metrics, our GT-DLA-dsHFF is generally outperformed all the reproduced methods on the HRF datasets.

Besides the comparison of commonly used performance indicators such ad Acc, Se, Sp, and AUC, we also compared the MCC and F1 of the above approaches on the above datasets. From Fig.6, both the MCC and F1 of the proposed GT-DLA-dsHFF are significantly better than other methods, which indicates that the proposed GT-DLA-dsHFF has satisfactory performance in dealing with the problem of category imbalance. Furthermore, we perform the Wilcoxon rank-sum test on the Acc obtained from the experiment, where the results are also displayed in Table II and Table III. All the *p*-values do not exceed 0.05, which demonstrates the improvement of the segmentation performance compared with recent methods of our GT-DLA-dsHFF is statistically significant.

D. Visual Performance

In terms of the visual performance comparison, it can be observed from Fig. 7, the baseline approach cannot recognize the fine vessels effectively. Moreover, it can be seen from the detailed illustration of blood vessels in Fig.7 that the edge of the vessel in the results of U-Net is seriously missing. The AttU-Net captures the local weight map through the gating mechanism and obtains better vessel probability in the attention map than U-Net. However, the network structure of the AttU-Net is too simple to obtain the local and global information, and the overall performance is still unsatisfactory. Although the CE-Net, CS-Net, and Iter-Net are significantly better than the above-mentioned two schemes in the fine blood vessel segmentation, the inferior feature fusion method makes it impossible to effectively integrate the information of deep and shallow features in the network, resulting in the inefficient alleviation of external noises. The U-Net++ has poor segmentation performance at the intersection of blood vessels, and our proposed GT-DLA-dsHFF with edge detection can effectively avoid this limitation. The ROC curve, which is

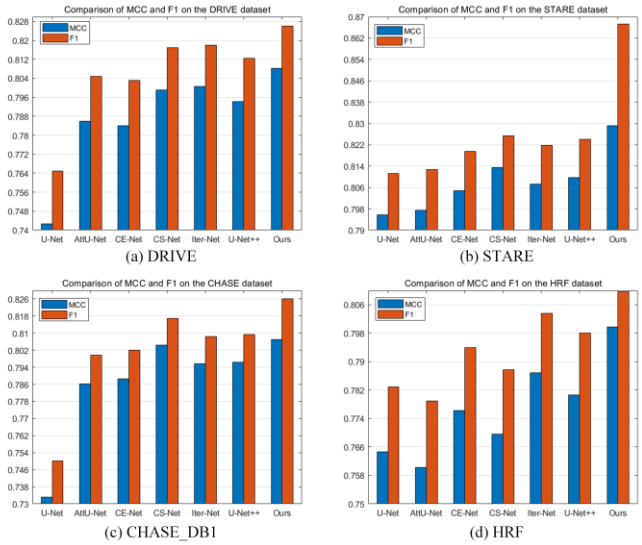


Fig. 6. Comparison of MCC and F1 on the four datasets.

widely used for performance comparison, is also employed in the comparison of visual performance in this paper. As shown in Fig. 8, the key regions of the ROC curve of the proposed GT-DLA-dsHFF surround that of the remaining six schemes, which indicates that the performance of our GT-DLA-dsHFF is pixel-level classification is better than the other methods. Due to the low contrast in some regions of fundus images, some fine vessels are difficult to be directly observed by doctors, resulting in the omission of vessels. From Fig. 9, vessels in the low contrast regions in the fundus image, which are unlabeled in the ground truth by the clinical experts, are also detected by our GT-DLA-dsHFF. Thus, it can be inferred that the proposed GT-DLA-dsHFF can assist physicians in clinical analysis, avoiding fine vessels that from ignored by physicians.

IV. Discussions

A. Ablation Study

To demonstrate the efficacy of the proposed GT, DLA, and dsHFF approaches, we perform ablation studies on DRIVE. In this section, the U-Net is utilized as the original method to explore the contribution of the three proposed modules. The visual comparison of the segmentation results generated by the U-Net, U-Net+GT, U-Net+GT+DLA, and U-Net+GT+DLA+dsHFF are displayed in Fig. 10. The statistical indicators are displayed in Table IV. For comprehensive comparisons in different perspectives, we calculate six evaluation metrics (Acc, Se, Sp, AUC, F1, and MCC) of all the models.

1) *Efficacy of the Global Transformer*

To illustrate the efficiency of the proposed GT, the GT is combined with the baseline method to observe the improvement. From Table IV, compared with the baseline, the U-Net with GT outperforms all six metrics on the DRIVE dataset. The Acc is improved from 0.9641 to 0.9661, and the AUC increased by 0.27%. The improvement of Se is 0.72%, indicating that the global information by the GT can significantly improve the sensitivity of the algorithm on fine

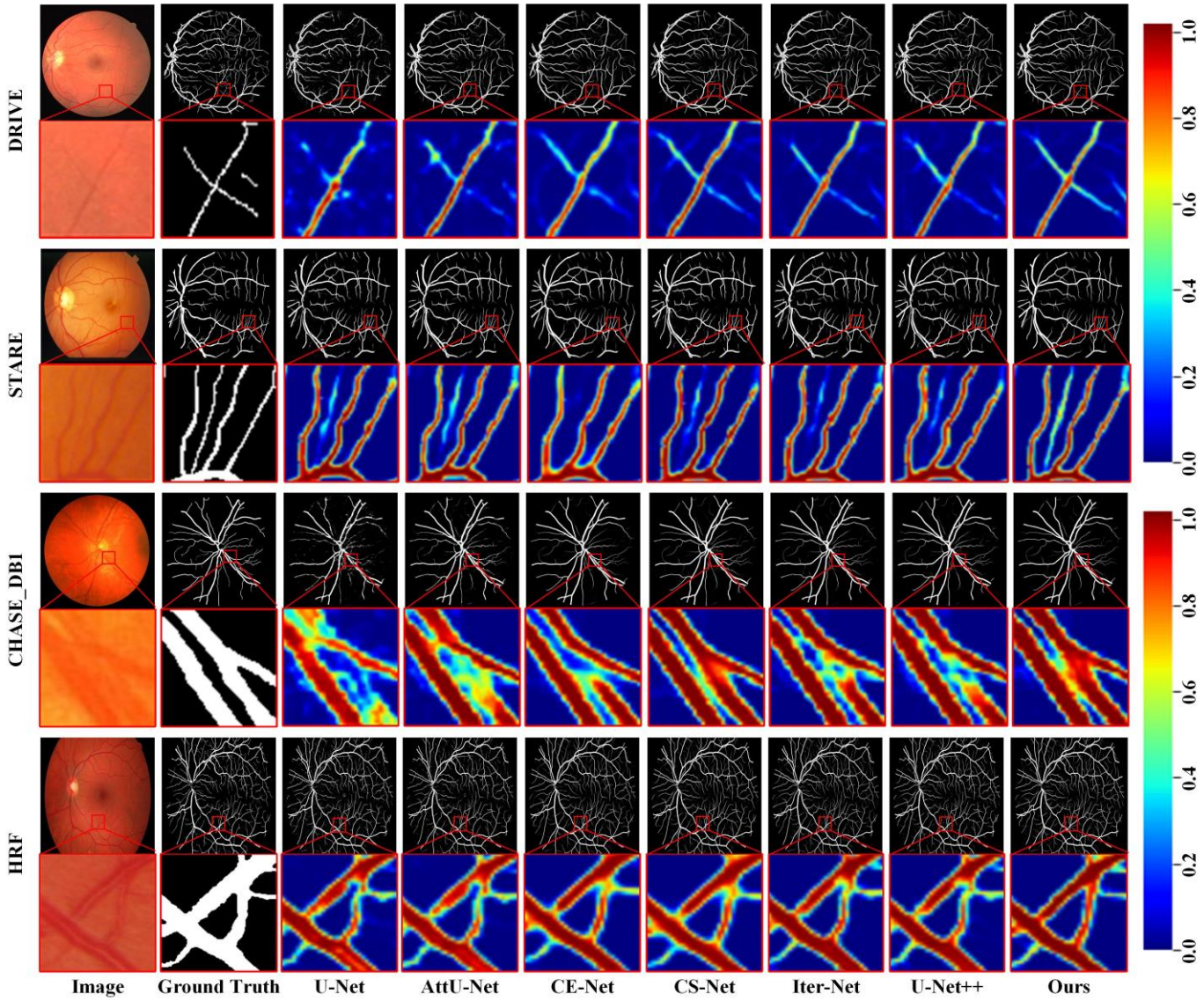
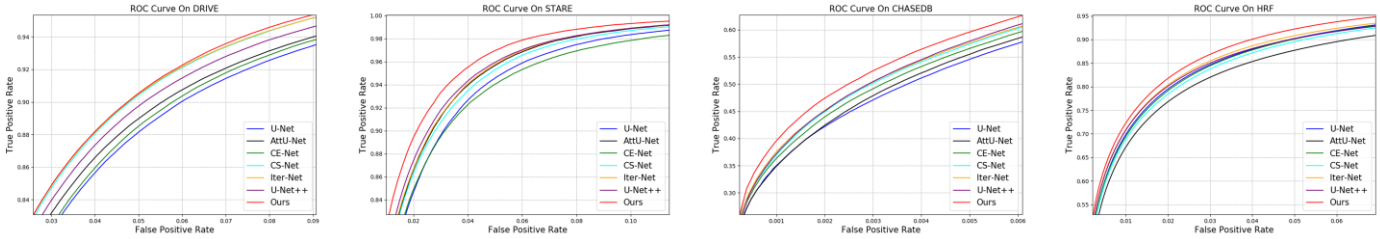


Fig. 7. The visualization comparison of segmentation results on the four datasets.



(a) DRIVE (b) STARE (c) CHASE_DB1 (d) HRF

Fig. 8. The vital region of ROC curves on the four datasets.

blood vessels. In addition, the Sp, F1, and MCC are increased by 0.26%, 0.36%, and 1.39% respectively. In Fig.10, compared with the baseline method, the architecture with the proposed GT connects the breakpoints of fine vessels and alleviates the influence of the external noises, demonstrating the effective integration of global information is beneficial to fine vessel segmentation.

2) *Efficacy of the Dual Local Attention*

Compared with other methods, the network combined with DLA not only obtains local information as much as possible through multiple receptive fields but also highlights the image channel with possesses abundant blood vessel information through the unsupervised edge detection scheme and squeeze excitation module. In order to prove that the DLA has a positive effect on vessel segmentation, we add the DLA based on the above experiment. From Table IV, the DLA improves Acc

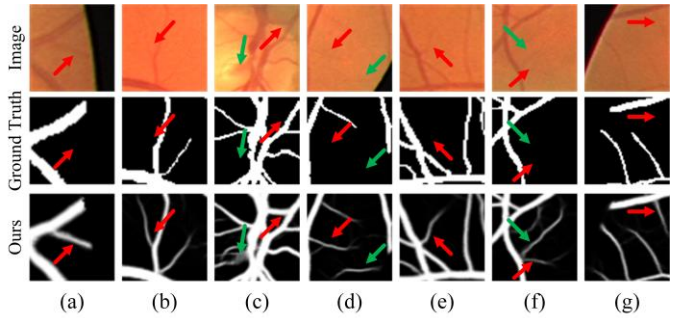


Fig. 9. Segmentation results for the vessels in low contrast regions.

(0.9661 to 0.9682) and Se (0.8139 to 0.8224) under the premise of sacrificing 0.01% a small amount of Sp, which is acceptable. In Fig. 10, the probability of the vessel pixels is enhanced, so the Se increases. The other metrics such as AUC, F1, and MCC are roughly unchanged.

3) *Efficacy of the Deep-Shallow Hierarchical Feature Fusion*

The dsHFF integrates features at different levels in the network, which fully considers the differences between deep and shallow features in the deep learning network, maintaining its unique characteristics while alleviating the attenuation of valid information in the process of feature fusion. We add the dsHFF to the above two experiments to evaluate the effect of the dsHFF. From Table IV, after embedding the dsHFF in our model, all of the evaluation metrics on DRIVE datasets have reached the best results. Since shallow feature fusion preserves most of the details of the image, our algorithm can identify more small categories of elements, such as vessel pixels, and thus the SE increases by 1.31%. Deep feature fusion can identify larger categorical elements, such as background pixels, and improves the SP from 0.9821 to 0.9827. Since the Se and Sp are improved, the Acc and AUC, which measure the overall segmentation performance, have also improved by 0.21% and 0.13% respectively. Moreover, the F1 and MCC are higher than other methods, which manifests that the dsHFF can retard the loss of valid information under the condition of data imbalance.

TABLE IV
STATISTICAL RESULTS OF ABLATION STUDIES ON THE DRIVE DATASET

Method	ACC	SE	SP	AUC	F1	MCC
U-Net	0.9641	0.8067	0.9796	0.9813	0.8108	0.7827
U-Net+GT	0.9661	0.8139	0.9822	0.9840	0.8144	0.7966
U-Net+GT+DLA	0.9682	0.8224	0.9821	0.9850	0.8189	0.8015
Ours	**0.9703**	**0.8355**	**0.9827**	**0.9863**	**0.8257**	**0.8104**

where the bold fonts denote the best results.

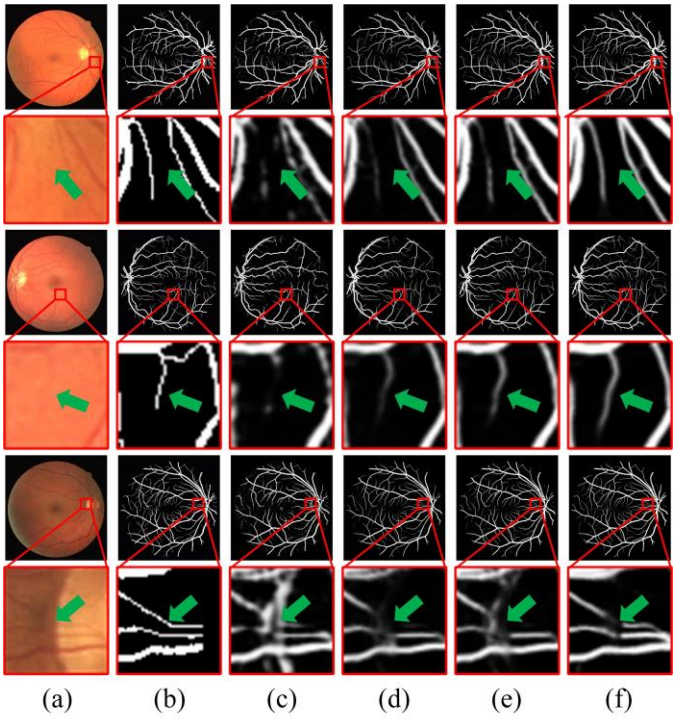


Fig. 10. Visualization of the segmentation results in ablation study of the proposed GT-DLA-dsHFF on the DRIVE dataset. (a) Original retinal image. (b) Ground truth. (c) U-Net. (d) U-Net+GT. (e) U-Net+GT+DLA. (f) U-Net+GT+DLA+dsHFF.

B. *Comparison with the State-of-the-Art Methods*

To demonstrate the superiority of our GT-DLA-dsHFF on retinal vessel segmentation, we compared the GT-DLA-dsHFF with the mainstream methods which were used in the past three years on four datasets Table V and Table VI. The methods which are utilized for comparison include Joint-Loss [22], BTS-DSN [23], CC-Net [24], DA-Net[37], DEU-Net [15], DU-Net [3], NFN+ [18], DDNet [4], STD-Net [25], RVSeg-Net [19], HA-Net [20], LA-Net [14], GLUE [21], and SCS-Net [5]. The performance comparison of the three datasets DRIVE, STARE, and CHASE_DB1 are listed in Table V. Since there are fewer methods that employ the HRF dataset, we list them separately in Table VI. On the DRIVE and STARE datasets, the proposed GT-DLA-dsHFF achieved the best performance on Acc, Se, and AUC. Although GLUE obtained the highest Sp (0.9861 on DRIVE, 0.9916 on STARE), the Acc and Se are significantly lower than the GT-DLA-dsHFF. It is worth mentioning that STD-Net adopts the approach of texture enhancement to achieve comparable AUC performance with our GT-DLA-dsHFF on DRIVE dataset. However, the STD-Net utilizes continuous convolution and sampling when acquiring texture information, resulting in the incomplete acquisition of blood vessel information and the Se score is only 0.8151, which is 2.04% lower than the GT-DLA-dsHFF proposed in the paper. As far as the CHASE_DB1 dataset, the GT-DLA-dsHFF achieves the best Acc and Se scores. Although NFN+ cascades two backbone networks to obtain multi-scale information, obtaining the optimal Sp and AUC scores on the CHASE_DB1, the improvement of these two indicators compared with GT-DLA-dsHFF is not significant (0.22% and 0.02% respectively). On the contrary, our GT-DLA-dsHFF is 0.72% and 3.37% higher than NFN+ on Acc and Se, which is acceptable for the minor sacrifice of Sp and AUC. On the HRF dataset, the Acc, Se, and AUC of GT-DLA-dsHFF are higher than all approaches used for comparison, which demonstrates that the proposed GT-DLA-dsHFF in this paper outperforms

TABLE V
PERFORMANCE COMPARISON ON THE DRIVE, STARE, AND CHAE_DB1 DATASETS

Method	Year	DRIVE				STARE				CHASE_DB1			
		Acc	Se	Sp	AUC	Acc	Se	Sp	AUC	Acc	Se	Sp	AUC
Joint-Loss	2018	0.9542	0.7653	0.9818	0.9752	0.9612	0.7581	0.9846	0.9801	0.9610	0.7633	0.9809	0.9781
BTS-DSN	2018	0.9551	0.7800	0.9806	0.9796	0.9660	0.8201	0.9828	0.9872	0.9627	0.7888	0.9801	0.9840
CC-Net	2018	0.9528	0.7626	0.9809	0.9678	0.9633	0.7709	0.9848	0.9700	-	-	-	-
DA-Net	2019	0.9615	0.8075	0.9841	0.9808	0.9679	0.7705	0.9873	0.9781	-	-	-	-
DEU-Net	2019	0.9567	0.7940	0.9816	0.9772	-	-	-	-	0.9661	0.8074	0.9821	0.9812
DU-Net	2019	0.9566	0.7963	0.9800	0.9802	0.9641	0.7595	0.9878	0.9832	0.9610	0.8155	0.9752	0.9804
NFN+	2020	0.9582	0.7996	0.9813	0.9830	0.9672	0.7963	0.9863	0.9875	0.9688	0.8003	**0.9880**	**0.9894**
DDNet	2020	0.9607	0.8132	0.9783	-	0.9685	0.8391	0.9769	0.9858	0.9648	0.8275	0.9768	-
STD-Net	2020	0.9695	0.8151	0.9846	**0.9863**	-	-	-	-	-	-	-	-
RVSeg-Net	2020	0.9681	0.8107	0.9845	0.9817	-	-	-	-	0.9726	0.8069	0.9848	0.9833
HA-Net	2020	0.9581	0.7991	0.9813	0.9823	0.9673	0.8186	0.9844	0.9881	0.9670	0.8239	0.9813	0.9871
LA-Net	2021	0.9568	0.7921	0.9810	0.9806	0.9678	0.8352	0.9823	0.9875	0.9635	0.7818	0.9819	0.9810
GLUE	2021	0.9692	0.8278	**0.9861**	-	0.9740	0.8342	**0.9916**	-	-	-	-	-
SCS-Net	2021	0.9697	0.8289	0.9838	0.9837	0.9736	0.8207	0.9839	0.9844	0.9744	0.8365	0.9839	0.9867
Ours	**2021**	**0.9703**	**0.8355**	0.9827	**0.9863**	**0.9760**	**0.8480**	0.9864	**0.9905**	**0.9760**	**0.8440**	0.9858	0.9892

where the bold fonts denote the best results.

TABLE VI
PERFORMANCE COMPARISON ON THE HRF DATASET

Method	Year	Acc	Se	Sp	AUC
Joint-Loss	2018	0.9437	0.8084	0.9417	-
DU-Net	2019	0.9651	0.7464	**0.9874**	0.9831
HA-Net	2020	0.9654	0.7803	0.9843	0.9837
SCS-Net	2021	0.9687	0.8114	0.9823	0.9842
Ours	**2021**	**0.9698**	**0.8178**	0.9828	**0.9853**

where the bold fonts denote the best results.

the existing methods in the blood vessel segmentation task in retinal images with large resolution. The DU-Net adopts deformable convolution to capture the curved feature of blood vessels, thereby achieving the highest Sp score (0.9874). Nevertheless, the DU-Net is 7.14% lower than our GT-DLA-dsHFF on Se, indicating that our GT-DLA-dsHFF has more ideal segmentation performance than DU-Net in the classification of blood vessel pixels. In general, after comprehensively comparing the four indicators on the four datasets, the conclusion of this section is that our GT-DLA-dsHFF achieves the best performance against the methods proposed in the last three years.

C. Segmentation on Diseased Images

Clinically, fundus diseases cause structural changes in the retinal, which makes the fine vessels are harder to segment. STARE dataset is a challenging blood vessel segmentation dataset because it contains several retinal images of the diseased fundus. There are 10 abnormal images in STARE. In addition, the DRIVE dataset also contains seven unhealthy fundus images. In order to verify our GT-DLA-dsHFF adapts to the diseased fundus images, we test the proposed GT-DLA-dsHFF on abnormal images in DRIVE and STARE. The experimental results are displayed in Table VII and Fig. 11. Among all the methods adapted for comparison, the six indicators of the GT-DLA-dsHFF have reached the best on the two datasets (DRIVE: Acc: 0.9684; Se: 0.8259; Sp: 0.9821; AUC: 0.9861; F1; 0.8208; MCC: 0.0835. STARE: Acc: 0.9724; Se: 0.8328; Sp: 0.9835; AUC: 0.9876; F1: 0.8023; MCC; 0.8171.). Compared with the evaluation metrics obtained on all images in the test sets of the

TABLE VII
PERFORMANCE COMPARISON OF DISEASED IMAGES IN TWO DATASETS.

Dataset	Method	Acc	Se	Sp	AUC	F1	MCC
DRIVE	U-Net	0.9573	0.7546	0.9790	0.9731	0.7649	0.7427
	AttU-Net	0.9656	0.8096	0.9806	0.9828	0.8049	0.7860
	CE-Net	0.9650	0.8149	0.9794	0.9833	0.8033	0.7842
	CS-Net	0.9675	0.8261	0.9811	0.9857	0.8169	0.7991
	Iter-Net	0.9679	0.8234	0.9818	0.9860	0.8182	0.8006
	U-Net++	0.9671	0.8130	0.9819	0.9845	0.8124	0.7944
	Ours	**0.9684**	**0.8259**	**0.9821**	**0.9861**	**0.8208**	**0.8035**
STARE	U-Net	0.9631	0.7783	0.9779	0.9776	0.7576	0.7380
	AttU-Net	0.9637	0.7654	0.9796	0.9787	0.7579	0.7384
	CE-Net	0.9689	0.7945	0.9829	0.9840	0.7912	0.7744
	CS-Net	0.9695	0.8107	0.9822	0.9832	0.7977	0.7813
	Iter-Net	0.9665	0.8087	0.9791	0.9825	0.7816	0.7639
	U-Net++	0.9690	0.8093	0.9818	0.9852	0.7947	0.7781
	Ours	**0.9724**	**0.8328**	**0.9835**	**0.9876**	**0.8023**	**0.8171**

where the bold fonts denote the best results.

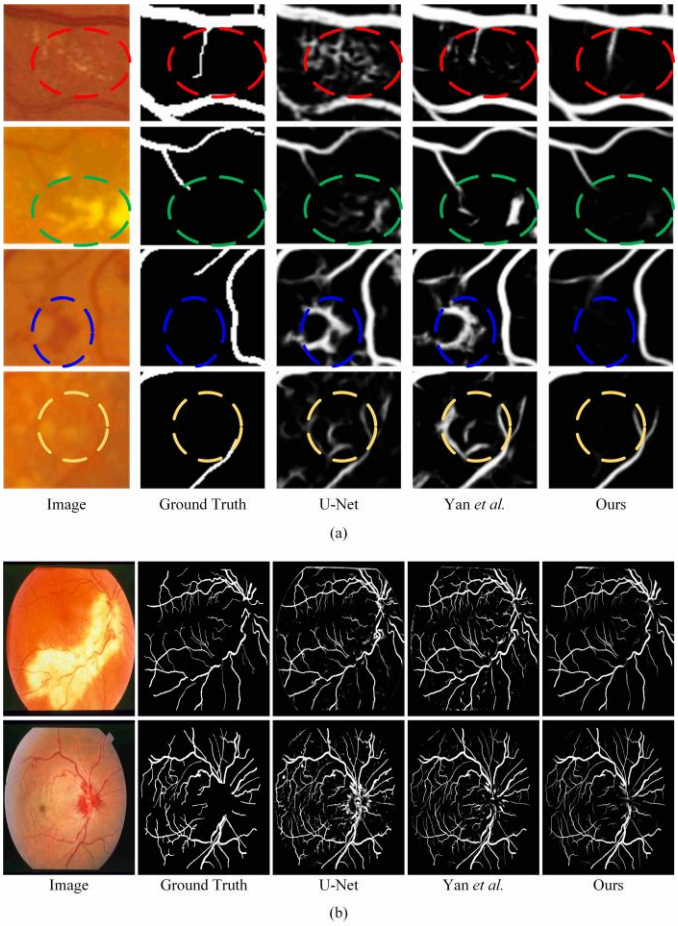


Fig. 11. Visualization of the vessel segmentation of diseased fundus on the DRIVE and STARE. (a) regional lesions. (b) extensive lesions.

above two datasets, the metrics achieved by the proposed GT-DLA-dsHFF on diseased retinal images have little fluctuation. Taking Acc as an example, when the GT-DLA-dsHFF focuses on the diseased retinal image, the Acc index of GT-DLA-dsHFF in the DRIVE dataset is only reduced by 0.19%, and the Acc index in the STARE dataset is only reduced by 0.36%, which is an acceptable fluctuation range, indicating that the proposed GT-DLA-dsHFF has the satisfactory anti-interference ability. The major reason why the decrease of the indicators is that the diseased area affects the detection of tiny vessels, which is mistaken for the background. As the network deepens, because of the pooling and convolution operations, the proportion occupied by the diseased area in the fundus image will shrink. The existence of dsHFF enables the GT-DLA-dsHFF to minimize the influence of the lesion area. From Fig. 11, compared with the other methods, our GT-DLA-dsHFF minimizes the impact of lesions on blood vessel segmentation and preserves the continuation of vessels in the diseased area, which demonstrates that our GT-DLA-dsHFF has satisfied performance on the diseased fundus vessel segmentation task, indicating the potential clinical application prospect of our GT-DLA-dsHFF.

V. Limitations and Future Work

Although the proposed method is highly competitive with current methods in terms of the segmentation performance, it still has some limitations which need to be focused on in future work. The first issue is the limited number of images used for model training and testing, which is a drawback that commonly exists in most retinal segmentation methods. Even we applied data enhancement and other ways to expand the number of images in the original datasets in the data preprocessing stage, the generalization ability of the deep learning model is still restricted because of the limited number of the images. In the future, we will aim to expand the number of fundus images for model validation and use multiple data types including ultra-wide fundus photograph, optical coherence tomography (OCT), and so on. Secondly, the long acquisition time and the high-quality requirement of the medical images result in the limited amount of the available data. In recent years, the transfer learning has made it possible for deep learning models to learn features from medical images from natural images. Therefore, in the future, we will also focus on adopting the transfer learning in different retinal vessel datasets and the transfer learning mode from natural image to retinal image.

VI. Conclusion

In this paper, we propose a novel global transformer and dual local attention network via deep-shallow hierarchical feature fusion (GT-DLA-dsHFF) for retinal vessel segmentation. Our GT-DLA-dsHFF involves three innovative modules: the global transformer (GT), the dual local attention (DLA), and the deep-shallow hierarchical feature fusion (dsHFF). The GT which adopts the transformer architecture captures the global information within the encoder, solving the problem that low efficiency in extracting long-distance dependence between pixels, strengthening the connection of the breakpoints of fine vessels, and obtaining more continuous segmentation results. The DLA employs the dilated convolution, edge detection method, and squeeze-excitation block to extract local information in different scales, resolving the problem of the single receptive field in current methods and preserving the edge characterization of vessels. In terms of the difference between deep and shallow features, the dsHFF is developed to perform a hierarchical fusion of deep-shallow features, which maximizes the retention of valid information. We perform experiments on four public retinal image datasets to evaluate our GT-DLA-dsHFF and reproduce several deep-learning-based methods in the same experimental environment. Experimental results demonstrate that our GT-DLA-dsHFF can achieve better performance against the reproduced methods. The efficacy of the GT, DLA, and dsHFF are carefully verified in this article. Detail discussions illustrate the segmentation performance of our proposed GT-DLA-dsHFF is superior to the current methods and prove that the proposed GT-DLA-dsHFF is adapted to the segmentation of diseased fundus images, indicating the potential clinical application prospect of our GT-DLA-dsHFF.

References

[1] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Rev. Biomed. Eng.,* vol. 3, pp. 169-208 Dec. 2010.
[2] B. Sheng *et al.*, "Retinal Vessel Segmentation Using Minimum Spanning

Superpixel Tree Detector," *IEEE Transactions on Cybernetics,* vol. 49, no. 7, pp. 2707-2719, Jul. 2019.
[3] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "DUNet: A deformable network for retinal vessel segmentation," *Knowledge-Based Syst.,* vol. 178, pp. 149-162, Aug. 2019.
[4] L. Mou, L. Chen, J. Cheng, Z. W. Gu, Y. T. Zhao, and J. Liu, "Dense Dilated Network With Probability Regularized Walk for Vessel Detection," *IEEE Trans. Med. Imaging,* vol. 39, no. 5, pp. 1392-1403, May 2020.
[5] H. Wu, W. Wang, J. Zhong, B. Lei, Z. Wen, and J. Qin, "SCS-Net: A Scale and Context Sensitive Network for Retinal Vessel Segmentation," *Med. Image Anal.,* vol. 70, p. 102025, May 2021.
[6] S. Y. Shin, S. Lee, I. D. Yun, and K. M. Lee, "Deep vessel segmentation by learning graphical connectivity," *Med. Image Anal.,* vol. 58, p. 101556, Dec. 2019.
[7] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, "Detection of blood vessels in retinal images using two-dimensional matched filters," *IEEE Trans. Med. Imaging,* vol. 8, no. 3, pp. 263-269, Sept. 1989.
[8] A. Can, H. Shen, J. N. Turner, H. L. Tanenbaum, and B. Roysam, "Rapid automated tracing and feature extraction from retinal fundus images using direct exploratory algorithms," *IEEE Trans. Inf. Technol. Biomed.,* vol. 3, no. 2, pp. 125-138 Jun. 1999.
[9] A. M. Mendonca and A. Campilho, "Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction," *IEEE Trans. Med. Imaging,* vol. 25, no. 9, pp. 1200-1213 Sept. 2006.
[10] Y. A. Tolias and S. M. Panas, "A fuzzy vessel tracking algorithm for retinal images based on fuzzy clustering," *IEEE Trans. Med. Imaging,* vol. 17, no. 2, pp. 263-273 Apr. 1998.
[11] A. Salazar-Gonzalez, D. Kaba, Y. Li, and X. Liu, "Segmentation of the blood vessels and optic disk in retinal images," *IEEE J. Biomed. Health Inform.,* vol. 18, no. 6, pp. 1874-1886 Jan. 2014.
[12] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, "Scene Segmentation With Dual Relation-Aware Attention Network," *IEEE Trans. Neural Netw. Learn. Syst.,* pp. 1-14, Aug. 2020.
[13] Y. Li, H. Yang, B. Lei, J. Liu, and C. Y. Wee, "Novel Effective Connectivity Inference Using Ultra-Group Constrained Orthogonal Forward Regression and Elastic Multilayer Perceptron Classifier for MCI Identification," *IEEE Trans. Med. Imaging,* vol. 38, no. 5, pp. 1227-1239, May 2019.
[14] X. Li, Y. Jiang, M. Li, and S. Yin, "Lightweight Attention Convolutional Neural Network for Retinal Vessel Image Segmentation," *IEEE Trans. Ind. Inform.,* vol. 17, no. 3, pp. 1958-1967, Mar. 2021.
[15] B. Wang, S. Qiu, and H. He, "Dual Encoding U-Net for Retinal Vessel Segmentation," in *Medical Image Computing Computer Assisted Intervention*, 2019, pp. 84-92.
[16] Y. Li, Y. Zhang, W. Cui, B. Lei, X. Kuang, and T. Zhang, "Dual Encoder-based Dynamic-Channel Graph Convolutional Network with Edge Enhancement for Retinal Vessel Segmentation," *IEEE Trans. Med. Imaging,* early access, Feb. 2022, doi: 10.1109/TMI.2022.3151666.
[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing Computer Assisted Intervention*, 2015, pp. 234-241.
[18] Y. Wu, Y. Xia, Y. Song, Y. Zhang, and W. Cai, "NFN + : A novel network followed network for retinal vessel segmentation," *Neural Netw.,* vol. 126, pp. 153-162, Jun. 2020.
[19] W. Wang, J. Zhong, H. Wu, Z. Wen, and J. Qin, "RVSeg-Net: An Efficient Feature Pyramid Cascade Network for Retinal Vessel Segmentation," in *Medical Image Computing Computer Assisted Intervention*, 2020, pp. 796-805.
[20] D. Wang, A. Haytham, J. Pottenburgh, O. Saeedi, and Y. Tao, "Hard Attention Net for Automatic Retinal Vessel Segmentation," *IEEE J. Biomed. Health Inform.,* vol. 24, no. 12, pp. 3384-3396, Dec. 2020.
[21] S. Lian, L. Li, G. Lian, X. Xiao, Z. Luo, and S. Li, "A Global and Local Enhanced Residual U-Net for Accurate Retinal Vessel Segmentation," *IEEE/ACM Trans. Comput. Biol. Bioinform.,* vol. 18, no. 3, pp. 852-862, Jun. 2021.
[22] Z. Yan, X. Yang, and K.-T. Cheng, "Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation," *IEEE Trans. Biomed. Eng.,* vol. 65, no. 9, pp. 1912-1923 Apr. 2018.
[23] S. Guo, K. Wang, H. Kang, Y. Zhang, Y. Gao, and T. Li, "BTS-DSN: Deeply supervised neural network with short connections for retinal vessel segmentation," *Int. J. Med. Inform.,* vol. 126, pp. 105-113, Jun. 2019.
[24] S. Feng, Z. Zhuo, D. Pan, and Q. Tian, "CcNet: A cross-connected convolutional network for segmenting retinal vessels using multi-scale features," *Neurocomputing,* vol. 392, pp. 268-276, Jun. 2020.
[25] S. Zhang, H. Fu, Y. Xu, Y. Liu, and M. Tan, "Retinal Image Segmentation with a Structure-Texture Demixing Network," in *Medical Image Computing Computer Assisted Intervention*, 2020, pp. 765-774.
[26] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," *IEEE Trans. Med. Imaging,* vol. 39, no. 6, pp. 1856-1867, Jun. 2020.
[27] L. Li, M. Verma, Y. Nakashima, H. Nagahara, R. Kawasaki, and I. C. Soc, "IterNet: Retinal Image Segmentation Utilizing Structural Redundancy in Vessel Networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. Workshops*, ed, 2020, pp. 3645-3654.
[28] S. Feng *et al.*, "CPFNet: Context Pyramid Fusion Network for Medical Image Segmentation," *IEEE Trans. Med. Imaging,* vol. 39, no. 10, pp. 3008-3018, Oct. 2020.
[29] Y. Liu, H. Yuan, Z. Wang, and S. Ji, "Global Pixel Transformers for Virtual Staining of Microscopy Images," *IEEE Trans. Med. Imaging,* vol. 39, no. 6, pp. 2256-2266, 2020.
[30] Y. Li, Y. Liu, Y. Guo, X. Liao, B. Hu, and T. Yu, "Spatio-Temporal-Spectral Hierarchical Graph Convolutional Network with Semi-Supervised Active Learning for Patient-Specific Seizure Prediction," *IEEE Trans. Cybern.,* early access, May 2021, doi: 10.1109/TCYB.2021.3071860.
[31] A. Vaswani *et al.*, "Attention is all you need," presented at the Proc. Int. Conf. on Neural Inform. Process. Syst., Long Beach, California, USA, 2017.
[32] S. Huang, J. Li, Y. Xiao, N. Shen, and T. Xu, "RTNet: Relation Transformer Network for Diabetic Retinopathy Multi-lesion Segmentation," *IEEE Trans. Med. Imaging,* pp. 1-1, 2022.
[33] J. Cheng, J. Liu, H. Kuang, and J. Wang, "A Fully Automated Multimodal MRI-based Multi-task Learning for Glioma Segmentation and IDH Genotyping," *IEEE Trans. Med. Imaging,* pp. 1-1, 2022.
[34] D. Guo and D. Terzopoulos, "A Transformer-Based Network for Anisotropic 3D Medical Image Segmentation," in *Proceedings of International Conference on Patern Recognition (ICPR)*, 2021, pp. 8857-8861.
[35] F. Yu, V. Koltun, and T. Funkhouser, "Dilated Residual Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 636-644.
[36] Ozan Oktay *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas," *arXiv:1804.03999,* Apr. 2018.
[37] J. Fu *et al.*, "Dual Attention Network for Scene Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3141-3149.
[38] L. Mou *et al.*, "CS-Net: Channel and Spatial Attention Network for Curvilinear Structure Segmentation," in *Medical Image Computing Computer Assisted Intervention*, 2019, pp. 721-730.
[39] Z. Gu *et al.*, "CE-Net: Context Encoder Network for 2D Medical Image Segmentation," *IEEE Trans. Med. Imaging,* vol. 38, no. 10, pp. 2281-2292, Oct. 2019.
[40] S. Graham *et al.*, "MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images," *Med. Image Anal.,* vol. 52, pp. 199-211, Feb. 2019.
[41] M. Zhang, F. Yu, J. Zhao, L. Zhang, and Q. Li, "BEFD: Boundary Enhancement and Feature Denoising for Vessel Segmentation," in *Medical Image Computing Computer Assisted Intervention*, 2020, pp. 775-785.
[42] Q. Yan *et al.*, "An Attention-guided Deep Neural Network with Multi-scale Feature Fusion for Liver Vessel Segmentation," *IEEE J. Biomed. Health Inform.,* pp. 1-1, 2020.
[43] Y. Li, J. Liu, Z. Tang, and B. Lei, "Deep Spatial-Temporal Feature Fusion From Adaptive Dynamic Functional Connectivity for MCI Identification," *IEEE Trans. Med. Imaging,* vol. 39, no. 9, pp. 2818-2830, Sept. 2020.
[44] J. Kittler, "On the accuracy of the Sobel edge detector," *Image Vis. Comput.,* vol. 1, pp. 37-42, Feb. 1983.
[45] F. Ghadiri, M. Akbarzadeh-T, and S. Haddadan, "Vessel segmentation based on Sobel operator and fuzzy reasoning," in *Proc. Int. eConf. Comput. Knowl. Eng. (ICCKE)*, 2011, pp. 189-194.
[46] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating Fully Convolutional Networks With Spatial and Channel "Squeeze and Excitation" Blocks," *IEEE Trans. Med. Imaging,* vol. 38, no. 2, pp. 540-549, Feb. 2019.
[47] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. v. Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imaging,* vol. 23, no. 4, pp. 501-509, Apr. 2004.
[48] A. D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Trans. Med. Imaging,* vol. 19, no. 3, pp. 203-210, Mar.

2000.
[49] M. M. Fraz *et al.*, "An Ensemble Classification-Based Approach Applied to Retinal Blood Vessel Segmentation," *IEEE Trans. Biomed. Eng.,* vol. 59, no. 9, pp. 2538-2548, Sept. 2012.
[50] V. Cherukuri, V. K. B.G, R. Bala, and V. Monga, "Deep Retinal Image Segmentation With Regularization Under Geometric Priors," *IEEE Trans. Image Process.,* vol. 29, pp. 2552-2567, 2020 2020.
[51] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," presented at the Int. Conf. Learn. Represent., Toulon, France, 2017.
[52] Y. Li *et al.*, "Multimodal hyper-connectivity of functional networks using functionally-weighted LASSO for MCI classification," *Med. Image Anal.,* vol. 52, pp. 80-96, Feb. 2019.
[53] Y. Li, J. Liu, Y. Jiang, Y. Liu, and B. Lei, "Virtual Adversarial Training-Based Deep Feature Aggregation Network From Dynamic Effective Connectivity for MCI Identification," *IEEE Trans. Med. Imaging,* vol. 41, no. 1, pp. 237-251, 2022.