scientific reports

OPEN

TCDDU-Net: combining transformer and convolutional dual-path decoding U-Net for retinal vessel segmentation

Nianzu Lv[1], Li Xu[1]✉, Yuling Chen[2], Wei Sun[3], Jiya Tian[1] & Shuping Zhang[1]

Accurate segmentation of retinal blood vessels is crucial for enhancing diagnostic efficiency and preventing disease progression. However, the small size and complex structure of retinal blood vessels, coupled with low contrast in corresponding fundus images, pose significant challenges for this task. We propose a novel approach for retinal vessel segmentation, which combines the transformer and convolutional dual-path decoding U-Net (TCDDU-Net). We propose the selective dense connection swin transformer block, which converts the input feature map into patches, introduces MLPs to generate probabilities, and performs selective fusion at different stages. This structure forms a dense connection framework, enabling the capture of long-distance dependencies and effective fusion of features across different stages. The subsequent stage involves the design of the background decoder, which utilizes deformable convolution to learn the background information of retinal vessels by treating them as segmentation objects. This is then combined with the foreground decoder to form a dual-path decoding U-Net. Finally, the foreground segmentation results and the processed background segmentation results are fused to obtain the final retinal vessel segmentation map. To evaluate the effectiveness of our method, we performed experiments on the DRIVE, STARE, and CHASE datasets for retinal vessel segmentation. Experimental results show that the segmentation accuracies of our algorithms are 96.98, 97.40, and 97.23, and the AUC metrics are 98.68, 98.56, and 98.50, respectively. In addition, we evaluated our methods using F1 score, specificity, and sensitivity metrics. Through a comparative analysis, we found that our proposed TCDDU-Net method effectively improves retinal vessel segmentation performance and achieves impressive results on multiple datasets compared to existing methods.

Keywords Segmentation, Retinal blood vessels, Transformer, TCDDU-Net, Convolution

Glaucoma, cataracts, diabetes, and other diseases not only cause physical pain to patients but also continue to affect their normal life and work[1]. Research has shown that diabetes is closely related to the structure of retinal blood vessels[2]. Doctors diagnose patients' conditions by observing the structure of the blood vessels in the retina of the patient's eye fundus image. However, directly observing the fundus image or segmented image through manual means can not only take more time but also result in misjudgment due to image quality issues. Therefore, automatic segmentation of retinal blood vessels can be used for early diagnosis of certain diseases, improve the efficiency of doctor diagnosis, and to some extent, improve the diagnosis results. This has led to the proposal of many retinal blood vessel segmentation algorithms, which has achieved good research results.

However, retinal images have low contrast, uneven lighting, and high levels of noise, and retinal blood vessels are small in size with complex shape and structure, which makes retinal blood vessel segmentation very challenging[3]. To address the problem of retinal blood vessel detection, numerous excellent scholars have proposed unique insights from different perspectives. Zhao et al.[4] proposed a non-local total variation model to solve the problem of intensity inhomogeneity and low contrast and segmented the retinal blood vessel image into superpixels to locate the region of interest. Wang et al.[5] determined the wavelet kernel based on the relationship between blood vessels and edges and iteratively segmented the blood vessels. You et al.[6] proposed a radial

[1]College of Information Engineering, Xinjiang Institute of Technology, No.1 Xuefu West Road, Aksu 843100, Xinjiang, China. [2]School of Information Engineering, Mianyang Teachers' College, No. 166 Mianxing West Road, High Tech Zone, Mianyang 621000, Sichuan, China. [3]CISDI Engineering Co., LTD, Chongqing 401120, China. ✉email: 1184294752@qq.com

projection and semi-supervised method to locate the centerline of low-contrast and narrow retinal blood vessels, which helps with vessel segmentation. Yan et al.[7] proposed a retinal blood vessel segmentation method based on the hessian-based filter and random walk algorithm to enhance vessel structures and improve segmentation results to some extent. Imani et al.[8] innovatively proposed morphological component analysis based on sparse representation to separate retinal blood vessels and lesions, and finally obtained the final segmentation result by threshold processing. Although these traditional retinal blood vessel methods have solved some problems, there is still a certain gap between the segmentation results and the gold standard label, and the algorithm's scalability is not strong, requiring a significant number of manually adjusted parameters.

With the advent of big data and rapid development of computer hardware, convolutional neural network technology has rapidly developed in the field of medical image segmentation and achieved good results. In the field of semantic segmentation, many researchers have proposed end-to-end segmentation networks such as FCN[9], SegNet[10], DenseASPP[11], etc. In the field of medical image segmentation, many lesion segmentation algorithms are variants of the U-Net[12] structure. Mlynarski et al.[13] proposed a CNN-based tumor segmentation method that effectively combines 2D and 3D contextual distances. Murugesan et al.[14] proposed Psi-Net, which has a single encoder and three parallel decoders, as a universal medical image segmentation architecture. In the context of retinal vascular segmentation, Kamran et al.[15] introduced a multi-scale generation structure called RV-GAN. This approach utilizes two generators and two multi-scale encoders aiming to address the issue of information loss during the encoding process in automatic segmentation methods, ultimately enhancing the accuracy of retinal vascular segmentation.

In the field of NLP, transformer has been used for modeling sequential data, capturing long-range dependencies in text, and achieving remarkable results[16]. Inspired by the transformer algorithm, Dosovitskiy et al.[17] introduced transformer into image recognition tasks, transforming an image into a sequence of image patches for recognition. transformer has also become one of the research hotspots in the computer vision field. For example, there are successive algorithms such as swin transformer[18], Swin-Unet[19], and TransUNet[20].

Although many of the segmentation algorithms mentioned above have achieved good results, they still face problems of large-scale segmentation discontinuity and undetectable small blood vessels. Inspired by both the encoder-decoder network and the transfomer architecture, we proposed a combined transformer and convolutional dual path decoder U-Net (TCDDU-Net), whose overall network is shown in Fig. 1. TCDDU-Net is an end-to-end retinal vessel segmentation algorithm that can effectively solve the problem of vessel fragmentation or undetectability. In this paper, by selectively densely connecting swin transformer blocks, long-distance dependence can be captured, the network's receptive field can be improved, and contextual information can be modeled. At the same time, features from different stages are collected and feature selection is performed to effectively fuse different stage features. In addition, we take the background as a segmentation object and design the background decoder to learn the knowledge of the retinal vessel background, which forms a dual-path encoder with the foreground encoder. Lastly, the background decoder segmentation results are converted into retina segmentation results, which are fused with the foreground decoder segmentation results to solve the problem of poor background learning of the foreground decoder and improve the segmentation performance. To summarize the contribution of our work, we can outline the following three points:

(1) We propose a retinal vessel segmentation algorithm, TCDDU-Net, for the segmentation of small retinal vessels.

(2) We designed selective dense connection swin transformer block to capture the long-distance dependencies of blood vessels for selective dense connection and effective fusion of multi-stage features .

(3) We utilize deformable convolution to segment the background, and design a background decoder to learn the background knowledge and form a dual-path decoder with the foreground decoder, followed by converting the background segmentation results into retinal vessel segmentation results, and then fusing the

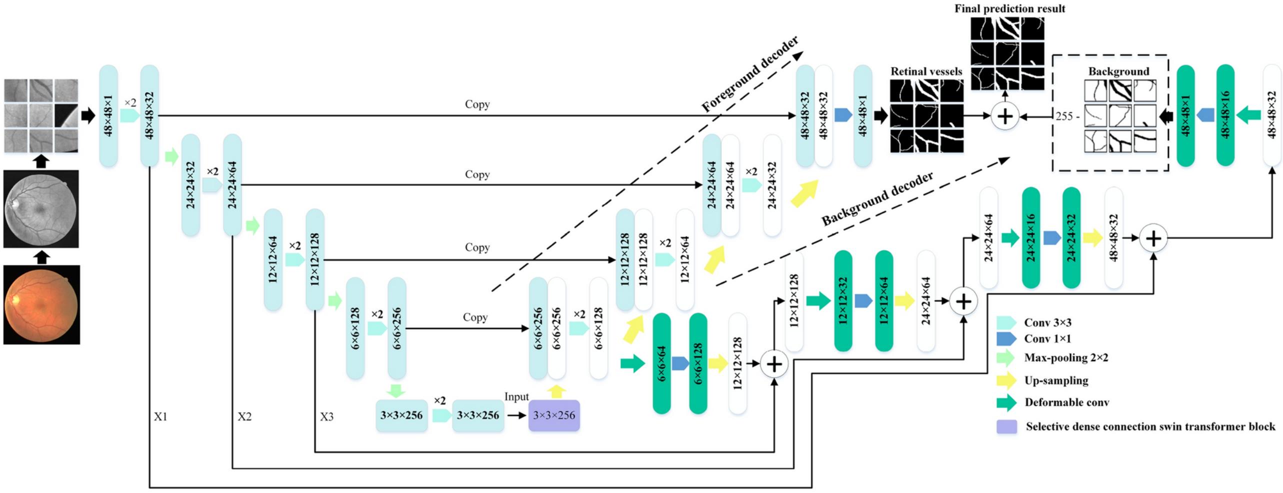

Fig. 1. TCDDU-Net overall network structure.

results with the foreground decoder segmentation results to improve the retinal vessel segmentation performance.The rest of this paper is organized as follows. Section "Related work" discusses related work in the field. Section "Methodology" describes the design and principles of TCDDU-Net. Section "Experiments" presents the experimental results and analysis, comparing the proposed method with other segmentation methods. Section "Conclusion" concludes the paper and provides prospects for future work.

Related work

CNN structure: research on convolutional neural networks (CNN) has been very popular in recent years. Convolutional neural network (CNN)-based classification algorithms have been widely utilized in segmentation networks for the purpose of extracting high-dimensional features, exemplified by Res-Unet[21] and H-DenseUNet[22]. Therefore, CNNs are driving the development of computer vision and also advancing research on retinal vessel segmentation algorithms. Yan et al.[23] proposed a segment-level loss to balance fine and coarse vessels, which was combined with pixel-wise loss for better balance. Xia et al.[24] proposed a coarse-to-fine segmentation network (CTF-Net) to address noise in retinal vessel segmentation in a cascade manner. Guo et al.[25] were inspired by DropBlock and introduced Dropout into the U-Net network to mine local features of retinal vessels, achieve end-to-end training and prediction, and propose the SD-Unet network structure. Li et al.[26] proposed a lightweight retinal vessel segmentation network with an attention mechanism that captures global information through the attention module and enhances feature representation in the feature fusion process while reducing model complexity. Wu et al.[27] proposed a scale-aware feature aggregation module (SFA) that dynamically adjusts the receptive field to extract features of different scales. They also proposed an adaptive feature fusion module (AFF) to effectively fuse features and a multi-level semantic supervision (MSS) to refine retinal vessel segmentation results. Xu et al.[28] proposed the SPNet, a retinal vessel segmentation network that shares a decoder and uses a pyramid-like loss to capture multiscale semantic information and achieve fine segmentation of retinal vessel edges at different scales. Zhang et al.[29] used Sobel to obtain edge prior knowledge, enhanced segmentation boundary in an unsupervised way, denoised the features, and finally integrated them into an encoder-decoder architecture for end-to-end retinal vessel segmentation. Yang et al.[30] analyzed the pixel ratio of coarse and fine retinal vessels and proposed a multi-task segmentation network and fusion network, and designed a loss function to solve the problem of sample imbalance. Tariq et al.[31] proposed MRC-Net, which learns contextual dependencies between different semantic features through multi-scale feature extraction and models these dependencies using bidirectional recurrent learning. This method effectively captures retinal vessel information at varying scales, pays particular attention to tiny blood vessels, and thereby enhances retinal vessel segmentation performance. Zhu et al.[32] proposed the DSeU net, which is based on deformable convolution and a squeeze-excitation residual module. This network dynamically adjusts the receptive field of retinal vascular features, scales the feature weights, and effectively learns the relationships between different features.

Transformer for computer vision: transformer is an important method in the field of natural language processing (NLP), initially used to solve machine translation problems, and now it has become one of the main methods for various NLP tasks[33]. However, some outstanding scholars has applied the transformer to computer vision tasks, successively proposing methods such as ViT[17], swin-transformer[18], Swin-Unet[19], which have promoted the development of the transformer algorithm in the field of vision. ViT converts images into multiple 16 16 size patches, and projects the patches into fixed-length vectors as the input of the transformer. Finally, image classification is completed through an MLP head. However, ViT has a large number of parameters and requires pre-training on large datasets. Therefore, Touvron et al.[34] proposed a distillation learning-based training strategy for the transformer, which achieved excellent results in training on the ImageNet dataset. swin-transformer solves the problem of large-scale changes and high resolutions in visual tasks through a hierarchical transformer composed of shifted windows, which not only improves the computational efficiency but also achieves high experimental indicators in various tasks[18], providing inspiration for the design of the algorithm in this paper. Swin-Unet uses swin-transformer as the encoding layer of the U-Net network to extract features from medical images, and designs a decoder based on swin-transformer to restore the spatial resolution of feature maps[19]. Similarly, there are also methods inspired by the transformer, such as[35,36]. Yuan et al.[37] proposed the novel cross-scale attention transformer (CAT), which utilizes a shared attention mechanism and integrates useful information from retinal vessel features. Additionally, they designed an edge refinement module (ERM) to refine the foreground and background edges of retinal vessels, thus enabling accurate segmentation of blood vessels. However, there are still significant gaps in retinal vessel segmentation methods based on the transformer structure, and relevant research is still scarce. Therefore, we propose a transformer-based retinal segmentation algorithm, which has important research significance and value.

Methodology

We combine transformer and convolutional neural network to propose the TCDDU-Net algorithm, which consists of three components: encoder,selective dense connection swin transformer block and dual path decoder. This section will describe the overall algorithm implementation idea, including selective dense connection swin transformer block, select swin block, and design of background decoder.

Combining transformer and convolutional dual-path decoding U-Net

Inspired by the convolutional neural network and the transformer structure, we combine the advantages of the two algorithms and design the TCDDU-Net, whose overall structure is shown in Fig. 1. TCDDU-Net input and output sizes are both 48 × 48, with an overall u-shaped structure consisting of an encoder, selective dense connection swin transformer block, and dual-path decoder. The encoder consists of a series of convolutional and max-pooling operations, which is used to extract deep semantic features of the image. The foreground

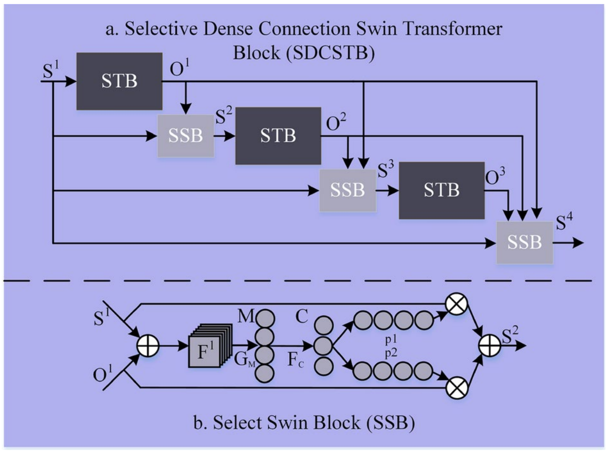
a. Selective Dense Connection Swin Transformer Block (SDCSTB)
S1
STB
O1
SSB
S2
STB
O2
SSB
S3
STB
O3
SSB
S4
S1
O1
F1
M
GM
C
Fc
p1
p2
S2
b. Select Swin Block (SSB)

Fig. 2. Selective dense connection swin transformer block and select swin block.

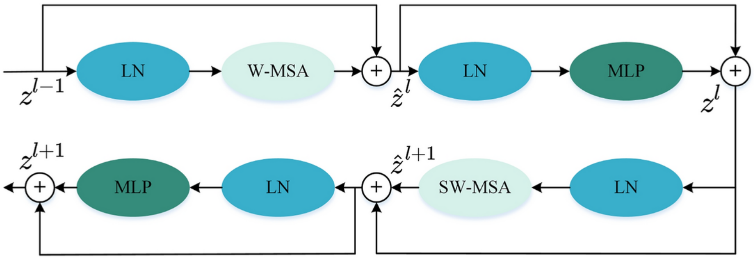

Fig. 3. Swin transformer block (STB).

decoder consists of an upsampling operation and a convolution operation, which is responsible for recovering the retinal vessel image resolution. The foreground decoder fuses the features extracted by the encoder with the up-sampled feature maps to fully utilize the low and high level features. The background decoder consists of deformable convolution, convolution and up-sampling operations that are responsible for restoring the resolution of the background image of the retinal vessels and learning about the background. At the same time, this paper introduces the selective dense connection swin transformer block in the last stage of feature extraction in the encoder to learn long-range dependencies between pixels, model contextual information and increase the network receptive field. In addition, the select swin block is introduced to adaptively fuse different sources of features, avoiding the introduction of redundant information in the dense connection process. The principles of the selective dense connection swin transformer block, the select swin block and the background decoder are described in detail in Parts 4, 5 and 6 of this section.

Selective dense connection swin transformer block

In retinal images, the retinal blood vessels are small and complex in structure. For the U-Net algorithm, the downsampling operation can lead to the loss of many details, resulting in the segmentation of small retinal blood vessels being incomplete or undetectable. To address this issue, this paper proposes the selective dense connection swin transformer block, which is added to the final stage of the feature extraction process in the encoder. The aim is to fully utilize information from different sources, strengthen the capture of long-range dependency relationships, model contextual information, and remove redundant information while focusing on useful information fusion. Specifically, the structure of the selective dense connection swin transformer block is shown in Fig. 2a, which includes a swin transformer block, a select swin block, and a dense connection structure. The swin transformer block is illustrated in Fig. 3, and the select swin block will be introduced in detail in the fifth section.

The swin transformer block[18] is a multi-head attention module designed based on the sliding window approach. It consists of 4 LayerNorm (LN) layers, 2 MLPs, 1 window-based multi-head self-attention, and 1 sliding window-based multi-head self-attention. The swin transformer block can be represented by the following formula:

$$\hat{Z}^{l} = W - MSA\left(LN\left(Z^{l-1}\right)\right) + Z^{l-1}, \tag{1}$$

$$Z^l = MLP\left(LN\left(\hat{Z}^l\right)\right) + \hat{Z}^l, \tag{2}$$

$$\hat{Z}^{l+1} = SW - MSA\left(LN(Z^l)\right) + Z^l, \tag{3}$$

$$Z^{l+1} = MLP\left(LN\left(\hat{Z}^{l+1}\right)\right) + \hat{Z}^{l+1}, \tag{4}$$

the input and output of the swin transformer block are represented by Z^{l-1} and Z^{l+1} , respectively. $\hat{Z}^l$ and $\hat{Z}^{l+1}$ are the outputs of $W - MSA$ and $SW - MSA$, respectively. The calculation formula for self-attention is similar to the method[38] and is shown as follows:

$$Attention(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d}} + B\right)V, \tag{5}$$

where $Q, K, V \in R^{M^2 \times d}$ denote the query, key and value matrices. M^2 represents the number of patches in the window, and d represents the dimension of the key and query. B indicates relative position bias.

Selective dense connection swin transformer block is composed of 3 swin transformer blocks and 3 select swin blocks connected densely. Dense connection can promote the effective flow of information, enhance the expressive ability of the model, and solve the problem of feature representation of small blood vessels in the deep layers of the network to the maximum extent. Specifically, it can be described by the following formula:

$$O^i = STB\left(S^i\right), \tag{6}$$

$$S^j = SSB\left(j, S^1, \left[O^1, ..., O^{j-1}\right]\right), \tag{7}$$

where O^i represents the output of the i-th STB, S^i represents the input of the i-th STB, j represents the number of branches in the input SSB, and [...] represents the input set of the SSB.

Select swin block

Different stages of features contain different information, and direct dense connections can lead to too much redundant information. Therefore, it is crucial to select information. Using the filtered information as the input of STB can yield a purer output, which helps to model contextual information and capture effective long-range dependencies. Inspired by the sknet[39] algorithm, we designed a select swin block. Moreover, the select swin block is highly flexible and simple, and can be extended to multiple branch inputs. The specific structure is shown in Fig. 2b, where GM represents global average pooling and FC represents fully connected operation. The figure only shows the case of two branch inputs. In the case of two branch inputs, first, S^1 and O^1 are added for fusion:

$$F^1 = S^1 + O^1, \tag{8}$$

next, the global representative information is obtained by performing global average pooling on F^1:

$$M = G_M\left(F^1\right) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F^1(i, j), \tag{9}$$

next, a fully connected operation is applied to *M* to guide the generation of probability feature maps and to filter the features from different sources, followed by the fusion of the adaptively selected features. The specific description is as follows:

$$C = F_C(M), \tag{10}$$

$$p^1 = \frac{e^{C^1}}{e^{C^1} + e^{C^2}}, \tag{11}$$

$$p^2 = \frac{e^{C^2}}{e^{C^1} + e^{C^2}}, \tag{12}$$

where $C^1, C^2 \in R^{2 \times d}$ represents the data of the first and second channels of feature map *C*.

Background decoder

For segmentation algorithms such as U-Net, Swin-Unet and TransUNet, they take the foreground as the segmentation object and ignore the auxiliary role of background information. Therefore, we designed the

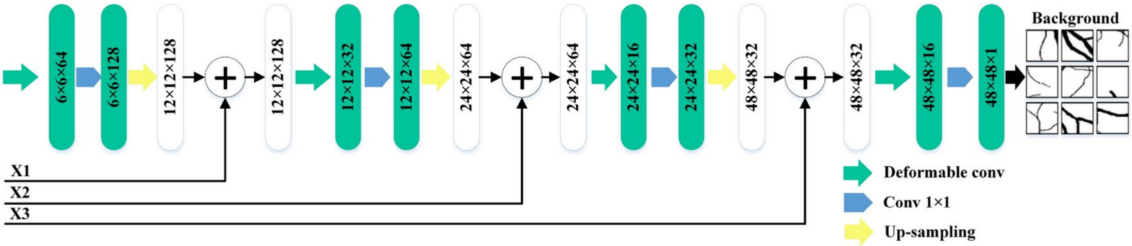

Fig. 4. Background decoder.

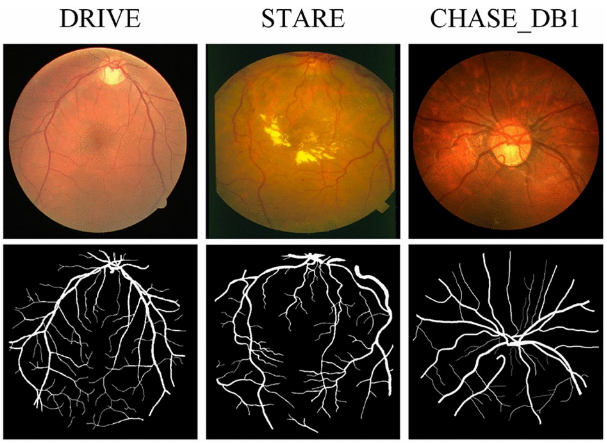

Fig. 5. Partial fundus images and corresponding label images of the three datasets.

background decoder to learn the background knowledge to help the network better understand and model the background information, understand the contextual information of the background, and avoid incorrectly segmenting similar regions into the foreground so as to improve the accuracy of retinal blood vessel segmentation, and the structure of the background decoder is shown in Fig. 4. X1, X2 and X3 are the feature maps obtained after 2 convolutions in the encoder. The background decoder consists of deformable convolution, convolution (convolution kernel size 1), upsampling and feature fusion. The overall process of the background decoder can be summarized as follows: firstly, deformable convolution is employed to extract features to better adapt to the deformation and spatial variations of the retinal vascular background, followed by increasing the number of channels of the feature map using the convolution with up-sampling, and then summing and fusing them with the encoder feature map Xi, where i = [1, 2, 3], and ultimately obtaining the background segmentation result.

Experiments

Based on the above proposed algorithm, we experimentally demonstrate the performance of the algorithm. In this section, we first describe the dataset and data processing methods, followed by the implementation details of the experiments. Then, we analyze the effects of different numbers of selective dense connection swin transformer blocks and the background decoder on the metrics. Finally, we compare our algorithm with other approaches.

Datasets introduction

This study conducted experiments on three widely recognized retinal vessel segmentation datasets, namely DRIVE[40], STARE[41], and CHASE[42]. The DRIVE dataset has an image resolution of 565 × 584 and was established by Staal et al. using fundus images of diabetic retinopathy patients. The STARE dataset was publicly released in 2000 and has an image resolution of 700 × 605, half of which is images of patients with retinal lesions. The CHASE dataset has a resolution of 999 × 960 and was obtained by capturing images of children's eyes. Figure 5 shows the original images and corresponding label images of the three datasets.

Data preprocessing

By appropriately preprocessing the data, the pressure on the model training can be alleviated and the training efficiency can be improved. Therefore, this article performs the following preprocessing operations on the data. First, the original color fundus images are converted to grayscale images. Second, the data is zero-mean-normalized to make the pixel values of the image follow a normal distribution. Third, the contrast limited adaptive histogram equalization (CLAHE)[43] algorithm is used to enhance the contrast of the image and highlight the location of small or tiny vessels. Fourth, gamma transformation is used to adjust the brightness of images

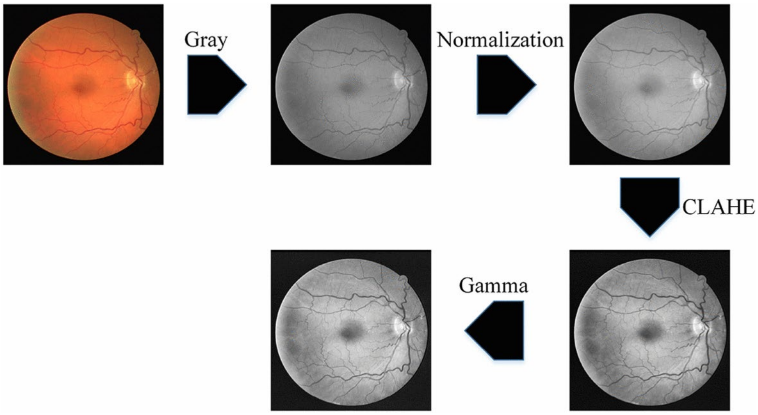

Fig. 6. Intermediate results of the preprocessing process.

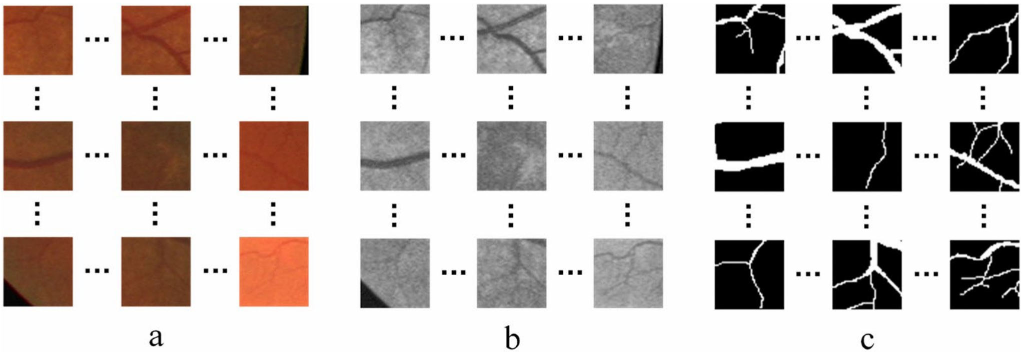

Fig. 7. Patches with a size of 48 × 48. (**a**) Is the original RGB image, (**b**) is the preprocessed image, and (**c**) is the corresponding segmentation label image.

that are too bright or too dark. Fifth, the pixel values of the image are divided by 255 for normalization, which can alleviate model overfitting and accelerate model training speed. The entire processing process and its effects are shown in Fig. 6.

In addition, due to the limited ability to collect data, the number of images in the datasets is relatively small and the sample diversity is insufficient, making it challenging to train the model. Therefore, in this paper, the images of the three datasets were randomly cropped into patches of size 48 × 48. The DRIVE and STARE datasets are cut by 300,000 patches, and CHASE is cut by 280,000 patches. The cropped patches were then divided into training and validation sets in a 1:1 ratios. Figure 7 shows some of the retinal vessel patch images.

Evaluation metrics

To validate the effectiveness of the proposed algorithm, this paper adopts Sensitivity (Se), Specificity (Sp), Accuracy (Acc), F1-score (F1), and Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) as evaluation metrics. Their calculation formulas are shown below:

$$Se = \frac{TP}{TP + FN}, \tag{13}$$

$$Sp = \frac{TN}{TN + FP}, \tag{14}$$

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

$$Pr = \frac{TP}{TP + FP}, \tag{16}$$

$$F1 = \frac{2 \times Pr \times Se}{Pr + Se}, \tag{17}$$

STB+SSB numbers	F1	Acc	AUC	Sp	Se
0 (baseline)	79.74 ± 2.01	96.51 ± 0.31	98.03 ± 0.50	**98.20** ± 0.53	79.03 ± 5.66
1	82.05 ± 1.34	96.81 ± 0.24	98.53 ± 0.40	98.10 ± 0.48	83.56 ± 4.97
2	82.26 ± 1.44	96.86 ± 0.23	98.57 ± 0.41	98.16 ± 0.45	83.49 ± 4.99
3	**82.48** ± 1.38	**96.88** ± 0.23	**98.66** ± 0.37	98.12 ± 0.47	**84.21** ± 4.88

Table 1. DRIVE dataset ablation experiment. Significant values are in bold.

STB+SSB numbers	F1	Acc	AUC	Sp	Se
0 (baseline)	79.56 ± 6.23	97.17 ± 0.7	98.18 ± 1.04	**98.89** ± 0.39	75.75 ± 10.84
1	80.12 ± 6.31	97.24 ± 0.69	98.32 ± 0.98	**98.89** ± 0.35	76.60 ± 10.31
2	80.50 ± 6.50	**97.27** ± 0.65	98.44 ± 0.98	98.82 ± 0.34	77.83 ± 11.06
3	**81.10** ± 4.41	97.18 ± 0.49	**98.60** ± 0.73	98.37 ± 0.50	**82.24** ± 8.15

Table 2. STARE dataset ablation experiment. Significant values are in bold.

STB+SSB numbers	F1	Acc	AUC	Sp	Se
0 (baseline)	74.37 ± 2.58	96.62 ± 0.62	97.86 ± 0.34	97.72 ± 0.73	79.64 ± 3.83
1	75.32 ± 2.53	96.78 ± 0.44	98.05 ± 0.32	97.84 ± 0.51	80.39 ± 3.64
2	76.66 ± 2.01	97.02 ± 0.38	98.18 ± 0.31	98.15 ± 0.38	79.85 ± 3.54
3	**78.04** ± 1.84	**97.19** ± 0.42	**98.43** ± 0.27	**98.21** ± 0.47	**81.42** ± 3.45

Table 3. CHASE dataset ablation experiment. Significant values are in bold.

where *TP* represents the number of correctly predicted positive samples, *TN* represents the number of correctly predicted negative samples, *FP* represents the number of negative samples predicted as positive, *FN* represents the number of positive samples predicted as negative.

Experimental details

The experiments in this paper were conducted on an NVIDIA GeForce RTX 3090 graphics card and the neural network structure was implemented using the PyTorch framework with a version of 1.7.1. The learning rate was set to 0.003, and Adam was used as the optimizer to train the model with default values for the parameters. For gamma transformation, the gamma coefficient was set to 1.2. When applying the CLAHE method, the threshold for contrast limiting parameter was set to 2.0 and the size of the grid for histogram equalization was set to 8 × 8.

Ablation experiments

To investigate the effect of incorporating the selective dense connection swin transformer block (STB) into the original model, we conducted ablation experiments on the DRIVE, STARE, and CHASE datasets, and subsequently analyzed and summarized the experimental findings. Tables 1, 2, and 3 showed the effect of the number of STB and SSB on the test results of different datasets. The experimental data in the tables indicate that the best performance was achieved when the number of STB+SSB was 3. Compared with the original U-Net network, F1, Acc, AUC, and Se were improved by 2.74%/1.54%/3.67%, 0.37%/0.01%/0.57%, 0.63%/0.42%/0.57%, and 5.18%/6.49%/1.78% respectively in the DRIVE/STARE/CHASE datasets. The performance of the proposed method was significantly improved, especially in F1 and Se metrics. In the DRIVE and STARE datasets, the Sp metric was slightly lower than that of the original U-Net algorithm, but in the CHASE dataset, the Sp metric was improved by 0.49%. When the number of STB and SSB modules was 1 or 2, the performance of the U-Net network was also improved, which demonstrated the effectiveness of the selective dense connection swin transformer block.

We integrated the background decoder after the selective dense connection swin transformer block and conducted experiments on the DRIVE, STARE, and CHASE datasets. The experimental results in Table 4 demonstrate a significant improvement in algorithmic metrics when the background decoder is incorporated. Specifically, on the DRIVE dataset, the F1 metric improved by 0.17%, the Acc metric by 0.1%, the Sp metric by 0.26%, and there was a slight improvement in the AUC metric. On the STARE dataset, the F1 metric showed an improvement of 0.53%, the Acc metric improved by 0.22%, and the Sp metric improved by 0.47%. Similarly, on the CHASE dataset, the F1 metric improved by 0.34% and the Se metric showed a 0.45% improvement, in addition to improvements in the Acc, AUC, and Sp metrics. By treating the background as a learning object and leveraging the acquired background information, the integration of the background decoder leads to a substantial enhancement in the performance of retinal vessel segmentation.

Datasets	F1	Acc	AUC	Sp	Se
DRIVE	82.65 ± 1.57	96.98 ± 0.25	98.68 ± 0.37	98.38 ± 0.42	82.58 ± 5.18
STARE	81.63 ± 5.53	97.40 ± 0.64	98.56 ± 0.88	98.84 ± 0.32	79.20 ± 9.41
CHASE	78.38 ± 1.68	97.23 ± 0.40	98.50 ± 0.23	98.23 ± 0.45	81.87 ± 2.99

Table 4. Background decoder test results on DRIVE, STARE and CHASE datasets.

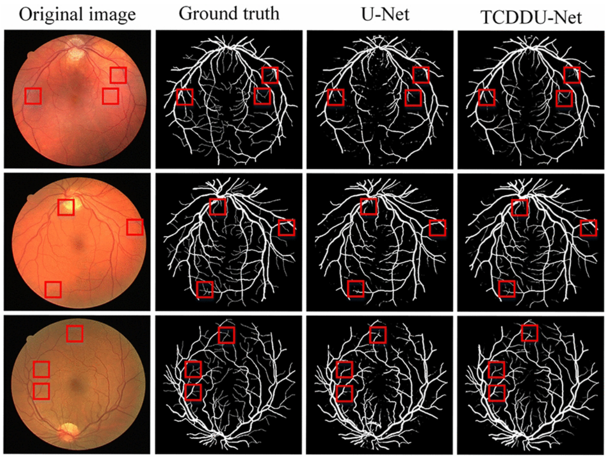

Fig. 8. Prediction results of U-Net and TCDDU-Net on DRIVE dataset.

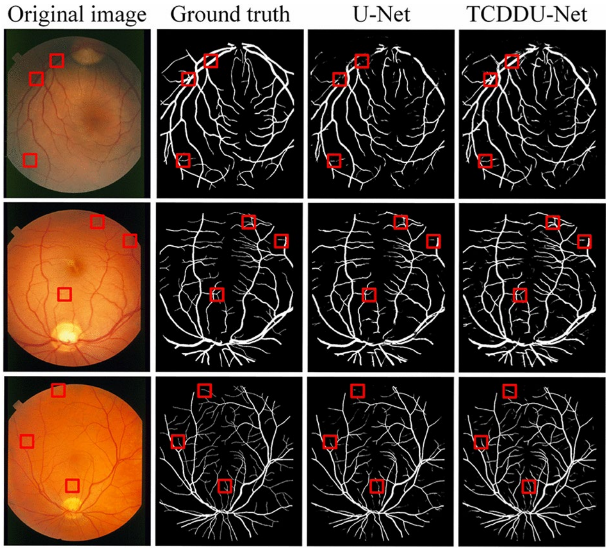

Fig. 9. Prediction results of U-Net and TCDDU-Net on STARE dataset.

In addition, we compares the U-Net and TCDDU-Net prediction results with real labels, and the comparison results are shown in Figs. 8, 9 and 10. By observing the prediction results and real label images, it is found that TCDDU-Net pays more attention to medium-sized and smaller blood vessels, and effectively utilizes contextual information to model long-distance dependencies, modeling and information fusion through background to foreground to alleviate the problem of retinal blood vessel segmentation breakage, and then improve segmentation effect.

Comparison of evaluation metrics for different algorithms

To provide a more comprehensive comparison of the performance of TCDDU-Net, we compared it with several state-of-the-art approaches from recent years, as depicted in Table 5, 6, and 7.

In the DRIVE dataset, TCDDU-Net performs optimally in F1, Acc, AUC, and Sp metrics, with the F1 metric reaching 82.65. For the STARE dataset, although the Acc and AUC metrics do not reach the optimal levels,

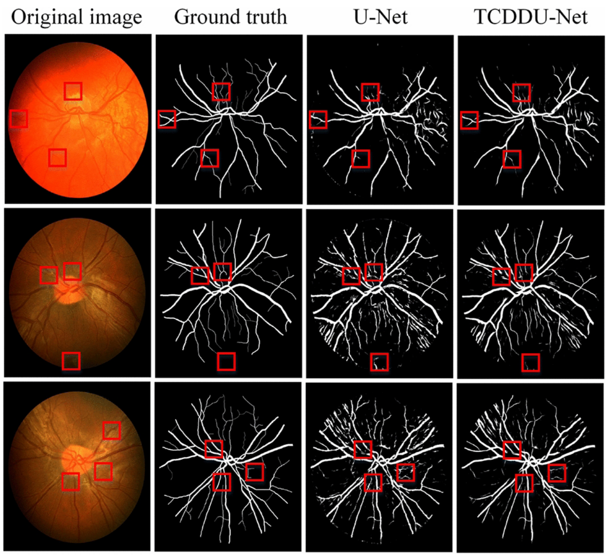

Fig. 10. Prediction results of U-Net and TCDDU-Net on CHASE dataset.

Methods	F1	Acc	AUC	Sp	Se
Yu[44]	–	95.24	97.23	98.03	76.43
DUNet[45]	82.37	95.66	98.02	98.00	79.63
Yan[46]	–	95.38	97.50	98.20	76.31
U-Net++[47]	81.92	96.88	98.12	–	–
CTF-Net[48]	82.41	95.67	97.88	–	–
CcNet[49]	–	95.28	96.78	98.09	76.25
Yang[30]	–	95.79	–	97.51	83.53
Xu[28]	–	96.64	98.28	98.02	82.43
Ours (w/o background decoder)	82.48 ± 1.38	96.88 ± 0.23	98.66 ± 0.37	98.12 ± 0.47	**84.21** ± 4.88
Ours (TCDDU-Net)	**82.65** ± 1.57	**96.98** ± 0.25	**98.68** ± 0.37	**98.38** ± 0.42	82.58 ± 5.18

Table 5. Comparison results of different segmentation algorithms in DRIVE,[Key: **Best**, Second Best]. Significant values are in bold and underline.

Methods	F1	Acc	AUC	Sp	Se
Yu[44]	–	96.13	97.87	98.22	78.37
DUNet[45]	81.43	96.41	98.32	98.78	75.95
Yan[46]	–	96.38	98.33	98.57	77.35
U-Net++[47]	78.59	**97.57**	97.63	–	–
CTF-Net[48]	–	–	–	–	–
CcNet[49]	–	96.33	97.00	98.48	77.09
Yang[30]	–	96.26	–	98.21	79.46
Xu[28]	–	96.92	98.12	97.90	**85.04**
Ours (w/o background decoder)	81.10 ± 4.41	97.18 ± 0.49	**98.60** ± 0.73	98.37 ± 0.50	82.24 ± 8.15
Ours (TCDDU-Net)	**81.63** ± 5.53	97.40 ± 0.64	98.56 ± 0.88	**98.84** ± 0.32	79.20 ± 9.41

Table 6. Comparison results of different segmentation algorithms in STARE,[Key: **Best**, Second Best]. Significant values are in bold and underline.

TCDDU-Net is ranked second. Moreover, the F1 and Sp metrics of TCDDU-Net outperform other comparative methods. In the experimental results on the CHASE dataset, TCDDU-Net demonstrates superiority with AUC and Sp metrics of 98.50 and 98.23, respectively, surpassing other algorithms. TCDDU-Net also ranks second in terms of Acc and Se metrics among the compared algorithms, with scores of 97.23 and 81.87, respectively.

Methods	F1	Acc	AUC	Sp	Se
Yu[44]	–	–	–	–	–
DUNet[45]	78.83	96.10	98.04	97.52	81.55
Yan[46]	–	96.07	97.76	98.06	76.40
U-Net++[47]	**81.34**	**97.62**	98.35	–	–
CTF-Net[48]	–	–	–	–	–
CcNet[49]	–	–	–	–	–
Yang[30]	–	96.32	–	97.76	81.76
Xu[28]	–	96.85	98.40	97.60	**86.19**
Ours (w/o background decoder)	78.04 ± 1.84	97.19 ± 0.42	98.43 ± 0.27	98.21 ± 0.47	81.42 ± 3.45
Ours (TCDDU-Net)	78.38 ± 1.68	97.23 ± 0.40	**98.50** ± 0.23	**98.23** ± 0.45	81.87 ± 2.99

Table 7. Comparison results of different segmentation algorithms in CHASE,[Key: **Best**, Second Best]. Significant values are in bold and underline.

Based on the above analysis and Tables 5, 6 and 7, it is evident that the proposed algorithm in this paper outperforms the others on the three datasets, achieving AUC scores of 98.68/98.56/98.50 and Acc scores of 96.98/97.40/97.23. Overall for the 3 datasets with 5 metrics, many of the metrics are ranked in the 1st and 2nd position, which fully demonstrates that TCDDU-Net shows superior performance.

Conclusion

Accurately segmenting retinal blood vessels in fundus images can help doctors improve the efficiency of eye disease diagnosis. In this paper, we propose the TCDDU-Net for retinal vessel segmentation, aiming to improve segmentation accuracy. Based on the U-Net network, we combine transformer and convolution modules to propose the TCDDU-Net, which focuses on segmenting small vessels and alleviates issues such as vessel segmentation discontinuity and undetectability. We introduce the selective dense connection swin transformer block to capture the long-distance dependencies of retinal blood vessels and model the context of surrounding information, thereby expanding the network's receptive field. We also design a select swin block to selectively fuse the outputs of swin blocks, focusing on the fusion of important features. We propose the background decoder, which takes the background as the segmentation target, learns the background knowledge, and fuses it with the foreground segmentation results to assist retinal vessel segmentation and improve the segmentation accuracy. Ablation experiments were conducted on the DRIVE, STARE, and CHASE datasets to compare the proposed approach with contemporary state-of-the-art segmentation methodologies. The results of these experiments unequivocally demonstrate the superior effectiveness of the proposed approach.

In future work, we plan to include more retinal blood vessel data to train more powerful models and optimize model parameters for practical diagnosis applications to improve doctor's diagnostic efficiency. Furthermore, we aim to extend the TCDDU-Net to three-dimensional medical image segmentation tasks, such as brain tumor segmentation and spleen segmentation, to improve the segmentation accuracy of 3D medical images and help more patients with disease diagnosis.

Data availability

All datasets are publicly available. DRIVE-https://drive.grand-challenge.org/, STARE-https://cecas.clemson.edu/~ahoover/stare/probing/index.html, CHASE DB1-https://github.com/LvNianZu/CHASE_DB1/tree/main

Received: 6 March 2024; Accepted: 22 October 2024
Published online: 29 October 2024

References

1. Medert, C. M., Sun, C. Q., Vanner, E., Parrish, R. K. & Wellik, S. R. The influence of etiology on surgical outcomes in neovascular glaucoma. *BMC Ophthalmol.* **21**(1), 440. https://doi.org/10.1186/s12886-021-02212-x (2021).
2. Sidhu, R. K., Sachdeva, J. & Katoch, D. Segmentation of retinal blood vessels by a novel hybrid technique- principal component analysis (PCA) and contrast limited adaptive histogram equalization (CLAHE). *Microvasc. Res.* **148**, 104477. https://doi.org/10.1016/j.mvr.2023.104477 (2023).
3. Han, Z., Yin, Y., Meng, X., Yang, G. & Yan, X.: Blood vessel segmentation in pathological retinal image. in *2014 IEEE International Conference on Data Mining Workshop* 960–967 (IEEE, 2014). https://doi.org/10.1109/ICDMW.2014.16
4. Zhao, Y. et al. Saliency driven vasculature segmentation with infinite perimeter active contour model. *Neurocomputing* **259**, 201–209. https://doi.org/10.1016/j.neucom.2016.07.077 (2017).
5. Wang, Y., Ji, G., Lin, P. & Trucco, E. Retinal vessel segmentation using multiwavelet kernels and multiscale hierarchical decomposition. *Pattern Recogn.* **46**(8), 2117–2133. https://doi.org/10.1016/j.patcog.2012.12.014 (2013).
6. You, X., Peng, Q., Yuan, Y., Cheung, Y.-M. & Lei, J. Segmentation of retinal blood vessels using the radial projection and semi-supervised approach. *Pattern Recogn.* **44**(10–11), 2314–2324. https://doi.org/10.1016/j.patcog.2011.01.007 (2011).
7. Li, Y., Gong, H., Wu, W., Liu, G. & Chen, G. An automated method using hessian matrix and random walks for retinal blood vessel segmentation. in *2015 8th International Congress on Image and Signal Processing (CISP)* 423–427 (IEEE, 2015). https://doi.org/10.1109/CISP.2015.7407917
8. Imani, E., Javidi, M. & Pourreza, H.-R. Improvement of retinal blood vessel detection using morphological component analysis. *Comput. Methods Programs Biomed.* **118**(3), 263–279. https://doi.org/10.1016/j.cmpb.2015.01.004 (2015).

9. Shelhamer, E., Long, J. & Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 640–651. https://doi.org/10.1109/TPAMI.2016.2572683 (2017).
10. Badrinarayanan, V., Kendall, A. & Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615 (2017).
11. Yang, M., Yu, K., Zhang, C., Li, Z. & Yang, K. DenseASPP for semantic segmentation in street scenes. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3684–3692 (IEEE, 2018). https://doi.org/10.1109/CVPR.2018.00388
12. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional networks for biomedical image segmentation*. arXiv:1505.04597 (2015).
13. Mlynarski, P., Delingette, H., Criminisi, A. & Ayache, N. 3D convolutional neural networks for tumor segmentation using long-range 2D context. *Comput. Med. Imaging Graph.* **73**, 60–72. https://doi.org/10.1016/j.compmedimag.2019.02.001 (2019).
14. Murugesan, B., Sarveswaran, K., Shankaranarayana, S. M., Ram, K., Joseph, J. & Sivaprakasam, M. Psi-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation. in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 7223–7226 (IEEE, 2019). https://doi.org/10.1109/EMBC.2019.8857339
15. Kamran, S. A., Hossain, K. F., Tavakkoli, A., Zuckerbrod, S. L., Sanders, K. M. & Baker, S. A. RV-GAN: Segmenting retinal vascular structure in fundus photographs using a novel multi-scale generative adversarial network. in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part VIII 24* vol. 12908 34–44 (2021). https://doi.org/10.1007/978-3-030-87237-3_4
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I. *Attention is all you need*. arXiv:1706.03762 (2023).
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv:2010.11929 (2021).
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. *Swin transformer: Hierarchical vision transformer using shifted windows*. arXiv:2103.14030 (2021).
19. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. & Wang, M. *Swin-Unet: Unet-like pure transformer for medical image segmentation*. arXiv:2105.05537 (2021).
20. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L. Zhou, Y. *TransUNet: Transformers make strong encoders for medical image segmentation*. arXiv:2102.04306 (2021).
21. Xiao, X., Lian, S., Luo, Z. & Li, S.: Weighted Res-UNet for high-quality retina vessel segmentation. in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)* 327–331 (IEEE, 2018). https://doi.org/10.1109/ITME.2018.00080
22. Li, X. et al. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging* **37**(12), 2663–2674. https://doi.org/10.1109/TMI.2018.2845918 (2018).
23. Yan, Z., Yang, X. & Cheng, K.-T. Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation. *IEEE Trans. Biomed. Eng.* **65**(9), 1912–1923. https://doi.org/10.1109/TBME.2018.2828137 (2018).
24. Xia, H., Zhuge, R. & Li, H. Retinal vessel segmentation via a coarse-to-fine convolutional neural network. in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 1036–1039. (IEEE, 2018). https://doi.org/10.1109/BIBM.2018.8621352
25. Guo, C., Szemenyei, M., Pei, Y., Yi, Y. & Zhou, W. SD-Unet: A structured dropout U-Net for retinal vessel segmentation. in *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)* 439–444 (IEEE, 2019). https://doi.org/10.1109/BIBE.2019.00085
26. Li, X., Jiang, Y., Li, M. & Yin, S. Lightweight attention convolutional neural network for retinal vessel image segmentation. *IEEE Trans. Ind. Inf.* **17**(3), 1958–1967. https://doi.org/10.1109/TII.2020.2993842 (2021).
27. Wu, H. et al. SCS-Net: A scale and context sensitive network for retinal vessel segmentation. *Med. Image Anal.* **70**, 102025. https://doi.org/10.1016/j.media.2021.102025 (2021).
28. Xu, G.-X. & Ren, C.-X.. *SPNet: A novel deep neural network for retinal vessel segmentation based on shared decoder and pyramid-like loss*. arXiv:2202.09515 (2022).
29. Zhang, M., Yu, F., Zhao, J., Zhang, L. & Li, Q. *BEFD: Boundary enhancement and feature denoising for vessel segmentation*. arXiv:2104.03768 (2021).
30. Yang, L., Wang, H., Zeng, Q., Liu, Y. & Bian, G. A hybrid deep segmentation network for fundus vessels via deep-learning framework. *Neurocomputing* **448**, 168–178. https://doi.org/10.1016/j.neucom.2021.03.085 (2021).
31. Khan, T. M., Naqvi, S. S., Robles-Kelly, A. & Razzak, I. Retinal vessel segmentation via a multi-resolution contextual network and adversarial learning. *Neural Netw.* **165**, 310–320. https://doi.org/10.1016/j.neunet.2023.05.029 (2023).
32. Zhu, X., Li, W., Zhang, W., Li, D. & Li, H. A deformable network with attention mechanism for retinal vessel segmentation. *J. Beijing Inst. Technol.* **33**(3), 186–193. https://doi.org/10.15918/j.jbit1004-0579.2024.050 (2024).
33. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K.: *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv:1810.04805 (2019).
34. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. & Jégou, H. *Training data-efficient image transformers & distillation through attention*. arXiv:2012.12877 (2021).
35. Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P. & Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 548–558 (IEEE, 2021). https://doi.org/10.1109/ICCV48922.2021.00061
36. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C. & Wang, Y. *Transformer in transformer* (2020). arXiv:2103.00112
37. Yuan, Y., Zhang, Y., Zhu, L., Cai, L. & Qian, Y. Exploiting cross-scale attention transformer and progressive edge refinement for retinal vessel segmentation. *Mathematics* **12**(2), 264. https://doi.org/10.3390/math12020264 (2024).
38. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J. *Exploring the limits of transfer learning with a unified text-to-text transformer*. arXiv:1910.10683 (2020).
39. Li, X., Wang, W., Hu, X. & Yang, J. *Selective kernel networks*. arXiv:1903.06586 (2019).
40. Staal, J., Abramoff, M. D., Niemeijer, M., Viergever, M. A. & Van Ginneken, B. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging* **23**(4), 501–509. https://doi.org/10.1109/TMI.2004.825627 (2004).
41. Hoover, A. D., Kouznetsova, V. & Goldbaum, M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Trans. Med. Imaging* **19**(3), 203–210. https://doi.org/10.1109/42.845178 (2000).
42. Fraz, M. M. et al. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Trans. Biomed. Eng.* **59**(9), 2538–2548. https://doi.org/10.1109/TBME.2012.2205687 (2012).
43. Pizer, S. M. et al. Adaptive histogram equalization and its variations. *Comput. Vis. Graph. Image Process.* **39**(3), 355–368. https://doi.org/10.1016/S0734-189X(87)80186-X (1987).
44. Yu, L. et al. A framework for hierarchical division of retinal vascular networks. *Neurocomputing* **392**, 221–232. https://doi.org/10.1016/j.neucom.2018.11.113 (2020).
45. Jin, Q. et al. DUNet: A deformable network for retinal vessel segmentation. *Knowl. Based Syst.* **178**, 149–162. https://doi.org/10.1016/j.knosys.2019.04.025 (2019).
46. Yan, Z., Yang, X. & Cheng, K.-T. A three-stage deep learning model for accurate retinal vessel segmentation. *IEEE J. Biomed. Health Inform.* **23**(4), 1427–1436. https://doi.org/10.1109/JBHI.2018.2872813 (2019).
47. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **39**(6), 1856–1867. https://doi.org/10.1109/TMI.2019.2959609 (2020).

48. Wang, K., Zhang, X., Huang, S., Wang, Q. & Chen, F. CTF-Net: Retinal vessel segmentation via deep coarse-to-fine supervision network. in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* 1237–1241 (IEEE, 2020). https://doi.org/10.1109/ISBI45749.2020.9098742
49. Feng, S., Zhuo, Z., Pan, D. & Tian, Q. CcNet: A cross-connected convolutional network for segmenting retinal vessels using multi-scale features. *Neurocomputing* **392**, 268–276. https://doi.org/10.1016/j.neucom.2018.10.098 (2020).

Acknowledgements

This study was supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (Grant No. 2022D01C461), the Natural Science Foundation of Xinjiang Uygur Autonomous Region (Grant No. 2022D01C349), and the school-level project of Xinjiang Institute of Technology (Grant No. ZZ202202).

Author contributions

Conceptualization, N.L. and L.X.; methodology, N.L. and Y.C.; validation, N.L. and W.S.; formal analysis, J.T.; investigation, L.X. and S.Z.; data curation, N.L. and Y.C.; writing—original draft preparation, N.L and Y.C.; writing—review and editing, N.L.; visualization W.S. and Y.C.; supervision, Y.C. and J.T.; project administration, N.L. and L.X.; funding acquisition, N.L. and S.Z. All authors have read and agreed to the published version of manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.