diagnostics

MDPI

Article

Landmark-Assisted Anatomy-Sensitive Retinal Vessel Segmentation Network

Haifeng Zhang †, Yunlong Qiu † and Chonghui Song * and Jiale Li

College of Information Science and Engineering, Northeastern University, Shenyang 110819, China; haif_zh@163.com (H.Z.); qyl6016@163.com (Y.Q.); li_jiale2020@163.com (J.L.)
* Correspondence: songchonghui@mail.neu.edu.cn
† These authors contributed equally to this work.

Abstract: Automatic retinal vessel segmentation is important for assisting clinicians in diagnosing ophthalmic diseases. The existing deep learning methods remain constrained in instance connectivity and thin vessel detection. To this end, we propose a novel anatomy-sensitive retinal vessel segmentation framework to preserve instance connectivity and improve the segmentation accuracy of thin vessels. This framework uses TransUNet as its backbone and utilizes self-supervised extracted landmarks to guide network learning. TransUNet is designed to simultaneously benefit from the advantages of convolutional and multi-head attention mechanisms in extracting local features and modeling global dependencies. In particular, we introduce contrastive learning-based self-supervised extraction anatomical landmarks to guide the model to focus on learning the morphological information of retinal vessels. We evaluated the proposed method on three public datasets: DRIVE, CHASE-DB1, and STARE. Our method demonstrates promising results on the DRIVE and CHASE-DB1 datasets, outperforming state-of-the-art methods by improving the F1 scores by 0.36% and 0.31%, respectively. On the STARE dataset, our method achieves results close to the best-performing methods. Visualizations of the results highlight the potential of our method in maintaining topological continuity and identifying thin blood vessels. Furthermore, we conducted a series of ablation experiments to validate the effectiveness of each module in our model and considered the impact of image resolution on the results.

Keywords: retinal vessel segmentation; TransUNet self-supervised landmark; contrastive learning

Citation: Zhang, H.; Qiu, Y.; Song, C.; Li, J. Landmark-Assisted Anatomy-Sensitive Retinal Vessel Segmentation Network. *Diagnostics* **2023**, *13*, 2260. https://doi.org/10.3390/diagnostics13132260

Academic Editors: Jaafar M. Alghazo and Ghazanfar Latif

Received: 31 May 2023
Revised: 28 June 2023
Accepted: 30 June 2023
Published: 4 July 2023

1. Introduction

Retinal vessel segmentation is an important diagnostic method for detecting hypertension, arteriosclerosis, and retinal diseases [1]. However, the retinal vascular structure is extremely complex, and the distribution of vascular pixel intensity is unbalanced. Furthermore, due to the low contrast between the blood vessel pixels and the background, the thin blood vessels located at the ends of the vascular structures are difficult to completely segment from the background. Accurate retinal vessel segmentation has always been an extremely challenging task.

In recent years, a great deal of work has focused on automatically segmenting retinal blood vessels. The methods used are broadly classified into two groups: unsupervised and supervised methods. Unsupervised methods are suitable for image segmentation with little annotation information. Commonly used algorithms include the matched filtering method [2], multi-threshold blood vessel detection method [3], mathematical morphology method [4], and so on. However, due to the absence of supervision from prior knowledge, unsupervised methods can easily detect false edges and achieve lower performance. In contrast to unsupervised methods, supervised methods utilize human-annotated data to train networks to learn feature information hidden in images. Currently, state-of-the-art semantic segmentation methods employ deep learning methods for pixel-level prediction.

U-Net has shown excellent performance in medical image segmentation due to its unique encoder-decoder structure. Many U-Net variants have been designed for retinal vessel segmentation. Jin et al. [5] proposed a method combining deformable convolution and U-Net to detect retinal blood vessels. Wu et al. [6] incorporated U-Net into a generative adversarial network. Retinal vessel segmentation is performed in an end-to-end manner. Although these methods have improved the accuracy of retinal vessel segmentation to a certain extent, the connectivity of vessels is difficult to guarantee due to insufficient use of contextual information in the structure, and the segmentation of thin vessels is still difficult. Clinically, thin blood vessels and vascular connectivity provide an indispensable reference for diagnosing vascular diseases. Therefore, it is imperative to explore new retinal vessel segmentation techniques.

To tackle the above-mentioned problem, this paper proposes an anatomy-sensitive retinal vessel segmentation framework that can jointly improve the performance of retinal vessel segmentation by exploiting the latent association among multiple modules. The backbone network adopts the improved U-Net network. To take full advantage of semantic information, we design a context relation module, which effectively combines the strong local modeling ability of convolution and the advantages of transformers in long-range modeling, and maps the features of various scales of the encoder to the decoder through skip pathways. In addition, we also design a sub-network for landmark detection, which learns a set of landmarks from retinal images using heatmap regression, to guide the network segmentation direction. The main contributions of this paper are as follows.

- TransUNet is more in line with anatomical retinal vessel segmentation due to its special structure. We use transformers as the segmentation backbone to benefit from the advantages of convolutional layers in extracting local features and multi-head self-attention in modeling global relations. Meanwhile, we reform the skip connections in TransUNet to decode deep semantics more easily and accurately.
- A self-supervised landmark-assisted segmentation framework is proposed to further improve the accuracy of retinal vessel segmentation. In particular, we propose a strategy for contrastive learning to improve the plausibility and accuracy of landmark representations of anatomical topology. We utilize landmarks that sparsely represent retinal vessel morphology to guide the model towards learning the content, rather than the style that is not conducive to segmentation. Furthermore, landmarks enhance the richness of explicit descriptions of retinal vascular anatomy, which is friendly for the model to learn based on fewer samples.
- We implement the proposed network on the DRIVE, CHASE-DB1, and STARE datasets, and extensive experimental results show that our method achieves state-of-the-art performance in most cases.

2. Related Work

Deep convolutional neural networks have become the most popular method for retinal vessel segmentation due to their excellent performance in medical image segmentation tasks. Among them, U-Net [7] and its variants are the most widely used as the backbone. A symmetric encoder-decoder structure and skip-connected architecture from encoding paths to decoding paths lead U-Net to achieve efficient information flow. Benefiting from an architecture that integrates local and global information from low-level and high-level feature maps, U-Net exhibits better performance in medical image analysis. However, although U-Net achieves multi-scale contextual information aggregation, it is still insufficient to cope with thin and irregular retinal vascular structures. Multiple studies have been devoted to addressing this issue.

Wang et al. [8] improved on the standard U-Net network and designed a two-channel encoder to extract information about retinal blood vessels. The improved encoder includes a context channel and a spatial channel to capture more receptive field and spatial information. The design of the backbone network of Li et al. [9] adopts the iterative principle to cascade multiple small U-Net networks to learn the structural features of retinal blood

vessels. The input of each small U-Net network is the coarse segmentation probability map output by its previous U-Net network, and the vessel segmentation accuracy is improved by iterating from coarse to fine. Despite the excellent representational power of convolution, CNN-based methods often exhibit limitations in modeling explicit long-term relationships because of the inherent locality of convolution operations. The transformer module shows outstanding performance in capturing long-distance dependencies in the field of natural language processing, and is gradually being introduced into image processing. Cao et al. [10] designed the Swin-Unet network for medical image segmentation. The proposed network adopts a symmetrical structure similar to the U-Net network, and both the encoder and the decoder use pure transformer modules. However, the construction of a pure transformer network requires a large amount of computation, and the network is difficult to train. Xia et al. [11] proposed a combined CNN and transformer method to segment the optic cup and optic disc in the retina. First, the local features of the retina are obtained by convolution, and the extracted features are respectively passed through the multi-scale convolution module and the transformer module to obtain multi-scale feature information and global feature information. Finally, the segmentation performance of the optic cup and optic disc can be improved by fusing these two parts of the feature information. Chen et al. [12] integrated the transformer module into the U-Net network to achieve multi-organ segmentation. Convolutions are first utilized to extract low-level features, and then global interactions are modeled through the transformer module. The framework effectively combines the powerful local modeling capabilities of convolutions and the advantages of transformers in long-range modeling, enabling finer organ detail segmentation.

Accurate detection of landmark points is a critical step in medical imaging, as it provides quite valuable information for subsequent medical image analysis. Coordinate regression is the most typical method. The landmark coordinates are used as the target for the network regression to predict a set of landmark locations directly from the image space. Sun et al. [13] proposed a cascade of deep convolutional networks to improve the detection accuracy of face landmarks through coarse-to-fine regression. Zhang et al. [14] combined multi-task learning with a regression model for face landmark detection and used cascaded deep convolutional networks to predict face and landmark locations in a coarse-to-fine manner. However, the direct mapping from original images to landmark coordinates is a complex nonlinear problem that is not easily learned by the network. Compared with the numerical value of landmark coordinates, heatmaps can provide more abundant supervision information in space, which also improves the accuracy of landmark detection to a certain extent. Kowalski et al. [15] proposed a heatmap-based cascaded deep convolutional network DAN. The detected landmark positions are refined by each stage and passed to the next stage to correct the landmark positions iteratively. Shi et al. [16] designed a superimposed hourglass network and introduced offset learning to refine the predicted landmarks. The network effectively combines heatmap information and coordinate information to achieve accurate facial landmark detection.

Our proposed method focuses on improving the ability of the model to learn anatomical structures, thus achieving higher segmentation accuracy.

3. Methods and Materials

Our objective is to develop a deep learning model for segmenting blood vessel pixels in retinal images. To achieve this, we propose a framework, as depicted in Figure 1, which comprises two main sections: (i) An enhanced version of the U-Net is employed for precise segmentation of fundus blood vessels. (ii) Additionally, landmark detection is used as an auxiliary task to further enhance the accuracy of segmentation.

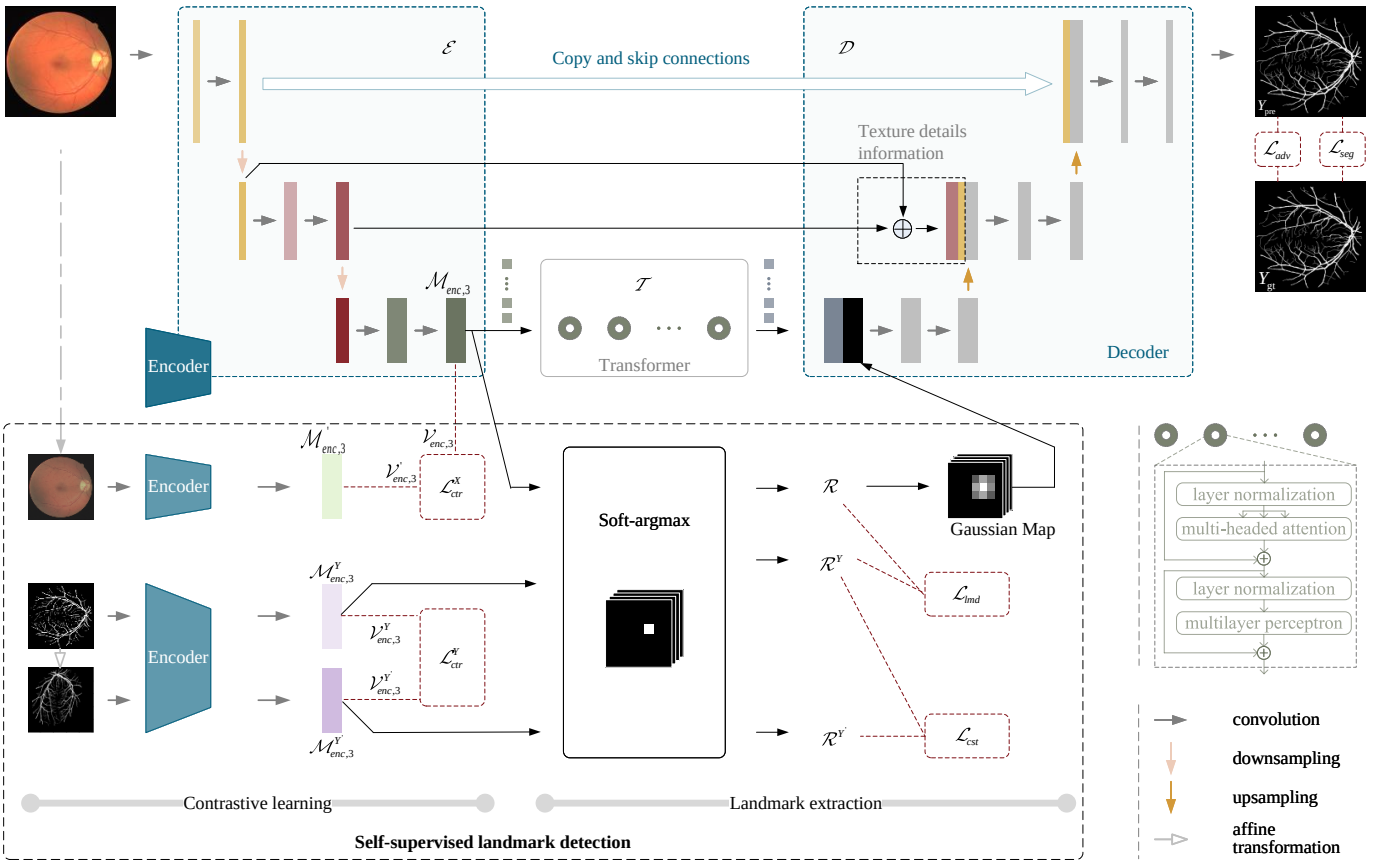

Figure 1. The pipeline of the proposed method.

3.1. Datasets

We use three public datasets for experiments, namely DRIVE, CHASE-DB1, and STARE. To improve the accuracy of segmentation, we implemented a data augmentation technique that utilized random flipping, rotation, and scaling.

The DRIVE dataset includes 40 fundus retinal color images, 7 of which are pathologically abnormal. The dimensions of each image are 584 × 565 pixels. The last twenty images of this dataset are used to train the network, and the first twenty images are used to test the network. All images in the test set consist of the results of manual segmentation by two professionals. We chose to use the result of the first professional manual segmentation as the label of the retinal blood vessels.

The CHASE-DB1 dataset contains 28 retinal images. They were taken from the eyes of 14 children. All images in the dataset are 996 × 960 pixels. Unlike the DRIVE dataset, there are no fixed training and test set partitions for CHASE-DB1. We randomly placed 20 retinal images in the training set and 8 images in the test set.

The STARE dataset has a total of 20 images. All images are 700 × 605 pixels. Since the STARE dataset does not have a pre-separated training set and test set, we employed leave-one-out cross-validation to verify the feasibility of our proposed method.

We improved upon the common approach of completely random data augmentation. First, we defined a sliding window with dimensions 0.6 times the width and height of the original image (i.e., the window area is 0.36 times that of the original image). Then, using this sliding window, we extracted 9 slices of the image with a stride of 1. Next, we selected the slice with the highest proportion of foreground from these 9 slices and performed other operations (such as flipping, contrast adjustment, brightness modification) before adding it to the training data. This approach helps to alleviate the issues of class imbalance or foreground–background imbalance to some extent.

3.2. TransUNet

Medical images have the unique advantage of having explicit contextual priors, due to the anatomical properties of tissues. Therefore, we propose to consider the long-range dependencies of pixels while also extracting local features. As illustrated in Figure 1, we introduce a transformer into the U-Net architecture. The convolutional layer of U-Net ensures that the model remains locally sensitive to the image, while the transformer module allows the model to capture global features of the image.

First, some symbols are defined. The convolutional encoder is $\mathcal{E} = \{\mathcal{E}_{\frac{H}{2^{n_d}} \times \frac{W}{2^{n_d}}}\}_{n_d=0}^{N_d}$ where H and W are the height and width of the input of the convolution operator, respectively. N_d is the number of down-sampling operations f_d. That is, convolution operators are grouped by the resolution of their input. Similarly, the convolutional decoder is $\mathcal{D} = \{\mathcal{D}_{\frac{H \cdot 2^{n_u}}{2^{N_d}} \times \frac{W \cdot 2^{n_u}}{2^{N_d}}}\}_{n_u=0}^{N_u}$. The transformer module is denoted by $\mathcal{T}$. The feature map is denoted by $\mathcal{M}$ with channel C.

3.3. Convolutional Encoder

The convolutional encoder of our method is the same as that of the standard U-Net encoder. Considering the missing information of tiny blood vessels caused by down-sampling and the over-fitting problem caused by too deep model layers, N_d is set to 2. That is, the convolution operators are divided into three groups, i.e., the feature maps have three resolutions. The original image is denoted as X. Then,

$$\begin{aligned} \mathcal{M}_{enc,1} &= \mathcal{E}_{H \times W}(X) \in \mathbb{R}^{H \times W \times C_1}, \\ \mathcal{M}_{enc,2} &= \mathcal{E}_{\frac{H}{2} \times \frac{W}{2}}(f_d(\mathcal{M}_{enc,1})) \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C_2}, \\ \mathcal{M}_{enc,3} &= \mathcal{E}_{\frac{H}{4} \times \frac{W}{4}}(f_d(\mathcal{M}_{enc,2})) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_3}. \end{aligned} \tag{1}$$

3.4. Transformer Module

To address the challenges of training transformers and their resource-intensive nature, we propose to connect the transformer module behind the convolutional encoder. This approach allows the transformer to receive input of smaller resolution, thereby reducing the equipment resources required. Moreover, the feature maps that are fed into the transformer already contain deep semantic information, making it easier to train. By incorporating the transformer module in this way, we can ensure that the model captures both local and global information, as the transformer mines long-range dependencies based on feature maps that have already extracted local features.

First, $\mathcal{M}_{enc,3}$ is decomposed into ${N_P}^2$ patches, i.e., $\mathcal{M}_{enc,3} \mapsto \{\mathcal{M}_{enc,3}^{n_P} \in \mathbb{R}^{\frac{H}{N_P} \times \frac{W}{N_P} \times C_3}\}_{n_P=1}^{N_P}$. The input of the transformer is

$$Z_0 = \{z_{pos}^{n_p} + z_{pat}^{n_p}\}_{n_p=1}^{{N_p}^2}, \tag{2}$$

where $z_{pos}^{n_p}$ and $z_{pat}^{n_p}$ are the position embedding and feature embedding of $\mathcal{M}_{enc,3}^{n_P}$, respectively. $z_{pat}^{n_p} = f_{pf}(\mathcal{M}_{enc,3}^{n_P})$, where f_{pf} is patch-wise flatten.

The transformer module $\mathcal{T}$ is composed of N_t transformer layers; each of them $\mathcal{T}_{n_t}$ consists of a multi-head self-attention (MHSA) block, multi-layer perceptron (MLP) block, and layer normalization (LN) blocks. The output of the n_t-th transformer layer is $Z_{n_t} = \mathcal{T}_{n_t}(Z_{n_t-1})$, specifically,

$$\begin{aligned} Z_{n_t}^{'} &= MHSA(LN(Z_{n_t-1})) + Z_{n_t-1}, \\ Z_{n_t} &= MLP(LN(Z_{n_t}^{'})) + Z_{n_t}^{'}. \end{aligned} \tag{3}$$

Finally, the output sequence Z_{N_t} of $\mathcal{T}$ is reconstructed into $\mathcal{M}_t$ by the patch merging layer f_{pm}, i.e., $\mathcal{M}_t = f_{pm}(Z_{N_t}) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_3}$.

3.5. Convolutional Decoder

We elaborately design a convolutional decoder $\mathcal{D}$ for progressive decoding. Similar to the convolutional encoder, the convolutional decoding layer is also divided into three groups according to the resolution, i.e., $N_u = 2$. Up-sampling f_u is bilinear interpolation. $\mathcal{D}_{\frac{H}{4} \times \frac{W}{4}}$ is fed by $\mathcal{M}_{dec,0}$ channel-wise connected by $\mathcal{M}_t$ and $\mathcal{M}_{gm}$, where $\mathcal{M}_{gm}$ is the Gaussian map of the landmark. In a traditional UNet decoder, $\mathcal{D}_{\frac{H}{2} \times \frac{W}{2}}$ is fed by the channel-wise connection of $f_u(\mathcal{D}_{\frac{H}{2} \times \frac{W}{2}}(\mathcal{M}_{dec,0}))$ and $\mathcal{M}_{enc,2}$. Considering the local detailed information lost due to the transformer modeling global relations, we further fuse $f_d(\mathcal{M}_{enc,1})$, which contains more texture information for $\mathcal{D}_{\frac{H}{2} \times \frac{W}{2}}$. In particular, we enhance the sensitivity of convolutional encoders to anatomical topology through contrastive learning; $\mathcal{M}_{enc,1}$ is considered to represent dense local shape details. In this way, when the global information and local information are fused in $\mathcal{D}_{\frac{H}{2} \times \frac{W}{2}}$, they are constrained by the texture information of the shape, which can avoid decoding information that violates the anatomical topology. Formalized,

$$\begin{aligned} &\mathcal{M}_{dec,1} = \mathcal{D}_{\frac{H}{4} \times \frac{W}{4}}(\mathcal{M}_t, \mathcal{M}_{gm}) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_3}. \\ &\mathcal{M}_{dec,2} = \mathcal{D}_{\frac{H}{2} \times \frac{W}{2}}(f_u(\mathcal{M}_{dec,1}), \mathcal{M}_{enc,2}, f_d(\mathcal{M}_{enc,1})) \\ &\in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C_2}, \\ &Y_{pre} = \mathcal{D}_{H \times W}(f_u(\mathcal{M}_{dec,2}), \mathcal{M}_{enc,1}) \in \mathbb{R}^{H \times W \times 1}, \end{aligned} \tag{4}$$

where Y_{pre} is the label predicted by the model.

3.6. Self-Supervised Landmark Detection

To address the difficulty of segmenting thin blood vessels from the background in retinal images due to their high complexity and low contrast, we propose a novel approach that incorporates landmark points to assist the network in segmentation. This approach represents a departure from previous methods that relied solely on implicit feature vectors learned from images by the network for pixel-by-pixel segmentation. The introduction of landmark detection represents a critical component of our segmentation network. Considering the small amount of retinal vessel data and the high cost of manual annotation, we propose unsupervised learning of a set of ordered landmarks from dense retinal vessel images under the framework of contrastive learning to guide the model for segmentation. The detected landmarks sparsely characterize the key features of dense anatomical topology and thus can represent the intrinsic structure of fundus vessels. For the model to extract accurate and robust landmark points, we propose a contrastive learning strategy and introduce a series of optimization objectives to train the model. Landmarks are generated based on the heatmap of the convolutional encoder, as shown in the landmark detector part of Figure 1.

Coordinate Extraction for Landmarks

We extract the landmarks in a way that activates the highest weighted pixel in the feature map. We extract landmarks in the feature space $\mathcal{M}^*_{enc,3}$ spanned by $\mathcal{E}$. First, we adopt the spatial softmax normalization method to convert all channels of $\mathcal{M}^*_{enc,3}$ to the probability response map $\mathcal{M}^*_{prob}$. Then, the site with the highest weight in the probability map $\mathcal{M}^*_{prob}$ is activated by soft-argmax as a landmark. Formally, the feature map $\mathcal{M}^{*,}_{enc,3}[c]$ of the c-th channel is probabilized as

$$\mathcal{M}^{*}_{prob}[c] = \frac{\exp(\mathcal{M}^{*,}_{enc,3}[c,r])}{\sum_{r\in(\frac{H}{4}\times\frac{W}{4})}\exp(\mathcal{M}^{*,}_{enc,3}[c,r])}\Bigg|_{r=1}^{\frac{H}{4}\times\frac{W}{4}}. \tag{5}$$

The set of landmark coordinates is

$$\mathcal{R}^{*} = \left\{soft-argmax(\mathcal{M}^{*}_{prob}[c])\right\}_{c=1}^{C_3}, \tag{6}$$

where $\mathcal{R}^{*}$ is the landmarks.

We utilize consistency loss $\mathcal{L}_{cst}$ to guarantee the quality of landmarks. $\mathcal{L}_{cst}$ is defined as

$$\mathcal{L}_{cst} = dist_{cst}(\mathcal{R}^{Y}, \mathcal{A}_{Y}^{-1}(\mathcal{R}^{Y'})), \tag{7}$$

where $dist_{cst}$ is the L2 distance. The landmarks are stable and reliable when the landmarks extracted in $Y^{'}$ can be consistent with the landmarks extracted in Y by inverse affine transformation. This is as described in [17].

3.7. Landmark Auxiliary Guided Segmentation

The total loss for model training is

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{seg} + \lambda_2\mathcal{L}_{adv} + \lambda_3(\mathcal{L}_{ctr} + \mathcal{L}_{cst}) + \lambda_4\mathcal{L}_{lmd}, \tag{8}$$

where λ_1, λ_2, λ_3 and λ_4 are the balance coefficients of corresponding loss. $\mathcal{L}_{seg}$ is the pixel-level loss for the segmentation task.

$$\mathcal{L}_{seg} = BCE(Y_{pre}, Y) + DICE(Y_{pre}, Y) \tag{9}$$

where BCE and $DICE$ are the binary cross-entropy (BCE) loss and dice loss, respectively. $\mathcal{L}_{adv}$ is the adversarial loss as the global loss for segmentation.

$$\mathcal{L}_{adv} = \mathbb{E}_{Y}[\log \mathcal{D}(Y)] + \mathbb{E}_{Y'}\left[\log(1-\mathcal{D}(Y^{'}))\right] \tag{10}$$

where $\mathcal{D}$ is the discriminator. $\mathcal{L}_{seg}$ and $\mathcal{L}_{adv}$ constrain the segmentation of the model locally and globally, respectively. $\mathcal{L}_{lmd}$ is the landmark-based auxiliary loss based on optimal transport theory. We use the obtained landmarks based on ground truth Y as pseudo-labels, i.e., $\mathcal{R}^{Y}$. $\mathcal{L}_{lmd}$ is defined as

$$\mathcal{L}_{lmd} = \|\mathcal{R} - \mathcal{R}^{Y}\|_2^2 \tag{11}$$

where $\mathcal{R}$ is the landmark-obtained base on X. $\mathcal{L}_{lmd}$ can guide the convolutional encoder to learn more effective information.

Further, we map the landmark information into a Gaussian map $\mathcal{M}_{gm}$ that is easier to embed in the network, and feed it to the convolutional decoder in order to boost the performance of the decoder. The Gaussian map is defined as

$$\mathcal{M}_{gm} = \exp\left(-\frac{1}{2\sigma^2}\left\|\mathcal{R} - \mathcal{R}^{Y}\right\|^2\right), \tag{12}$$

where the standard deviation σ is set to 0.7 for all the experiments. Then, $\mathcal{M}_{gm}$ is connected with $\mathcal{M}_t$ channel-wise as the input of $\mathcal{D}$.

As $\mathcal{D}$ accepts the input composed of $\mathcal{M}_t$ and $\mathcal{M}_{gm}$, it benefits from both global and local high-level semantic information extracted earlier. In particular, $\mathcal{M}_{gm}$ is an explicit sparse representation of the anatomical topology. Additionally, $\mathcal{M}_{gm}$ is a further disentangled representation of the anatomical topology. In addition, $\mathcal{M}_{gm}$ provides the model with prior topological constraints, which enrich the semantics of the data.

3.8. Implementation Details

Considering that there are many tiny blood vessels in the retinal vascular structure, excessively deep convolutional layers may cause some features that are beneficial for segmentation to be ignored. Therefore, we only take the first three layers of the U-Net network and integrate the transformer module into the network in the third layer. Apart from that, the encoders in the semantic segmentation network share the same weights as those under the contrastive learning framework. After conducting experiments, we determined that the values of λ_1, λ_2, λ_3 and λ_4 should be set to 0.2, 0.3, 0.4, and 0.1, respectively.

During the training, instead of patches, we input the entire image into the model to generate the retinal vessel prediction map. We adopt Adam to optimize the deep model with an initial learning rate of 0.001 and a weight decay of 0.0005. Due to GPU memory constraints, we only input one retinal vessel image per iteration and resize all training images to 512×512 pixels. All models used in the experiments are implemented using pytorch-based python programs. They run on a computer configured with RTX3090 GPU.

4. Results and Discussions

4.1. Evaluation Metrics

The retinal vessel segmentation problem can be viewed as a binary classification. All pixels in retinal images can be classified into vascular and non-vascular pixels. Therefore, four definitions are derived according to the classification results of blood vessels. Those correctly classified as vascular pixels are regarded as true positives (TP). Those correctly detected as non-vascular pixels are counted as true negatives (TN). Those misclassified as non-vascular pixels are recorded as false positives (FP). Non-vascular pixels falsely detected as vascular pixels are counted as false negatives (FN).

To validate the feasibility of our designed network, we introduce four metrics of accuracy (Acc), sensitivity (Se), specificity (Sp), and F1 score to evaluate our network. Among them, the F1 score, as a trade-off between sensitivity and specificity, dominates the performance evaluation.

4.2. Comparison with the State-of-the-Art Methods

4.2.1. Quantitative Analysis

We compare our method with other state-of-the-art methods on the DRIVE, CHASE, and STARE datasets. The experimental evaluation indicators are shown in Table 1. It is evident that our method achieves leading F1 scores in all three datasets. For the DRIVE dataset, the Se, Sp, Acc, and F1 scores obtained by our proposed method are 0.9577, 0.8147, 0.9862, and 0.8329, respectively. Jiang et al.'s method [18] obtained the highest Acc and Sp scores, but only 0.7839 and 0.8246 for Se and F1. These are far lower than our results, and our Sp is only 0.0028 lower than theirs, which can be negligible. In the CHASE dataset, we obtain Acc, Se, Sp, and F1 of 0.9754, 0.8110, 0.9881, and 0.8222, respectively. The best performance metrics obtained by other methods are 0.9670, 0.8329, 0.9813, and 0.8191, respectively. In contrast, our F1 reaches the peak of existing methods. Although the Acc, Se and Sp scores produced by our network are not optimal, these three metrics are also at high levels compared with other methods. On the STARE dataset, our method achieves high Acc, Sp, and F1 results while maintaining the highest Se score. Compared with other methods, these results show that our network has stronger vessel detection ability and stronger generalization ability across different databases.

Table 1. Performance comparison with state-of-the-art methods on the DRIVE, CHASE-DB1 and STARE datasets.

		DRIVE				CHASE-DB1				STARE			
Method	**Year**	**Acc**	**Se**	**Sp**	**F1**	**Acc**	**Se**	**Sp**	**F1**	**Acc**	**Se**	**Sp**	**F1**
U-Net [7]	2015	0.9536	0.7653	0.9811	0.8078	0.9604	0.7870	0.9777	0.7828	0.9588	0.7639	0.9796	0.7817
Orlando et al. [19]	2017	0.9454	0.7897	0.9684	0.7857	0.9467	0.7565	0.9655	0.7332	0.9519	0.7680	0.9738	0.7644
Zhang et al. [20]	2017	0.9466	0.7861	0.9712	0.7953	0.9502	0.7644	0.9716	0.7581	0.9547	0.7882	0.9729	0.7815
Srinidhi et al. [21]	2018	0.9589	0.8644	0.9667	0.7607	0.9474	0.8297	0.9663	0.7189	0.9502	0.8325	0.9746	0.7698
Yan et al. [22]	2018	0.9542	0.7653	0.9818	-	0.9610	0.7633	0.9809	-	0.9612	0.7581	0.9846	-
Xu et al. [23]	2018	0.9557	0.8026	0.9780	0.8189	0.9613	0.7899	0.9785	0.7856	0.9499	0.8196	0.9661	0.7982
Zhuang et al. [24]	2018	0.9561	0.7856	0.9810	0.8202	0.9536	0.7978	0.9818	0.8031	-	-	-	-
Alom et al. [25]	2019	0.9556	0.7792	0.9813	0.8171	0.9634	0.7756	0.9820	0.7928	0.9712	0.8292	0.9862	0.8475
Jin et al. [5]	2019	0.9566	0.7963	0.9800	0.8237	0.9610	0.8155	0.9752	0.7883	0.9641	0.7595	0.9878	0.8143
Jiang et al. [18]	2019	0.9709	0.7839	0.9890	0.8246	0.9721	0.7839	0.9894	0.8062	0.9781	0.8249	0.9904	0.8482
Guo et al. [26]	2019	0.9561	0.7891	0.9804	0.8249	0.9627	0.7888	0.9801	0.7983	-	-	-	-
Wang et al. [27]	2019	0.9567	0.7940	0.9816	0.8270	0.9661	0.8074	0.9821	0.8037	-	-	-	-
Zhou et al. [28]	2020	0.9535	0.7473	0.9835	0.8035	0.9506	0.6361	0.9894	0.7390	0.9605	0.7776	0.9832	0.8132
Xu et al. [29]	2020	0.9557	0.7953	0.9807	0.8252	0.9650	0.8455	0.9769	0.8138	0.9590	0.8378	0.9741	0.8308
Wang et al. [30]	2020	0.9581	0.7991	0.9813	0.8293	0.9670	0.8329	0.9813	0.8191	0.9673	0.8186	0.9844	-
Li et al. [9]	2020	0.9573	0.7735	0.9838	0.8205	0.9760	0.7969	0.9881	0.8072	0.9701	0.7715	0.9886	0.8146
Mou et al. [31]	2021	0.9553	0.8154	0.9757	0.8228	0.9651	0.8329	0.9784	0.8141	0.9670	0.8396	0.9813	0.8420
Zhang et al. [32]	2022	0.9565	0.785	0.9618	0.82	-	-	-	-	0.9668	0.8002	0.9864	0.8289
Liu et al. [33]	2023	0.9561	0.7985	0.9791	0.8229	0.9672	0.8020	0.9794	0.8236	0.9635	0.8039	0.9836	0.8315
Proposed	2023	0.9577	0.8147	0.9862	**0.8329**	0.9754	0.8110	0.9881	**0.8222**	0.9635	0.8518	0.9829	**0.8450**

4.2.2. Qualitative Analysis

Figure 2 shows the results of retinal vessel segmentation using several representative methods and our proposed method. The results show that our proposed method preserves almost all the structures of retinal vessels and guarantees the connectivity of the vessel tree. In addition, the model can clearly segment from the background thin blood vessels that cannot be segmented by other methods, especially at the retinal edge and vessel ends. To more clearly show the difference between the prediction results of other network models and our network model, we visualize the local segmentation results of the model and color-label the different segmentation cases. Blue pixels in the image represent false negatives from undetected vessel regions. Red pixels represent false positives, indicating over-segmentation of blood vessels. It is evident from the patches in Figure 2 that the predicted segmentation maps of other methods show more blue pixels. This further proves that our proposed model has certain advantages in detecting thin blood vessels.

Some segmentation examples are given in Figure 3, which contains locally enlarged images of the original retinal images, the corresponding ground truth values, and segmentation prediction maps obtained by several other methods and our proposed method. As can be seen from Figure 3, our algorithm can detect thin blood vessels more clearly and ensure connectivity between blood vessels.

These experimental data demonstrate that our model can more accurately distinguish vascular and non-vascular pixels and preserve vascular structure better.

4.3. Ablation Experiments

In this paper, we introduce the TransUNet structure and self-supervised landmark detection to improve retinal vessel segmentation performance. To test the effectiveness of these modules, ablation experiments are performed on DRIVE, STARE and CHASE-DB1. We start with the original U-Net method to evaluate how these modules affect segmentation performance. The self-supervised landmark detection is denoted by SLD. The results are shown in Table 2. For simplicity, we only visualize a few of the most representative instance images.

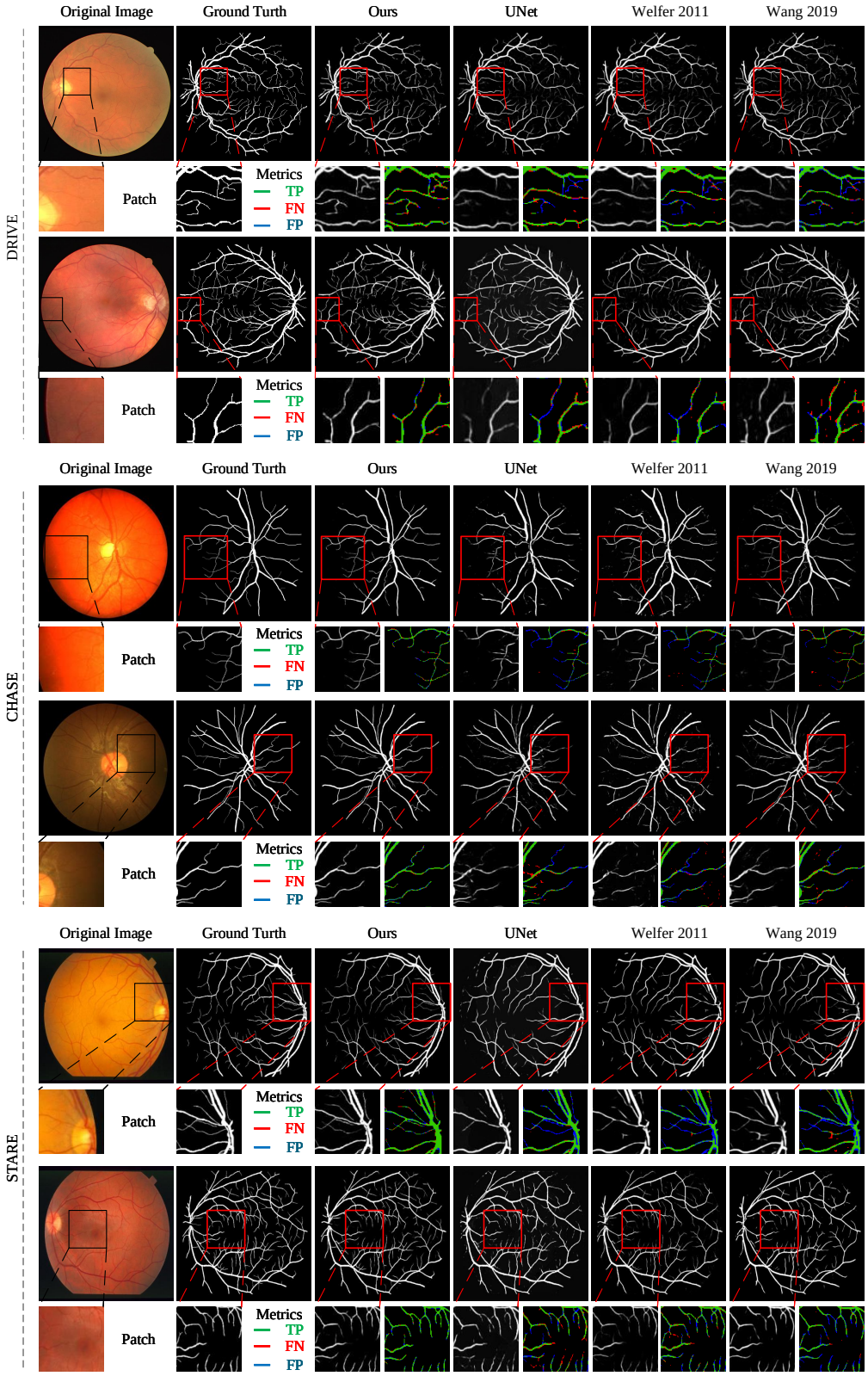

Figure 2. Examples of retinal vessel segmentation for three datasets (Welfer 2011 [4]; Wang 2019 [27]).

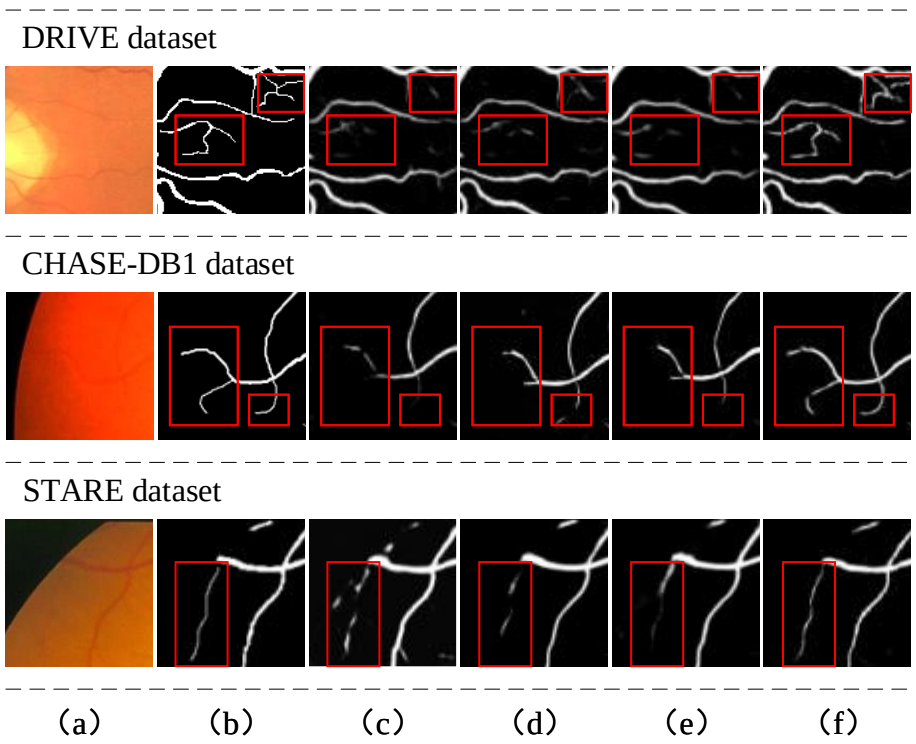

Figure 3. Locally magnified view of the segmentation results: (**a**) raw fundus image, (**b**) ground truth, (**c**) U-Net, (**d**) Jin 2019 [5], (**e**) Zhou 2020 [28], (**f**) our method.

Table 2. Ablation studies on the DRIVE, CHASE-DB1 and STARE datasets.

	DRIVE				CHASE-DB1				STARE			
Method	**Acc**	**Se**	**Sp**	**F1**	**Acc**	**Se**	**Sp**	**F1**	**Acc**	**Se**	**Sp**	**F1**
U-Net	0.9536	0.7653	0.9811	0.8078	0.9604	0.7870	0.9777	0.7828	0.9588	0.7639	0.9796	0.7817
TransUNet	0.9543	0.7874	0.9860	0.8148	0.9681	0.7994	0.9878	0.8079	0.9610	0.7670	0.9879	0.8057
TransUNet + SLD	0.9577	0.8147	0.9862	0.8329	0.9754	0.8110	0.9881	0.8222	0.9635	0.8518	0.9829	0.8450

4.3.1. Effect of TransUNet

To demonstrate the feasibility of the proposed TransUNet structure, we compare the U-Net network with the U-Net with transformer embedded. The same configuration and environment were used for both experiments. The results show that we achieve 0.9543, 0.7874, 0.9860, and 0.8148 on the DRIVE dataset for Acc, Se, Sp, and F1, respectively, and 0.9536, 0.7653, 0.9811, and 0.8078 on the baseline model for Acc, Se, Sp, and F1, respectively. At the same time, the performance on the other two datasets is also improved. Additionally, from the visualization in Figure 4, we can observe that the TransUNet structure can fully help the network to learn more feature information that ensures the connectivity of blood vessels.

4.3.2. Effect of Self-Supervised Landmark Detection

To justify the use of landmark points to guide network segmentation, in Figure 5, we show an example visualization including the original retinal image and the style-transformed image, ground truth, and the affine-transformed ground-truth image of the DRIVE dataset.

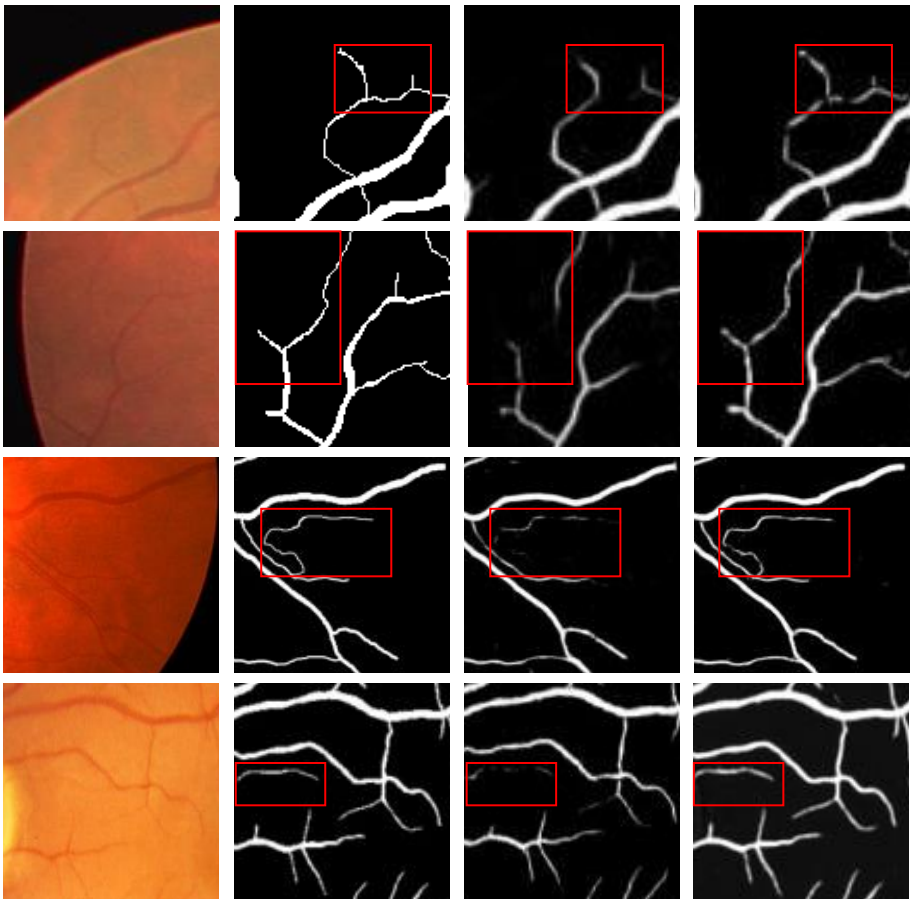

Figure 4. Illustration of vessel connectivity: (**a**) the retinal fundus patches, (**b**) ground truth, (**c**) segmentation output from U-Net, (**d**) segmentation output from TransUNet. First row and second row: DRIVE dataset, third row: CHASE-DB1 dataset, fourth row: STARE dataset.

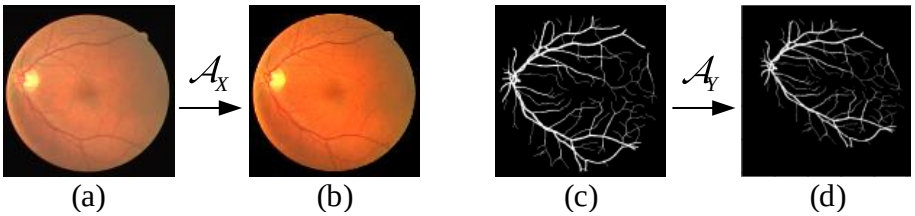

Figure 5. Example of transformation: (**a**) original retinal image, (**b**) style-transformed retinal image, (**c**) ground truth, (**d**) ground truth image after affine transformation.

The affine transformation matrix is shown in (13).

$$\mathcal{A} = \begin{pmatrix} 0.90411 & 0.17613 & 0 \\ 0.05871 & 0.82583 & 0 \end{pmatrix}, \tag{13}$$

According to Table 2, it can be observed that the segmentation results with the addition of the self-supervised cues show improvements on all three datasets to varying degrees. Furthermore, in the visualization results shown in Figure 6, the segmentation guided by the self-supervised cues demonstrates superior performance in segmenting small blood vessels.

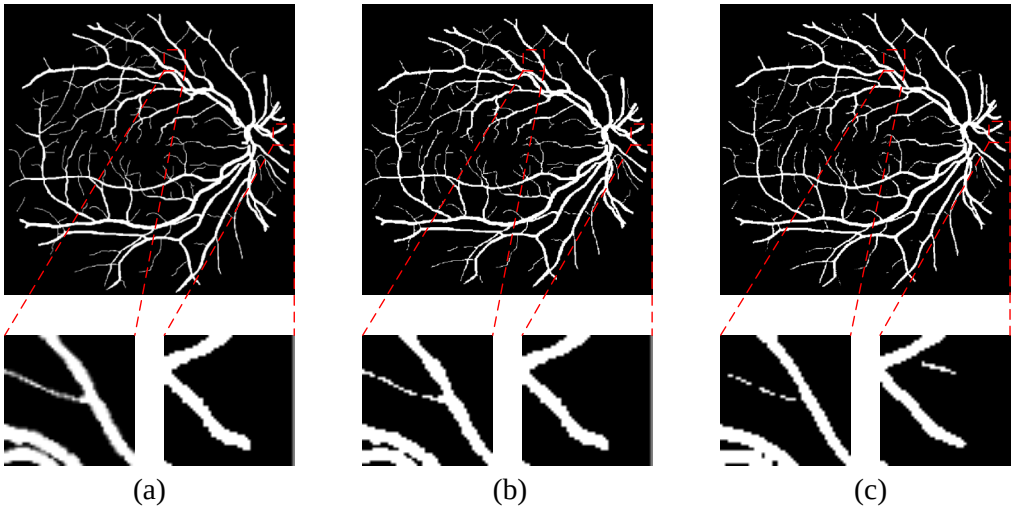

Figure 6. Illustration of thin vessel segmentation results: (**a**) ground truth, (**b**) segmentation results of the network without the self-supervised landmark detection module, (**c**) segmentation results of our method.

Therefore, our proposed landmark detection module can help us detect thin blood vessels more accurately.

4.4. Effect of Image Size

As is customary in most works, we initially resized all training images to dimensions of 512 × 512 pixels. However, inspired by the findings in work [34] regarding the impact of image size on deep learning, we conducted an additional evaluation. We resized the images to a dimension of 256 × 256 pixels and performed training accordingly. As shown in Table 3, the adjusted F1 scores and other metrics exhibited improvements. Moreover, as illustrated in Figure 7, the visualizations demonstrate that the segmented vessels became more intact.

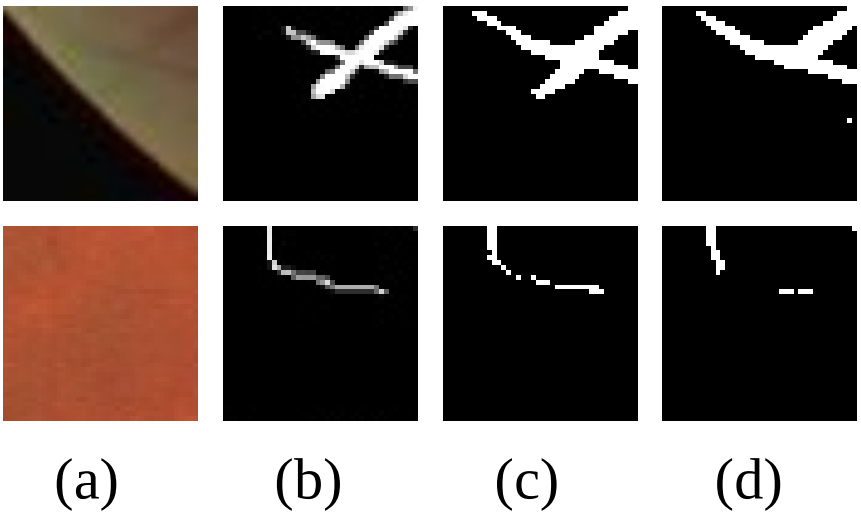

Figure 7. Sample segmentation results for small blood vessels in images of different sizes: (**a**) original retinal image, (**b**) ground truth, (**c**) segmentation results for an input image of size 256 × 256 pixels, (**d**) segmentation results for an input image of size 512 × 512 pixels.

Table 3. Segmentation results for images of different sizes on the DRIVE, CHASE-DB1 and STARE datasets.

	DRIVE				CHASE-DB1				STARE			
Size	**Acc**	**Se**	**Sp**	**F1**	**Acc**	**Se**	**Sp**	**F1**	**Acc**	**Se**	**Sp**	**F1**
512×512	0.9577	0.8147	0.9862	0.8329	0.9754	0.8110	0.9881	0.8222	0.9635	0.8518	0.9829	0.8450
256×256	0.9688	0.8188	0.9869	0.8455	0.9685	0.8155	0.9889	0.8243	0.9641	0.8188	0.9888	0.8466

5. Conclusions

In this paper, we construct a novel retinal vessel segmentation framework, aiming to address the problems of vessel breakage and low accuracy of thin vessels in segmentation. The U-Net acts as the basic network. The designed TransUNet structure combines context information of different scales in the process of encoding and decoding, which effectively ensures the connectivity of blood vessels. The detected landmarks sparsely represent the anatomical features of retinal blood vessels, and segmentation guided by landmarks can help the network better detect thin blood vessels. Experimental results on three public datasets demonstrate that our constructed network outperforms the existing mainstream networks. In the future, we will conceive more methods to integrate into the retinal segmentation network.

Author Contributions: Conceptualization, H.Z. and Y.Q. and C.S.; methodology, H.Z. and Y.Q. and J.L.; software, H.Z.; validation, H.Z., Y.Q. and J.L.; formal analysis, H.Z., Y.Q. and C.S.; investigation, H.Z.; resources, H.Z.; data curation, Y.Q.; writing—original draft preparation, J.L.; writing—review and editing, H.Z.; visualization, Y.Q.; supervision, C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No additional data are available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Grélard, F.; Baldacci, F.; Vialard, A.; Domenger, J.P. New methods for the geometrical analysis of tubular organs. *Med. Image Anal.* **2017**, *42*, 89–101. [CrossRef] [PubMed]
2. Saroj, S.K.; Kumar, R.; Singh, N.P. Fréchet PDF based Matched Filter Approach for Retinal Blood Vessels Segmentation. *Comput. Methods Programs Biomed.* **2020**, *194*, 105490. [CrossRef] [PubMed]
3. Mapayi, T.; Owolawi, P.A. Automatic Retinal Vascular Network Detection using Multi-Thresholding Approach based on Otsu. In Proceedings of the 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC), Vanderbijlpark, South Africa, 21–22 November 2019; pp. 1–5.
4. Welfer, D.; Scharcanski, J.; Marinho, D.R. Fovea center detection based on the retina anatomy and mathematical morphology. *Comput. Methods Programs Biomed.* **2011**, *104*, 397–409. [CrossRef] [PubMed]
5. Jin, Q.; Meng, Z.; Pham, T.D.; Chen, Q.; Wei, L.; Su, R. DUNet: A deformable network for retinal vessel segmentation. *Knowl. Based Syst.* **2019**, *178*, 149–162. [CrossRef]
6. Wu, C.; Zou, Y.; Yang, Z. U-GAN: Generative Adversarial Networks with U-Net for Retinal Vessel Segmentation. In Proceedings of the 2019 14th International Conference on Computer Science & Education (ICCSE), Toronto, ON, USA, 19–21 August 2019; pp. 642–646.
7. Ronneberger, O.; Fischer, P.; Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*; Springer: Cham, Switzerland, 2015.
8. Wang, B.; Wang, S.; Qiu, S.; Wei, W.; Wang, H.; He, H. CSU-Net: A Context Spatial U-Net for Accurate Blood Vessel Segmentation in Fundus Images. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 1128–1138. [CrossRef] [PubMed]
9. Li, L.; Verma, M.; Nakashima, Y.; Nagahara, H.; Kawasaki, R. IterNet: Retinal Image Segmentation Utilizing Structural Redundancy in Vessel Networks. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 3645–3654.

10. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. In *Computer Vision—ECCV 2022 Workshops*; Springer: Cham, Switzerland, 2023.
11. Xia, X.; Huang, Z.; Huang, Z.; Shu, L.; Li, L. A CNN-Transformer Hybrid Network for Joint Optic Cup and Optic Disc Segmentation in Fundus Images. In Proceedings of the 2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI), Shijiazhuang, China, 22–24 July 2022; pp. 482–486.
12. 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* **2021**, arXiv:2102.04306.
13. Sun, Y.; Wang, X.; Tang, X. Deep Convolutional Network Cascade for Facial Point Detection. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3476–3483.
14. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
15. Kowalski, M.; Naruniec, J.; Trzcinski, T. Deep Alignment Network: A Convolutional Neural Network for Robust Face Alignment. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 2034–2043.
16. Shi, H.; Wang, Z. Improved Stacked Hourglass Network with Offset Learning for Robust Facial Landmark Detection. In Proceedings of the 2019 9th International Conference on Information Science and Technology (ICIST), Kopaonik, Serbia, 10–13 March 2019; pp. 58–64.
17. Siarohin, A.; Lathuiliere, S.; Tulyakov, S.; Ricci, E.; Sebe, N. Animating arbitrary objects via deep motion transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2377–2386.
18. Jiang, Y.; Tan, N.; Peng, T.; Zhang, H. Retinal Vessels Segmentation Based on Dilated Multi-Scale Convolutional Neural Network. *IEEE Access* **2019**, *7*, 76342–76352. [CrossRef]
19. Orlando, J.I.; Prokofyeva, E.; Blaschko, M.B. A Discriminatively Trained Fully Connected Conditional Random Field Model for Blood Vessel Segmentation in Fundus Images. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 16–27. [CrossRef]
20. Zhang, J.; Chen, Y.; Bekkers, E.; Wang, M.; Dashtbozorg, B.; ter Haar Romeny, B.M. Retinal vessel delineation using a brain-inspired wavelet transform and random forest. *Pattern Recognit.* **2017**, *69*, 107–123. [CrossRef]
21. Srinidhi, C.L.; Aparna, P.; Rajan, J. A visual attention guided unsupervised feature learning for robust vessel delineation in retinal images. *Biomed. Signal Process. Control.* **2018**, *44*, 110–126. [CrossRef]
22. 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.* **2018**, *65*, 1912–1923. [CrossRef] [PubMed]
23. Xu, R.; Jiang, G.; Ye, X.; Chen, Y. Retinal vessel segmentation via multiscaled deep-guidance. In *Pacific Rim Conference on Multimedia*; Springer: Berlin, Germany, 2018; pp. 158–168.
24. Zhuang, J. LadderNet: Multi-Path Networks Based on U-Net for Medical Image Segmentation. *arXiv* **2018**, arXiv:1810.07810.
25. Alom, M.Z.; Yakopcic, C.; Hasan, M.; Taha, T.M.; Asari, V.K. Recurrent residual U-Net for medical image segmentation. *J. Med. Imaging* **2019**, *6*, 14006. [CrossRef] [PubMed]
26. Guo, S.; Wang, K.; Kang, H.; Zhang, Y.; Gao, Y.; Li, T. BTS-DSN: Deeply Supervised Neural Network with Short Connections for Retinal Vessel Segmentation. *Int. J. Med. Inform.* **2018**, *126*, 105–113. [CrossRef]
27. Wang, B.; Qiu, S.; He, H. Dual encoding U-Net for retinal vessel segmentation, Medical Image Computing and Computer Assisted Intervention. *Med. Image Comput. Comput. Assist. Interv.* **2019**, *22*, 84–92.
28. 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* **2020**, *39*, 1856–1867. [CrossRef]
29. Xu, R.; Ye, X.; Jiang, G.; Liu, T.; Tanaka, S. Retinal Vessel Segmentation via a Semantics and Multi-Scale Aggregation Network. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020.
30. Wang, D.; Haytham, A.; Pottenburgh, J.; Saeedi, O.; Tao, Y. Hard attention net for automatic retinal vessel segmentation. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 3384–3396. [CrossRef]
31. Mou, L.; Zhao, Y.; Fu, H.; Liu, Y.; Cheng, J.; Zheng, Y.; Su, P.; Yang, J.; Chen, L.; Frangi, A.F.; et al. CS2-Net: Deep learning segmentation of curvilinear structures in medical imaging. *Med. Image Anal.* **2021**, *67*, 101874. [CrossRef]
32. Zhang, Y.; He, M.; Chen, Z.; Hu, K.; Li, X.; Gao, X. Bridge-Net: Context-involved U-Net with patch-based loss weight mapping for retinal blood vessel segmentation. *Exp. Syst. Appl.* **2022**, *195*, 116526. [CrossRef]
33. Liu, Y.; Shen, J.; Yang, L.; Bian, G.; Yu, H. ResDO-UNet: A deep residual network for accurate retinal vessel segmentation from fundus images. *Biomed. Signal Process. Control.* **2023**, *79*, 104087. [CrossRef]
34. Rukundo, O. Effects of image size on deep learning. *Electronics* **2023**, *12*, 985. [CrossRef]