electronics

MDPI

Article

A Multi-Task Dense Network with Self-Supervised Learning for Retinal Vessel Segmentation

Zhonghao Tu [1], Qian Zhou [1], Hua Zou [1,*] and Xuedong Zhang [2]

[1] School of Computer Science, Wuhan University, Wuhan 430072, China
[2] School of Information Engineering, Tarim University, Alaer 843300, China
* Correspondence: zouhua@whu.edu.cn

Abstract: Morphological and functional changes in retinal vessels are indicators of a variety of chronic diseases, such as diabetes, stroke, and hypertension. However, without a large number of high-quality annotations, existing deep learning-based medical image segmentation approaches may degrade their performance dramatically on the retinal vessel segmentation task. To reduce the demand of high-quality annotations and make full use of massive unlabeled data, we propose a self-supervised multi-task strategy to extract curvilinear vessel features for the retinal vessel segmentation task. Specifically, we use a dense network to extract more vessel features across different layers/slices, which is elaborately designed for hardware to train and test efficiently. Then, we combine three general pre-training tasks (i.e., intensity transformation, random pixel filling, in-painting and out-painting) in an aggregated way to learn rich hierarchical representations of curvilinear retinal vessel structures. Furthermore, a vector classification task module is introduced as another pre-training task to obtain more spatial features. Finally, to make the segmentation network pay more attention to curvilinear structures, a novel dynamic loss is proposed to learn robust vessel details from unlabeled fundus images. These four pre-training tasks greatly reduce the reliance on labeled data. Moreover, our network can learn the retinal vessel features effectively in the pre-training process, which leads to better performance in the target multi-modal segmentation task. Experimental results show that our method provides a promising direction for the retinal vessel segmentation task. Compared with other state-of-the-art supervised deep learning-based methods applied, our method requires less labeled data and achieves comparable segmentation accuracy. For instance, we match the accuracy of the traditional supervised learning methods on DRIVE and Vampire datasets without needing any labeled ground truth image. With elaborately training, we gain the 0.96 accuracy on DRIVE dataset.

Keywords: self-supervised learning; vessel segmentation; multi-task network; aggregation task

check for updates

Citation: Tu, Z.; Zhou, Q.; Zou, H.; Zhang, X. A Multi-Task Dense Network with Self-Supervised Learning for Retinal Vessel Segmentation. *Electronics* **2022**, *11*, 3538. https://doi.org/10.3390/electronics11213538

Academic Editor: Jichai Jeong

Received: 13 September 2022
Accepted: 11 October 2022
Published: 30 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Fundus image analysis during clinical trials can effectively assist ophthalmologists in the diagnosis and treatment of diseases [1]. During the analysis, retinal vessel segmentation plays an important role in the prediction, diagnosis, and treatment. Clinicians can analyze the patient's condition by observing certain characteristics of retinal vessels, such as the thickness of the arterial wall, the diameter of the arteries and veins, the crossing angle of the vessels, and the density of new capillaries. However, due to the large intra-individual variation and inter-individual morphological variation, precise retinal vessel segmentation has always been a challenge in the field of medical image processing.

With the success of deep learning in medical image analysis, medical image segmentation algorithms based on deep neural networks have also gained an advantage over traditional methods. Most existing medical image segmentation approaches are based on the U-net [2] structure. U-net obtains a hierarchical feature set by encoding the original image in a down-sampling process. Then, U-net adopts decoding operations to recover the hierarchical feature set into the original image. One of the most important characteristics

of the U-net structure is the skipping connection mechanism, in which the features from the encoder layers are fed into the corresponding decoder layers. The skip connections can fuse features from both deep and shallow layers to retain detailed information.

Most deep neural networks, including U-net, require massive high-quality labeled data during the training stage. However, very few fundus image datasets can offer extensive pixel-level labels, and the quality of labels is highly variable. High-quality annotations require many experts to spend a lot of time on them, which is high-cost and time-consuming. In order to solve this problem, using self-supervised learning to learn from unlabeled data has attracted great interest. However, due to the complex distribution of retinal capillaries in fundus images, the morphology of retinal fundus vessels varies greatly from patient to patient in terms of size, thickness, and orientation. This also degrades the accuracy of existing self-supervised learning-based segmentation methods when applied to retinal vessels.

Overall, in this paper, we propose a retinal vessel segmentation network based on an aggregation task model of self-supervised learning for fundus images. The main contributions are listed as follows:

(1) We improve the U-net structure by passing different layers of feature maps to each other to meet the requirements of fundus image detail feature extraction. Benefiting from this improved network structure, network pruning can also be used to improve the segmentation speed during the process, and the network will automatically fit the capabilities of the hardware.

(2) In order to address the problem that the features acquired by self-supervised learning methods are not rich enough, an aggregation task strategy is designed to acquire more global and detailed features of the fundus image during the pre-training process. Specifically, three pretext tasks including intensity transformation, random pixel filling, in-painting and out-painting are combined in an aggregated way.

(3) A vector classification task module is introduced to generate different vector routes, in which the encoder is trained to obtain spatially correlated features of fundus images by predicting vector routes from the network.

The rest of this paper is organized as follows: We further discuss the related research about retina vessel segmentation in Related works. The network details are discussed in Method alongside the training procedure. In Experiments and Results, experimental results on Drive, VAMPIRE, iChallenge, and STARE datasets demonstrate our methods can gain more vessel detail than other methods. Finally, we conclude this paper in Conclusion.

2. Related works

Exiting retina vessel segmentation algorithms can be divided into three categories, i.e., traditional machine learning segmentation methods, supervised deep learning segmentation methods, and self-supervised deep learning segmentation methods.

Vessel segmentation methods based on traditional machine learning.Traditional vessel segmentation methods usually follow these steps: firstly, a fundus image is pre-processed, then feature vector space is built based on the acquired blood vessel features, and a suitable classifier is selected to extract blood vessels. He et al. [3] first defined the feature vector of 39-dimensional blood vessels to train the AdaBoost classifier to obtain a strong classifier, and then realized the segmentation process of retinal blood vessels. Fraz et al. [4] introduced gradient vector field operations, morphological transformation, line intensity measure, and Gabor filter response to construct feature vectors and used a decision tree classifier for vessel segmentation. Roy et al. [5] discussed various issues related to the implementation of matched filters and constructed twelve vessel segmentation matchers for vessel segmentation.

Yue et al. [6] designed a network for retinal vessel segmentation by using an attention map to guide the vessel extraction process. Xu et al. [7] achieved segmentation by combining the Hessian matrix and Canny operator to track coronary arteries.

Vessel segmentation methods based on deep learning.The traditional machine learning segmentation methods are easy to implement, but their generalization ability is insufficient. Therefore, the results of traditional methods are usually worse than desired. With the facilitation of deep learning in computer vision, it has also achieved fruitful results in medical image segmentation. In 2015, Long et al. [8] proposed a Fully Convolutional Network (FCN) and made a breakthrough in segmentation. Since then, image segmentation using convolutional neural networks has gradually become a trend. Ronneberger et al. [2] proposed U-Net with excellent results for medical image segmentation. U-net follows the encoder–decoder structure and can be trained end to end. During the segmentation stage, U-net obtains lots of semantic features, which lead to state-of-the-art performance. Wang et al. [9] proposed the fusion net to divide retina images into easy and hard regions by combining attention networks and U-net, in which the two different regions are processed by two U-nets separately. Wang et al. [10] proposed a two-channel U-net network for retinal vessel segmentation. In the two-channel U-net network, one branch uses a large kernel to acquire spatial information and the other branch acquires the content feature of the retina image. Yue et al. [6] proposed an optimized U-Net model. The authors added multiple input layers into the traditional U-net model in order to learn rich details from medical images. Ma et al. [11] proposed a novel coarse-to-fine Optical Coherence Tomography Angiography (OCTA) image vessel segmentation network (OCTA-Net). The network generates an initial map of vessel location information by the coarse segmentation module and uses the refine segmentation module to optimize the shape/outline of tiny capillaries in fundus images.

Vessel segmentation methods based on self-supervised learning.Deep learning methods generally rely on a large number of labeled data, but most medical image data are unlabeled. Self-supervised learning methods design pre-training tasks to extract features from unlabeled data, reducing the reliance on labeled data. Zhu et al. [12] proposed a self-supervised learning framework called "Cube+" to pre-train the network. The authors defined three pre-training operations (i.e., cube sorting, cube rotation, and cube masking) to force the network to learn spatial and content features from the original 3D medical images. In 2021, Zhou et al. [13] proposed a self-supervised pre-training model based on 3D medical image metadata. The authors combined four optional graph transformation methods into a self-supervised learning framework: nonlinear transformation, local pixel reconstruction, out-painting, and in-painting. The original image is fed into a U-Net (or V-Net) structure after a series of transformations, then the network is trained to recover the original image and learn contextual features. In 2021, Chen et al. [14] proposed a method for the detection of COVID-19 by momentum contrast learning. Contrast learning classified images into positive pairs and negative images. The network is trained to minimize the distance between the positive pairs and maximize the distance between negative pairs.

3. Method

Self-supervised learning methods can reduce the reliance on labeled data as they can learn from unlabeled data. For retinal vessel segmentation, self-supervised learning can achieve even better performance than the supervised deep learning methods [15]. Existing supervised segmentation methods are weak in feature extraction due to the limited high-quality annotations. Directly applying self-supervised learning in retinal vessel segmentation may not be suitable, since medical images are quite different from natural images. For retinal images, there are many microvessels that are easily ignored by existing methods. To address this issue, we propose a dense network that uses the aggregation task model [16] to combine multiple self-supervised learning tasks. The learned spatial and semantic contextual information is aggregated through the self-supervised pre-training. Besides, the pre-training task uses an optimized U-net network structure to preserve more details during feature extraction. In such a design, our network can completely learn the ocular features and utilize more detailed information.

3.1. Network Architecture

The whole network architecture follows the U-Net [2]. Unlike existing U-net models which simply concatenate the feature maps of the same size on the channel dimension, our network aggregates feature maps across different layers through dense connections. As can be seen in Figure 1, our network consists of three modules: the encoder module, the fusion module, and the decoder module. The encoder module and decoder module from our network has the same structure from the traditional U-net, with four convolutional blocks in the encoder and four convolutional blocks in the decoder. Our dense network is an encoder and decoder structure, but each convolutional block contains a fusion module, which is different from original U-net. Our network encoder module takes color fundus or fluorescence fundus angiography (FFA) images as input, and the decoder output is the same size picture as the input. The fusion module transforms the feature maps at different layers into the same size by using convolution operation with different kernel sizes. Then, the transformed features are concatenated and fed into the next layer. The fusion network (denoted by F) has three inputs: the features of the previous layer, the pre-trained features of the same layer, and all features from previous convolutional blocks. The output of the fusion module is the concated feature map. The fusion network ensures that the encoder and decoder can leverage features of different layers, thus enhancing the robustness of the network.

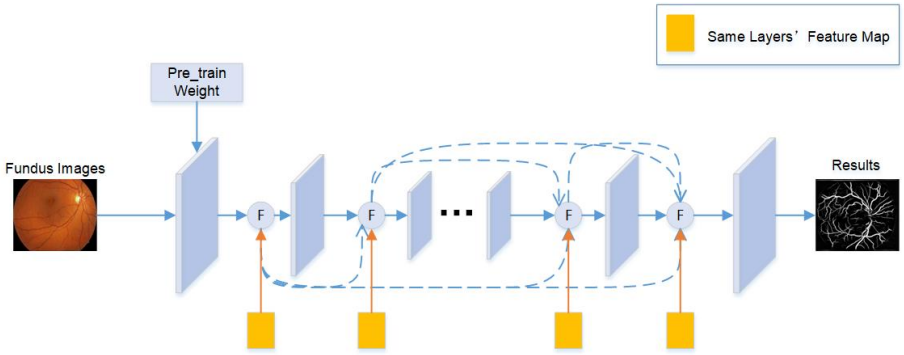


Figure 1. Vessel segmentation network for fundus images based on aggregation task.

The input of the whole network can be represented by Equation (1), in which CB represents the convolutional block, and the input of the network is retina image or FFA as G_{input}, the convolutional size we used is 3×3, the initial kernel size is 64. Then, three successive same size convolution block are used for processing retinal images, and convolutional kernels are halved layer by layer. The feature extraction module finally obtains the feature map G' by processing the convolutional layer with filter size 3×3 and the final convolutional kernels is 16. Then, the encoder stage is processed by the convolutional layer with the filter size of 3×3 and convolutional kernel is 16, and then the final image result is obtained by successively using the filter size of 3×3 and doubling the number of convolutional kernels with the number of layers, converting the number of image channels to 3, and the final result of the vessel segmentation image G_{result} is shown in Equation (2).

$$G' = f^{(3\times3)}(CB_{total}(CB_{(total-1)}(\ldots CB_1(f^{(3\times3)}(G_{input}))\ldots))) \quad (1)$$

$$G_{result} = f^{(3\times3)}(CB_{total}(CB_{total-1}(\ldots CB_1(f^{(3\times3)}(G'))\ldots))) \quad (2)$$

Original U-net only aggregates the feature map from the same size, in which the features from different layers can not be well inherited, reducing the encoder–decoder network structure segmentation ability. In this paper, we optimize the original U-net structure into a Dense U-net structure. Motivated by the idea of Dense-Net [17], we add dense connections to fuse the feature maps from different layers. Since the feature sizes are

different, we introduce another convolutional block to refine them. The network data flow diagram is shown in Figure 2.

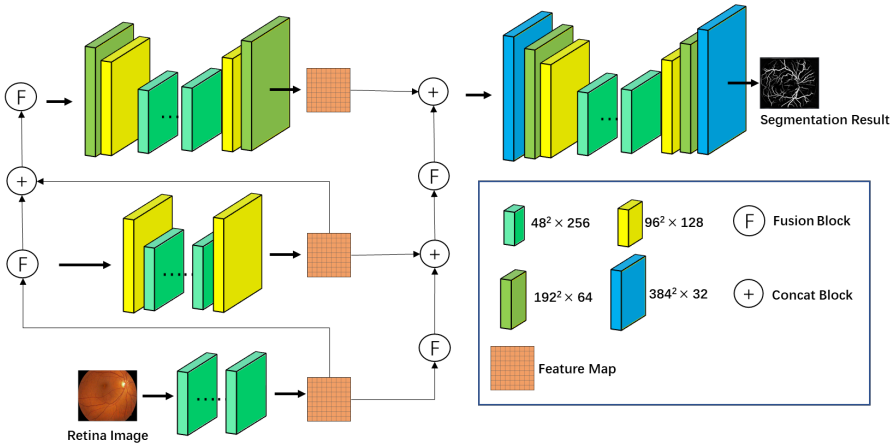


Figure 2. Pre-training network data flow diagram.

As can be seen in Figure 2, this network structure is designed to fill up the original U-net by using fusion modules, so that the feature maps can be aggregated between the different layers of the U-net. Suppose the input of the encoder of the fourth layer of the network is the feature map F, and the feature map F' is obtained by the 3×3 convolution. The feature map F' is then used as the input to the fusion network, and the new feature map is obtained through a 3×3 deconvolution process, and concat with the feature map F' obtained at the beginning of the same layer, and then passed through a convolution layer with the same parameter settings to obtain the processed feature map F'' as the input to the decoder of the fourth layer. The whole processing flow can be represented by Equations (3) and (4). In which, $f^{(3\times3)}(\cdot)$ denotes a convolutional layer with filter size 3×3.

$$F' = f^{(3\times3)}(F) \tag{3}$$

$$F'' = f^{(3\times3)}((F') \oplus F) \tag{4}$$

3.2. Pre-Training Task Aggregation

To improve the performance, we adopt four pre-training tasks in an aggregated way. How to aggregate multiple pre-training tasks is a big challenge, in which the network is supposed to combine multiple features to fit the retina image segmentation.

The first pre-training task is nonlinear intensity transformation. The retinal images are preprocessed with nonlinear intensity transformation and used as input. The network is trained to reconstruct original images. Since intensity information is the most significant feature information of FFA images, using the nonlinear intensity transformation enables the pre-training network to learn many global features. We use the Bézier curve to perform the nonlinear transformation process as Equation (5). This curve is a smooth and monotonic transformation function. During the transformation process, each pixel of the image is using the same Bézier transform to ensure a one-to-one mapping.

$$B(t) = (1-t)^3 P_0 + 3(1-t)^2 t P_1 + 3(1-t)t^2 P_2 + t^3 P_3, t \in [0,1] \tag{5}$$

in which, t is a constant value, the Bessel curve has two endpoints P_0, P_3 and two control endpoints P_1, P_2. The nonlinear intensity change is accomplished by picking a monotonic function interval. Both color fundus images and FFA images are converted to binary maps and then regularized to $[0, 1]$.

The second pre-training task is random pixel filling. This pre-training task will randomly select a fixed size $m \times n$ window in the real retina image for random pixel filling, then the processed image is used as input, and the output is still the retina image recon-

structed by the network. The random pixel filling process is shown in Equation (6), where $\widetilde{W}$ denotes the transformed window, P and P' denote windows of size $m \times m$ and $n \times n$. Unlike the traditional random pixel filling, the vessel images after transformation are completely different. Our transformation only blurs the edge of the vessel, and the pre-training task needs to accurately restore the edge and texture details of the retina image vessels during the reconstruction process. By local pixel random filling, the pre-training network learns the local details of the fundus image.

$$\widetilde{w} = P \times W \times P' \tag{6}$$

The third pre-training task is in-painting and out-painting. This task takes the randomly masked retinal images as the input and outputs the reconstructed retina images. The specific process of out-painting is that we first randomly generate several windows of different sizes, then overlap them to form a single window with a complex shape. We subsequently assign a random value to all pixels outside the window while preserving the intensity of the original image inside the window. For in-painting, a window in the retina image is randomly selected. The intensity outside the window is retained, and the intensity inside the window is modified. The pre-trained network learns the global features and spatial local features by reconstructing the original images during the out-painting process, and learns the continuity features of the retina vessels when in-painting.

The loss function used in the previous three pre-training tasks is L_1 loss, and the encoder target is to train the network to learn the features of fundus images by calculating the L_1 loss of the input and output of the three different pre-training tasks.

The fourth pre-training task is the vector classification task. Compared to the previous three pre-training tasks, this task focuses on the training of the encoder to improve the feature extraction ability of the encoder [16]. The process of this pre-training task starts with cropping the original image into small 3×3 blocks. As Figure 3 shows that there are a total of $4 \times 6 = 24$ different vectors. Five small blocks from the complete retina image are used as input, and then the Dense U-net is followed by a fully connected layer. The network output is the vector prediction result. The output result is compared with the real vector, thus enabling the network to learn the spatial features of the fundus images.

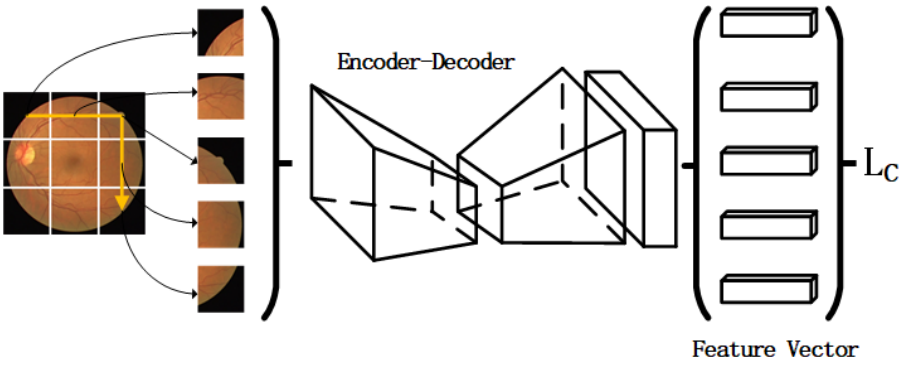


Figure 3. Vector prediction pre-training.

The input process is shown as Equation (7):

$$z_i = f(x_i), i = 1, \ldots, m \tag{7}$$

Specific vectors are selected as Figure 4, and these routes are chosen randomly from the top left corner to the bottom right corner, where equation z_i represents the output feature vector. We take 2D retina images so the value of m is taken as 5. The loss function is shown as Equation (8):

$$L_C = -\frac{\sum_i^{(z_i \neq z_m)} z_m \log z'_m + (1 - z_m) logz(1 - z_m')}{z_i logz_m' + (1 - z_i) logz(1 - z_m')} \tag{8}$$

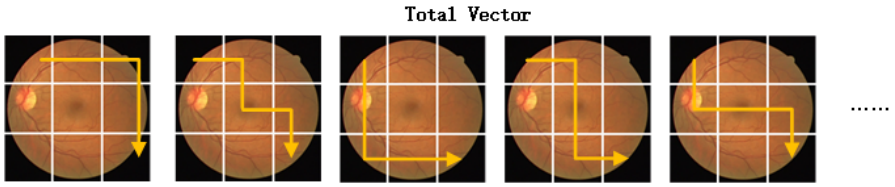


Figure 4. Total vector route.

In summary, the pre-training process is as follows: the images in the dataset are selected, and we begin to use the vector classification tasks to train the encoder then combine the previous three pre-training tasks to train our network, and during this process we train the encoder and decoder in one turn. We designed these four tasks that totally fit the retinal vessel segmentation. Unlike traditional self-supervised tasks such as rotation, we will immediately understand that rotating a retinal image cannot let the pre-training process learn the precise vessel feature from the existing dataset but noisy feature. In the specific training process, we use the same dataset, and each round of training uses the complete network parameters trained in the previous round to continue training, so that the final network can learn all the fundus image features learned by the pre-training tasks in aggregate. After the network is trained, the weights of the encoder and decoder are retained as the initial training weights of the target segmentation network.

3.3. Target Segmentation Tasks

The segmentation model shares the same architecture with the pre-training network and is initialized with the pre-trained weights.

The target segmentation network can not only share the network weights but also obtain the feature maps of the pre-training network, and share the pre-training network feature maps of the same layers to the target segmentation network at the same time. In the actual process of segmentation, the gradient disappearance problem may occur if the layers of encoder–decoder are deeper. Due to the strategy of encoder–decoder network using deep encoder–decoder may lose a lot of fundus image details in the process of encoding, which leads to the effect of fundus image vessel segmentation not well. From the convolution formula, it is known that the feature map length and width will be reduced to two times of the original one after each encoder operation of U-net.

$$w_{out} = \frac{w_{in} + 2 \times padding - F + 1}{stride} \tag{9}$$

The loss function of the target segmentation task is shown in Equation (10), which calculates the differences between predictions and ground truth.

$$L = \sum_{i=1}^{W} \sum_{j=1}^{H} \| G_{result}(i,j) - G_{groundtruth}(i,j) \|_1 \tag{10}$$

where $G_{groundtruth}$ and G_{result} represent manual annotations and predicted results, W and H represent the length and width of the image, $\| \cdot \|_1$ denotes the L_1 loss, and (i, j) is the pixel location.

4. Experiments and Results

4.1. Implements Details

We implement our algorithm using Pytorch and train the model on NVIDIA RTX 3090 GPU. The network is tested on the same environment. We train our network for 100 epochs in total. The batch size is set to 16 and the initial learning rate is set to 0.001 with a decay of

0.96 for every 100 iterations. In order to avoid overfitting, we adopt random rotation and flip as data augmentation.

4.2. Datasets

We take DRIVE [18], VAMPIRE [19], iChallenge, and STARE [20] as the experimental datasets. VAMPIRE is ultra-wide-angle fluorescein fundus angiography (UWFFA) dataset, which contains 8 high-resolution ultra-wide-angle images. The DRIVE, iChallenge, and STARE datasets are all color fundus image datasets. The Drive and STARE datasets have provided manual labels of retinal vessel segmentation.

The DRIVE dataset contains a collection of 40 retinal iamges, we use 20 images as training data the others as testing data. The VAMPIRE dataset provides eight high-resolution ultra-wide-angle fundus image, and we use all images during pre-training stage.

When pre-training the network with self-supervised learning tasks, all images from four datasets are used. During the training stage of the segmentation model, 40 images from the DRIVE dataset, 400 images from the iChallenge dataset, and 8 UWFFA images from VAMPIRE are used as the training dataset. The remaining 20 images from the DRIVE dataset and 4 images from the VAMPIRE dataset are used for testing. Overall, the pre-trained dataset includes 480 color fundus images and 8 FFA images, the segmentation training dataset contains 20 color fundus images and 4 FFA images, and the test dataset consists of 20 color fundus images and 4 FFA images.

4.3. Evaluation Metrics

We evaluate our methods both subjectively and objectively. The subjective evaluation reflects whether the results are admitted by ophthalmologists, and the objective evaluation measures the segmentation results quantitatively. Subjective evaluation criteria.The Mean Opinion Score (MOS) is taken as the subjective quality evaluation metric. Five levels of MOS score are assigned. Each level and the corresponding score are as follows.

- Score 5, the segmentation results are excellent (the clarity of the retinal vascular structure is between 80% and 100%, and the vessel segmentation results are not inferior to human eye observation);
- Score 4, the segmentation results are favorable (the clarity of retinal vascular structure is between 60% and 80%, and the vessel segmentation result has few differences with human eye observation, and a few capillaries can not be distinguished);
- Score 3, the segmentation results are borderline (the clarity of retinal vascular structure is between 40–60%, there is a gap between the vascular segmentation result and human eye observation, and a few capillaries are missing);
- Score 2, the segmentation results are weakly poor (retinal vascular structure clarity between 20–40%, the disparity between vascular segmentation results and human eye observation, and a large number of capillaries are missing);
- Score 1, the segmentation results are strongly poor (the clarity of retinal vascular structure is between 0–20%, there is a big gap between the vascular segmentation result and human eye observation, and the main vessels are missing).

Supposing we need L participants to rate T segmentation results in total, the MOS is calculated as Equation (11).

$$\overline{MOS_j} = \frac{\frac{1}{L}\sum_{i=1}^{L}\sum_{k=1}^{T} M_{i,j,k}}{T} \tag{11}$$

where $M_{i,j,k}$ denotes the score of the k-th segmentation results given by the i-th person, in which the image is segmented by the j-th algorithm.

Objective evaluation criteria.Objective evaluation criteria are numerical indicators that are calculated by a fixed formula to judge the segmented results of the algorithm. We take Dice Score (DS), Accuracy (AC), Area Under Receiver Operating Characteristic

Curve (AUC), and Average Precision (AP) as the objective quality evaluation metrics of the segmentation task. We denote true positive predictions as TP, true negative predictions as TN, false positive predictions as FP, and false negative predictions as FN. We calculate the objective metrics as follows:

Dice Score (DS):

$$DS = \frac{2TP}{2TP + FP + FN} \tag{12}$$

Accuracy (AC):

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

Area Under Receiver Operating Characteristic Curve (AUC):

$$AUC = \frac{\sum_{i \in positiveClass} rank_i - \frac{M(1+M)}{2}}{M \times N} \tag{14}$$

in which, $positiveClass(\cdot)$ denotes the set of real vessel pixels, $rank_i$ denotes the serial number of the i-th real vascular pixel point in the ranking result after sorting all pixel points in the segmentation result from smallest to largest probability of being classified as vessels, $M = TP + FN$, $N = TN + FP$.

Average Precision (AP):

$$AP = \sum_{i=1}^{n-1} (r_{i+1} - r_i) p_{interp}(r_{i+1}) \tag{15}$$

in which, $r_1, r_2, \ldots, r_{n-1}$ are the $Recall(\cdot)$ values corresponding to the first interpolation of the precision interpolation segment in ascending order.

4.4. Results and Analysis

We compare our aggregation self-supervised model with several common self-supervised segmentation algorithms, including COL (Colorization) [21] as intensity transformation, RO (Rotation) [22] as image rotation, SimCLR [23] and Model Genesis [24]. The segmentation results are also compared with the supervised D2D-CNNs [9]. All models are tested on the DRIVE dataset.

Table 1 shows the testing results on the DRIVE dataset. It can be seen that the proposed method in this paper occupies advantages in all metrics compared with existing methods, which fully demonstrates the effectiveness and superiority of our method. Some segmented results are shown in Figures 5 and 6.

The COL [21] and RO [22] methods have the worst segmentation results since the single-task self-supervised learning task does not learn enough features when processing unlabeled data. The combination of COL [21] and RO [22] tasks showed a significant improvement over the single task, which shows that using a combination of tasks can improve the segmentation accuracy. SimCLR [23] has an improvement in accuracy, but is inferior to our method for capillary processing due to the existing details of vessels in fundus images.

Table 1. Retinal vessel segmentation results (DRIVE Dataset).

Methods	Subjective Metrics	Objective Metrics		
	MOS	DS	AUC	AC
COL [21]	3.7	0.734	0.7234	0.7907
RO [22]	2.8	0.735	0.7217	0.8783
COL + RO	3.5	0.947	0.8061	0.8158
SimCLR [23]	3.3	0.721	0.8826	0.9397
Our Method	4	0.962	0.9494	0.9274
Model Genesis [24]	3.9	0.96	0.9796	0.9246
D2D-CNNs [9]	3.5	0.917	0.9602	0.9867

Both our method and Model Genesis [24] are multi-task self-supervised learning frameworks. The capillary extraction is greatly improved by multi-task pre-training of the encoder. We can see that in terms of accuracy and subjective perception, Model Genesis [24] is slightly inferior to our method in all objective evaluation metrics except for the AUC evaluation metric, showing that the encoder have learned more detailed features in the process of feature extraction. Compared with the supervised D2D-CNNs [9], the segmentation results of our method are more accurate, and the subjective visual segmentation results of the tiny capillary structures are much better than D2D-CNNs [9]. It can be seen that the self-supervised pre-training methods outperform the supervised learning method, since the accuracy of the labeled data limits the performance of the D2D-CNNs [9] in blood vessel segmentation.

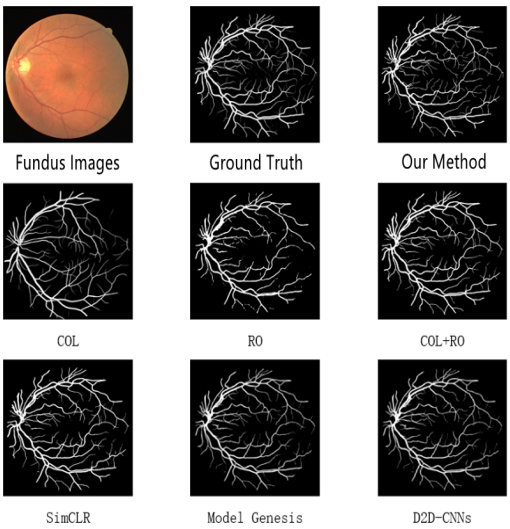


Figure 5. Experimental results.

Table 2. Crossover experiments.

Pre-Train Dataset	Target Dataset	ACC
Drive	Vampire	0.9067
Drive	Drive	0.9620
Vampire	Drive	0.9527
Vampire	Vampire	0.9130

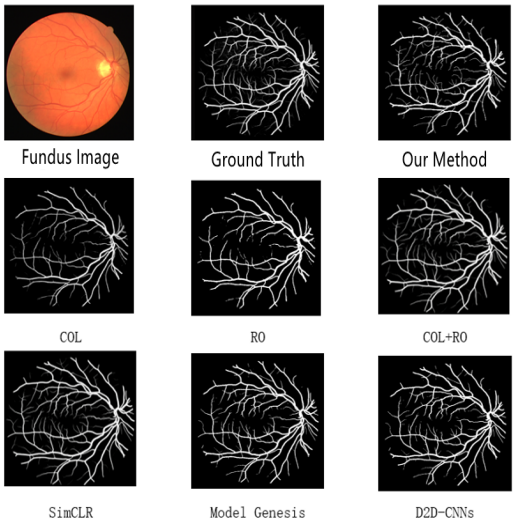


Figure 6. Experimental results.

To demenstrate our network can gain retial vessel feature and have the generalisation ability, we also perform crossover experiments by using different datasets to train pre-training network and traget network. The results shows in Table 2.

5. Conclusions

Retinal vessel segmentation is a fundamental task in the automated diagnosis of retinal diseases, which still remains a big challenge despite considerable research efforts. In this paper, we develop a multi-task strategy with self-supervised pre-training for retinal vessel segmentation. Specifically, we combine three pretext tasks (i.e., intensity transformation, random pixel filling, in-painting and out-painting) in an aggregated way to automatically learn features related to vessel morphological and functional changes from unlabeled data. The aggregation strategy is elegantly designed for retinal vessel segmentation and benefit the network greatly. Besides, a vector classification task module is introduced to improve the segmentation abilities for curvilinear structures during the pre-training stage. The pre-trained dense network can effectively obtain robust blood vessel structures. After that, we use a dynamic loss to improve performance during the segmentation process. The self-supervised pre-training and dynamic loss make our network effectively learn the ocular organ vessel feature with great generalization capabilities. The experimental results show that our method is superior both objectively and subjectively compared with other self-supervised and supervised learning methods. Moreover, compared to traditional methods, we use less labeled data and our network will perform better with the increase of the image volume.

Author Contributions: Funding acquisition, H.Z.; Methodology, Z.T. and H.Z.; Software, Z.T.; Validation, Q.Z. and X.Z.; Writing—original draft, Z.T.; Writing—review and editing, Q.Z. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Bingtuan Science and Technology Program (Nos. 2022DB005, 2019BC008).

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

References

1. Balyen, L.; Peto, T. Promising artificial intelligence-machine learning-deep learning algorithms in ophthalmology. *Asia-Pac. J. Ophthalmol.* **2019**, *8*, 264–272.
2. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.

3. He, Q.; Zou, B.; Zhu, C.; Liu, X.; Fu, H.; Wang, L. Multi-Label Classification Scheme Based on Local Regression for Retinal Vessel Segmentation. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018.
4. Fraz, M.M.; Remagnino, P.; Hoppe, A.; Uyyanonvara, B.; Rudnicka, A.R.; Owen, C.G.; Barman, S.A. Blood vessel segmentation methodologies in retinal images—A survey. *Comput. Methods Programs Biomed.* **2012**, *108*, 407–433. [CrossRef] [PubMed]
5. Roy, K.; Chaudhuri, S.S.; Roy, P.; Chatterjee, S.; Banerjee, S. Transfer Learning Coupled Convolution Neural Networks in Detecting Retinal Diseases Using OCT Images. In *Intelligent Computing: Image Processing Based Applications*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 153–173.
6. Yue, K.; Zou, B.; Chen, Z.; Liu, Q. Retinal vessel segmentation using dense U-net with multiscale inputs. *J. Med. Imaging* **2019**, *6*, 034004. [CrossRef] [PubMed]
7. Xu, Y.; Hu, G.; Shang, L.; Geng, J. Adaptive tracking extraction of vessel centerlines in coronary arteriograms using Hessian matrix. *J.-Tsinghua Univ.* **2007**, *47*, 889.
8. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
9. 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] [PubMed]
10. Wang, B.; Qiu, S.; He, H. Dual encoding u-net for retinal vessel segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 84–92.
11. Ma, Y.; Hao, H.; Xie, J.; Fu, H.; Zhang, J.; Yang, J.; Wang, Z.; Liu, J.; Zheng, Y.; Zhao, Y. ROSE: A retinal OCT-angiography vessel segmentation dataset and new model. *IEEE Trans. Med. Imaging* **2020**, *40*, 928–939. [CrossRef] [PubMed]
12. Zhu, J.; Li, Y.; Hu, Y.; Ma, K.; Zhou, S.K.; Zheng, Y. Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. *Med. Image Anal.* **2020**, *64*, 101746. [CrossRef] [PubMed]
13. Zhou, Z.; Sodha, V.; Pang, J.; Gotway, M.B.; Liang, J. Models genesis. *Med. Image Anal.* **2021**, *67*, 101840. [CrossRef] [PubMed]
14. Chen, X.; Yao, L.; Zhou, T.; Dong, J.; Zhang, Y. Momentum contrastive learning for few-shot COVID-19 diagnosis from chest CT images. *Pattern Recognit.* **2021**, *113*, 107826. [CrossRef] [PubMed]
15. Doersch, C.; Zisserman, A. Multi-task self-supervised visual learning. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2051–2060.
16. Zhu, J.; Li, Y.; Zhou, S.K. Aggregative Self-Supervised Feature Learning from a Limited Sample. *arXiv* **2020**, arXiv:2012.07477.
17. He, Y.; Yang, G.; Yang, J.; Chen, Y.; Kong, Y.; Wu, J.; Tang, L.; Zhu, X.; Dillenseger, J.L.; Shao, P.; et al. Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation. *Med. Image Anal.* **2020**, *63*, 101722. [CrossRef] [PubMed]
18. Staal, J.; Abràmoff, M.D.; Niemeijer, M.; Viergever, M.A.; Van Ginneken, B. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging* **2004**, *23*, 501–509. [CrossRef] [PubMed]
19. Ballerini, L.; Fetit, A.E.; Wunderlich, S.; Lovreglio, R.; McGrory, S.; Valdes-Hernandez, M.; MacGillivray, T.; Doubal, F.; Deary, I.J.; Wardlaw, J.; et al. Retinal Biomarkers Discovery for Cerebral Small Vessel Disease in an Older Population. In *Communications in Computer and Information Science*; Springer International Publishing: Cham, Switzerland, 2020; pp. 400–409. [CrossRef]
20. Hoover, A.; Kouznetsova, V.; Goldbaum, M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Trans. Med. Imaging* **2000**, *19*, 203–210. [CrossRef] [PubMed]
21. Zhang, R.; Isola, P.; Efros, A.A. Colorful image colorization. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 649–666.
22. Gidaris, S.; Singh, P.; Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv* **2018**, arXiv:1803.07728.
23. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*; PMLR: Toronto, ON, Canada, 2020; pp. 1597–1607.
24. Zhou, Z.; Sodha, V.; Siddiquee, M.M.R.; Feng, R.; Tajbakhsh, N.; Gotway, M.B.; Liang, J. Models genesis: Generic autodidactic models for 3d medical image analysis. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 3–17 October 2019; Springer: Cham, Switzerland, 2019; pp. 384–393.