Retinal Vessel Segmentation by a Transformer-U-Net Hybrid Model With Dual-Path Decoder

Yishuo Zhang and Albert C. S. Chung

***Abstract*—This paper introduces an effective and efficient framework for retinal vessel segmentation. First, we design a Transformer-CNN hybrid model in which a Transformer module is inserted inside the U-Net to capture long-range interactions. Second, we design a dual-path decoder in the U-Net framework, which contains two decoding paths for multi-task outputs. Specifically, we train the extra decoder to predict vessel skeletons as an auxiliary task which helps the model learn balanced features. The proposed framework, named as TSNet, not only achieves good performances in a fully supervised learning manner but also enables a rough skeleton annotation process. The annotators only need to roughly delineate vessel skeletons instead of giving precise pixel-wise vessel annotations. To learn with rough skeleton annotations plus a few precise vessel annotations, we propose a skeleton semi-supervised learning scheme. We adopt a mean teacher model to produce pseudo vessel annotations and conduct annotation correction for roughly labeled skeletons annotations. This learning scheme can achieve promising performance with fewer annotation efforts. We have evaluated TSNet through extensive experiments on five benchmarking datasets. Experimental results show that TSNet yields state-of-the-art performances on retinal vessel segmentation and provides an efficient training scheme in practice.**

***Index Terms*—Annotation correction, dual-path decoder, retinal vessel segmentation, transformer.**

I. Introduction

OPHTHALMOLOGIC diseases, such as Diabetic Retinopathy (DR), have become common causes of illness blinding and attracted an increasing concern over the world [1]. In the ophthalmologic examination, the morphological and topographical appearances of retinal vessels are used to evaluate and grade ophthalmologic diseases [2]. Manually delineating retinal vessels in the retinal images is tedious, time-consuming, and error-prone. Developing automatic vessel segmentation techniques can significantly help ophthalmology diagnosis and thus is urgently needed. However, the task is inherently challenging due to the poor image quality and complicated structures of blood vessels [3].

During the past decade, various methods have been proposed for retinal vessel segmentation. Unsupervised methods detect blood vessels by manually designed rules which are summarized by analyzing existing samples, such as [4], [5], [6], [7], and [8]. Supervised methods, such as [3], [9], [10], and [11], require vessel annotations as ground truth and automatically learn to segment vessels under the supervision of ground truth. Generally, supervised methods achieve better performances than unsupervised methods as supervised methods have more information from the ground truth. In recent years, with the emergence of deep learning, Convolutional Neural Networks (CNNs) have been successfully applied for retinal vessel segmentation, such as [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], and [23]. CNNs based methods rapidly become dominant in the task as they exceed traditional methods in terms of performance.

Despite the promising progress made by CNN-based methods, it is worth noting that several challenges in deploying retinal vessel segmentation techniques remain not fully solved. First, CNNs based methods inherently fail to learn long-range interactions, which misses crucial information about vessels. Capturing long-range interactions helps the model infer the vessel from a large view and improves segmentation performance. As a solution, Transformer [24], [25], which recently emerged, is verified to be an effective structure for learning long-range interactions in vision tasks. A Transformer-CNN hybrid model can be applied to capture both local and global information. The second challenge is that the segmentation results of tiny vessels are not good, and tiny vessels tend to be ignored by the model. The solution to this issue can be used to guide the model to pay more attention to the tiny vessels.

Besides the above two challenges caused by the property of the task, it is worth noting two challenges in the annotation process as deep learning-based methods highly rely on the data and annotations. On the one hand, obtaining low-quality annotations, such as vessel annotations with positional errors, is unavoidable. The errors in annotations can severely affect the training process. This problem can be formulated as a label noise learning task. Here, labels indicate whether pixels belong to the foreground. On the other hand, acquiring precise vessel annotations is expensive. The solution can be semi-supervised learning methods.

Manuscript received 1 June 2023; revised 26 December 2023 and 3 April 2024; accepted 18 April 2024. Date of publication 26 April 2024; date of current version 6 September 2024. *(Corresponding author: Albert C. S. Chung.)*

The authors are with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong (e-mail: ys.zhang@connect.ust.hk; achung@ust.hk).

Digital Object Identifier 10.1109/JBHI.2024.3394151

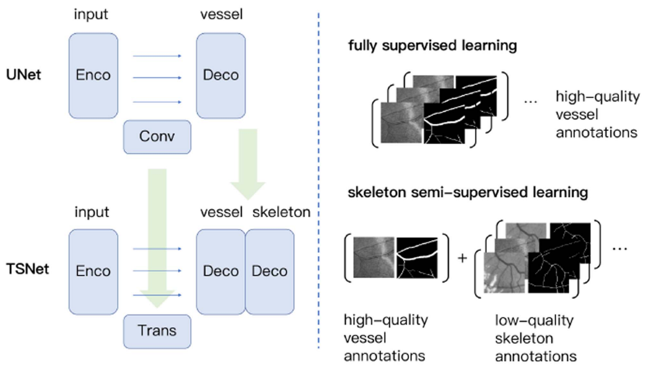
input
vessel
UNet
Enco
Deco
Conv
input
vessel
skeleton
TSNet
Enco
Deco
Deco
Trans
fully supervised learning
high–quality vessel annotations
skeleton semi–supervised learning
high–quality vessel annotations
low–quality skeleton annotations

Fig. 1. Illustration of the proposed TSNet (left) and skeleton semi-supervised learning (right) for retinal vessel segmentation.

This paper aims to focus on tackling the issues mentioned above and propose an effective learning framework for retinal vessel segmentation. The proposed model has a novel Transformer-CNN hybrid architecture. A Transformer module is inserted inside a U-Net model, which enables the model to capture long-range interactions. We design a dual-path decoder with two decoding paths for multi-task outputs. Specifically, we set predicting vessel skeletons as an auxiliary task which helps the model learn rich and balanced feature representations. The proposed framework also enables a rough skeleton annotation process in which the annotators do not need to provide precise vessel annotations. Instead, they only need to delineate vessel skeletons roughly. We introduce a skeleton semi-supervised learning framework to leverage such a type of rough skeleton annotations plus a few precise vessel annotations. We adopt a mean-teacher framework to produce pseudo-vessel annotations for samples without vessel annotations. We conduct annotation correction for rough skeletons annotations to correct unavoidable positional errors of vessel skeletons. This learning scheme can achieve satisfactory performance with fewer annotation efforts, which is beneficial in practice. The proposed TSNet has been evaluated by extensive experiments on five public benchmarking datasets in two different imaging modalities, including STARE, CHASE, DRIVE, HRF, and RECOVERY. Experimental results show that TSNet achieves state-of-the-art performances for retinal vessel segmentation and provides an efficient learning scheme.

The contributions of this paper are summarized as follows (as illustrated in Fig. 1).

1) This paper presents a novel and effective framework for retinal vessel segmentation. It has a Transformer-CNN hybrid design to capture global and local information. Besides, a dual-path decoder learns balanced feature representations of vessels.
2) The proposed method is evaluated on five benchmarking datasets in two retinal imaging modalities. It outperforms current methods and provides a new state-of-the-art benchmark for retinal vessel segmentation.
3) The proposed model enables a rough skeleton annotation process. It achieves promising performances with many rough vessel skeleton annotations and a few precise vessel annotations. We introduce a skeleton semi-supervised learning scheme to utilize such types of data.

The paper is organized as follows. Section II reviews related works. Section III introduces the architecture of the proposed TSNet. Section IV introduces the rough skeleton annotation and the skeleton semi-supervised learning process. Section V gives the experiment's design. Section VI gives the results of the experiments. A conclusion is drawn in Section VII.

II. Related Work

A. Retinal Vessel Segmentation

Existing methods for retinal vessel segmentation can be categorized into unsupervised and supervised methods. Unsupervised methods are mainly based on manually designed rules. For example, Frangi et al. [4] proposed the Hessian filter, which was based on analyzing the Hessian matrix. Zhang et al. [5] proposed a matched filtering approach where an image was convolved with a multi-scale template in the orientation domain. Bekkers et al. [6] proposed a multi-orientation vessel connectivity analysis method where vessels were tracked based on local patterns. Roychowdhury et al. [8] proposed a method that used an iterative vessel segmentation approach with global adaptive thresholding followed by novel stopping criteria.

Supervised methods for retinal vessel segmentation adopt machine learning techniques. Early supervised methods use hand-crafted features and simple models. For example, Fraz et al. [3] proposed a novel method where a decision tree was adopted with a combination of multiple feature extraction techniques. Roychowdhury et al. [9] used a combination of first-order and second-order gradient features along with a Gaussian mixture model classifier. Orlando et al. [10] trained a fully connected conditional random field model with a structured output support vector machine. Zhang et al. [11] adopted a neural network classifier with a multi-scale texton dictionary where the initial key points for k-means clustering were extracted from a Gabor filter bank.

Compared with traditional methods, deep learning methods achieve more promising performances for retinal vessel segmentation nowadays. Liskowski et al. [12] designed a deep neural network that predicted the center pixel of each patch. Zhang et al. [14] proposed to learn thin vessels, thick vessels, and their boundary regions separately and designed a U-Net model with a boundary-aware mechanism. Wu et al. [16] designed a lightweight U-shaped model equipped with inception modules and residual modules to improve feature representation. Yan et al. [18] designed a novel segment-level loss that paid more attention to the thickness consistency of thin vessels. Guo et al. [19] presented the BTS-DSN model which was a multi-scale CNNs model with short connections and deep supervision learning low-level and high-level information. Oliveira et al. [20] presented a novel method that adopted stationary wavelet transform and a multi-scale fully convolutional neural network to cope with varying widths and directions of vessels. Jin et al. [21] designed the deformable U-Net in which they adopted deformable convolution to exploit local features of retinal vessels. Wang et al. [22] designed a novel dual encoding U-Net with two separate encoders and adopted the channel attention module to select useful features. Shin et al. [23]

proposed to model the vasculature as graphs and learn the graphical structures of vessels with GATs.

B. Transformer

The Transformer was first proposed by Vaswani et al. [24] in the natural language processing task and became dominating due to its good capacity. The Vision Transformer (ViT) was proposed by Dosovitskiy et al. [25], which first provided a novel insight into utilizing Transformers in vision tasks. ViT learns from sequences of patches that are obtained by splitting the input image. The Shifted Windows Transformer (Swin Transformer), proposed by Liu et al. [26], further developed the Transformer using the shifted windowing scheme. These attempts have shown that Transformers work effectively on the vision tasks. Moreover, a hybrid model that combines convolution and Transformer could inherit both merits. Some attempts have been made in this direction. We mainly focus on attempts in the medical image analysis field. For example, Chen et al. [27] proposed the TransUNet model for the multi-organ segmentation task. Two recent works, [28], [29], also adopted a Transformer for retinal vessel segmentation. Compared with them, our work not only achieves competitive performances but also explores utilizing low-quality annotations.

C. Semi-Supervised Learning

Semi-supervised learning methods aim to leverage a small amount of labeled data plus a large amount of unlabeled data, which can address the issue of data limitation [30]. Mean-Teacher, which was proposed by Tarvainen and Valpola [31], is a common solution for semi-supervised learning. The key idea of the mean-teacher is to form a target-generating teacher model by averaging the model weights of the student model. Chen et al. [32] have explored semi-supervised techniques for retinal vessel segmentation. In this work, they trained a U-Net model with a mean-teacher paradigm. Hou et al. [33] also proposed a GAN-based method for semi-supervised retinal vessel segmentation. This work used a mean-teacher mechanism to regularize the discriminator to overcome image variations. Different from the mentioned semi-supervised methods, we present a learning scheme with a small amount of precisely annotated samples and a large amount of roughly annotated samples. We leverage mean-teacher to learn from samples with low-quality annotations.

D. Label Noise Learning

Label noise learning studies the case where labels or annotations contain considerable errors [34]. Several techniques are proposed to counter the adverse effect of label errors, as summarized in [35]. For example, Song et al. [36] proposed to selectively exploit unclean samples that can be corrected with confidence and gradually increase the number of clean samples. Patrini et al. [37] added a noise adaptation layer following a deep model to learn the label transition process. Huang et al. [38] applied the exponential moving average to refurbish noisy labels, which prevented overfitting false labels. In this paper, we consider a rough skeleton annotation process that enables annotators to give low-quality skeleton annotations and significantly saves annotation efforts. The proposed rough skeleton annotation process introduces spatial errors in skeleton annotations. We formulate it as a label noise learning task and tackle it by label adaptation.

III. Methodology

In this section, we first present the overview of the proposed model. Then, we introduce the Transformer module. Finally, we introduce the dual-path decoder.

A. Architecture Overview

The overview of the proposed TSNet is illustrated in Fig. 2. TSNet is built based on U-Net [39], which has been regarded as a de-facto choice in medical image segmentation tasks. Generally, it contains an encoder, a decoder, and skip connections. Both the encoder and decoder contain several convolutional blocks which learn features at different resolutions. The encoder takes raw image patches as input and extracts hierarchical semantic information, while the decoder reconstructs spatial information and produces final segmentation. Skip connections link the corresponding blocks of the encoder and the decoder at the same resolutions. Features learned from the encoder are fed into the decoder by concatenation, which helps the decoder reconstruct spatial information.

There is a concern that down-sampling operations can lose spatial information about tiny objects, e.g., blood vessels. Increasing the number of feature maps before each down-sample operation can enrich feature representations to keep spatial information. In practice, the number of channels will double after each convolutional block in the encoder and halve after each convolutional block in the decoder.

To effectively learn rich features, we adopt densely connected convolutional blocks (Dense ConvBlock) [40] as the basic component of the encoder and the decoder. A densely connected convolutional block contains 2 convolutional layers with skip connections from former positions to all subsequent positions by addition. The densely connected design effectively reuses features and helps gradient back-propagation. Each convolutional layer has a 3×3 kernel, followed by a Batch Normalization layer and a ReLU layer. We adopt the max-pooling layer for down-sampling in the encoder and the deconvolution layer for up-sampling.

Please note that the down-sampling will not affect the segmentation of vessels. The down-sampling operation only causes the loss of spatial information, while the model can keep the category information of the vessel in the deep layers. Besides, the skip connections in the U-Net model will retain the spatial information from the previous layers. Therefore, the final output layer can segment the tiny vessels by summarizing the above information.

B. Transformer

Transformers can capture interactions among sequence data. It can be applied in vision tasks by transforming images into sequences of patches [25], which inherently has a global view.

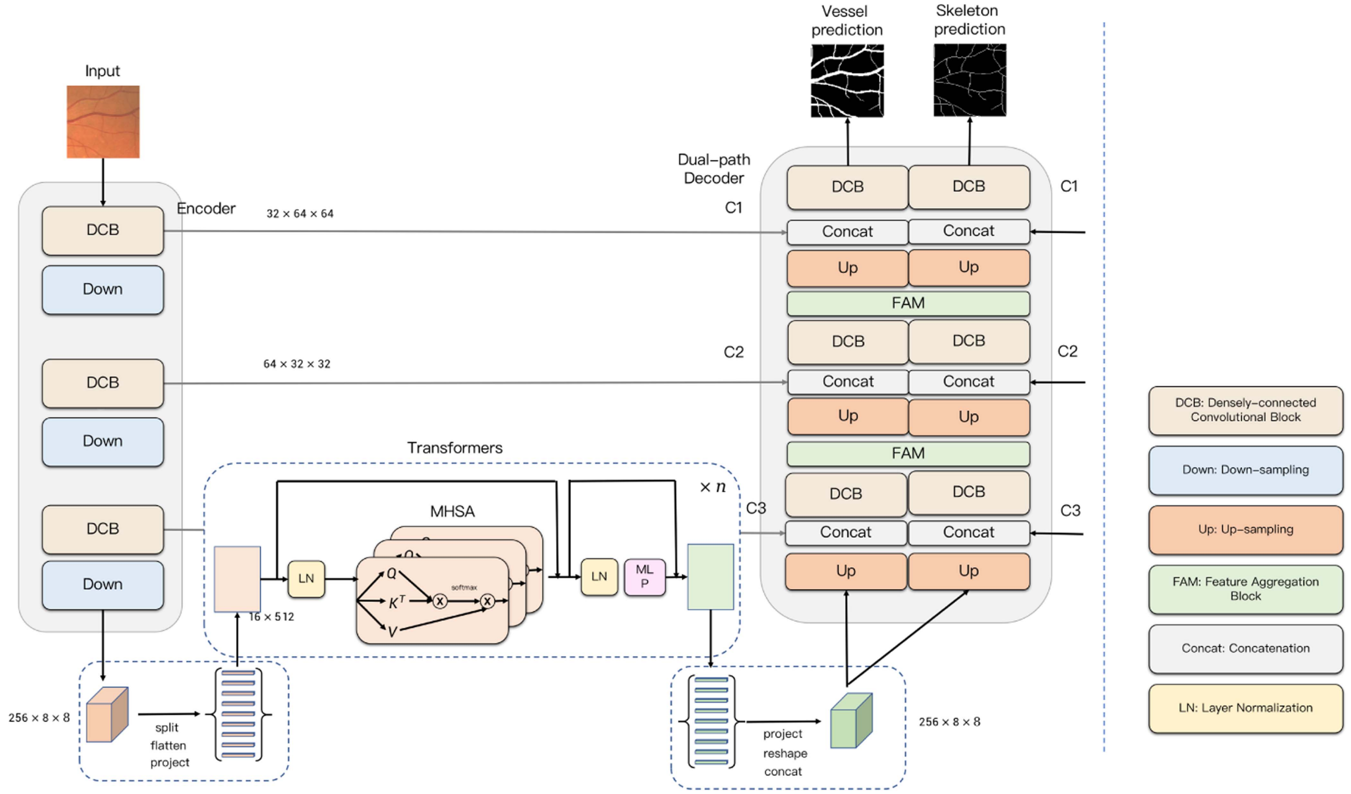

Fig. 2. Overview of the proposed TSNet. Rounded rectangles denote operations (blocks). Shapes of feature maps are represented in the format of ’channel, width, shape’.

Given an input feature map $X \in \mathbb{R}^{H\times W\times C_0}$ with a resolution $H \times W$ and C_0 channels, it is split into N patches (tokens) with a size of $P \times P$. Next, patches are flattened into vectors $p_1, p_2, \ldots, p_N$ and then projected by a trainable linear projection $E \in \mathbb{R}^{P^2 \cdot C_0 \times d}$. Position information of each patch is also embedded as $e^{pos} \in \mathbb{R}^d$ and added to the patch embedding. Finally, a sequence of embeddings from the input feature map is obtained, as shown in (1).

$$Z_0 = [p_1E + e_1^{pos}; p_2E + e_2^{pos}; \ldots; p_NE + e_N^{pos}]. \tag{1}$$

The Transformer captures interactions among sequences of embeddings by the Self-Attention (SA) module. Concretely, an input feature matrix $\mathbf{Z} \in \mathbb{R}^{N\times C}$ denotes a previously obtained sequence of N embeddings which have C numbers of channels. Three trainable linear functions, $E^Q, E^K, E^V \in \mathbb{R}^{C\times d}$, are adopted to project $\mathbf{Z}$ into to Query, Key, and Value embeddings: $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N\times d}$. The output of self-attention is calculated as follows:

$$SA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d}}\right)\mathbf{V}. \tag{2}$$

For each query vector, (2) computes the normalized pair-wise dot production between the query vector and each key vector. Then the computed results are used as the weights to aggregate the information of values.

A Transformer model contains several Transformer layers, while one Transformer layer contains a Multi-Head Self-Attention (MHSA) layer and a Multi-Layer Perceptron (MLP). The operations in the l^{th} layer can be written as (3).

$$\begin{aligned} Z_l' &= MHSA(LN(Z_l)) + Z_l, \\ Z_{l+1} &= MLP(LN(Z_l')) + Z_l', \end{aligned} \tag{3}$$

where $LN(\cdot)$ denotes the layer normalization. Residual connections are added to the MHSA layer and MLP layer for residual learning. Multi-head self-attention is the extension of Self-Attention, which concatenates outputs from h independent self-attention modules, SA_1,..., SA_h, and projects them with a linear operation $\mathbf{W} \in \mathbb{R}^{hd\times C}$, as shown in the (4).

$$MHSA(X) = concat(SA_1(X), SA_2(X), \ldots, SA_h(X))\mathbf{W}. \tag{4}$$

We utilize the Transformer’s capacity to capture long-range interactions in the proposed model. We integrate Transformers inside the U-Net model at the position between the encoder and the decoder. The input of the Transformers module is feature maps extracted by the encoder, while the output of the Transformers module is fed into the decoder. This Transformer-convolution hybrid design takes full advantage of convolution and Transformer. Local patterns are extracted by convolution at low levels, and Transformers learn long-range interactions at high levels. Besides, we want to mention that employing Transformers at a low-resolution level only increases the computation burden to an acceptable extent.

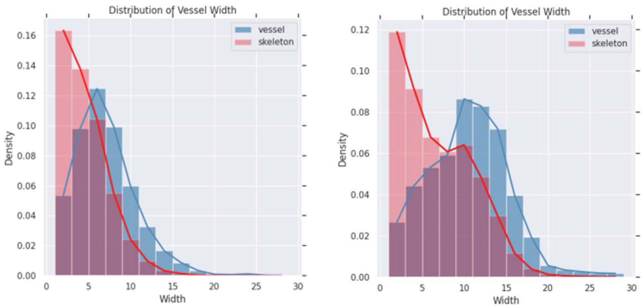

Fig. 3. Statistics of vessels in the STARE (left) and CHASE (right) dataset.

C. Dual-Path Decoder

In the proposed framework, we design a Dual-Path Decoder (DPD), an extension of a normal decoder. The dual-path decoder has two parallel decoding paths, each containing several convolutional blocks to lean hierarchical feature representation. In the dual-path decoder, two decoding paths can learn rich feature representation and are blended to make predictions. Besides, it can be used for multi-task learning, in which the model is trained to pursue two relevant tasks. Specifically, we introduce vessel skeleton segmentation as an auxiliary task in the retinal vessel segmentation task. The motivation is to tackle the imbalance issue among vessels introduced by vessel widths. The imbalance issue refers to that the quantities of thick vessel pixels and thin vessel pixels are significantly unbalanced. As most foreground pixels belong to thick vessels, the model will focus on the segment of thick vessels while ignoring the segment of thin vessels, failing for overall accurate vessel segmentation. In contrast, regarding the lengths of two types of vessels, namely the numbers of pixels on the vessel skeletons, thin vessels are more than thick vessels. (We group vessels by widths and show the distributions of vessel pixels and skeleton pixels in Fig. 3). In other words, when the model is trained to predict vessel skeletons, it will pay more attention to thin vessels. When the model is trained to predict vessels and vessel skeletons, it will pay balanced attention to vessels with different widths.

We define two paths in the dual-path decoder the vessel decoding path and the skeleton decoding path. The vessel decoding path is trained to predict vessels, while the skeleton decoding path is trained to predict vessel skeletons. Features from two paths are aggregated at each resolution, which enables features learned from different distributions to exchange sufficiently. At each stage, information exchange is performed via the designed Feature Aggregation Module (FAM). In the feature aggregation module, feature maps from two paths are first concatenated together and then learned with a squeeze-and-excitation module [41], which will selectively enhance the useful features and suppress less useful features. The learned feature maps are split and fed into two paths, as shown in Fig. 4.

Although skeleton annotations are not usually provided, we can simply obtain them by skeletonizing vessel annotations. To make the single-pixel skeleton more obvious, an option is to conduct a morphological dilation operation to obtain dilated skeleton annotation without causing vessels to mix up.

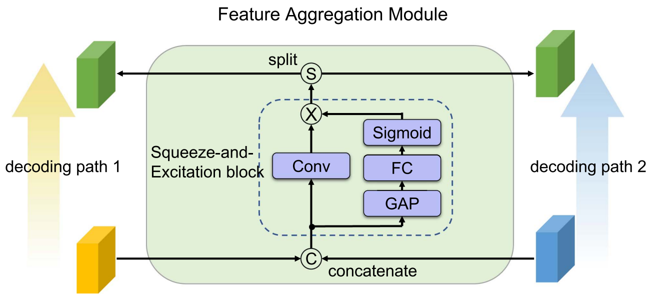

Fig. 4. Illustration of the proposed Feature Aggregation Module.

Finally, we mathematically introduce the training process. The model, parameterized by Θ, produces vessel prediction p_i^v and skeleton prediction p_i^s given the input image x_i.

$$p_i^v, p_i^s = f(x_i|\Theta). \tag{5}$$

Given vessel annotation v_i, we can obtain the skeleton annotation s_i by morphological skeletonization operation. We calculate binary cross-entropy (CE) loss between the prediction and the ground truth. The loss for vessel prediction and the loss for skeleton prediction is defined as follows.

$$\mathcal{L}_{ves} = \sum_i -v_i log p_i^v - (1 - v_i) log(1 - p_i^v), \tag{6}$$

$$\mathcal{L}_{skel} = \sum_i -s_i log p_i^s - (1 - s_i) log(1 - p_i^s). \tag{7}$$

The total loss for the dual-path decoder contains two above terms, which can be written as (8). The weights of two terms can be adjusted according to the specific circumstances. Moderately increasing the weight of $\mathcal{L}_{skel}$ can enhance the performance in segmenting fine vessels. However, it is advisable that the weight increase of $\mathcal{L}_{skel}$ should not be excessive and $\mathcal{L}_{ves}$ should be considered as a major component.

$$\mathcal{L}_{sup} = \mathcal{L}_{ves} + \mathcal{L}_{skel}. \tag{8}$$

IV. Rough Skeleton Annotation

The proposed TSNet enables a rough skeleton annotation process, significantly saving the annotation effort. The annotators only need to give a rough delineation of vessel skeletons. In contrast with giving pixel-wisely precise vessel annotations, delineating skeletons can be pursued by simply drawing lines, which can be much more convenient. We can utilize this type of rough annotation with only one piece of precise vessel annotation to train the model. We formulate it as a Skeleton Semi-Supervised Learning task with label errors (S3L).

Mathematically, we have a small amount of high-quality samples $\{(x_i, v_i)|i \in \mathbb{D}_{\mathbb{L}}\}$ plus many low-quality samples $\{(x_i, s_i^n)|i \in \mathbb{D}_{\mathbb{W}}\}$. $\mathbb{D}_{\mathbb{L}}$ denotes the index set of samples with precise vessel annotations, and $\mathbb{D}|_{\mathbb{W}}$ denotes the index set of samples with rough skeleton annotations. x_i, v_i, and s_i^n denote input image, vessel annotation, and rough (noised) skeleton annotation. As vessel annotations can be converted into skeleton annotations by morphological skeletonization operation, having v_i means having s_i, but not vice versa. Under a fully supervised learning scheme, only a small amount of samples ($\{(x_i, v_i, s_i)|i \in \mathbb{D}_{\mathbb{L}}\}$) with vessel annotations can be leveraged. The model, which is parameterized by Θ, aims to predict vessel p_i^v and its skeleton p_i^s. The model is trained to minimize the supervised loss $\mathcal{L}_{sup}$, which measures gaps between its predictions and annotations, as shown in (6), (7), and (8).

A. Mean Teacher

To leverage the samples without vessel annotation, semi-supervised learning can be used. Inspired by the mean-teacher model [31], we define two TSNet models, parameterized by Θ_T and Θ_S, respectively. The student model S learns with precise vessel annotations and their precise vessel skeleton annotations. For samples with only rough skeleton annotations $\{x_i|i \in \mathbb{D}_{\mathbb{W}}\}$. the teacher model T will generate pseudo vessel annotations v_i' for the student model.

$$v_i', s_i' = f(x_i|\Theta_T), i \in \mathbb{D}_{\mathbb{W}}. \tag{9}$$

The student model will be supervised by generated pseudo vessel annotations and rough vessel skeleton annotations. The mean-teacher loss between vessel predictions and pseudo vessel annotations is defined as $\mathcal{L}_{mt}$.

$$\mathcal{L}_{mt} = \sum_{i \in \mathbb{D}_{\mathbb{W}}} ||p_i^v - v_i'||^2. \tag{10}$$

The teacher model will not be optimized by back-propagation. The parameters in the teacher model will be updated by the corresponding parameters in the student model by exponential moving average, as shown in (11),

$$\Theta_T \leftarrow \Theta_T \times \alpha_{mt} + \Theta_S \times (1 - \alpha_{mt}), \tag{11}$$

where α_{mt} is a smoothing coefficient hyper-parameter in mean-teacher and controls the moving rate of the teacher model. In the experiments, it is decayed with the epoch e, i.e. $\alpha_{mt} = 1 - 1/(1 + e)$, following the common practices of the mean-teacher method [31].

B. Annotation Correction

Rough vessel skeleton annotations can contain numerous spatial errors. The spatial errors of skeletons mean that skeletons can be shifted away from the actual position. As introduced by Huang et al. [38], we can formulate it as a label noise learning task and correct the errors by the label adaptation learning scheme. Under the assumption that the deep model learns easy and clean patterns before noisy patterns [42], we can use predictions from the model to correct annotation errors during the training process.

Concretely, a skeleton can be wrongly annotated on both sides of the real position with roughly equal chance. Suppose that the wrong annotation process repeats independently. The expected result of the annotated skeleton is the real position. Considering the whole data, similar vessel structures repeat. Therefore, the wrong annotation process for a similar case repeats. The model that fits the whole data can provide suggestive information about the correct position. In such an annotation correction process, regions with correct annotation and high certainty for making correct predictions will remain unchanged. The ambiguous regions will gradually be smoothed. The regions with spatial errors will gradually be corrected.

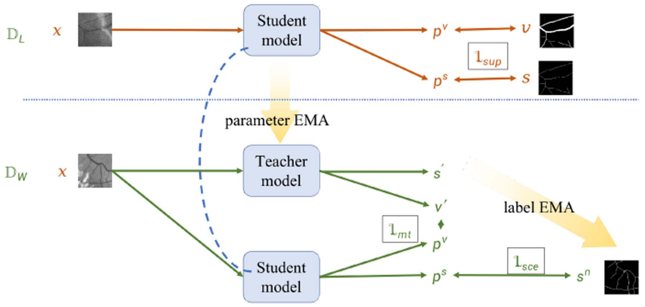

Fig. 5. Illustration of skeleton semi-supervised learning.

It is worth noting that the predictions of the teacher model are used to correct rough annotations, as the ensemble teacher model usually performs better than the student model. Mathematically, given a rough skeleton annotation s_i^n, it can be corrected by the prediction of the teacher model p_i^s. We adopt an Exponential-Moving-Average (EMA) scheme for label adaptation, as shown in (12).

$$s_i^n \leftarrow s_i^n \times \beta_{la} + s_i' \times (1 - \beta_{la}), i \in \mathbb{D}_{\mathbb{W}}. \tag{12}$$

The momentum weight β_{la} controls the progress of adaptation. In the experiments, the momentum weight β_{la} is set to 0.99 and we have noted that a higher or lower momentum weight can affect the progress of learning from our empirical experience. Before adopting the label adaptation learning scheme, the model is first warmed in the initial epoch. Then for each batch, we update annotations by (12). The values in s_i^n will become decimals in [0,1]. We regard the adapted annotation as smoothed annotations and calculate the smoothed cross-entropy loss for skeleton predictions.

$$\mathcal{L}_{sce} = \sum_{i \in \mathbb{D}_{\mathbb{W}}} -s_i^n log p_i^s - (1 - s_i^n) log(1 - p_i^s). \tag{13}$$

The final loss in the rough skeleton annotation scheme is defined by summarizing (8), (10), and (13).

$$\mathcal{L}_{s3l} = \mathcal{L}_{sup} + \mathcal{L}_{mt} + \mathcal{L}_{sce}. \tag{14}$$

The learning process is illustrated in Fig. 5.

V. Experiments

A. Datasets

For evaluation, we chose five benchmarking datasets, including four fundus photography datasets (STARE,[1] CHASE,[2] DRIVE,[3] and HRF)[4] and one fluorescein angiography dataset (RECOVERY). Brief introductions to the datasets are given below.

[1] https://cecas.clemson.edu/~ahoover/stare/
[2] https://blogs.kingston.ac.uk/retinal/chasedb1/
[3] https://drive.grand-challenge.org
[4] https://www5.cs.fau.de/research/data/fundus-images/

1) Fundus Photography: Fundus photography is the most common imaging modality in ophthalmology examination. It records conditions of the eye's interior surface with a fundus camera taking color photographs of the fundus. The STARE dataset was collected to assist the ophthalmologist in diagnosing eye diseases [43]. It contains 20 fundus photographs with a resolution of 700 × 605. The CHASE dataset is collected from two eyes of 14 children [3]. It contains 28 images with a high resolution of 960 × 999. The DRIVE dataset contains 40 photographs with a resolution of 565 × 584 [44]. The HRF dataset contains 15 images of healthy patients, 15 images of patients with diabetic retinopathy, and 15 images of patients with glaucomatous [45]. The resolution of HRF is 3504 × 2336.

Regarding the training-test partitions, we followed the settings in previous works [19], [21], [23]. For the STARE/CHASE dataset, we used the first 10/20 images as the training set and the remaining 10/8 images as the test set. The DRIVE dataset is officially divided into the training/test set, each containing 20 images. For the HRF dataset, we used the first 5 images in three groups as the training set (15 images) and the remaining images as the test set (30 images). Following previous works, we only counted pixels inside the Field Of View (FOV) when calculating evaluation metrics. Note that evaluation scores will be higher without using FOVs as the task is highly unbalanced. Determining the appropriate patch size requires considering the specific resolution of the image. A higher patch size causes much more computational burden while a small patch size may lose neighboring information of vessels. In our experiments, the input patch size was empirically set to 64/128/64/128 for the STARE/CHASE/DRIVE/HRF dataset. The patch (token) size inside the Transformer was set to 1/2/1/2 for the STARE/CHASE/DRIVE/HRF dataset with the consideration of the concrete resolution of each dataset.

2) Fluorescein Angiography: Fluorescein Angiography (FA) records fluorescence intensity images under blue illumination after intravenous injection of sodium fluorescein dye. It provides a larger view of the retina beyond the macula with a higher resolution, making vessel segmentation more challenging. The RECOVERY dataset contains 8 high resolution (3900 × 3072 pixels) ultra-widefield fluorescein angiography images acquired by Optos California and 200Tx cameras with a 200°FOV of the retina [46]. Following the original work by Ding et al. [46], we performed leave-one-out cross-validation. We sampled 1000 patches with a size of 128 × 128 from each raw image for training.

B. Evaluation Metrics

For a comprehensive evaluation of vessel segmentation, we followed the previous works [21], [46], [47], calculating and comparing two commonly used metrics, including Accuracy (ACC) and Dice score (DC), which are defined as follows.

$$ACC = \frac{TN + TP}{TN + FN + TP + FP}, DC = \frac{2TP}{2TP + FP + FN}, \tag{15}$$

where TP, TN, FP, FN denote the numbers of pixels corresponding to true positive, true negative, false positive, and false negative, respectively. The other two commonly used metrics are the Area Under ROC Curve (AUC ROC) and the Area Under the Precision-Recall Curve (AUC PR). AUC ROC measures the area underneath the ROC curve, which plots SE versus $1-$SP for a varying threshold. AUC PR measures the area underneath the precision-recall curve, which plots precision versus recall with respect to a varying threshold. AUC ROC and AUC PR are not sensitive to the threshold and are adopted to measure the overall performances of models. Furthermore, we considered an evaluation metric applicable to imbalanced data, the balanced accuracy (BAcc). We also calculated the CAL score which evaluates retinal vasculature and connectivity [48].

C. Implementation Details

The proposed method has been implemented in PyTorch [49]. All experiments were conducted on one NVIDIA GeForce GTX 1080Ti GPU card. It took around 4 hours to train a single model. We used AdamW as the optimization algorithm with a learning rate of 0.0001 and a weight decay rate of 0.0005.

Regarding image pre-processing operation, the input image was first converted into a single-channel gray-scale image [12]. Then we normalized the values of pixels into [0, 1] and conducted gamma adjustment. The data augmentation methods include image rotation, random flipping, random noise, random blur, random contrast, and random sharpening.

We cropped raw images into patches to form the training and validation sets at the beginning. In the training phase, we randomly sampled 1000 patches from each raw image. In the testing phase, we sampled patches by a sliding window, scanning the whole image with a stride of 8. Extracting more patches can lead to better performance while increasing the computation burden. The final prediction of the whole image was obtained by aggregating the predictions of local patches. We randomly excluded one raw image from the training set to form the evaluation set. Some processed images are visualized in Fig. 6. The first column shows original images; the second column shows cropped patches; the third column shows vessel annotations in patches; the fourth column shows vessel skeleton which is obtained from vessel annotations by the morphological skeletonization operation; the fifth column shows dilated vessel skeletons which are dilated from vessel skeleton; the sixth column shows rough skeleton annotations which are simulated by distorting vessel skeleton; the last column compares positional differences between the precise skeleton and rough skeleton annotation.

VI. Results

The former groups of experiments were conducted in a fully supervised manner, which means all the samples were provided with precise vessel annotations. Through these groups of experiments, We will show that the proposed TSNet outperforms state-of-the-art methods in a fully supervised manner. The experiments in the last Section VI-E were considered with the rough skeleton annotation process, in which most of the samples only have low-quality skeleton annotations instead of vessel annotations. Although vessel annotations are available in these

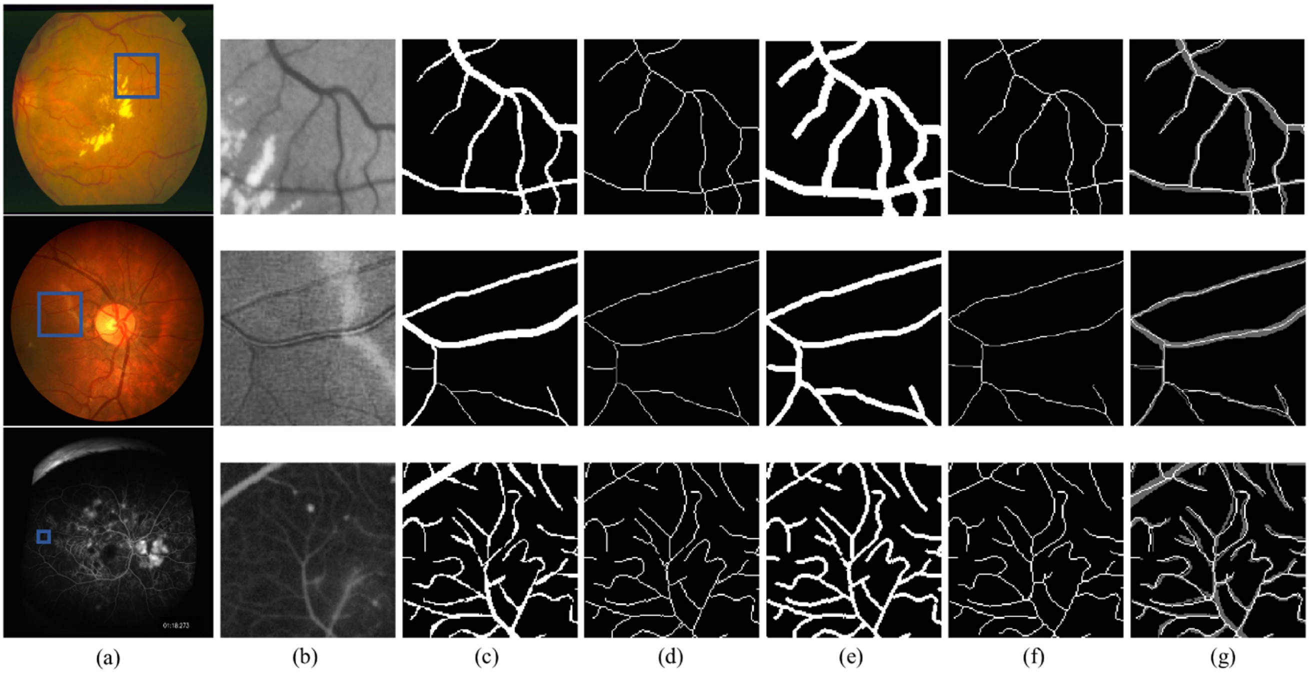

Fig. 6. Visualization of processed fundus images. From left to right: (a) original image, (b) processed patch, (c) vessel annotation (patch), (d) vessel skeleton (patch),(e) dilated vessel skeleton (patch), (f) rough skeleton annotation (patch), and (g) positional comparison between the precise skeleton and rough skeleton annotation (patch).

TABLE I
Ablation Study on the STARE and the CHASE Dataset

	STARE				CHASE			
Methods	AUC ROC	AUC PR	DICE	p-value	AUC ROC	AUC PR	DICE	p-value
baseline	0.9881	0.9213	0.8280	-	0.9868	0.9043	0.8174	-
baseline+Trans	0.9883	0.9237	0.8225	0.0093	0.9873	0.9066	0.8200	0.0007
baseline+DPD	0.9887	0.9242	0.8282	0.0016	0.9876	0.9067	0.8206	0.0002
baseline+Trans+DPD	**0.9889**	**0.9252**	**0.8300**	0.0018	**0.9878**	**0.9080**	**0.8211**	0.0005

The last column gives p-values of paired t-tests between the baseline model and the compared model in terms of AUC PR.

'The best results are highlighted in bold. The same below.'

datasets, we can simulate the case of the rough skeleton annotation process by degrading the current annotations. Through these groups of experiments, We will show that the proposed TSNet cooperates well with the rough skeleton annotation process.

A. Ablation Studies

We first conducted ablation experiments to study the effects of each component in the proposed TSNet, i.e., the Transformer module (recorded as Trans) and the Dual-Path Decoder (recorded as DPD). We took a standard U-Net model with the densely connected convolutional blocks as the baseline model. By inserting the Transformer module at the intermediate position, we had a model referred to as baseline+Trans. By replacing the normal decoder sub-network with the proposed dual-path decoder, we had a model referred to as baseline+DPD. The proposed TSNet can be built by combining the two above replacements. Experiments were conducted on the STARE and the CHASE datasets, and the results are reported in Table I.

According to Table I, the effects of the two modules have been separately verified by the experimental results on the two datasets. Adopting the Transformer module or the dual-path decoder boosted the baseline model in terms of AUC ROC and AUC PR. Moreover, adopting both modules on the baseline model yielded a greater boost. As the vessel segmentation task is severely unbalanced, changes in the predictions only yield minor gaps in the metrics, e.g., Accuracy.

We conduct paired t-tests to verify that the improvement can be taken as a significant gap. The last column gives p-values of paired t-tests between the baseline model and the compared model in terms of AUC PR. All $p < 0.05$ demonstrate that the compared model (last row) performs significantly better than the baseline model. Besides, we visualize the box plots in Fig. 7. Visualization of the result on the STARE and CHASE dataset can be found in Fig. 8, which shows that TSNet yielded a better-detailed segmentation compared with the baseline model.

B. Study on the Loss Function

To study the effect of the proposed skeleton loss function in (8), we further conducted a series of experiments with different settings on the loss function. We first considered the focal loss and the weighted BCE loss, which are commonly used to deal with the imbalance issue. Accuracy, AUC ROC, AUC PR, and

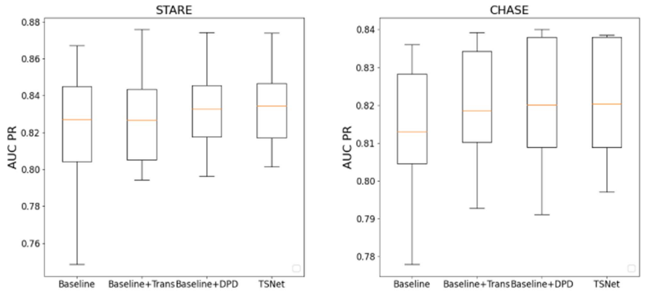

Fig. 7. Box plot of ablation experiments on the STARE (left) and CHASE (right) dataset.

TABLE II
STUDY ON THE LOSS FUNCTION

Method	Acc	AUC ROC	AUC PR	CAL
Weighted BCE	0.9412	0.9867	0.9123	0.7880
Focal loss	0.9654	0.9874	0.9189	0.8010
Proposed (α=0.1)	0.9668	0.9884	0.9235	0.8086
Proposed (α=1)	0.9664	0.9886	0.9244	0.8048
Proposed (α=10)	0.9665	0.9886	0.9238	0.8074

α Denotes the weight of the skeleton loss.

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE STARE, CHASE, DRIVE, AND HRF DATASETS

Dataset	Methods	Accuracy	AUC ROC	Dice
STARE	[18]	0.9612	0.9801	N.A.
	[23]	0.9378	0.9877	N.A.
	[21]	0.9641	0.9832	0.8143
	[50]	0.9641	0.9620	0.8230
	Proposed	**0.9675**	**0.9889**	**0.8300**
CHASE	[18]	0.9610	0.9781	N.A.
	[23]	0.9373	0.9830	N.A.
	[21]	0.9610	0.9804	0.7883
	[51]	0.9639	0.9832	N.A.
	Proposed	**0.9665**	**0.9878**	**0.8211**
DRIVE	[18]	0.9542	0.9752	N.A.
	[23]	0.9271	0.9802	N.A.
	[21]	0.9566	0.9802	0.8237
	[52]	N.A.	0.9810	0.8279
	Proposed	**0.9579**	**0.9835**	**0.8280**
HRF	[18]	0.9437	N.A.	N.A.
	[23]	0.9349	0.9838	N.A.
	[21]	0.9651	0.9831	0.7989
	[52]	N.A.	0.9825	**0.8103**
	Proposed	**0.9662**	**0.9859**	0.8014

the CAL value are adopted for comparisons (Table II). Specifically, we separated thick vessels (more than 5-pixel width) and thin vessels and evaluated them respectively. According to the results of the experiments, the focal loss and the weighted BCE loss yielded worse performances compared with the proposed loss, especially in terms of the thin vessel. It also verifies that the proposed loss is beneficial for handling the imbalance among vessels. To study the effect on the skeleton loss, we also choose different weights for the skeleton loss $\mathcal{L}_{skel}$ in (12). According to the results, we find that the performance of the model changed slightly as the weight of the skeleton loss changed. In the following experiments, we didn't consider justifying the weights of these two losses to avoid over-fitting and set the equal weights for two terms.

C. Comparison With State-of-the-Art Methods

1) Fundus Photography: To verify the superiority of the proposed TSNet model, we evaluated it on four benchmarking fundus photography datasets and compared it with state-of-the-art methods, as shown in Table III. We have ensured that all the compared methods use the same training-test partition as ours for fairness. Following previous papers ([23]), We report Accuracy, AUC ROC, and Dice score for a comprehensive comparison. On the STARE dataset, TSNet achieved an Accuracy of 0.9675 and an AUC ROC of 0.9889. On the CHASE dataset, TSNet achieved an Accuracy of 0.9665 and an AUC ROC of 0.9878. On the HRF dataset, TSNet achieved an Accuracy of 0.9662 and an AUC ROC of 0.9859. In conclusion, TSNet yielded promising performances on four benchmarking fundus photography datasets and outperformed state-of-the-art methods in terms of Accuracy and AUC scores.

2) Fluorescein Angiography: We further considered evaluating the proposed TSNet on the newly emerged OCT-A data. As few works report experiments on the RECOVERY dataset [46], we compared the TSNet model with our re-implementation of commonly compared models, which includes U-Net and its variations, e.g., UNet [39], UNet++ [53], Attention UNet [54], R2UNet [55], DUNet [21], CSNet [56], TransUNet [27], and MISSFormer [57]. We followed the original implementations of these models and trained them on the RECOVERY dataset from scratch. The quantities of parameters for these models are roughly close.

Performances of evaluated models are listed in Table IV. The last row shows p-values by paired t-tests between the TSNet model and the compared model in terms of AUC ROC. All $p < 0.05$ demonstrate that TSNet yielded a significantly higher AUC ROC than other models. In terms of the CAL score and balanced accuracy, TSNet also achieved higher scores compared with other models, which shows that TSNet has a clear advantage on this task and dealing with the issue of imbalanced data. Second, we compared TSNet with the results reported in the original paper [46]. TSNet yielded an AUC ROC of 0.9891, higher than the AUC ROC reported in [46]. In conclusion, TSNet reached state-of-the-art performance on the fluorescein angiography dataset.

We showcase the segmentation results from TSNet and compared models in Fig. 9. For better visualization, red denotes false positives, which are background but misclassified as vessels; blue denotes false negatives, which are vessels but misclassified as background. In the results from other compared models, thin vessels are likely to be ignored. In contrast, TSNet achieved more precise segmentation, especially in detecting thin vessels.

D. Cross-Training

To verify the generalization performance of our method, we conducted cross-training experiments. We choose two papers [19], [58] that report cross-training results for comparison. The results are shown in Table V. Concretely, we trained the model on the training subset in the STARE/CHASE dataset

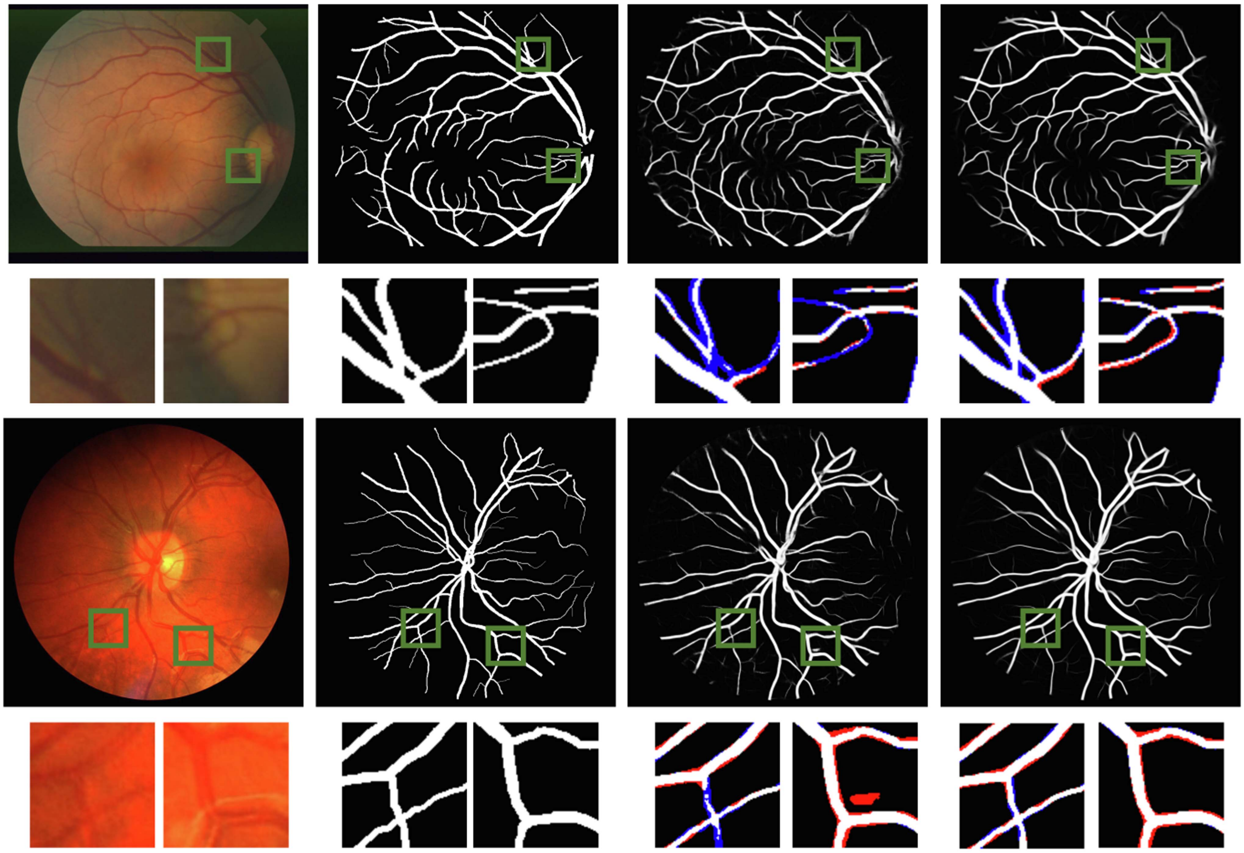

Fig. 8. Visualization of results on the STARE and CHASE dataset. From left to right: original image, ground truth, baseline, and TSNet. Red denotes false positives and blue denotes false negatives.

TABLE IV
Comparison With the State-of-the-Art Methods on the RECOVERY Dataset

Methods	# Para	Acc	AUC ROC	AUC PR	CAL	BAcc	DICE	p-value
Ding et al.	-	-	0.987	0.930	-	-	-	-
UNet	4.86M	0.7623	0.9831	0.8988	0.6042	0.6824	0.9606	<0.0001
UNet++	9.16M	0.7870	0.9862	0.9128	0.6333	0.7215	0.8318	0.0004
DUNet	5.54M	0.7364	0.9716	0.8578	0.5567	0.7020	0.9527	<0.0001
AttUNet	8.65M	0.7986	0.9874	0.9213	0.6482	0.7317	0.9658	0.0248
R2UNet	9.72M	0.6092	0.9475	0.8134	0.3487	0.4647	0.9454	0.0011
CSNet	8.86M	0.7881	0.9853	0.9115	0.6291	0.7223	0.9639	0.0015
TransUNet	6.15M	0.7953	0.9861	0.9163	0.6428	0.7311	0.9600	0.0041
MISSFormer	42.46M	0.8046	0.9876	0.9201	0.6739	0.7528	0.9661	0.0275
TSNet	10.63M	**0.8143**	**0.9891**	**0.9258**	**0.7340**	**0.9011**	**0.9611**	-

p-values are calculated by paired t-tests between the AUC ROC values from the compared models.

TABLE V
Cross-Training on the STARE and CHASE Dataset

Training dataset	Test dataset	Methods	Accuracy	AUC ROC
STARE	CHASE	[58]	0.9417	0.9553
		[19]	0.9411	0.9511
		proposed	**0.9626**	**0.9862**
CHASE	STARE	[58]	**0.9536**	0.9620
		[19]	0.9501	0.9517
		proposed	0.9490	**0.9868**

and tested it on the test subset in the CHASE/STARE dataset. Under both settings, TSNet exceeded the compared methods in terms of AUC scores and yielded comparable performance compared with the performance under non-cross-training, as listed in Table III. In conclusion, cross-training experiments show that the proposed TSNet can generalize well to real-world applications. V.

E. Performances With Skeleton Semi-Supervised Leaning

It is easy to notice that the proposed rough skeleton annotation process can significantly reduce the annotation workload. In our experiments, only 10% samples are provided with pixel-wise annotations, while the other 90% samples are provided with rough skeleton annotations. We further verify that the TSNet can cope with the rough skeleton annotation process and achieve promising performances with low-quality annotations. We

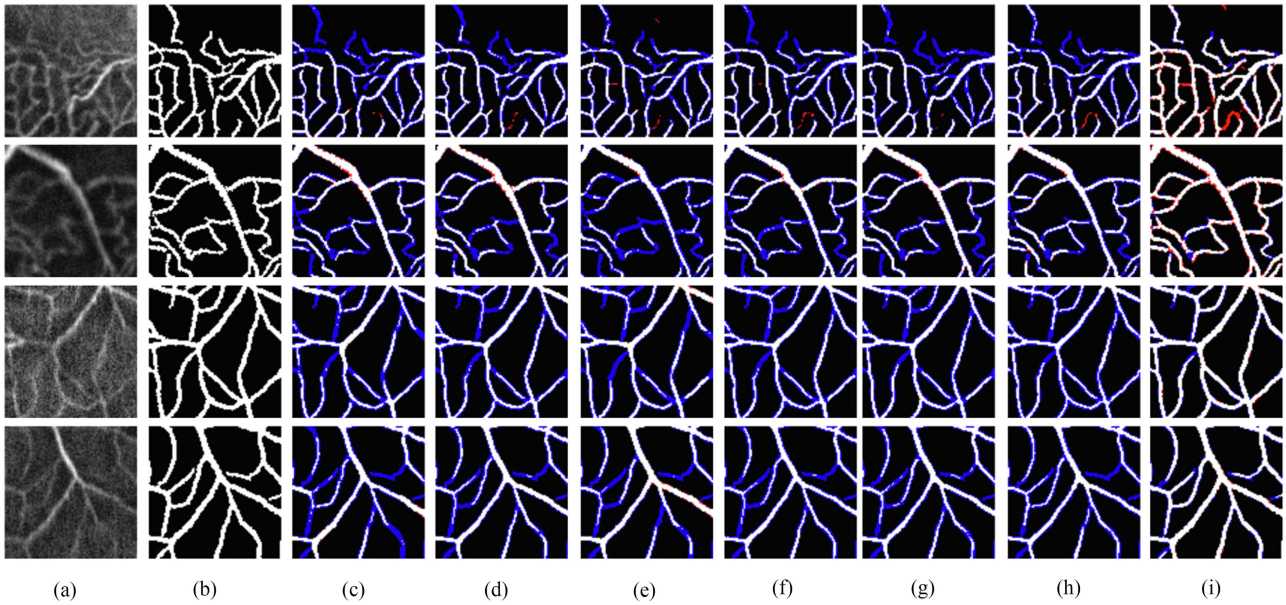

Fig. 9. Visualization of results on the RECOVERY dataset. (a) original image; (b) ground truth; (c) UNet; (c) UNet++; (e) DUNet; (f) AttUNet; (g) CSNet; (h) TransUNet; (i) TSNet. Red denotes false positives and blue denotes false negatives.

TABLE VI
Performances Under Rough Skeleton Annotation, We Conduct Ablation Experiments on the STARE and the CHASE Dataset

		STARE		CHASE	
Model	Training scheme	AUC ROC	AUC PR	AUC ROC	AUC PR
U-Net	10% vessel	0.9789	0.8944	0.9810	0.8748
U-Net	10% vessel + MT	0.9797	0.8934	0.9816	0.8748
TSNet	10% vessel	0.9797	0.8934	0.9839	0.8815
TSNet	10% vessel + MT	0.9800	0.8889	0.9839	0.8845
TSNet	10% vessel + skeleton	0.9798	0.8951	0.9838	0.8742
TSNet	10% vessel + skeleton+MT	0.9824	0.8970	0.9858	0.8918
TSNet	10% vessel + skeleton+MT+AC	**0.9826**	**0.8974**	**0.9862**	**0.8922**

Mt denotes mean teacher; ac denotes annotation correction.

design ablation experiments to study the effects of each component. The training strategy is similar to previous experiments in Section VI-C1.

We start by training the model in a fully supervised manner, in which only 10% samples with precise vessel annotations are used for supervision (recorded as 10% vessel). It means only $\mathcal{L}_{sup}$ in (14) is optimized. Next, considering other samples without vessel annotations, we can use the mean-teacher learning scheme to leverage these data in a semi-supervised manner. It means the first two terms in the (14), $\mathcal{L}_{sup}$ and $\mathcal{L}_{mt}$, are optimized (recorded as 10% vessel+MT). We train a baseline model U-Net and the proposed model TSNet for comparison. The advantages of TSNet are learning capacity attributes to the Transformer module and dual-path decoder. Besides, the dual-path decoder enables the model to leverage skeleton annotations. We further consider feeding rough skeleton annotations for the TSNet (recorded as 10% vessel+skeleton). Furthermore, applying mean teacher corresponds to (14) without (12) (recorded as 10% vessel+skeleton+MT). Finally, the whole skeleton semi-supervised learning scheme is conducted (recorded as 10% vessel+skeleton+MT+AC).

Table VI shows the performances under different experiment settings. According to the results, adopting a mean teacher enables the model to utilize more data without annotations. Therefore, mean-teach brought improvements to U-Net and TSNet. Next, TSNet can be trained under additional skeleton annotation, which provides extra information for the model. As a result, the adoption of skeleton annotation brought further improvement. Finally, considering positional errors in the skeleton annotations, annotation correction can correct these errors and yield the best performance. These experiments verify the effectiveness of the proposed skeleton semi-supervised learning scheme.

Please note that we aim to propose a general learning scheme that can utilize rough skeleton annotation in a semi-supervised manner. This learning scheme has significance for practical applications by saving annotation efforts. To learn in a semi-supervised manner, we adopted a classical and effective method, the mean-teacher model. To learn with rough annotations, we adopted the annotation correction method. With the advancement of research on semi-supervised learning and label noise learning, both modules can also be replaced by more effective methods in future work.

VII. Conclusion

In this paper, we have proposed a novel method, namely TSNet, for retinal vessel segmentation. In a U-Net model, TSNet is built by integrating two novel modules, Transformers and the dual-path decoder. As demonstrated in the paper, the introduction of Transformers enables the model to learn long-range interactions. The designed dual-path decoder is adopted to tackle the issue of imbalance between thick vessels and thin vessels. The contributions of this work are not only proposing a novel model with promising performance but also enabling a rough skeleton annotation process which saves annotation efforts. We have proposed a skeleton semi-supervised learning method to leverage rough skeleton annotations. Experiments show that the proposed method outperforms state-of-the-art methods on five public datasets in two imaging modalities and provides an efficient training scheme.

The contributions for clinical applications can be summarized as follows. First, the proposed method achieves promising performances on retinal vessel segmentation compared with state-of-the-art methods. Specifically, it provides precise segmentation on the tiny vessels, which plays a vital role in diagnosis. Second, the proposed model enables a novel rough skeleton annotation process. The annotators can roughly delineate vessel skeletons instead of giving precise pixel-wise vessel annotations, which takes fewer annotation efforts. The proposed skeleton semi-supervised learning scheme can achieve promising performance with the rough skeleton annotation. This property significantly saves the annotation efforts, which can benefit practical applications.

Although the proposed skeleton semi-supervised learning method has significantly reduced annotation efforts, the limitation of this scheme is that it does not consider annotation mistakes, e.g., missing a vessel in the annotation. Future work can explore this direction to enable more annotation mistakes, which can further decrease the high requirements for annotators.

References

[1] J. L. Leasher et al., “Global estimates on the number of people blind or visually impaired by diabetic retinopathy: A meta-analysis from 1990 to 2010,” *Diabetes Care*, vol. 39, no. 9, pp. 1643–1649, 2016.

[2] B. Bowling, *Kanski’s Clinical Ophthalmology*. Edinburgh, U.K.: Elsevier, 2016.

[3] M. M. Fraz et al., “An ensemble classification-based approach applied to retinal blood vessel segmentation,” *IEEE Trans. Biomed. Eng.*, vol. 59, no. 9, pp. 2538–2548, Sep. 2012.

[4] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, “Multiscale vessel enhancement filtering,” in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 1998, pp. 130–137.

[5] J. Zhang, B. Dashtbozorg, E. Bekkers, J. P. Pluim, R. Duits, and B. M. t. H. Romeny, “Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores,” *IEEE Trans. Med. Imag.*, vol. 35, no. 12, pp. 2631–2644, Dec. 2016.

[6] E. Bekkers, R. Duits, T. Berendschot, and B. t. H. Romeny, “A multi-orientation analysis approach to retinal vessel tracking,” *J. Math. Imag. Vis.*, vol. 49, no. 3, pp. 583–610, 2014.

[7] M. M. Fraz, A. Basit, and S. Barman, “Application of morphological bit planes in retinal blood vessel extraction,” *J. Digit. Imag.*, vol. 26, no. 2, pp. 274–286, 2013.

[8] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, “Iterative vessel segmentation of fundus images,” *IEEE Trans. Biomed. Eng.*, vol. 62, no. 7, pp. 1738–1749, Jul. 2015.

[9] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, “Blood vessel segmentation of fundus images by major vessel extraction and subimage classification,” *IEEE J. Biomed. Health Inform.*, vol. 19, no. 3, pp. 1118–1128, May 2015.

[10] J. I. Orlando, E. Prokofyeva, and M. B. Blaschko, “A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 1, pp. 16–27, Jan. 2017.

[11] L. Zhang, M. Fisher, and W. Wang, “Retinal vessel segmentation using multi-scale textons derived from keypoints,” *Computerized Med. Imag. Graph.*, vol. 45, pp. 47–56, 2015.

[12] P. Liskowski and K. Krawiec, “Segmenting retinal blood vessels with deep neural networks,” *IEEE Trans. Med. Imag.*, vol. 35, no. 11, pp. 2369–2380, Nov. 2016.

[13] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. V. Gool, “Deep retinal image understanding,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2016, pp. 140–148.

[14] Y. Zhang and A. C. Chung, “Deep supervision with additional labels for retinal vessel segmentation task,” in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2018, pp. 83–91.

[15] Y. Wu, Y. Xia, Y. Song, Y. Zhang, and W. Cai, “Multiscale network followed network model for retinal vessel segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2018, pp. 119–126.

[16] Y. Wu et al., “Vessel-net: Retinal vessel segmentation under multi-path supervision,” in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2019, pp. 264–272.

[17] Z. Yan, X. Yang, and K.-T. Cheng, “A three-stage deep learning model for accurate retinal vessel segmentation,” *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1427–1436, Jul. 2019.

[18] Z. Yan, X. Yang, and K.-T. Cheng, “Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation,” *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1912–1923, Sep. 2018.

[19] S. Guo, K. Wang, H. Kang, Y. Zhang, Y. Gao, and T. Li, “BTS-DSN: Deeply supervised neural network with short connections for retinal vessel segmentation,” *Int. J. Med. Inform.*, vol. 126, pp. 105–113, 2019.

[20] A. Oliveira, S. Pereira, and C. A. Silva, “Retinal vessel segmentation based on fully convolutional neural networks,” *Expert Syst. Appl.*, vol. 112, pp. 229–242, 2018.

[21] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, “DUNet: A deformable network for retinal vessel segmentation,” *Knowl.-Based Syst.*, vol. 178, pp. 149–162, 2019.

[22] B. Wang, S. Qiu, and H. He, “Dual encoding u-net for retinal vessel segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2019, pp. 84–92.

[23] S. Y. Shin, S. Lee, I. D. Yun, and K. M. Lee, “Deep vessel segmentation by learning graphical connectivity,” *Med. Image Anal.*, vol. 58, 2019, Art. no. 101556.

[24] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[25] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” 2020, *arXiv:2010.11929*.

[26] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[27] J. Chen et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” 2021, *arXiv:2102.04306*.

[28] X. Shen et al., “Self-attentional microvessel segmentation via squeeze-excitation transformer Unet,” *Computerized Med. Imag. Graph.*, vol. 97, 2022, Art. no. 102055.

[29] D. Chen, W. Yang, L. Wang, S. Tan, J. Lin, and W. Bu, “PCAT-UNET: UNet-like network fused convolution and transformer for retinal vessel segmentation,” *PLoS One*, vol. 17, no. 1, 2022, Art. no. e0262689.

[30] J. E. V. Engelen and H. H. Hoos, “A survey on semi-supervised learning,” *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[31] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[32] D. Chen, Y. Ao, and S. Liu, “Semi-supervised learning method of U-net deep learning network for blood vessel segmentation in retinal images,” *Symmetry*, vol. 12, no. 7, 2020, Art. no. 1067.

[33] J. Hou, X. Ding, and J. D. Deng, “Semi-supervised semantic segmentation of vessel images using leaking perturbations,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 2625–2634.

[34] G. Algan and I. Ulusoy, “Image classification with deep learning in the presence of noisy labels: A survey,” *Knowl.-Based Syst.*, vol. 215, 2021, Art. no. 106771.

[35] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8135–8153, Nov. 2023.

[36] H. Song, M. Kim, and J.-G. Lee, "SELFIE: Refurbishing unclean samples for robust deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5907–5915.

[37] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1944–1952.

[38] L. Huang, C. Zhang, and H. Zhang, "Self-adaptive training: Beyond empirical risk minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 19365–19376.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2015, pp. 234–241.

[40] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[42] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Commun. ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[43] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Trans. Med. Imag.*, vol. 19, no. 3, pp. 203–210, Mar. 2000.

[44] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. V. Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.

[45] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson, "Robust vessel segmentation in fundus images," *Int. J. Biomed. Imag.*, vol. 2013, 2013, Art. no. 154860.

[46] L. Ding, M. H. Bawany, A. E. Kuriyan, R. S. Ramchandran, C. C. Wykoff, and G. Sharma, "A novel deep learning pipeline for retinal vessel detection in fluorescein angiography," *IEEE Trans. Image Process.*, vol. 29, pp. 6561–6573, 2020.

[47] Y. Ma et al., "ROSE: A retinal OCT-angiography vessel segmentation dataset and new model," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 928–939, Mar. 2021.

[48] M. E. Gegúndez-Arias, A. Aquino, J. M. Bravo, and D. Marín, "A function for quality evaluation of retinal vessel segmentations," *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 231–239, Feb. 2012.

[49] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[50] C. Wang, Z. Zhao, and Y. Yu, "Fine retinal vessel segmentation by combining nest u-net and patch-learning," *Soft Comput.*, vol. 25, no. 7, pp. 5519–5532, 2021.

[51] J. Zhang, Y. Zhang, and X. Xu, "Pyramid u-net for retinal vessel segmentation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 1125–1129.

[52] A. Galdran, A. Anjos, J. Dolz, H. Chakor, H. Lombaert, and I. B. Ayed, "State-of-the-art retinal vessel segmentation with minimalistic models," *Sci. Rep.*, vol. 12, no. 1, pp. 1–13, 2022.

[53] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Berlin, Germany: Springer, 2018, pp. 3–11.

[54] O. Oktay et al., "Attention u-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[55] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," 2018, *arXiv:1802.06955*.

[56] L. Mou et al., "CS-net: Channel and spatial attention network for curvilinear structure segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2019, pp. 721–730.

[57] X. Huang, Z. Deng, D. Li, and X. Yuan, "MISSFormer: An effective medical image segmentation transformer," 2021, *arXiv:2109.07162*.

[58] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, and T. Wang, "A cross-modality learning approach for vessel segmentation in retinal images," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 109–118, Jan. 2016.