Computer Methods and Programs in Biomedicine 233 (2023) 107454

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

OCT^{2}Former: A retinal OCT-angiography vessel segmentation transformer

Xiao Tan[a], Xinjian Chen[a,b], Qingquan Meng[a], Fei Shi[a], Dehui Xiang[a], Zhongyue Chen[a], Lingjiao Pan[c], Weifang Zhu[a,*]

[a] MIPAV Lab, the School of Electronic and Information Engineering, Soochow University, Jiangsu, China
[b] The State Key Laboratory of Radiation Medicine and Protection, Soochow University, Jiangsu, China
[c] School of Electrical and Information Engineering, Jiangsu University of Technology, Jiangsu, China

ARTICLE INFO

Article history:
Received 14 October 2022
Revised 25 January 2023
Accepted 27 February 2023

Keywords:
Optical coherence tomography angiography
Retinal vessel segmentation
Transformer
Dynamic token aggregation
Deep learning

ABSTRACT

Background and objective: Retinal vessel segmentation plays an important role in the automatic retinal disease screening and diagnosis. How to segment thin vessels and maintain the connectivity of vessels are the key challenges of the retinal vessel segmentation task. Optical coherence tomography angiography (OCTA) is a noninvasive imaging technique that can reveal high-resolution retinal vessels. Aiming at make full use of its characteristic of high resolution, a new end-to-end transformer based network named as OCT^{2}Former (OCT-a Transformer) is proposed to segment retinal vessel accurately in OCTA images.

Methods: The proposed OCT^{2}Former is based on encoder-decoder structure, which mainly includes dynamic transformer encoder and lightweight decoder. Dynamic transformer encoder consists of dynamic token aggregation transformer and auxiliary convolution branch, in which the multi-head dynamic token aggregation attention based dynamic token aggregation transformer is designed to capture the global retinal vessel context information from the first layer throughout the network and the auxiliary convolution branch is proposed to compensate for the lack of inductive bias of the transformer and assist in the efficient feature extraction. A convolution based lightweight decoder is proposed to decode features efficiently and reduce the complexity of the proposed OCT^{2}Former.

Results: The proposed OCT^{2}Former is validated on three publicly available datasets i.e. OCTA-SS, ROSE-1, OCTA-500 (subset OCTA-6M and OCTA-3M). The Jaccard indexes of the proposed OCT^{2}Former on these datasets are 0.8344, 0.7855, 0.8099 and 0.8513, respectively, outperforming the best convolution based network 1.43, 1.32, 0.75 and 1.46%, respectively.

Conclusion: The experimental results have demonstrated that the proposed OCT^{2}Former can achieve competitive performance on retinal OCTA vessel segmentation tasks.

1. Introduction

A large number of studies have pointed out that diseases such as diabetic retinopathy and cardiovascular diseases are related to the structural and morphological changes of retinal vessels [1–3]. As shown in Fig. 1, retinal optical coherence tomography angiography (OCTA) can reveal high-resolution vessels with various sizes (large vessels, small vessels and capillaries) surrounding the fovea and parafovea regions with its noninvasive depth-resolved imaging, which is a new clinical tool for diagnosis and treatment of retinal vascular disease, choroidal neovascularization, macular degeneration, idiopathic macular fovea telangiectasia and other fundus vascular diseases [4,5]. Recent studies have found that the changes in retinal microvascular structure revealed by OCTA, including microvascular perfusion density, vessel calibers and alterations of vascular network organization, are associated with some neurodegenerative diseases such as Alzheimer's and Parkinson's disease [6–8]. Therefore, the automatic and accurate retinal vessel segmentation in OCTA images is not only a crucial step in the severity evaluation of vascular diseases, but also plays a significant role in the assessment of disease progression and therapeutic effects [9]. However, there are many challenges in retinal vessel segmentation such as difficulties in detail structure (thin vessel, vessel edge and vessel bifurcation) identification and inconsistency of segmented vessels.

There are many previous studies focused on retinal vessel segmentation in fundus images. In traditional algorithms, different thresholding based methods were used for automatic blood vessel segmentation [10–15], which is difficult to determine the op-

* Corresponding author.
E-mail address: wfzhu@suda.edu.cn (W. Zhu).

https://doi.org/10.1016/j.cmpb.2023.107454
0169-2607/

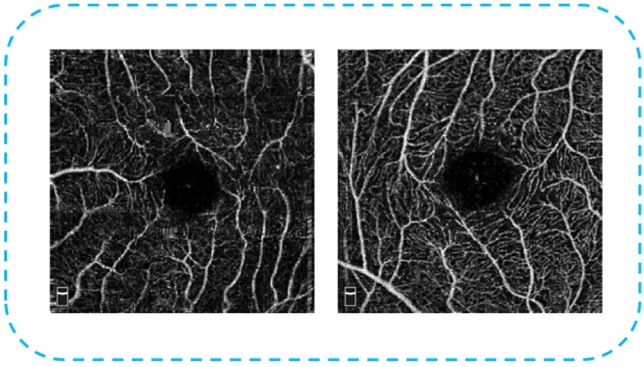

Fig. 1. OCTA images.

timal thresholding and make full use of the geometrical information of the vessel. Various of traditional trainable classifiers such as KNN-classifier [16], Bayesian classifier [17] and Adaboost-classifier [18] were used for retinal vessel segmentation, whose robustness is relatively poor. With the rapid development of deep learning, convolutional neural networks (CNN) have been widely used in medical segmentation tasks [19–23]. Wang et al. applied the UNet for retinal vessel segmentation in fundus images [24].Jin et al. proposed a modified U-Net for retinal vessel segmentation, which introduced deformable convolution into the structure [25]. Wu et al. cascaded two U-shape encoder-decoder structures together to get further refined retinal vessel segmentation results [26]. Mou et al. proposed CS^2-Net for the automatic detection of curvilinear structure including blood vessel and nerve fiber, in which channel and spatial attention modules are utilized to aggregate the local and global curvilinear structure features [27]. Fu et al. integrated CNN and conditional random field to learn both hierarchical representation and long-range information for retinal vessel segmentation [28]. Gu et al. proposed a U-shaped network called context encoder network (CE-Net), which used dense atrous convolution block and residual multi-kernel pooling block to capture more high-level information and preserve spatial information for vessel segmentation [29]. Ye et al. proposed a multiscale feature interaction network (MFI-Net) for retinal vessel segmentation, which is equipped with pyramid squeeze-and-excitation (PSE) module to learn multiscale features and handle vessels with variable width and coarse-to-fine (C2F) module to preserve vessel details during the decoding process [30]. Wu et al. proposed a scale and context-sensitive network (SCS-Net) for retinal vessel segmentation, in which the scale-aware feature aggregation (SFA) module is designed for multi-scale feature extraction and aggregation, the adaptive feature fusion (AFF) module is designed for the guidance of efficient feature fusion and the multi-level semantic supervision (MSS) module is employed to learn distinctive semantic representation for refining the vessel maps [31]. Yuan et al. proposed the AACA-MLA-D-UNet for retinal vessel segmentation, which aims to fully utilize the low-level detailed information and the complementary information encoded in different layers to accurately distinguish the vessels from the background with low model complexity [32]. Wu et al. proposed a lightweight deep learning model named as Vessel-Net for retinal vessel segmentation, which combines the advantages of the inception model and residual module for improved feature representation, and uses four deep supervision paths to preserve multi-scale deep features during model optimization [33].

Due to the superiority of OCTA in visualizing the retinal plexuses, many researchers have devoted their attention to the retinal vessel segmentation in OCTA images in recent years. Yousefi et al. combined multi-scale Hessian filters and intensity-based method for retinal vessel segmentation in OCTA images [34]. Eladawi et al. proposed a joint Markov-Gibbs random field model to segment the retinal blood vessels, which integrates both appearance and spatial information in OCTA images [35]. Sarabi et al. used a segmentation pipeline which consists of curvelet-based denoising, optimally oriented flux vessel enhancement and projection artifact removal for OCTA retinal vessel segmentation [36]. Ma et al. proposed a split-based coarse-to-fine vessel segmentation network for OCTA images [37], in which a split-based coarse segmentation module was applied to generate the preliminary confidence map of vessels and a split-based refined segmentation module was used to optimize the shape/contour of the vessels. Li et al. proposed an end-to-end image projection network (IPN) to achieve 3D-to-2D retinal vessel and foveal avascular zone segmentation in OCTA images [38]. Based on [38], Li et al. proposed the image projection network V2 (IPN-V2) [39], which extends IPN by adding a plane perceptron to enhance the perceptron ability in the horizontal direction. Wu et al. proposed a progressive attention-enhanced network (PAENet) for 3D to 2D retinal vessel segmentation, which consists of the 3D feature learning path and the 2D segmentation path [40]. Li et al. proposed a novel image magnification network (IMN) for vessel segmentation in OCTA images, which consists of an up-sampled encoding path and a down-sampled decoding path to capture more image details and reduce the omission of thin-and-small structures [41]. Menten et al. presented a pipeline to synthesize large amounts of realistic OCTA images with intrinsically matching ground truth labels, including a physiology-based simulation that models the various retinal vascular plexuses and a suite of physics-based image augmentations that emulate the OCTA image acquisition process, which can improve the vessel segmentation performance of several segmentation networks such as U-Net, CS-Net and CE-Net on the publicly available ROSE-1 dataset [42]. Li et al. proposed a retinal image projection segmentation network (RPS-Net) to achieve retinal vessel and foveal avascular zone segmentation in OCTA images, in which a dual-way projection learning module is designed to extract global planar features and local detail supplements simultaneously [43]. Pissas et al. proposed a recurrent CNN named as iUNet for vessel segmentation in OCTA images, which iteratively refines the quality of the produced vessel with weight sharing coupled with a perceptual loss [44].

Although above CNN based segmentation networks have achieved good performances in retinal vessel segmentation, the local and limited receptive field of CNN, which is critical for semantic segmentation, is still one of its shortcomings. Different from CNN, the forms of the global input embeddings allow transformer to have a global receptive field, solving the limited receptive field problem of CNN in each layer [45]. Therefore, many works have attempted to introduce transformer into semantic segmentation tasks [46–48]. Zheng et al. first introduced transformer into semantic segmentation and achieved promising results on ADE20K, Pascal Context and Cityscapes datasets [45]. Xie et al. proposed a simple, efficient and powerful semantic segmentation framework named as SegFormer, which contains a positional-encoding-free, hierarchical transformer encoder and a lightweight All-MLP (Multi-Layer Perceptron) decoder. SegFormer achieved good segmentation performance on ADE20K and Cityscapes datasets [48]. Chen et al. first used transformer for multi-organ segmentation in CT images, which laid the foundation for the transformer in many medical image segmentation tasks later [49]. Zhang et al. combined the convolutional network with transformer from a local and global perspective for ultrasound image segmentation [50]. Ji et al. replaced the skip connection in U-Net with a transformer structure and attained significant segmentation improvements on six biomedical image benchmarks including Pannuke, CVC-Clinic, CVC-Colon, Etis, Kavirs and ISIC2018 datasets [51]. Shen et al. developed a CNN-Transformer hybrid network for micro-vessel segmentation and outperformed several state-of-the-art vessel segmentation net-

works on DRIVE, STARE, HRF and CHASE-DB1 datasets [52]. Because the retinal vessels are characterized by the tree-like topological structure that the thick vessel are usually connected with several thin vessels, the global receptive field property of the transformer will be applied for the retinal vessel segmentation in OCTA images in this paper, which is crucial for segmenting thin vessel and maintaining the connectivity of vessels. However, there are still two challenges to be overcome. One is that the calculation of transformer is huge. The other is that the convergence of the transformer is slow, which may be difficult to converge on commonly small medical image datasets such as retinal OCTA image datasets. Therefore, can transformer be applied to retinal OCTA vessel segmentation task? How to solve the problem of the expensive calculation in transformer? How to solve the problem of the slow convergence of the transformer? These three questions are the focuses of this paper.

In summary, the main contributions of this paper can be highlighted as:

(1) A novel hybrid transformer OCT^{2}Former (OCT-a Transformer) is proposed for retinal OCTA vessel segmentation.
(2) Multi-head dynamic token aggregation attention based dynamic token aggregation transformer is proposed to capture global retinal vessel information and reduce the expensive calculation in both time and space perspectives.
(3) Auxiliary convolution branch is designed to compensate for the lack of inductive bias of the transformer, which can speed up the convergence of the proposed OCT^{2}Former with negligible increase of parameters.
(4) Comprehensive experiments on three publicly available retinal OCTA datasets including OCTA-SS [53], ROSE-1[38] and OCTA-500[40] (subset OCTA-6M and OCTA-3M) indicate the effectiveness of the proposed OCT^{2}Former.

2. Methods

2.1. Network architecture

Fig. 2 shows the architecture of the proposed OCT^{2}Former, which is based on the U-shape encoder-decoder structure with skip connections. The encoder path consists of the dynamic transformer encoder and the group embedding module. Two convolution based lightweight decoders and a 1×1 convolution constitute the decoder path. The novelly proposed dynamic transformer encoder consists of dynamic token aggregation transformer and auxiliary convolution branch. As is shown in Fig. 2, when the OCTA image is fed into the proposed OCT^{2}Former, a convolutional stem module consisted of two 3×3 convolutions is first applied to the original OCTA image to obtain the primary feature maps and increase the number of channels without resolution change. Then, these features are fed into three consecutive dynamic transformer encoders to obtain semantic tokens with rich global information level by level. In order to capture multi-scale information and make up for the loss of position information of the dynamic transformer encoders, the semantic tokens from the first two dynamic transformer encoders are then fed into the group embedding modules, respectively. In the decoder path, the multi-scale semantic tokens from dynamic transformer encoders are fed into the lightweight decoder to recover original resolution. At the end of the decoder path, a 1×1 convolution is applied to obtain the final segmentation map.

2.2. Dynamic transformer encoder

2.2.1. Vision transformer

Typically, as shown in Fig. 3(a), a vision transformer (ViT) consists of a stack of self-attention (SA) layers followed by a feed-forward network, with the main idea of processing the images in a sequence-to-sequence manner and taking self-attention mechanism between each sequence [54]. For a given entity in the sequence, SA layer essentially consists of the dot product of the query and all keys and the normalized attention score acquisition via the *softmax* operator. Through the SA layer, the entity becomes a weighted sum of all entities in the sequence, in which the weights are determined by the normalized attention score. The function of SA layer between different input embeddings can be calculated as follows:

$$SA(Q,\ K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \tag{1}$$

where *K, Q, V* are the projections of key, query and value, respectively. d_k is the dimension of the key projection K and $\sqrt{d_k}$ provides a normalization to make the gradient more stable.

To encapsulate multiple complex relationships between different positions in the sequence, multi-head self-attention (MHSA) consisted of multiple self-attention blocks is taken in transformer. Different from single-head self-attention, MHSA is aimed to learn sequence-to-sequence information in different representation subspaces. MHSA divides the input into M heads ($head_1$,...$head_i$,..., $head_M$), calculates the self-attention of each head in parallel, and concatenates them to get the final output. The formula of MHSA can be written as follows:

$$head_i = SA\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{2}$$

$$MHSA(Q, K, V) = Concat\left((head_1, \ldots head_i, \ldots, head_M)W^O\right) \tag{3}$$

where W_i^Q W_i^K $W_i^V (i = 1, 2, \ldots, M)$ W^O are the independently learnable weight matrices.

Each MHSA layer is followed by a multi-layer perceptron (MLP), which is consisted a linear layer and a GeLU activation. Similar to residual networks, both MHSA and MLP have a skip-connection and a normalization layer. The output of operations mentioned above can be described as:

$$\hat{x}_l = MHSA(LN(x_{l-1})) + x_{l-1} \tag{4}$$

$$x_l = MLP\left(LN\left(\hat{x}_l\right)\right) + \hat{x}_l \tag{5}$$

where *LN* represents the layer normalization, x_l and x_{l-1} represent the output of l_{th} and $(l-1)_{th}$ layer, respectively.

2.2.2. Dynamic token aggregation transformer

Although transformer has the advantage of global receptive field capture capability, the expensive calculation may limit its application in semantic segmentation tasks. Some previous works took $4 \times$ or even $8 \times$ down-sampling operation to reduce the calculation of transformer, which can be hardly adopted in the retinal vessel segmentation task, because the thin vessels will disappear with the high ratio down-sampling operation. How to design a transformer to overcome the huge computational complexity and the resolution change issue is crucial for retinal vessel segmentation. In vision transformers, the tokens usually contain a large amount of redundant information, with only a subset of most informative tokens contributing to the final prediction [55]. To prune redundant tokens, inspired by deep learning based superpixel sampling [56], we propose a novel approach called multi-head dynamic token aggregation attention (MDTAA), which attempts to dynamically aggregate relevant tokens in the embedding and remove the redundant information. Fig. 3 (b) shows the architecture of the proposed dynamic token aggregation transformer, in which the proposed MDTAA module replaces with the MHSA module in the original transformer.

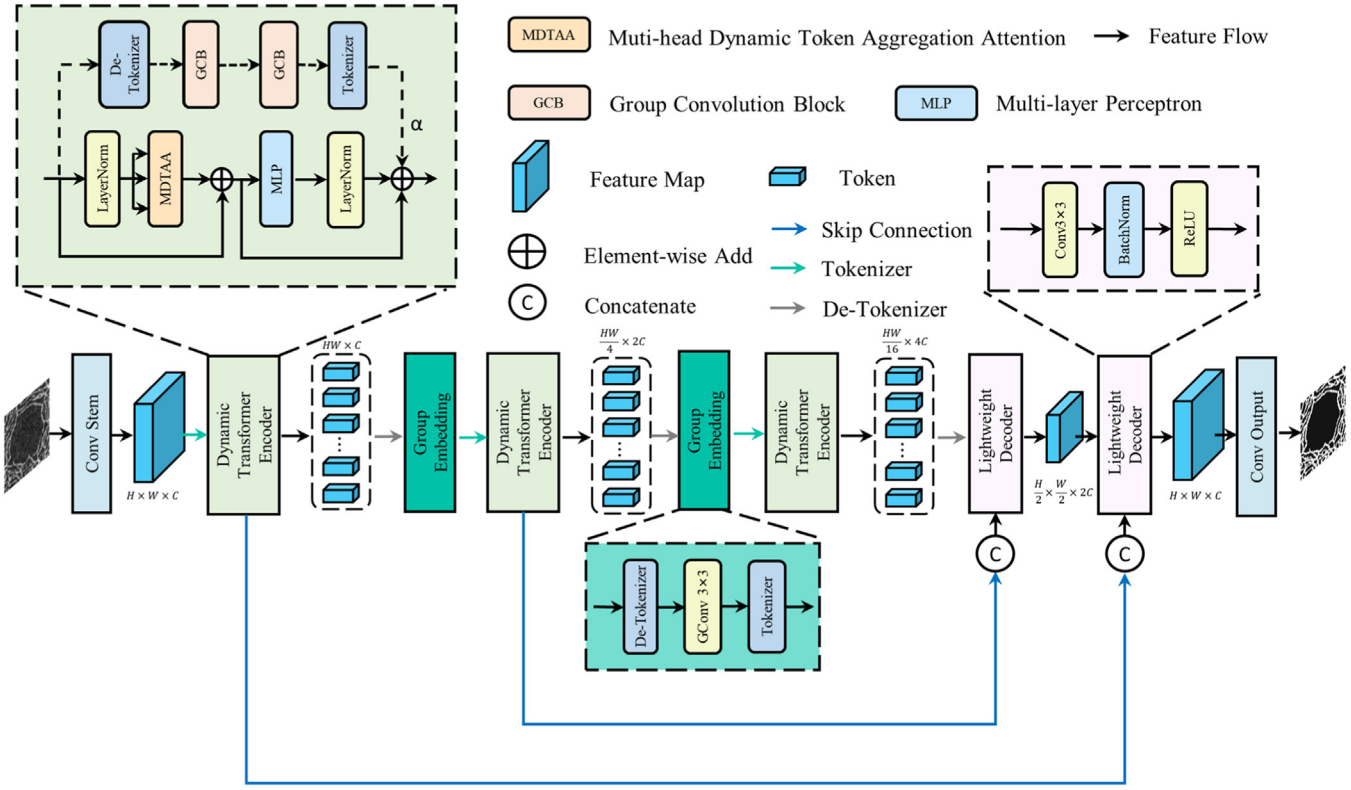

Fig. 2. An overview of the proposed OCT^{2}Former.

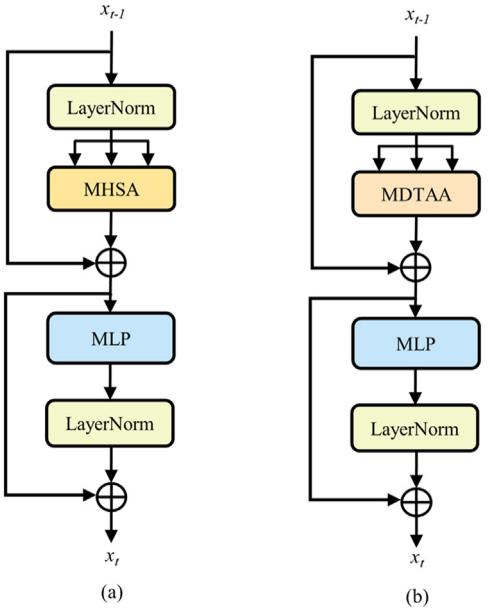

Fig. 3. Structure of vision transformer (ViT) and the proposed dynamic token aggregation transformer. (a) ViT, (b) The proposed dynamic token aggregation transformer. MHSA represents multi-head self-attention, MDTAA represents multi-head dynamic token aggregation attention, and MLP represents multi-layer perceptron.

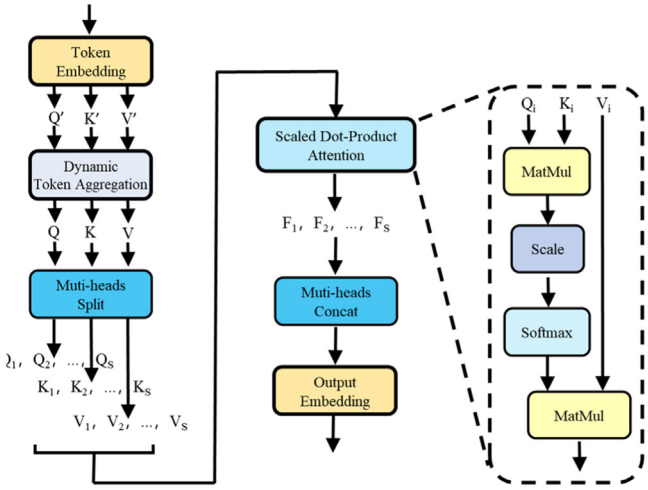

Fig. 4. The architecture of multi-head dynamic token aggregation attention (MDTAA).

Fig. 4 shows the detailed architecture of the proposed MDTAA module. In order to effectively utilize the input token embedding $\mathcal{T} \in \mathbb{R}^{D\times N}$, where D denotes the dimension of embedding and N is the length of embedding, our main idea is to create a new, condensed token embedding named as aggregation embedding $\mathcal{T}_{ag} \in \mathbb{R}^{D\times k}(k << N$, k is set to 128 in this paper), which can effectively aggregate the important information and reduce the redundant information from the original embedding. The input token embedding $\mathcal{T}$ is first fed into initialization unit to generate the initial aggregation embedding $\mathcal{T}_{ag} \in \mathbb{R}^{D\times k}$, which initially aggregates tokens that are adjacent in pixel location via an adaptive average pooling layer.

$$\mathcal{T}_{ag} = Init(\mathcal{T}) \tag{6}$$

The close proximity between tokens in pixel location may not always be reflected in high-dimensional semantic space. As a result, the $\mathcal{T}_{ag} \in \mathbb{R}^{D\times k}$ captures some of the global information from the original embedding, but its retained semantic information is limited. To dynamically retain more abstract information and elim-

inate redundant information, the cosine distance map $\mathcal{M}(\mathcal{T}_{ag}, \mathcal{T}) \in \mathbb{R}^{N\times k}$ is adopted to measure the similarity between $\mathcal{T}_{ag}$ and $\mathcal{T}$ in a high-dimensional vector space, where the higher value indicates lower similarity, and vice versa. For each $(i,\ j)$ candidate representation pair $\tau_{ag}^{i} \in \mathcal{T}_{ag}$ and $\tau^{j} \in \mathcal{T}$ in embedding dimension D, the cosine distance between them can be formulated as:

$$\cos\left(\tau_{ag}^{i},\ \tau^{j}\right) = \frac{\left(\tau_{ag}^{i}\right)^{T} \cdot \tau^{j}}{\tau_{ag}^{i} \cdot \tau^{j}} \tag{7}$$

The cosine distance map $\mathcal{M} \in \mathbb{R}^{N\times k}$ is constructed by each $\cos(\tau_{ag}^{i},\ \tau^{j})$, and can be defined as follows,

$$\mathcal{M}\left(\mathcal{T}_{ag}, \mathcal{T}\right) = \begin{bmatrix} \cos\left(\tau_{ag}^{0},\ \tau^{0}\right) & \cos\left(\tau_{ag}^{0},\ \tau^{1}\right) & \cdots & \cos\left(\tau_{ag}^{0},\ \tau^{N}\right) \\ \cos\left(\tau_{ag}^{1},\ \tau^{0}\right) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \cos\left(\tau_{ag}^{k},\ \tau^{0}\right) & \cdots & \cdots & \cos\left(\tau_{ag}^{k},\ \tau^{N}\right) \end{bmatrix} \tag{8}$$

If as expected, the maximum value in each N dimension of $\mathcal{M}(\mathcal{T}_{ag}, \mathcal{T})$ represents the most informative token that should be preserved. However, using the *argmax* function to identify these tokens is not desirable, because it is not differentiable and would result in other tokens being dropped which should be aggregated initially. To overcome this issue, a *softargmax* (a smooth version of *argmax*) function is used on $\mathcal{M}(\mathcal{T}_{ag}, \mathcal{T})$ to obtain a soft assignment map $\Omega \in \mathbb{R}^{N\times k}$, which is differentiable and can better reflect the correlation between $\mathcal{T}_{ag}$ and $\mathcal{T}$.

$$\Omega = softargmax(\mathcal{M}(\mathcal{T}_{ag}, \mathcal{T})) \tag{9}$$

Therefore, each aggregation token can be regarded as the weighted summation or aggregation of the original token embeddings, where the informative tokens have larger weights and vice versa. With the optimized Ω_{opt}, the final aggregation embedding $\mathcal{T}$ag can be represented as,

$$\mathcal{T}_{ag} = \mathcal{T} \cdot \Omega_{opt} \tag{10}$$

However, since the initial distance between the aggregation embedding and the original embedding is random, the token obtained via single aggregation is relatively rough. To optimize the Ω better, an iteration optimization algorithm named as dynamic token aggregation (DTA) is designed as the following Algorithm 1, in which the iteration number T is a super-parameter and selected according to the prior task-specific information and set to 3 in this paper as the iteration stop condition.

As can be seen from Fig. 4, the projections of query and key (Q',K') are sent into the dynamic token aggregation operation, respectively to get the aggregated projections of query, and key(Q, K). Then the dynamic token aggregation attention can be formulated as:

$$DTAA(Q,\ K, V) = softmax\left(\frac{DTA(Q) \cdot DTA(K)^{T}}{\sqrt{d_k}}\right) \cdot V \tag{11}$$

where *DTAA* represents the dynamic token aggregation attention, and *DTA* represents the dynamic token aggreagation operation.

Aiming to learn sequence-to-sequence information in the different representation subspaces, multi-head dynamic token aggregation attention (MDTAA) is adopted as follows:

$$head_i = DTAA\left(QW_i^{Q}, KW_i^{K}, VW_i^{V}\right) \tag{12}$$

$$MDTAA(Q, K, V) = Concat\left(head_1, \ldots, head_i,\ \ldots,\ head_M\right)W^{O} \tag{13}$$

where $W_i^{Q}, W_i^{K}, W_i^{V} (i = 1, 2, \ldots, M), W^{O}$ are independently learnable weight matrices.

Algorithm 1
Dynamic Token Aggregation (DTA).

Input: Projection token embedding $\mathcal{T} \in \mathbb{R}^{D\times N}$

Output: Aggregation token embedding $\mathcal{T}_{ag} \in \mathbb{R}^{D\times k}$
1.Initialize the aggregation embedding $\mathcal{T}_{ag}$,
$\mathcal{T}_{ag} = Init(\mathcal{T})$
2.For iteration from 1 to T do:

(1) Calculate cosine distance $\cos(\tau_{ag}^{i},\ \tau^{j})$ between each candidate representation pairs τ_{ag}^{i} and τ^{j},
$\cos(\tau_{ag}^{i},\ \tau^{j}) = \frac{(\tau_{ag}^{i})^{T} \cdot \tau^{j}}{\tau_{ag}^{i} \cdot \tau^{j}}$
(2) Construct cosine distance map,
$$\mathcal{M}(\mathcal{T}_{ag}, \mathcal{T}) = \begin{bmatrix} \cos(\tau_{ag}^{0},\ \tau^{0}) & \cdots & \cdots & \cos(\tau_{ag}^{0},\ \tau^{N}) \\ \cos(\tau_{ag}^{1},\ \tau^{0}) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \cos(\tau_{ag}^{k},\ \tau^{0}) & \cdots & \cdots & \cos(\tau_{ag}^{k},\ \tau^{N}) \end{bmatrix}$$
(3) Calculate soft assignment Ω,
$\Omega = softargmax(\mathcal{M}(\mathcal{T}_{ag}, \mathcal{T}))$
(4) Update $\mathcal{T}_{ag}$,
$\mathcal{T}_{ag} = \mathcal{T} \cdot \Omega$

3.End for
4.Return $\mathcal{T}_{ag}$

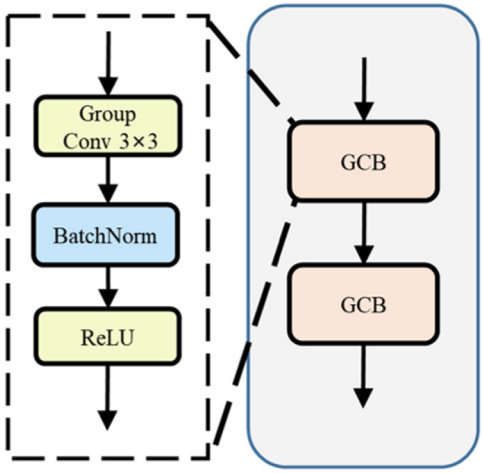

Fig. 5. The architecture of auxiliary convolution branch (ACB). GCB represents group convolution block.

With the dynamic token aggregation attention, the computational complexity of the proposed dynamic token aggregation transformer can be reduced from $O(N^2)$ to $O(kN)$. As k is set to 128 in this paper, the computational complexity can be approximately equal to $O(N)$.

2.2.3. Auxiliary convolution branch

Due to the lack of inductive bias in transformer, the proposed dynamic token aggregation transformer based model suffers from slow convergence problem, which is a typical challenge in transformer. To overcome this problem, an auxiliary convolution branch (ACB) is design to supplement the inductive bias for transformer, which is shown in Fig. 5. As shown in Fig. 5, the ACB consists of two cascaded group convolution blocks (GCB), each of which is composed of a 3 × 3 group convolution, a batch normalization and a ReLU activation. With the ACB, the proposed OCT^{2}Former can converge faster with smaller loss fluctuations. Although the ACB offers limited contribution to the overall performance improvement, it can speed up the convergence of the proposed OCT^{2}Former with negligible increase of parameters, which is of great significance to apply transformer in retinal vessel segmentation.

2.2.4. Token fusion

The token from dynamic token aggregation transformer X_t and the feature from auxiliary convolution branch X_c are fused to obtain the encoded tokens from dynamic token aggregation transformer encoder $X_{encoder}$ as follows:

$$X_{encoder} = \alpha \cdot f(X_c) + X_t \tag{14}$$

where $f(\cdot)$ is a feature tokenizer function converting 2D features to 1D sequence and α is a learnable parameter and is initialized to 0.1.

2.3. Group embedding module

In order to obtain multi-scale features, pooling operation is commonly inserted after the encoder in U-shape network. Following this principle, a group embedding module (GEM) is designed and inserted after the first two dynamic transformer encoders, respectively, which consists of a de-tokenizer, a 3 × 3 group convolution and a tokenizer (shown in Fig. 2). Each semantic token $X_{encoder}$ generated by dynamic token aggregation transformer is de-tokenized into 2D shapes, fed into a 3 × 3 group convolution with stride 2 and tokenized to obtain the output of GEM X_{GEM}, which can be formulated as:

$$X_{GEM} = f\left(GConv_{3x3}(f^{-1}(X_{encoder}))\right) \tag{15}$$

where $f(\cdot)$ is a feature tokenizer function, $f^{-1}(\cdot)$ is a feature de-tokenizer function and $GConv_{3x3}(\cdot)$ is a 3 × 3 group convolution with stride 2.

Different from the traditional pooling operation and linear embedding, GEM can not only compensate for the loss of position information after feature tokenization, but also prevent the loss of detailed information by storing it via multiple channels during the resolution decreasing, which is of great significance for the segmentation of thin retinal vessels with blurred boundaries.

2.4. Lightweight decoder

In encoder-decoder architecture, decoder is designed to recover the spatial resolution of semantic feature maps from the encoder. To recover the spatial resolution efficiently and reduce the parameters of the proposed OCT^{2}Former, convolution based lightweight decoder (LD) is designed to constitute the decoder path. LD consists of a lightweight convolution unit (a 3 × 3 convolution, a batch normalization and a ReLU activation) and a bilinear up-sampling unit.

Given the feature maps $F_{in} \in \mathbb{R}^{C_{in} \times H \times W}$ from encoder, the up-sampled feature maps $F_{out} \in \mathbb{R}^{C_{out} \times \sigma H \times \sigma W}$ via LD can be formulated as:

$$F_{out} = BU(ReLU(BN(Conv_{3x3}(F_{in})))) \tag{16}$$

where $Conv_{3x3}(\cdot)$ represents a 3 × 3 convolution, $BN(\cdot)$ is a batch normalization layer, $ReLU(\cdot)$ is a ReLU activation and $BU(\cdot)$ means a bilinear up-sampling unit.

2.5. Loss function

In this paper, binary cross-entropy loss (BCE) is used as the loss function for training the network, as it is a pixel-wise loss function that directly evaluates the distance between the label and the prediction. The *BCE* loss is defined as follows:

$$\mathcal{L}_{BCE} = -\frac{1}{M}\sum_{i=1}^{M} g_i \cdot \log(p_i) + (1 - g_i) \cdot \log(1 - p_i) \tag{17}$$

where $g_i \in \{0, 1\}$ indicates the ground truth, $p_i \in [0, 1]$ is the corresponding predicted value, and M is the number of pixels.

3. Experiment settings

3.1. Datasets

Three publicly available datasets including OCTA-SS [53], ROSE-1 [37] and OCTA-500 [39] (subset OCTA-6M and OCTA-3M) are adopted to evaluate the performance of the proposed OCT^{2}Former. Fig. 6 shows some examples of OCTA images and the corresponding pixel-level ground truth.

3.1.1. OCTA-SS

OCTA-SS is provided by Usher Institute, University of Edinburgh, UK, which is an open dataset of retinal parafoveal OCTA images with associated manual vasculature segmentations from 11 participants. Imaging was performed using the commercial RTVue-XR Avanti OCT system (Optovue, Fremont, CA). The superficial layer (containing the vasculature enclosed in the internal limiting membrane (ILM) and the inner plexiform layer (IPL)) with 3mm × 3mm field-of-view (FOV) from left and right eyes of 11 participants were selected to generate *en face* angiograms. For each of those images, five sub-images were extracted from each clinical region of interest (ROI): superior, nasal, inferior, temporal and fovea. Poor quality ROIs were discarded and from the remaining a dataset containing 55 ROIs was created. The size of each ROI image is 91 × 91. The dataset is splitted into training set (27 images), validation set (3 images) and testing set (25 images), which is consistent with Reference [53].

3.1.2. ROSE-1

Dataset ROSE-1 is provided by Cixi Institute of Biomedical Engineering, Ningbo Institute of Industrial Technology, Chinese Academy of Sciences, China. ROSE-1 contains 117 OCTA images from 39 subjects (including 26 with Alzheimer's disease and 13 healthy controls), which were captured by the RTVue XR Avanti SD-OCT system (Optovue, USA) equipped with AngioVue software. The OCTA scan area was 3 × 3 mm^2 centered on the fovea with the image resolution of 304 × 304 pixels. The superficial vascular complexes (SVC) *en face* angiograms (39 images) with pixel-wise ground truth are selected in our study. The dataset is splitted into training set (27images), validation set (3 images) and testing set (9 images), which is consistent with Reference [37].

3.1.3. OCTA-500

OCTA-500 is provided by School of Computer Science and Engineering, Nanjing University of Science and Technology, China, which contains two subsets including OCTA-6M from 300 subjects with 6mm × 6mm FOV and OCTA-3M from 200 subjects with 3mm × 3mm FOV. The data were collected using a commercial 70 kHz spectral-domain OCT system with a center wave-length of 840 nm (RTVue-XR, Optovue, CA). OCTA maximum projection between internal limiting membrane (ILM) and Bruch's membrane (BM), which can clearly reveal the vascular morphology of the inner retina and is the commonly used OCTA projection map for retinal vessel segmentation, is adopted in our study. Pixel-wise ground truth were manually drawn by five trained researchers and reviewed by three ophthalmologists. OCTA-6M and OCTA-3M are independently used to evaluate the retinal vessel segmentation performance of the proposed OCT^{2}Former and other networks. OCTA-6M is split into training set (NO.10001-NO.10180), validation set (NO.10181- NO.10200) and test set (NO.10201-NO.10300), and OCTA-3M is split into training set (NO.10301-NO.10440), validation set (NO.10441-NO.10450) and test set (NO.10451-NO.10500), which are consistent with References [38] and [39].

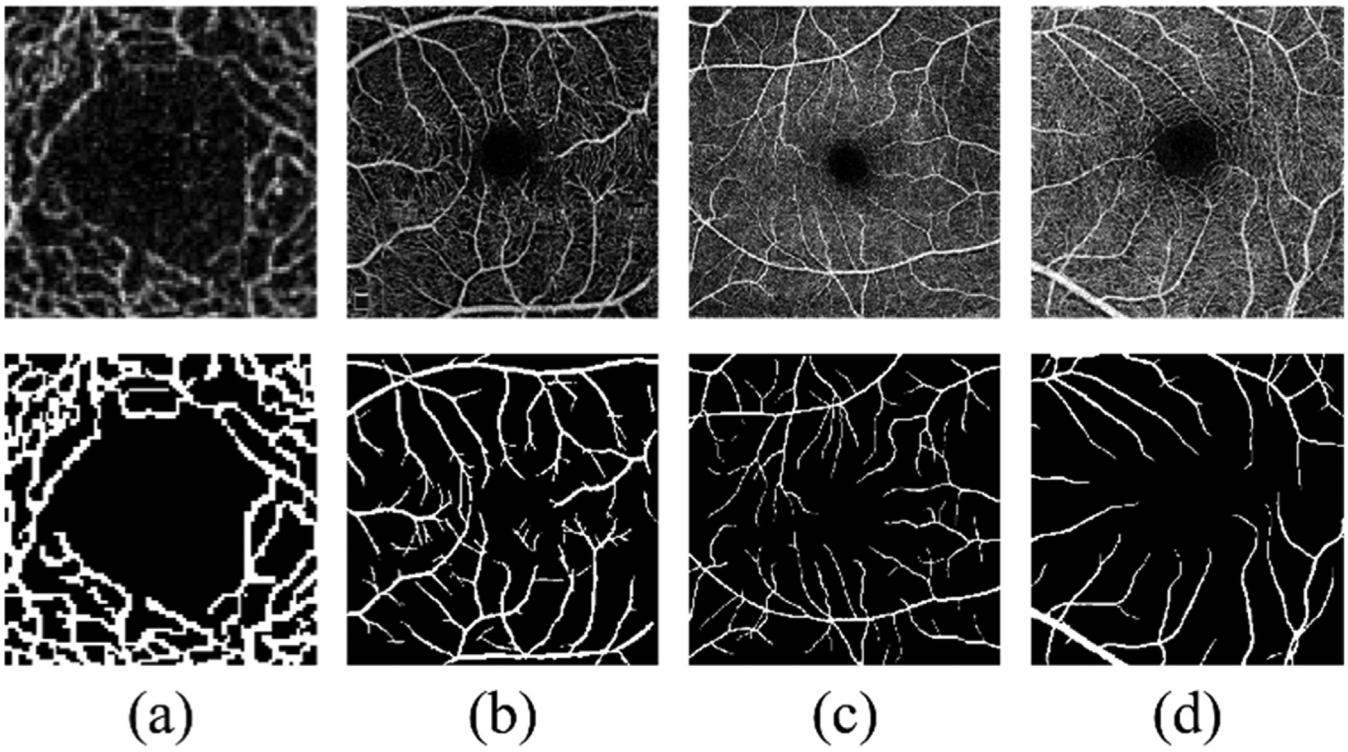

Fig. 6. Examples of OCTA image and pixel-wise ground truth from three datasets. (a)OCTA-SS, (b) ROSE-1, (c) OCTA-500: subset OCTA-6M, (d) OCTA-500: subset OCTA-3M.

3.2. Implementation details

The proposed OCT^{2}Former is implemented on the pyt-orch platform with one NVIDIA RTX3090 GPU with 24GB memory. For fairness, both OCT^{2}Former and other networks are trained with 100 epochs and batch size of 2. Adaptive moment estimation (Adam) optimization with momentum 0.9 and weight decay 0.001, and ploy learning rate $lr = base\_lr(1 - iter/(max\_iter))^{power}$ with power 0.9 are employed. Random left-right flipping, top-down flipping and rotation from $-10°$ to $10°$ are applied for data augmentation. We have released the codes on Github (https://github.com/coreeey/OCT2Former).

3.3. Evaluation metrics

To objectively evaluate the pixel-level segmentation performance of our proposed OCT^{2}Former, five metrics including Jaccard index (JAC), Dice coefficient (DICE), balanced accuracy (BACC), precision (PRE) and recall (REC) are adopted, which are defined as follows:

$$JAC = TP/(TP + FP + FN) \tag{18}$$

$$DICE = 2TP/(2TP + FP + FN) \tag{19}$$

$$BACC = \frac{(TPR + TNR)}{2} \tag{20}$$

$$PRE = TP/(TP + FP) \tag{21}$$

$$REC = TPR = TP/(TP + FN) \tag{22}$$

$$TNR = TN/(TN + FP) \tag{23}$$

where TP and FP represent true postive and false postive, respectively, TN and FN respresent true negative and false negative, respectively, TPR is true positive rate, and TNR is true negative rate.

Furthermore, for the evaluation of the topology-level segmentation performance, connectivity area length metric (CAL) and largest connected component ratio (LCC) [53] are also adopted. CAL metric is based on three descriptive features: (1) Connectivity (C), to assess the fragmentation degree between segmentations, (2) Area (A), to evaluate the degree of overlapping, (3) and Length (L), to capture the degree of coincidence, which are defined as follows:

$$C(S, G) = 1 - min\left(1, \frac{|\#_C G - \#_C S|}{\#G}\right) \tag{24}$$

$$A(S, G) = \frac{\#((\delta_\alpha(S) \cap G) \cup (S \cap \delta_\alpha(G))}{\#(S \cup G)} \tag{25}$$

$$L(S, G) = \frac{\#\left(\left(\varphi(S) \cap \delta_\beta G\right)\right) \cup \left(\delta_\beta(S) \cap \varphi(G)\right)}{\#(\varphi(S) \cup \varphi(G))} \tag{26}$$

$$CAL(S, G) = C \times A \times L \tag{27}$$

where $\#_C G$ and $\#_C S$ are the number of the connected components in the segmented image and ground truth, respectively. $\#G$ is the number of vessel pixels in the ground truth, $\delta_\alpha$and δ_β represent the morphological dilations using the disc of radiuses α and β, respectively, and φ is a skeletonization procedure. LCC is defined as:

$$LCC = 1 - min\left(1, \frac{|\#LCC_S - \#LCC_G|}{\#LCC_G}\right) \tag{28}$$

where $\#LCC_S$ and $\#LCC_G$ refer to the number of pixels in the longest connected component of the skeleton in the segmented image and ground truth, respectively.

4. Results and discussion

4.1. Comparison experiments

The proposed OCT^{2}Former is evaluated on three public OCTA datasets and compared with several state-of-the-art segmentation

Table 1
Comparison results on OCTA-SS dataset.

Methods	JAC	DICE	BACC	PRE	REC	CAL	LCC	p-value
U-Net [53]	-	89.00±0.00	-	97.00±0.00	87.00±0.00	90.00±0.00	93.00±0.00	-
CS-Net [53]	-	89.00±0.00	-	93.00±0.00	91.00±0.00	90.00±0.00	93.00±0.00	-
U-Net3+ [57]	81.44±2.22	89.75±1.34	91.65±1.65	91.92±1.65	92.33±2.64	90.12±3.39	91.89±1.08	<0.001
CE-Net [29]	73.31±2.17	84.58±1.44	87.25±2.13	88.00±3.10	85.85±2.82	86.55±3.36	79.90±2.18	<0.001
iUNet [44]	81.73±2.36	89.93±1.43	92.00±1.67	92.85±2.27	91.15±3.68	90.40±2.78	92.17±1.38	<0.001
MFI-Net [30]	79.79±2.74	88.74±1.71	90.69±1.51	90.99±3.12	90.39±2.88	86.82±5.56	77.18±2.86	<0.001
SCS-Net [31]	82.01±2.26	90.10±1.36	91.73±1.43	91.52±3.29	91.95±3.99	90.47±3.98	92.11±1.61	<0.001
AACA-MLA-D-UNet [32]	81.40±2.96	89.71±1.83	91.37±1.54	92.73±3.59	90.02±4.21	89.54±2.73	90.05±2.35	<0.001
Vessel-Net [33]	81.56±1.96	89.83±1.19	91.83±1.76	91.90±3.39	91.77±2.32	90.25±2.25	93.13±1.29	<0.001
TransUNet [49]	79.27±2.21	88.42±1.38	90.78±1.94	90.88±2.33	91.13±1.56	88.57±2.66	93.08±1.22	<0.001
UTNet [58]	81.52±2.49	89.80±1.51	91.62±1.65	92.06±2.18	91.34±3.16	89.21±3.83	87.27±1.80	<0.001
Swin-UNet [59]	78.46±2.48	87.91±1.55	90.22±1.81	90.53±2.39	90.48±2.50	87.63±3.93	81.29±2.45	<0.001
SegFormer [48]	80.20±2.08	88.93±1.69	90.79±4.74	90.91±6.35	90.68±3.68	89.47±3.87	89.93±1.58	<0.001
OCT^{2}Former	83.44±1.85	90.96±1.10	92.64±1.60	92.95±1.97	92.34±2.63	91.52±2.56	96.05±2.24	-

Table 2
Comparison results on ROSE-1 dataset.

Methods	JAC	DICE	BACC	PRE	REC	CAL	LCC	p-value
U-Net [60]	70.34±2.09	82.57±1.44	89.49±1.88	95.27±0.61	81.52±4.19	66.13±3.80	71.53±2.87	<0.001
CS-Net [27]	76.05±1.64	86.41±1.06	92.67±1.67	96.18±0.59	87.85±3.96	71.52±4.71	69.15±2.56	<0.001
U-Net3+ [57]	77.23±2.26	87.13±1.45	92.21±1.91	96.48±0.61	86.31±4.20	74.69±4.64	75.20±2.16	<0.001
CE-Net [29]	76.43±1.85	86.63±1.19	92.94±1.53	96.22±0.64	88.40±3.45	75.43±3.95	74.66±2.06	<0.001
iUNet [44]	76.61±1.92	86.74±1.53	93.08±1.77	96.44±0.85	88.71±4.13	75.12±3.94	80.17±1.52	<0.001
MFI-Net [30]	73.68±2.87	84.47±1.95	90.70±2.11	96.75±0.60	83.65±4.57	68.98±5.47	74.40±1.94	<0.001
SCS-Net [31]	76.06±1.60	86.39±1.04	93.33±1.51	97.11±0.94	89.55±3.66	75.29±3.34	81.16±1.38	<0.001
AACA-MLA-D-UNet [32]	75.56±2.09	86.06±1.35	91.93±1.82	97.76±0.57	86.11±4.02	73.30±3.60	79.37±1.52	<0.001
Vessel-Net [33]	75.75±1.90	86.19±1.23	93.18±1.46	97.10±0.94	89.26±3.53	73.31±3.74	75.37±1.95	<0.001
TransUNet [49]	73.36±2.24	84.61±1.49	90.67±1.80	95.81±0.61	83.51±3.78	69.32±3.51	81.00±1.53	<0.001
UTNet [58]	77.58±1.51	87.36±0.97	93.14±1.67	96.46±0.52	88.55±3.83	75.94±2.61	77.58±1.94	<0.001
Swin-UNet [59]	62.73±2.95	77.06±2.24	86.99±2.81	93.64±0.86	77.85±6.52	56.50±4.20	57.88±2.09	<0.001
SegFormer [48]	74.07±2.15	85.09±1.44	91.86±1.62	97.33±0.61	86.39±3.45	72.98±4.54	80.42±2.21	<0.001
OCT^{2}Former	78.55±2.15	87.97±1.35	93.12±1.65	96.67±0.55	88.20±3.55	76.86±4.23	81.70±1.23	-

Table 3
Comparison results on OCTA-6M dataset.

Methods	JAC	DICE	BACC	PRE	REC	CAL	LCC	p-value
U-Net [39]	79.11±3.97	88.28±2.59	92.98±1.62	-	-	-	-	-
CS-Net [27]	80.24±3.81	89.01±2.45	93.55±1.57	97.99±0.55	89.10±3.07	83.50±4.97	86.66±1.60	<0.001
U-Net3+ [39]	79.47±4.26	88.49±2.80	93.07±1.78	-	-	-	-	-
CE-Net [29]	80.21±3.91	88.96±2.56	93.74±1.43	97.98±0.67	88.53±2.78	83.27±4.19	85.82±1.61	<0.001
iUNet [44]	80.32±3.87	89.00±2.47	93.88±1.55	98.95±0.41	88.81±3.05	83.02±4.91	87.20±1.84	<0.001
MFI-Net [30]	79.59±3.77	88.62±2.57	93.51±1.60	98.39±0.51	88.65±3.12	83.21±4.99	85.52±1.61	<0.001
SCS-Net [31]	80.15±3.71	88.93±2.37	93.44±1.59	98.03±0.39	87.86±3.15	83.11±4.81	86.76±1.57	<0.001
AACA-MLA-D-UNet [32]	79.93±3.88	88.79±2.52	93.29±1.54	99.02±0.56	87.55±3.01	83.36±5.20	85.76±1.39	<0.001
Vessel-Net [33]	80.21±4.09	88.95±2.68	93.83±1.60	98.92±0.57	88.73±3.07	83.49±5.43	85.90±1.63	<0.001
TransUNet [49]	80.10±3.68	88.90±2.37	93.64±1.52	97.98±0.58	88.34±3.05	84.01±4.77	86.20±1.69	<0.001
UTNet [58]	80.41±1.51	89.08±2.53	93.91±1.50	98.00±0.65	88.90±2.92	83.69±5.15	87.65±1.65	<0.001
Swin-UNet [59]	70.16±3.78	82.40±2.68	90.58±1.73	96.77±0.59	83.01±3.43	75.91±5.68	69.73±1.80	<0.001
SegFormer [48]	77.68±3.89	87.38±2.57	92.95±1.51	98.75±0.55	87.16±2.94	82.34±5.02	77.79±1.54	<0.001
OCT^{2}Former	80.99±1.77	89.45±2.15	94.11±1.45	98.06±0.68	89.27±2.92	84.24±5.17	88.80±1.34	-

networks, including nine convolution based networks (U-Net, CS-Net, UNet3+, CE-Net, iUNet, MFI-Net, SCS-Net, AACA-MLA-D-UNet and Vessel-Net), two hybrid transformer networks (TransUNet and UTNet) and two pure transformer based network (Swin-UNet and SegFormer). Tables 1–4 show the results of different segmentation networks on three public OCTA datasets, respectively, where hybrid transformer networks including the proposed OCT^{2}Former and UTNet, outperform most of the convolution based networks and the pure transformer based network in Jaccard index, showing that transformer can be applied to the retinal vessel segmentation in OCTA images. In addition, our proposed network outperforms other competitive methods in almost all evaluation metrics, especially in Jaccard, Dice, CAL and LCC metrics

To evaluate if the improvement is statistically significant, the Wilcoxon signed-rank test is conducted on Jaccard index in all comparison experiments. It can be seen from Tables 1–4 that all p-values are less than 0.05, indicating that our OCT^{2}Former has achieved a significant improvement compared to other networks on all three datasets.

Fig. 7 shows the retinal vessel segmentation results of some of networks on three datasets. It can be seen from Fig. 7 that as a powerful convolution based network in medical image segmentation, U-Net, performs well on thick vessel segmentation, while having a poor performance on segmenting most of the thin vessels. Compared with U-Net, CS-Net achieves better segmentation results, but the connectivity of segmented vessels is poor. SCS-Net equipped with scale-aware feature aggregation (SFA) module, adaptive feature fusion (AFF) module and multi-level semantic supervision (MSS) module performs well on all three datasets. But its segmentation performance of thin vessels with blurred boundaries need to be improved, which is probably due to the decrease of receptive field during the decoding process. iUNet, which

Table 4
Comparison results on OCTA-3M dataset.

Methods	JAC	DICE	BACC	PRE	REC	CAL	LCC	p-value
U-Net [39]	82.88±3.47	90.60±2.16	94.91±1.38	-	-	-	-	-
CS-Net[27]	83.43±3.56	90.92±2.21	95.33±1.60	98.79±0.24	91.34±3.24	85.03±5.24	87.74±1.09	<0.001
U-Net3+ [39]	83.41±3.36	90.92±2.08	94.57±1.43	-	-	-	-	-
CE-Net [29]	83.33±3.34	90.87±2.06	94.41±1.65	98.80±0.23	89.34±3.34	84.28±5.46	88.84±0.98	<0.001
iUNet [44]	83.67±3.64	91.06±2.27	95.46±1.71	98.33±0.20	91.59±3.47	85.47±5.29	90.56±0.73	<0.001
MFI-Net [30]	82.15±2.57	90.15±1.59	93.68±1.71	99.10±0.31	89.01±2.04	83.40±5.66	88.41±1.28	<0.001
SCS-Net [31]	83.64±3.59	91.05±2.23	94.71±1.63	99.45±0.15	89.98±2.29	85.27±5.55	86.76±1.19	<0.001
AACA-MLA-D-UNet [32]	83.32±3.54	90.81±2.17	94.08±1.79	99.55±0.12	88.61±3.61	83.86±5.71	89.84±0.81	<0.001
Vessel-Net [33]	83.38±3.37	90.90±2.09	94.38±1.48	99.50±0.15	89.27±2.98	84.62±5.30	90.25±0.72	<0.001
TransUNet [49]	81.70±3.49	89.89±2.21	94.16±1.54	98.66±0.25	88.94±3.10	84.16±5.42	88.59±0.82	<0.001
UTNet [58]	83.76±2.44	91.12±2.11	95.47±1.37	98.81±0.24	91.59±2.76	85.10±5.13	90.04±0.91	<0.001
Swin-UNet [59]	73.14±4.19	84.42±2.87	90.96±1.69	97.96±0.34	82.84±3.47	72.30±6.33	65.30±1.57	<0.001
SegFormer [48]	80.67±3.57	89.25±2.33	93.51±1.83	99.38±0.15	87.63±3.71	83.74±5.86	82.21±1.09	<0.001
OCT^{2}Former	85.13±3.39	91.93±2.06	95.46±1.59	98.94±0.21	91.45±3.23	87.94±5.25	92.68±0.60	-

Missing indicators in cited literature are represented with '-'

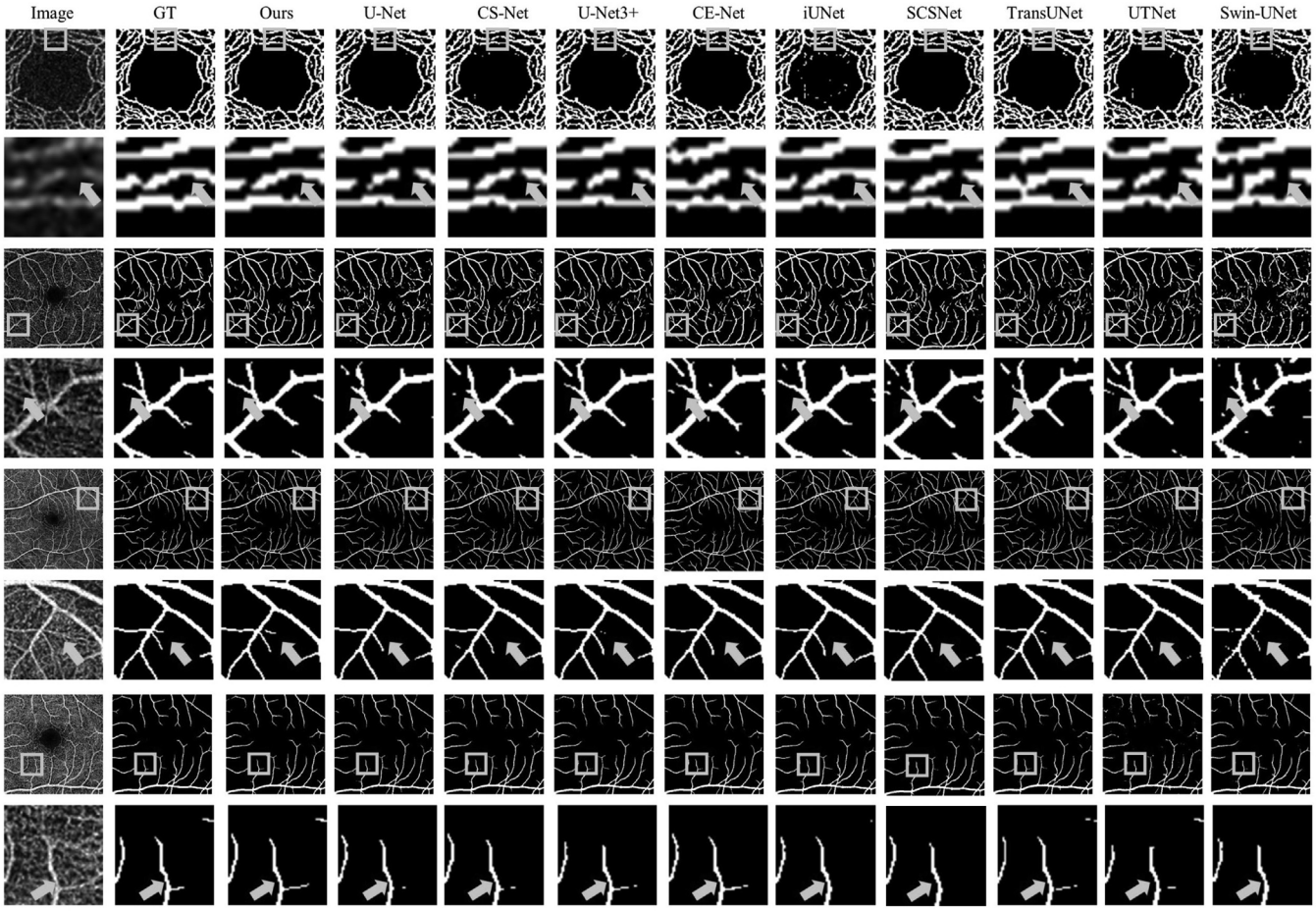

Fig. 7. Retinal vessel segmentation results of the proposed OCT^{2}Formerand other segmentation networks. From top to down, OCTA images of row1, 3, 5 and 7 come from OCTA-SS, ROSE-1, OCTA-6M and OCTA-3M, respectively. Row 2, 4, 6 and 8 show the corresponding locally zoomed-in OCTA images, ground truh and segmentation results.

utilizes an iterative approach to recurrently refine the segmentation result, shows promising performance in preserving vessel connectivity, but it still faces challenges in thin vessel segmentation. CE-Net and TransUNet have poor segmentation performance on thin vessels, because both of them use ResNet [61] as the backbone network, in which overmuch down-sampling operations lead to the lack of guidance from high-resolution feature and the challenge of thin vessel recovery. As a pure transformer based net, Swin-UNet performs poor because it could not converge the small datasets of OCTA images with limited training time (100 epochs). As a hybrid transformer network, the performance of UTNet has been greatly improved, but false negatives are still unavoidable due to the lack of high-resolution global information. Our proposed OCT^{2}Former, which is designed to obtain global information of each layer, achieves the best segmentation performance for thin vessel and keeps the best connectivity of segmented vessels on all three datasets. Table 5 shows the computational cost and the network scale of the above networks. As can be seen from Table 5, the proposed OCT^{2}Former needs the least computational cost and has least network parameters among the transformer based and hybrid networks, except for the SegFormer (B1) that uses MLPs as the decoder and not performs well in the retinal vessel segmenta-

Table 5
The FLOPs and parameters of different networks (with input size of 224 × 224).

Methods	Ours	U-Net	CS-Net	U-Net3+	CE-Net	iUNet	MFI-Net	SCS-Net	AACA-MLA -D-UNet	Vessel-Net	TransUNet	UTNet	Swin-UNet	SegFormer
FLOPs/G	7.34	3.35	8.40	27.01	29.22	89.45	7.16	6.00	2.13	1.27	93.19	57.45	27.12	2.60
Params/M	49.99	5.94	10.69	153.06	136.27	2.22	77.36	16.22	2.03	13.40	49.27	62.11	93.57	13.76

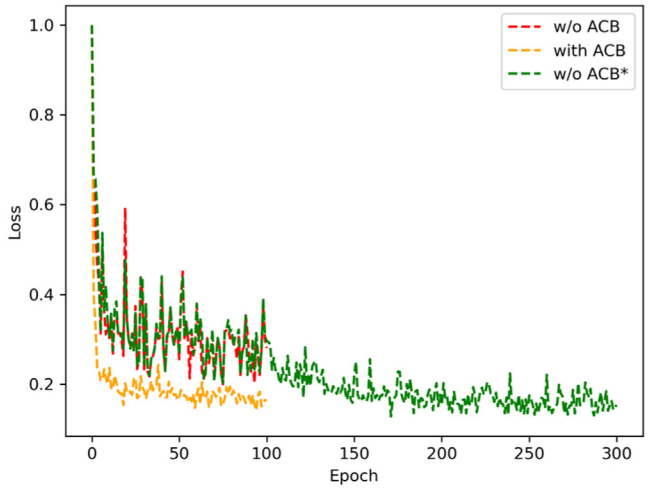

Fig. 8. The loss curve of OCT^{2}Former during the training process on OCTA-SS. * represents training for 300 epochs.

tion task yet. Furthermore, considering comprehensively both segmentation performance and computational cost and network scale, the proposed OCT^{2}Former is also competitive with convolution based networks such as CS-Net, which shows that the proposed OCT^{2}Former can process high-resolution information and obtain global information without excessive computational cost and parameters, and can be applied to retinal vessel segmentation in a more efficient way.

4.2. Ablation experiments

4.2.1. Ablation experiments about ACB

In this section, exhaustive ablation experiments about the auxiliary convolution branch (ACB) are designed and performed, which is proposed to speed up the convergence of OCT^{2}Former. To verify the convergence speed, two different training epochs (100 and 300) are adopted. Fig. 8 shows the loss curves of OCT^{2}Former during training with different settings on dataset OCTA-SS. As can be seen from Fig. 8, the training loss fluctuates greatly with the absence of ACB and can hardly descend to the optimum level even after 300 epochs, which in turn indicates the effectiveness and necessity of the ACB. The detailed ablation experimental results on all three datasets are shown in Table 6, which indicate that the insertion of ACB in the dynamic transformer encoder can not only speed up the convergence of the proposed OCT^{2}Former, but also improve the segmentation performances on all three datasets for the major index Jaccard (all the p-values of Wilcoxon signed-rank test on Jaccard index are less than 0.05, indicating that there is a statistically significant difference between OCT^{2}Former w/o ACB+100 epochs and OCT^{2}Former+100 epochs). The main reason is that transformer lacks inductive bias and requires a lot of training to learn the bias, which leads to the slow convergence on small datasets. In the case of limited training time (100 epochs), the network is difficult to converge to the global optimum. However, when trained for 300 epochs, the performers of OCT^{2}Former without ACB (OCT^{2}Former w/o ACB+300 epochs) are close to OCT^{2}Former +100 epochs (all the p-values of Wilcoxon signed-rank test on Jaccard index are great than 0.05, indicating that there is no statistically significant difference between OCT^{2}Former w/o ACB+300 epochs and OCT^{2}Former+100 epochs), which proves that the dynamic token aggregation transformer plays the leader role and ACB plays an auxiliary role in the proposed dynamic transformer encoder.

4.2.2. Ablation experiments about the structure of decoder path

Convolution based lightweight decoder (LD) is designed to construct the decoder path, which can not only recover the feature resolution efficiently, but also reduce the number of network parameters. In order to verify the rationality of this design, the ablation experiments about the structure of decoder path are conducted. Fig. 9 shows five different structures of decoder path. Structure (a) is the one adopted in our proposed OCT^{2}Former. Each decoder layer of (b) consists of two cascaded LDs (same as the decoder layer of U-Net [60]). (c) is a representative of dense connection decoder (DCD) path (same as the decoder path of U-Net3+ [57], which performs best in the convolution based networks). (d) is a representative of deep supervision decoder (DSD), which is similar with the decoder path in MFI-Net [30]. The decoder path of (e) is constructed by the symmetrical transformer (ST) structure in the encoder path (similar to Swin-UNet [59]). As shown in Table 7, there are no statistical significances between the segmentation performances of the OCT^{2}Former with these different five decoder paths (all the p-values of Wilcoxon signed-rank test on Jaccard index are greater than 0.05), while the proposed OCT^{2}Former with 1LD based decoder path (structure (a)) needs the least computational cost and parameters, which indicates that the structure of the decoder path has little influence on our proposed hybrid transformer network. The main reasons are: (1) the convolution based networks usually rely on down-sampling to enlarge the receptive field and acquire more advanced semantic features, and therefore need more complex decoders to fuse features from different layers with different receptive fields; (2) our OCT^{2}Former has the ability to acquire the global receptive field from each layer, which means that the features from each encoder layer are sufficient for the direct information recovery with simple decoder structure.

4.2.3. Ablation experiments about DTA algorithm

To further investigate the settings of iteration number T and aggregation embedding length k in DTA algorithm, the corresponding ablation experiments are conducted on OCTA-SS dataset. As shown in Table 8, the performance of our OCT^{2}Former is poor when T=1. This is because the limited learning ability of the network, which makes it difficult to aggregate embedding in single step to learn enough information from the original embedding. With the increase of number of iterations, the performance of our OCT^{2}Former is greatly improved, tending to be converged when T=3. Fig. 10 shows the changes of the main indexes including JAC and CAL with the number of iterations, which clearly indicates that T=3 is the optimal choice for our proposed DTA algorithm. Theoretically, when the number of iterations increases, the aggregation embedding information becomes more effective and abstract. However, due to the use of *softargmax* in DTA algorithm, the gradient tends

Table 6
Results of ablation study about ACB.

Dataset	Method	100 epochs	300 epochs	JAC	DICE	BACC	PRE	REC	CAL	LCC	p-value
OCTA-SS	OCT^{2}Former w/o ACB	√	×	8 2.87±3.01	90.61±2.34	92.17±1.93	92.72±3.05	91.33±2.59	90.37±3.18	93.11±1.89	<0.001
	OCT^{2}Former w/o ACB	×	√	83.15±1.74	90.79±1.11	92.52±1.41	92.71±2.12	93.24±2.54	91.69±3.21	95.80±2.10	0.162
	OCT^{2}Former	√	×	8 3.44±1.85	90.96±1.10	92.64±1.60	92.95±1.97	92.34±2.63	91.52±2.56	96.05±2.24	-
ROSE-1	OCT^{2}Former w/o ACB	√	×	7 7.54±2.41	87.33±1.87	92.87±1.74	96.48±0.59	87.86±3.76	74.79±4.54	79.64±1.93	<0.001
	OCT^{2}Former w/o ACB	×	√	78.28±2.01	87.81±1.29	93.22±1.67	96.60±0.62	88.53±3.61	76.78±4.12	81.10±1.30	0.088
	OCT^{2}Former	√	×	7 8.55±2.15	87.97±1.35	93.12±1.65	96.67±0.55	88.20±3.55	76.86±4.23	81.30±1.23	-
OCTA-6M	OCT^{2}Former w/o ACB	√	×	8 0.59±3.21	89.19±2.64	93.58±1.56	98.04±0.66	88.12±3.11	82.64±5.01	87.51±1.65	<0.001
	OCT^{2}Former w/o ACB	×	√	80.81±1.67	89.33±2.01	93.72±1.51	98.07±0.76	88.42±2.79	84.44±4.97	88.61±1.55	0.144
	OCT^{2}Former	√	×	8 0.99±1.77	89.45±2.15	94.11±1.45	98.06±0.68	89.27±2.92	84.24±5.17	88.80±1.34	-
OCTA-3M	OCT^{2}Former w/o ACB	√	×	8 4.32±3.53	91.45±2.23	95.55±1.51	98.86±0.30	91.74±2.98	86.21±5.12	90.87±1.02	<0.001
	OCT^{2}Former w/o ACB	×	√	84.77±3.41	91.72±2.12	95.49±1.63	98.91±0.16	91.55±3.17	88.11±4.95	92.78±0.71	0.091
	OCT^{2}Former	√	×	8 5.13±3.39	91.93±2.06	95.46±1.59	98.94±0.21	91.45±3.23	87.94±5.25	92.68±0.60	-

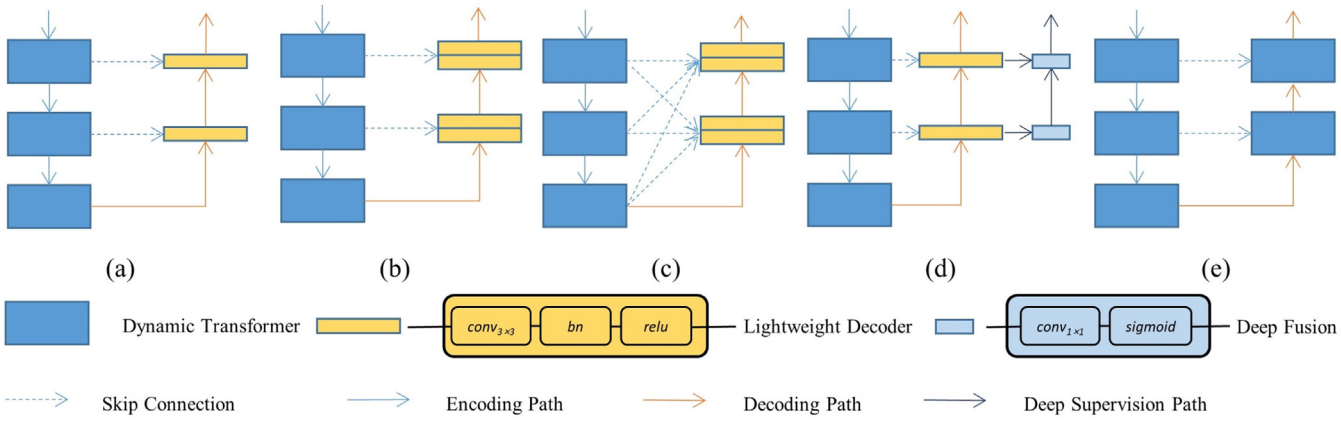

Fig. 9. OCT^{2}Former with different decoder structures. (a) single LD (default setting); (b) double LD; (c) decoder of UNet3+; (d) deep supervision decoder; (e) symmetrical decoder based on dynamic transformer.

Table 7
Results of ablation study about the structure of decoder path (LD: lightweight decoder, DCD: dense connection decoder, DSD: deep supervision decoder, ST: symmetrical transformer).

Dataset	Decoder setting	JAC	DICE	BACC	PRE	REC	CAL	LCC	p-value
OCTA-SS	1 LD(default setting)	83.44±1.85	90.96±1.10	92.64±1.60	92.95±1.97	92.34±2.63	91.52±2.56	96.05±2.24	-
	2 LDs	83.21±1.75	90.82±1.05	92.51±1.55	92.91±1.89	92.44±2.74	90.92±3.32	94.59±2.95	0.820
	DCD	83.11±1.95	90.75±1.19	92.37±1.72	92.91±1.95	91.68±2.91	91.63±2.72	96.13±2.41	0.810
	DSD	83.19±1.63	90.80±0.99	92.65±1.59	93.15±2.01	92.23±2.74	91.59±2.40	96.05±2.31	0.856
	ST	83.24±2.02	90.83±1.25	92.49±1.81	92.94±1.85	92.15±2.56	91.51±2.52	96.15±2.33	0.919
ROSE-1	1 LD(default setting)	78.55±2.15	87.97±1.35	93.12±1.65	96.67±0.55	88.20±3.55	76.86±4.23	81.30±1.23	-
	2 LDs	78.31±2.36	87.82±1.51	93.63±1.44	96.57±0.56	89.57±3.45	75.99±4.52	80.82±1.31	0.782
	DCD	78.39±2.21	87.87±1.49	93.41±1.63	96.63±0.49	88.98±3.51	76.79±4.19	81.21±1.31	0.813
	DSD	78.35±2.57	87.85±1.42	93.59±1.53	96.75±0.45	89.01±3.66	76.63±4.58	81.35±1.34	0.821
	ST	78.51±2.05	87.94±1.36	93.23±1.71	96.66±0.61	89.66±3.61	76.52±3.99	81.05±1.40	0.927
OCTA-6M	1 LD(default setting)	80.99±1.77	89.45±2.15	94.11±1.45	98.06±0.68	89.27±2.92	84.24±5.17	88.80±1.34	-
	2 LDs	80.96±1.73	89.42±2.30	93.96±1.67	98.07±0.71	88.92±2.81	84.14±4.98	88.56±1.41	0.899
	DCD	80.93±1.69	89.40±2.41	94.07±1.41	98.06±0.68	89.17±2.91	84.01±4.81	88.64±1.43	0.909
	DSD	80.79±1.54	89.21±2.25	94.21±1.77	98.10±0.61	88.71±3.15	84.08±5.12	88.69±1.52	0.784
	ST	80.84±1.80	89.34±2.23	93.91±1,59	98.06±0.58	88.84±2.92	84.19±5.03	88.77±1.37	0.765
OCTA-3M	1 LD(default setting)	85.13±3.39	91.93±2.06	95.46±1.59	98.94±0.21	91.45±3.23	87.94±5.25	92.68±0.60	-
	2 LDs	84.76±3.61	91.72±2.19	95.10±1.60	98.91±0.31	90.70±3.10	87.64±5.10	92.58±1.02	0.501
	DCD	84.94±3.41	91.81±2.09	95.40±1.55	98.91±0.29	91.39±3.15	87.85±5.42	92.50±1.05	0.883
	DSD	84.77±3.62	91.72±2.23	95.56±1.51	99.03±0.19	90.59±3.25	87.89±5.23	92.61±0.82	0.789
	ST	84.85±3.45	91.76±2.11	95.31±1.69	98.92±0.36	91.16±3.29	87.96±5.30	92.71±0.91	0.725

Table 8
Ablation experiments on iteration number T in DTA algorithm.

T	JAC	BACC	CAL
1	80.05	91.67	87.83
2	82.15	92.11	89.78
3	83.28	92.41	91.52
4	83.12	92.49	91.47
5	83.11	92.48	91.39
6	83.22	92.36	91.40

to decrease during iteration, which may lead to the amount of information retained by the aggregated embedding not increase continuously.

Table 9 and Fig. 11 show the ablation experiments on the length of the aggregation embedding k. When k =16, the performance of our OCT^{2}Former is poor, which is mainly because the aggregation embedding has insufficient aggregation capacity for the original embedding, resulting in the inability of the attention mechanism to capture important relationships between distant pixel locations.

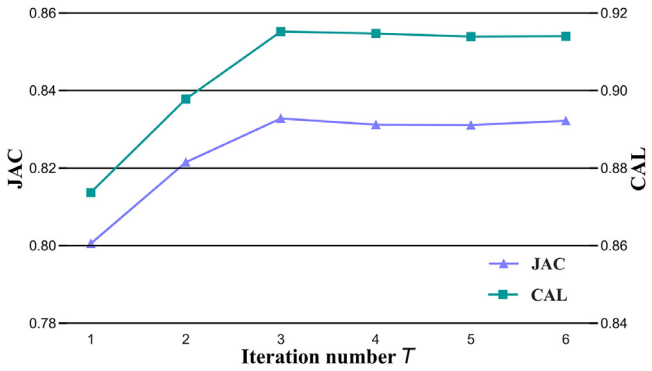

Fig. 10. Visualization of iteration number T in DTA algorithm.

Table 9
Ablation experiments on aggregation length k in DTA algorithm.

k	JAC	BACC	CAL
16	77.90	91.12	87.41
32	79.41	91.55	88.01
64	81.62	91.82	91.01
128	83.28	92.41	91.52
256	82.97	92.41	91.44
512	82.64	92.03	91.39

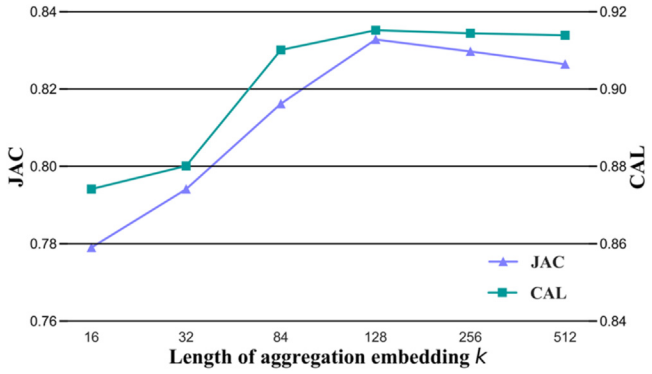

Fig. 11. Visualization of aggregation embedding length k in DTA algorithm.

The aggregation capability of the network increases gradually with the increase of k, and reaches an optimal value when k=128, indicating that aggregation embedding can effectively represent the original embedding now. The further increase of k does not improve the performance, possibly because the redundant information is already present in the aggregation embedding. Therefore, we conclude that the performance of the network will not be significantly improved even if k is increased to the length of the original embedding N, which is the main basis for our design of the DTA algorithm.

4.3. Generalization experiments

In order to evaluate the generalization ability of our proposed OCT^{2}Former, additional experiments on two datasets including ROSE-1 (DVC) and ROSE-2 [37] with centerline vessel ground truth are conducted. Area under ROC curve (AUC) and DICE are adopted to evaluate the performance of the centerline segmentation, which are the same ones in Reference [37]. Specifically, a three-pixel tolerance region around the manually traced centerlines is considered as a true positive in our evaluation which is consistent with Reference [37].

Tables 10 and 11 show the results of different segmentation networks on ROSE-1 (DVC) and ROSE-2. As shown in Tables 10 and 11, our OCT^{2}Former outperforms other state-of-the-art networks. Specifically, our OCT^{2}Former achieves a 1.25% improvement on ROSE-1(DVC) dataset and a 0.58% improvement on the ROSE2 dataset for DICE index, comparing to the second-best OCTA-Net, which indicates that our OCT^{2}Former is not only capable of pixel-level vessel segmentation but also suitable for the centerline segmentation.

Table 10
Comparison results on ROSE-1(DVC) dataset.

Methods	AUC	DICE
U-Net [37]	91.86	66.05
CE-Net [37]	95.05	57.83
CS-Net [37]]	96.71	58.84
CGNet [64]	96.41	67.36
OCTA-Net [37]	96.73	70.74
Ours	96.99	71.99

Table 11
Comparison results on ROSE-2 dataset.

Methods	AUC	DICE
U-Net [37]	83.70	65.64
CE-Net [37]	84.67	70.66
CS-Net [37]	85.42	70.10
CGNet [64]	85.62	68.86
OCTA-Net [37]	86.03	70.77
Ours	86.18	71.35

4.4. Discussion about the receptive field

The receptive field plays a crucial role in semantic segmentation, which allows the network to extract of both abstract semantic information and internal topological structures. In the retinal vessel segmentation task, a large receptive field is particularly important as the retinal vessels are characterized by the tree-like topological structure where the thick vessels are usually connected with several thin vessels. That is, the integrity and connectivity of thick and thin vessels should be considered during the segmentation, which requires a large receptive field.

Previous CNN-based retinal vessel segmentation networks typically utilize convolution layers as the encoder, which usually require to deepen the layers of the network or use dilated convolution to increase the receptive field. However, with the increase of the layers, the extracted features become more abstract and the features for thin vessels may disappear. Our proposed OCT^{2}Former can obtain global receptive field from each layer, allowing for comprehensive segmentation of both thick and thin vessels. Effective receptive field (ERF) [48] is adopted to show the receptive fields of U-Net, SCS-Net and our proposed OCT^{2}Former (average over 100 images from the test set of OCTA-6M). As shown in Fig. 12, the receptive field of U-Net increases with the increase of the encoder layers. SCS-Net adds a dilated convolution based SFA module at the bottleneck of the encoder, which increases the receptive field of the bottom feature. However, due to the layer-by-layer feature fusion via up-sampling in the decoder path, the receptive field of U-Net and SCS-Net at the output layer is still small. On the contrary, the proposed OCT^{2}Former has a global receptive field in each layer, especially in the output layer which is vital for the thin vessel segmentation and can improve the integrity and connectivity of the segmented vessels. As shown in Fig. 6(c), since the images in OCTA500-6M are centered on the fovea (non-perfusion area) and the retinal vessels are distributed around it, the global receptive field of our OCT^{2}Former can adapt to focus on the vessels instead of the fovea.

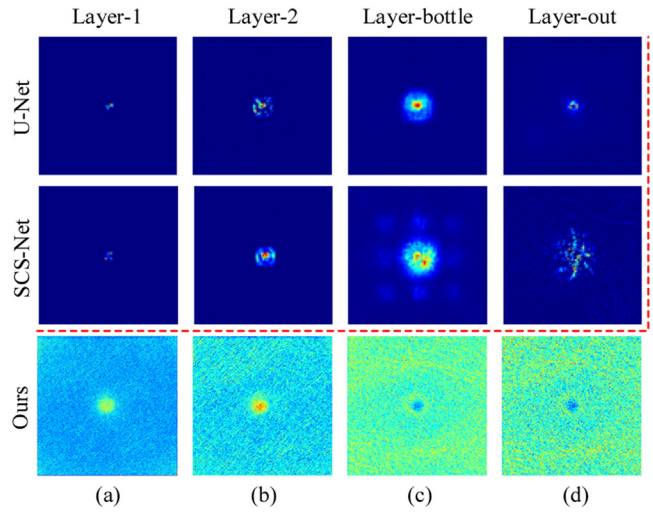

Fig. 12. Effective receptive field of U-Net, SCS-Net and the proposed OCT^{2}Former (average over 100 images from the test set of OCTA-6M).

5. Conclusion

In this paper, a novel end-to-end hybrid transformer based retinal vessel segmentation network named as OCT^{2}Former is proposed, in which transformer is first introduced and improved for retinal vessel segmentation in OCTA images. In order to efficiently apply transformer to retinal vessel segmentation, a dynamic token aggregation transformer is proposed to reduce the huge computational cost of the original transformer, and an auxiliary convolution branch is designed to speed up the convergence of the original transformer. The convolution based lightweight decoder is adopted to construct the decoder path, which can reduce the computational cost and parameters of the OCT^{2}Former while keeping the overall good performance. The proposed OCT^{2}Former is evaluated and on OCTA-SS, ROSE-1, OCTA-500 (subset OCTA-6M and subset OCTA-3M) datasets and the results indicate that the proposed OCT^{2}Former outperforms other state-of-the-art networks.

Although our proposed OCT^{2}Former performs well on the 2D OCTA images, there are still several limitations that should be addressed in future work. First, we have not fully exploited the spatial information in 3D OCTA volumes, which is of great significance to retinal vessel segmentation. To make full use of 3D spatial information and further improve the segmentation performance of the proposed framework, we will focus on exploring the combination of 3D feature extraction and 2D segmentation network, trying to use 3D information to guide the segmentation of 2D retinal vessels. Additionally, we will try to extend our network for OCTA images with quality degradation such as projection and motion artifacts, and other curvilinear structure segmentation such as nerve fiber segmentation in corneal confocal microscopy (CCM) images [62] and retinal linear lesion segmentation in indocyanine green angiography (ICGA) images [63].

Second, our OCT^{2}Former has not been fully explored in the incorporation of the multi-scale information, which can be beneficial to further improve the segmentation performance. In the future work, we will try to explore the integration of the multi-scale tokens into our proposed OCT^{2}Former.

Third, although the overall segmentation performance of our OCT^{2}Former is good, there are still false positives and false negatives for the vessels with very blurred boundaries, which is the common challenge in medical image segmentation. To alleviate the blurred boundary issue, we will try to explore the boundary-aware loss function as auxiliary supervision in the future work.

Furthermore, the limited amount of training data may lead to the overfitting of the proposed OCT^{2}Former. We will work on ways such as self-supervision and meta-learning strategies to reduce the network's dependence on large amounts of training data and further improve the performance of the proposed OCT^{2}Former. We will also explore our OCT^{2}Former for the subsequent clinical applications, e.g., the fractal dimension analysis on the detected vessels between disease groups and normal controls.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

Acknowledgments

This work is supported in part by the National Key R&D Program of China (Grant No. 2018YFA0701700), in part by the National Natural Science Foundation of China (NSFC) (Grant U20A20170 and Grant 62001196), and in part by the Natural Science Research of Jiangsu Higher Education Institutions of China (Grant 21KJB510021). We declare that this work is original research that has not been published previously and that we have no conflicts of interest regarding this work.

References

[1] J. Almotiri, K. Elleithy, A. Elleithy, Retinal vessels segmentation techniques and algorithms: a survey, Appl. Sci. 8 (2) (2018) 155.
[2] C. Srinidhi, P. Aparna, J. Rajan, Recent advancements in retinal vessel segmentation, J. Med. Syst. 41 (4) (2017) 1–22.
[3] C. Or, A. Sabrosa, O. Sorour, M. Arya, N. Waheed, Use of OCTA, FA, and ultra-widefield imaging in quantifying retinal ischemia: a review, Asia Pac. J. Ophthalmo. 7 (1) (2018) 46–51.
[4] Y. Shin, K. Nam, S. Lee, H. Lim, M. Lee, Y. Jo, J. Kim, Changes in peripapillary microvasculature and retinal thickness in the fellow eyes of patients with unilateral retinal vein occlusion: an OCTA study, Invest. Ophthalmol. Vis. Sci. 60 (2) (2019) 823–829.
[5] E. Novais, N. Waheed, OCT angiography in retinal and macular diseases, Am. Orthop J. 56 (2016) 132–138.
[6] L. Moons, L Groef, Multimodal retinal imaging to detect and understand Alzheimer's and Parkinson's disease, Curr. Opin. Neurobiol. 72 (2022) 1–7.
[7] A. Pujari, P. Sharma, P. Singh, S. Phuljhele, R. Saxena, S. Azad, Optical coherence tomography angiography in neuro-ophthalmology: current clinical role and future perspectives, Survey Ophthalmol. 66 (2021) 471–481 no.3.
[8] G. Tsokolas, K. Tsaousis, V. Diakonis, A. Matsou, S. Tyradellis, Optical coherence tomography angiography in neurodegenerative diseases: a review, Eye Brain 12 (2020) 73–87.
[9] D. Sampson, A. Dubis, F. Chen, R. Zawadzki, D. Sampson, Towards standardizing retinal optical coherence tomography angiography: a review, Light Sci. Appl. 11 (1) (2022) 1–22.
[10] H. Nugroho, R. Aras, T. Lestari, I. Ardiyanto, Retinal vessel segmentation based on frangi filter and morphological reconstruction, in: Proceedings of the of ICCREC, 2017, pp. 181–184.
[11] H. Aguirre-Ramos, J. Avina-Cervantes, I. Cruz-Aceves, J. Ruiz-Pinales, S. Ledesma, Blood vessel segmentation in retinal fundus images using gabor filters fractional derivatives and expectation maximization, Appl. Math. Comput. 339 (2018) 568–587.
[12] R. Annunziata, E. Trucco, Accelerating convolutional sparse coding for curvilinear structures segmentation by refining SCIRD-TS filter banks, IEEE Trans. Med. Imaging 35 (11) (2016) 2381–2392.
[13] A. Hoover, V. Kouznetsova, M. Goldbaum, Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response, IEEE Trans. Med. Imaging 19 (3) (2000) 203–210.
[14] X. Jiang, D. Mojon, Adaptive local thresholding by verification-based multithreshold probing with application to vessel detection in retinal images, IEEE Trans. Pattern Anal. Mach. Intell. 25 (1) (2003) 131–137.
[15] M. Saleh, C. Eswaran, A. Mueen, An automated blood vessel segmentation algorithm using histogram equalization and automatic threshold selection, J. Digital Imaging 24 (4) (2011) 564–572.
[16] J. Staal, M. Abràmoff, M. Niemeijer, M. Viergever, B. Ginneken, Ridge-based vessel segmentation in color images of the retina, IEEE Trans. Med. Imaging 23 (4) (2004) 501–509.
[17] J. Soares, J. Leandro, R. Cesar, H. Jelinek, M. Cree, Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification, IEEE Trans. Med. Imaging 25 (9) (2006) 1214–1222.
[18] C. Lupascu, D. Tegolo, E. Trucco, Fabc: retinal vessel segmentation using adaboost, IEEE Trans. Inf. Technol. Biomed 14 (5) (2010) 1267–1274.

[19] J. Song, X. Chen, Q. Zhu, F. Shi, D. Xiang, Z. Chen, W. Zhu, Global and local feature reconstruction for medical image segmentation, IEEE Trans. Med. Imaging 41 (9) (2022) 2273–2284.
[20] Z. Liu, L. Tong, L. Chen, F. Zhou, Z. Jiang, Q. Zhang, Y. Wang, C. Shan, L. Li, H. Zhou, CANet context aware network for brain glioma segmentation, IEEE Trans. Med. Imaging 40 (7) (2021) 1763–1777.
[21] Z. Zhou, M. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: a nested u-net architecture for medical image segmentation, Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018 (2018) 3–11 11045.
[22] C. Yao, M. Wang, W. Zhu, H. Huang, F. Shi, Z. Chen, X. Chen, Joint segmentation of multi-class hyper-reflective foci in retinal optical coherence tomography images, IEEE Trans. Biomed. Eng. 69 (2021) 1349–1358.
[23] S. Xia, H. Zhu, X. Liu, M. Gong, X. Huang, L. Xu, H. Zhang, J. Guo, Vessel segmentation of x-ray coronary angiographic image sequence, IEEE Trans. Biomed. Eng. 67 (2020) 1338–1348.
[24] X. Wang, W. Li, B. Miao, H. Jing, W. Zhang, X. Wen, Z. Ji, G. Hong, Z. Shen, Retina blood vessel segmentation using a u-net based convolutional neural network, in: Proceedings of the Conference Computing Data Science, 2018, pp. 8–9.
[25] Q. Jin, Z. Meng, T. Pham, Q. Chen, L. Wei, R. Su, Dunet: A deformable network for retinal vessel segmentation, Knowl. Based Syst. 178 (2019) 149–162.
[26] Y. Wu, Y. Xia, Y. Song, Y. Zhang, W. Cai, NFN+: a novel network followed network for retinal vessel segmentation, Neural Netw. 126 (2020) 153–162.
[27] L. Mou, Y. Zhao, H. Fu, Y. Liu, J. Cheng, Y. Zheng, P. Su, J. Yang, L. Chen, A. Frangi, M. Akiba, J. Liu, Cs^2-net: Deep learning segmentation of curvilinear structures in medical imaging, Med. Image Anal. 67 (2021) 101874.
[28] H. Fu, Y. Xu, S. Lin, D.W.K. Wong, J. Liu, DeepVessel: Retinal vessel segmentation via deep learning and conditional random field, Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, 2016, pp. 132–139.
[29] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, J. Liu, Ce-net: Context encoder network for 2d medical image segmentation, IEEE Trans. Med. Imaging 38 (10) (2019) 2281–2292.
[30] Y. Ye, C. Pan, Y. Wu, S. Wang, Y. Xia, MFI-Net: Multiscale feature interaction network for retinal vessel segmentation, IEEE J. Biomed. Health 26 (9) (2022) 4551–4562.
[31] H. Wu, W. Wang, J. Zhong, B. Lei, Z. Wen, J. Qin, Scs-net: A scale and context sensitive network for retinal vessel segmentation, Med. Image Anal. 70 (2021) 102025.
[32] Y. Yuan, L. Zhang, L. Wang, H. Huang, Multi-level attention network for retinal vessel segmentation, IEEE J. Biomed. Health 26 (1) (2021) 312–323.
[33] Y. Wu, Y. Xia, Y. Song, D. Zhang, D. Liu, C. Zhang, W. Cai, Vessel-net: retinal vessel segmentation under multi-path supervision, Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, 2019, pp. 264–272.
[34] S. Yousefi, T. Liu, R. Wang, Segmentation and quantification of blood vessels for oct-based micro-angiograms using hybrid shape/intensity compounding, Microvasc. Res. 97 (2015) 37–46.
[35] N. Eladawi, M. Elmogy, O. Helmy, A. Aboelfetouh, A. Riad, H. Sandhu, S. Schaal, A. El-Baz, Automatic blood vessels segmentation based on different retinal maps from octa scans, Comput. Biol. Med. 89 (2017) 150–161.
[36] M. Sarabi, J. Gahm, M. Khansari, J. Zhang, A. Kashani, Y. Shi, An automated 3d analysis framework for optical coherence tomography angiography, BioRxiv (2019) 655175.
[37] Y. Ma, H. Hao, J. Xie, H. Fu, J. Zhang, J. Yang, Z. Wang, J. Liu, Y. Zheng, Y. Zhao, Rose: A retinal oct-angiography vessel segmentation dataset and new model, IEEE Trans. Med. Imaging 40 (3) (2021) 928–939.
[38] M. Li, Y. Chen, Z. Ji, K. Xie, S. Yuan, Q. Chen, S. Li, Image projection network: 3D to 2D image segmentation in OCTA images, IEEE Trans. Med. Imaging 39 (11) (2020) 3343–3354.
[39] M. Li, K. Huang, Q. Xu, J. Yang, Y. Zhang, Z. Ji, K. Xie, S. Yuan, Q. Liu, Q. Chen, OCTA-500: A Retinal Dataset for Optical Coherence Tomography Angiography Study, arXiv preprint, arXiv:2012.07261, 2022 Dec 25.
[40] Z. Wu, Z. Wang, W. Zou, F. Ji, H. Dang, W. Zhou, M. Sun, PAENet: A progressive attention-enhanced network for 3D to 2D retinal vessel segmentation, Proceedings of International Conference on Bioinformatics and Biomedical, 2021, pp. 1579–1584.
[41] M. Li, W. Zhang, Q. Chen, Image magnification network for vessel segmentation in OCTA images, Proceedings of 5th Chinese Conference on Pattern Recognition and Computer Vision, 2022, pp. 426–435.
[42] M. Menten, J. Paetzold, A. Dima, B. Menze, B. Knier, D. Rueckert, Physiology-based simulation of the retinal vasculature enables annotation-free segmentation of OCT angiographs, Proceedings of the International Conference of Medical Image Computing and Computer Assisted Intervention, 2022, pp. 330–340.
[43] W. Li, H. Zhang, F. Li, L. Wang, RPS-Net: An effective retinal image projection segmentation network for retinal vessels and foveal avascular zone based on OCTA data, Med. Phys. 49 (6) (2022) 3830–3844.
[44] T. Pissas, E. Bloch, M. Cardoso, B. Flores, O. Georgiadis, S. Jalali, C. Bergeles, Deep iterative vessel segmentation in OCT angiography, Biomed. Opt. Express 11 (5) (2020) 2490–2510.
[45] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. Torr, Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, 2021, pp. 6881–6890.
[46] R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: Transformer for semantic segmentation, Proceedings of the IEEE/CVF international conference on computer vision, 2022, pp. 7262–7272.
[47] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
[48] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.
[49] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, arXiv preprint arXiv:2102.04306, 2021 Feb 8.
[50] Y. Zhang, H. Liu, Q. Hu, Transfuse: Fusing transformers and cnns for medical image segmentation, Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, 2021, pp. 14–24.
[51] Y. Ji, R. Zhang, H. Wang, Z. Li, L. Wu, S. Zhang, P. Luo, Multi-compound transformer for accurate biomedical image segmentation, Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, 2021, pp. 326–336.
[52] X. Shen, J. Xu, H. Jia, P. Fan, F. Dong, B. Yu, S. Ren, Self-attentional microvessel segmentation via squeeze-excitation transformer Unet, Comput. Med. Imaging Graph. 97 (2022) 102055.
[53] Y. Giarratano, E. Bianchi, C. Gray, A. Morris, T. MacGillivray, B. Dhillon, M.O. Bernabeu, Automated segmentation of optical coherence tomography angiography images: benchmark data and clinically relevant metrics, Transl. Vis. Sci. Technol. 9 (13) (2020) 5.
[54] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, An image is worth 16x16 words: transformers for image recognition at scale, Proceedings of International Conference on Learning Representations (ICLR), 2021, pp. 1–22.
[55] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, C. Hsieh, C, Dynamicvit: Efficient vision transformers with dynamic token sparsification, in: Proceedings of the Advances In Neural Information Processing Systems (Neurips), 34, 2021, pp. 13937–13949.
[56] V. Jampani, D. Sun, M.Y. Liu, M.H. Yang, J. Kautz, Superpixel sampling networks, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 352–368.
[57] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y. Chen, J. Wu, Unet 3+: a full-scale connected unet for medical image segmentation, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 2020, pp. 1055–1059.
[58] Y. Gao, M. Zhou, D. Metaxas, Utnet: a hybrid transformer architecture for medical image segmentation, Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, 2021, pp. 61–71.
[59] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," arXiv preprint arXiv:2105.05537, 2021.
[60] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, 2015, pp. 234–241.
[61] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition, 2016, pp. 770–778.
[62] C. Yang, X. Zhou, W. Zhu, D. Xiang, Z. Chen, Y. Jin, X. Chen, F. Shi, Multi-discriminator adversarial convolutional network for nerve fiber segmentation in confocal corneal microscopy images, IEEE J. Biomed. Health Inform. 26 (2) (2022) 648–659.
[63] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, X. Chen, CPFNet: Context pyramid fusion network for medical image segmentation, IEEE Trans. Med. Imaging 39 (10) (2020) 3008–3018.
[64] X. Yu, C. Ge, M. Aziz, M. Li, P. Shum, L. Liu, J. Mo, CGNet-assisted automatic vessel segmentation for optical coherence tomography angiography, J. BioPhotonics 15 (2022) e202200067.