Medical Image Analysis 89 (2023) 102929

Contents lists available at ScienceDirect

Medical Image Analysis

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

Stimulus-guided adaptive transformer network for retinal blood vessel segmentation in fundus images

Ji Lin [a], Xingru Huang [a], Huiyu Zhou [b], Yaqi Wang [c], Qianni Zhang [a,*]

[a] School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London, E1 4NS, United Kingdom
[b] School of Informatics, University of Leicester, University Road, Leicester, LE1 7RH, United Kingdom
[c] College of Media Engineering, Communication University of Zhejiang, Hangzhou, 310018, China

ARTICLE INFO

Keywords:
Retinal blood vessel segmentation
Visual cortex
Stimulus-guided adaptive pooling transformer
Stimulus-guided adaptive feature fusion
Receptive field

ABSTRACT

Automated retinal blood vessel segmentation in fundus images provides important evidence to ophthalmologists in coping with prevalent ocular diseases in an efficient and non-invasive way. However, segmenting blood vessels in fundus images is a challenging task, due to the high variety in scale and appearance of blood vessels and the high similarity in visual features between the lesions and retinal vascular. Inspired by the way that the visual cortex adaptively responds to the type of stimulus, we propose a Stimulus-Guided Adaptive Transformer Network (SGAT-Net) for accurate retinal blood vessel segmentation. It entails a Stimulus-Guided Adaptive Module (SGA-Module) that can extract local–global compound features based on inductive bias and self-attention mechanism. Alongside a light-weight residual encoder (ResEncoder) structure capturing the relevant details of appearance, a Stimulus-Guided Adaptive Pooling Transformer (SGAP-Former) is introduced to reweight the maximum and average pooling to enrich the contextual embedding representation while suppressing the redundant information. Moreover, a Stimulus-Guided Adaptive Feature Fusion (SGAFF) module is designed to adaptively emphasize the local details and global context and fuse them in the latent space to adjust the receptive field (RF) based on the task. The evaluation is implemented on the largest fundus image dataset (FIVES) and three popular retinal image datasets (DRIVE, STARE, CHASEDB1). Experimental results show that the proposed method achieves a competitive performance over the other existing method, with a clear advantage in avoiding errors that commonly happen in areas with highly similar visual features. The sourcecode is publicly available at: https://github.com/Gins-07/SGAT.

1. Introduction

Ocular diseases such as diabetic retinopathy (DR) (Sivaprasad et al., 2012), age-related macular degeneration (AMD) (Wong et al., 2014), and glaucoma (GC) (Thylefors and Negrel, 1994) as the important causes of vision impairment, and retinopathy-related abnormalities caused by anaemia (Mitani et al., 2020), hepatobiliary diseases (Xiao et al., 2021) and kidney disease (Penno et al., 2012), have a significant impact on people's health. For example, GC will affect around 120 million people in 2040 as reported in Tham et al. (2014). As most ocular diseases are chronic disorders, the symptoms such as micro aneurysms, and haemorrhages occur in the early stage and have a negligible effect on visual acuity until irrecoverable sight loss takes place (Kaur et al., 2021). Thus, early diagnosis and medical intervention are vital for achieving an enhanced outcome for ophthalmology patients and patients.

Compared with traditional diagnostic procedures like fluorescein angiography which injects the dye into intravenous, retinal fundus images provide an efficient and friendly method for patients. However, the lack of qualified physicians with adequate clinic experience impedes accurate diagnosis. Meanwhile, the time-consuming and tiresome process often drives ophthalmologists to misclassify cases. Moreover, subjectivity leads to diagnostic variance between physicians. In recent years, computer-aided technology is gradually accepted as an assistive solution for ophthalmology examination to overcome these limitations.

Deep learning (DL) techniques with effective data utilization have opened significant new research avenues in medical image analysis, and its representative, the convolution neural network (CNN) achieves remarkable performance gains in retinal image analysis tasks, including DR detection (Ayhan et al., 2020), AMD quantification (Szeskin et al., 2021), glaucoma screening (Pachade et al., 2021), etc. Particularly, the advent of U-Net (Ronneberger et al., 2015) introduces the encoder–decoder structure that paves the way for biomedical image segmentation and many variants of U-net emerge for retinal image

* Corresponding author.
E-mail address: qianni.zhang@qmul.ac.uk (Q. Zhang).

https://doi.org/10.1016/j.media.2023.102929
Received 22 February 2023; Received in revised form 15 June 2023; Accepted 7 August 2023
Available online 9 August 2023

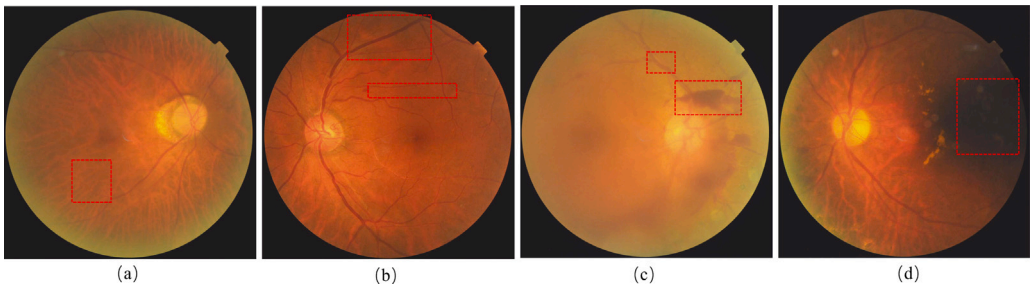


Fig. 1. Examples of the existing problems in fundus images. (a) the vague retinal blood vessels. (b) morphological divergence in blood vessels. (c) the high similarity between the blood vessels and lesions. (d) the artefacts caused by improper data acquisition.

segmentation. Deformable U-Net (DUNet) merges the deformable convolution into the U-Net to supply geometric offsets to blood vessels with various scales and shapes (Jin et al., 2019). Besides, scale and context sensitive network (SCS-Net) adopts dilated convolution for the multi-scale context and fuses high and low-level features of retinal vessels to improve the performance (Wu et al., 2021). Additionally, the multi-scale residual network (MSRNet) adds the residual connection to tackle the multi-scale retinal vessels (Xia et al., 2021). However, these CNN-based methods own the inherent disadvantage of locality, meaning the limited receptive field (RF).

Although transformer (Vaswani et al., 2017) with the long-range dependencies has been applied to mitigate the problem of limited RF from a holistic view (Cao et al., 2021; He et al., 2022), the full transformer is unable to provide effective fine-grained details of objects as CNN does. Thus, hybrid structures combining the strength of CNN and transformer, attract the attention of researchers working on various challenging medical image analysis tasks such as retinal vessel segmentation. Curvilinear structure segmentation network (CS2-Net) applies self-attention mechanisms to realize the dual attention in spatial and channel and enhance the long-range dependency and expression of the retinal vessel features (Mou et al., 2021). The relation transformer block places self-attention and cross-attention into two branches to establish the relationship between the lesion and retinal vessels and a global transformer block gives the specific information of lesion (Huang et al., 2022). Global transformer and dual local attention are inserted into U-Net to extract more global information and acquire the edge details respectively and combine the deep and shallow features to refine the retinal vascular segmentation (Li et al., 2022).

Nevertheless, several challenges still prevent the wide application of such technology. In retinal fundus image analysis, four main types of problems are often encountered, as illustrated in the highlighted regions in Fig. 1: (a) blood vessels may appear vague owing to the improper magnification level; (b) morphological divergence commonly exists in blood vessels including thick, thin, and filamentary retinal vessels; (c) some lesions and blood vessels may present highly similar visual features; and (d) the presence of artefacts due to improper data acquisition introduce additional challenges. In general, the intra-class morphological variances such as various scales and shapes and the high inter-class similarity such as the lesions and retinal vessels, require highly descriptive and distinctive features to represent these structures and thus enable accurate retinal blood vessel segmentation.

Besides, the usual encoder–decoder structure neglects the biological properties of the visual cortex in the human visual system, which is known to have the superior capability of delineating subtle foreground structures and background in images. Hubel et al. proved that the response of visual cortical cells heavily depends on the type of stimulus such as its shape, position, and orientation, and the visual cortex cells are divided into simple and complex categories according to the information flow (Hubel and Wiesel, 1962). As shown in Fig. 2, the explicit stimulus such as the object's colour and brightness in the scene tend to draw attention of the simple cells. Since fixed and relatively small receptive fields are more sensitive to low-level features with less

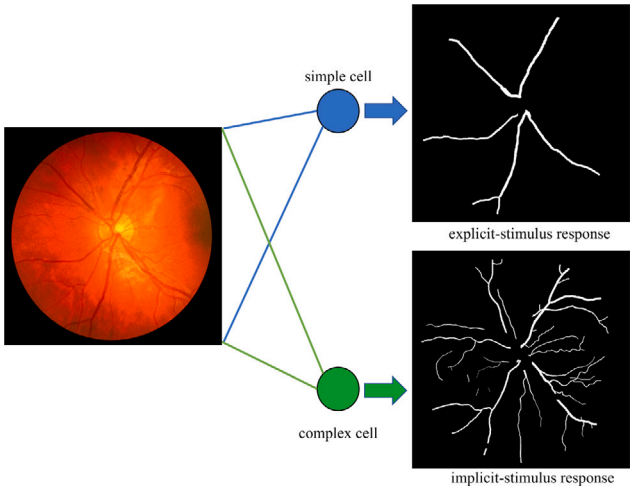


Fig. 2. The stimulus-guided adaptive RF mechanism. Although processing the same medical images, the RF of simple and complex cells relies on the type of stimulus. Simple cells are sensitive to the dark thick blood vessels (explicit stimulus) whereas complex cells are more capable of capturing light thin retinal vascular structures (implicit stimulus).

semantic information, eye vessels with visible and obvious features are clearly captured by simple cells. Thus, the thick and dark red vessels are illustrated in the visual cortex as shown in the top right of Fig. 2. On the contrary, the implicit stimulus such as the dependency between objects and the demand of the observer is of great importance in scene understanding by complex cells. The complex cells consist of a series of simple cells in a hierarchical structure, so they support more powerful inference abilities on long-range, implicit dependency, leading to reasonable inference and establishing the remote connections in relevant pixels to decrease the noise from the local background and artefacts. The bottom right of Fig. 2 shows an example of detected thin vessels which are not directly perceivable by the simple cells of the eye. However, the existing stimulus-based adaptive receptive field (RF) methods such as spatial transformer network are applied on only CNN (Jaderberg et al., 2015), which is insufficient for capturing the contextual information. In addition, the redundant information and noise from the global view hinder the transformer's application in medical images.

Inspired by the stimulus-driven approach to the visual cortex perception, a hybrid CNN-transformer network named Stimulus-Guided Adaptive Transformer Network (SGAT-Net) is proposed in this paper. Firstly, a Stimulus-Guided Adaptive Module (SGA-Module) based on inductive bias and long-range dependency is proposed to generate the descriptive and distinctive local–global compound features to extract the anatomy structure and the vascular details for effective segmentation. Then, to further differentiate truly descriptive features from redundant, misleading information, a Stimulus-Guided Adaptive Pooling Transformer (SGAP-Former) adaptively reweights the maximum

and average pooling outcomes to enhance the embedding representation capability under the limited resources. As a result, it eliminates the interference of background such as artefacts and distils the contextual information to help segment blurred retinal vascular. Finally, a Stimulus-Guided Adaptive Feature Fusion (SGAFF) module is designed to adaptively highlight the CNN-based features and transformer-based embeddings and mix them in the latent space to generate the appropriate RF according to the task. For example, morphological divergence in blood vessels requires more local details, such as subtle edges, that can be extracted using CNN-based features, whereas the long-range dependency supplied by transformer-based embeddings plays a more important role in distinguishing the blood vessels from lesions. In summary, there are mainly three contributions in this paper:

1. The SGA-Module combines CNN and transformer and provides the local–global compound features based on inductive bias and long-range dependency, resulting in descriptive and distinctive features for retinal blood vessel segmentation.
2. The SGAP-Former is introduced to extract the distinctive features using the adaptive fusion of average and maximum pooling. This fusion effectively focuses on the context information about anatomy and morphology and suppresses the redundant intervention.
3. The SGAFF module is proposed to selectively combine the features and embeddings in the latent space according to the importance of the features and the embeddings to produce the proper RF to meet the task requirement.

Experimental results demonstrate that the proposed model achieves competitive performance in multiple pubic datasets related to retinal fundus image segmentation in both qualitative and quantitative evaluations.

The remainder of this paper is organized as follows. Section 2 reviews the related work in image segmentation using CNN and transformer. Section 3 describes the methodology in detail. Section 4 presents experiment results and analysis, and Section 5 provides an in-depth discussion of the advantages and disadvantages of the proposed method. Section 6 concludes the paper.

2. Related work

This section reviews the existing research about CNN and the combination of CNN and transformer in semantic segmentation, medical image segmentation tasks and retinal blood vessel segmentation.

2.1. CNN in semantic segmentation

Since U-Net (Ronneberger et al., 2015) was proposed to integrate low-level and high-level information for image segmentation, many variants of U-Net emerged. Recurrent residual convolutional neural network based on U-Net (R2U-Net) empowers a larger RF and temporal relationships on the U-Net by residual connection and recurrent neural network (RNN) (Alom et al., 2018). Besides, Attention U-net amplifies the explicit features and removes the irrelevant information for the variety of object shapes (Oktay et al., 2018).

Apart from the U-Net and its variants, a few CNN-based models arose for more accurate segmentation. Efficient neural network (ENet) alternately uses the standard, dilated, and asymmetric convolution to make real-time segmentation possible (Paszke et al., 2016). Semantic pixel-wise segmentation neural network (SegNet) transforms the feature map in the contracting path to smooth segmentation by trainable pooled indices rather than classify a pixel as semantic segmentation by U-Net (Badrinarayanan et al., 2017). Peng et al. designed convolution in a classification and localization way reducing the impact of local disturbances in SegNet (Peng et al., 2017). Pyramid scene parsing network (PSPNet) parses the scene into sub-region with different pyramid pooling rates enhancing the scene understanding (Zhao et al., 2017). Deeplab V3+ supplies the miss details of boundaries caused by repeated convolution and pooling in PSPNet using atrous convolution at a certain computation cost (Chen et al., 2018).

2.2. CNN in retinal blood vessels segmentation

The progress of CNN in semantic segmentation paves the way to retinal blood vessel segmentation. A large number of researchers explore more accurate vascular segmentation methods based on U-Net.

Bottom–top and top–bottom short connections in deeply supervised network (BTS-DSN) bidirectionally connects the semantic and structure information to complement them mutually (Guo et al., 2019). Besides, pool-less residual segmentation network (PLRS-Net) reduces the times of pooling operations to retain the feature map size for small vessels and provides intermediate spatial information to make convergence faster (Arsalan et al., 2022). Cascaded residual attention U-Net (CRAU-Net) inserts DropBlock into the traditional residual block to avoid overfitting and treats the channels of low-level and high-level features equally in the decoder to stress the shallow features (Dong et al., 2022).

Meanwhile, some researchers utilize dilated convolution to enlarge the receptive field without the reduction of information. Mou et al. extracted a high-level feature by atrous convolutions placed in a dense way and fixed the broken vessels in the initial predictions by the probability regularized walk (Mou et al., 2019). Besides, context-aware network (CA-Net) utilizes the multi-dilated convolution with different rates to obtain the multi-scale contextual information and improves the fusion of multi-scale spatial information (Wang et al., 2022). Also, dilated convolutions U-Net (DilU-Net) places multi-dilated convolutions in parallel and concatenates them to form high-level features and upsamples the multi-scale outputs from different decoders to make the final predictions (Hussain et al., 2022).

In addition, the distribution of inter-class and intra-class is recognized as an important factor for vascular segmentation. The hybrid deep segmentation network (HD-Net) builds the relationships between thin and thick vessels based on the multi-task segmentation network and merges the thin and thick vessels segmentation by more skip connections (Yang et al., 2021). Context-involved U-Net (Bridge-Net) is proposed to capture the contextual information effectively by incorporating sequence information on the feature map and ensures the balance between the vascular and backgrounds by the patch-based loss weight mapping (Zhang et al., 2022).

Some other research focuses beyond the model design. Feedback attention network (FANet) combines the predicted mask provided by the previous epoch with current feature maps to add attention during the training phase and rectifies them to obtain the final predictions (Tomar et al., 2022). Besides, data-Driven Deep Supervision (DDS) determines the layer-wise efficient receptive field by the back-propagation error, calculates the objective perceptive field by the pixel-wise connectivity and vessel width and picks out the main layer of feature extraction (Mishra et al., 2022).

2.3. CNN complemented transformer in medical images segmentation

Influenced by the incredible success of transformer in natural language processing, more and more researchers attempt to evolve the transformer into computer vision. The first work is proposed to split the original image into non-overlapping image patches as tokens, embed patches with positional information, and feed them into the transformer (Dosovitskiy et al., 2020). However, the transformer-based methods achieve inferior performance in the fine-grained task due to the lack of details. Thus, CNN complements transformer is a feasible option for medical image segmentation.

Assembling the transformer module with CNN is a recent trend mainly for medical images. The first stream in this direction evolves around modifying the encoders. Transformers and U-Net (TransU-Net)

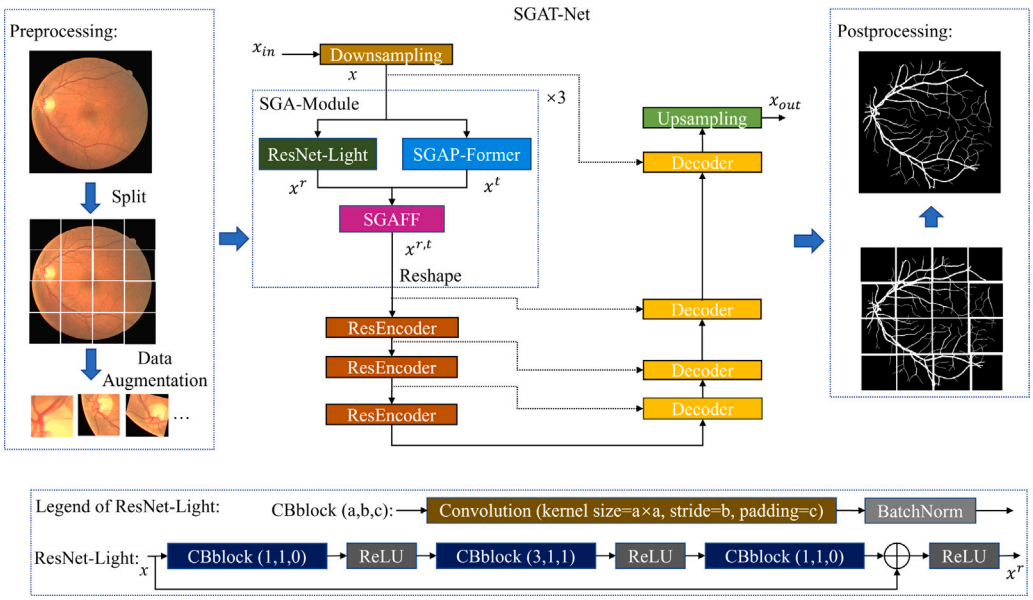


Fig. 3. The proposed retinal vessel segmentation framework. The main pipeline contains the preprocessing module, SGAT-Net, and a postprocessing module. The preprocessing module divides the original image into patches and performs augmentation. The SGAT-Net extracts the local–global compound features. The postprocessing module averages the multiple predictions on each original patch and replaces the patch masks back into the whole image. The details of ResNet-Light are shown at the bottom. Besides, the detailed design of the modules in SGAT-Net, including Downsampling, ResEncoder, Decoder, and Upsampling, is described after the overview of the main pipeline.

regards the CNN as information embedded and sets up long-term connections among these embedding vectors by the transformer for multi-organ segmentation and cardiac segmentation (Chen et al., 2021b). Correspondingly, claw U-Net with transformers (TransClaw U-Net) concatenates the various feature and tokens in each stage to ensure precise multi-organ segmentation (Chang et al., 2021).

Some research focuses on the modification of the connection between the encoder and decoder. Deng et al. fused the multi-scale features and global dependencies among them, leading to more accurate left ventricle segmentation (Deng et al., 2021). Chen et al. proposed transformer self-attention and global spatial attention to augment the multi-scale feature representation and aggregation for multi-modality segmentation (Chen et al., 2021a).

Another general idea is to modify the decoder of CNN structures to include the power of transformer. There is a low volume of research about modification in the decoder, Li et al. reconstructed the contract information based on window attention in pixel level and bilinear upsampled in the residual connection, attaching locality in the upsampled feature at lower computation cost for brain tumour segmentation (Li et al., 2021a). Besides, the squeeze-and-expansion transformer (SegTran) enlarges the RF size by first squeezing the single-head attention matrix and then stacking the dynamic attention calculated by softmax for colonoscopy segmentation (Li et al., 2021b).

2.4. CNN complemented transformer in retinal vascular segmentation

Retinal vascular segmentation as one of the branches of medical image segmentation also follows a similar model design. Group transformer network (GT-Net) contains a group of bottleneck structures that decrease the computational complexity of transformer to facilitate the combination of CNN and transformer (Li et al., 2021c). Yu et al. first predicted the gamma value by CNN to adjust the intensity distribution of retinal images and the channel attention vision transformer enhances the edge feature map spatially and in the channel (Yu et al., 2022).

Besides, retinal blood segmentation can be regarded as a coarse-to-fine process. Transformer Unet and local binary energy function model (TUnet-LBF) generates the coarse contour based on the vascular topology and increase the sensitivity of segmentation based on the energy parameters (Zhang et al., 2023). Similarly, the cascade hybrid transformer architecture (CasUTNet) first produces rough predictions by inserting self-attention mechanism into the U-Net and combines them with the input images, and rectifies the coarse results (Cai and Ma, 2022).

Moreover, incorporating spatial and channel attention is important for feature extraction. Triple attention network (TA-Net) establishes the long-range dependency and employs channel attention during downsampling and spatial attention during the upsampling to locate the useful features (Li et al., 2021d). Squeeze-Excitation Transformer U-net (SETUnet) extracts features efficiently by the residual block and selects the effective attention heads according to the spatial and channel attention to utilize the global information (Shen et al., 2022).

However, these existing methods still obtain sub-optimal results due to the insufficient combination of local and global information. Besides, the redundant information and noise from transformer also handicap its use.

3. Methodology

As depicted in Fig. 3, the proposed fundus retinal vessel segmentation framework consists of a preprocessing module, a Stimulus-Guided Adaptive Transformer Network (SGAT-Net), and a postprocessing module. The preprocessing module first resizes the original image to 2048×2048 and then divides it into 512×512 patches, augments them by resizing, mirroring, and cropping, and feeds them into the SGAT-Net. After feature extraction and pixel-level labelling in the SGAT-Net, the final segmentation is obtained by averaging the multiple results on each original patch, and replacing the patch masks back into the whole images through the postprocessing module.

The SGAT-Net follows the U-shape design of the classic U-Net, but it includes re-designed encoding and decoding branches to extract local-and-global compound features based on inductive bias and long-range dependency. As illustrated in Fig. 3, the left branch includes Downsampling module, Stimulus-Guided Adaptive Module (SGA-Module), ResEncoder module, and the right branch consists of Decoder module and Upsampling Module.

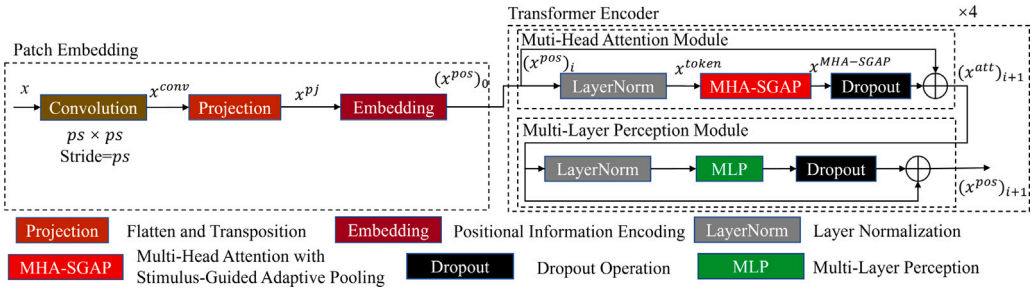


Fig. 4. Illustration of the basic structure of SGAP-Former. Patch Embedding module splits the features into patches and embeds them with positional information. Then, Transformer Encoder module enhances the embedding representation and establishes the long-range dependency.

3.1. The basic units of SGAT-Net

Downsampling Module: The downsampling module firstly reduces the resolution of the original image $x_{in} \in \mathbb{R}^{C_{in} \times H_{in} \times W_{in}}$, where H_{in}, W_{in} are the resolution of the images, C_{in} is the number of channels, to obtain $x \in \mathbb{R}^{C \times H \times W}$, with C, H, W being the one-quarter of C_{in}, H_{in} and W_{in}, respectively, the aim of the downsampling step is to reduce the computation load in the subsequent steps. Inspired by the ResNet, the downsampling module has two main parts: a 7×7 convolution kernel with strides 2 and paddings 3, and a 3×3 maxpooling operation with strides 2 and padding 1. Then, the downsampled images are fed into the SGA-Module, which will be described in detail in the next section.

ResEncoder Module: The SGA-Module is followed by three ResEncoder modules. Considering the computation complexity and the overfitting, ResEncoder has half of the expansion rate and the embedded dimension of the standard ResNet. There are multiple ResBlocks in the ResEncoder and the number of ResBlock for each ResEncoder is chosen from [3,4,6,3] according to the ResEncoder location. Each convolution block in ResBlock comprises convolution, batch normalization, and Rectified Linear Unit (ReLU) except for the last convolution block. The kernel size and strides are 1×1 and 1 for the first and last convolution blocks without padding, and 3×3 for the middle convolution block with 1 padding. The stride of the first ResBlock is 2 and the rest ResBlock is 1. Besides, the skip connection adds the original input and the highly encoded features to avoid vanish gradient due to the depth of convolution blocks.

Decoder Module: There are four decoder modules and one upsampling module. In the decoder module, an upsampling block increases the resolution of the features at the current stage to that of the previous stage. The upsampling blocks consist of upsampling with factor 2, 3×3 convolutions with stride 1 and padding 1, batch normalization, and ReLU operation. Then, two types of features are concatenated and go through the two convolution blocks to modify the channel of features. As depicted, each convolution block is built with a 3×3 convolution with stride 1 and padding 1, followed by batch normalization, and ReLU.

Upsampling Module: After the Decoder modules, the upsampling module upsamples the decoded feature to the original resolution, followed by two kinds of convolution, batch normalization and ReLU leading to the final prediction.

3.2. SGA-Module

As a result of the inductive bias, CNN is designed to focus on the target and the surrounding pixels in the RF but weakens the long-term relationship with the distant regions and even pixels. In contrast, the transformer neglects the inherent data structure and establishes the long-range relevance of each patch. Aiming to join the strength of both approaches and complement their weakness, the SGA module is proposed, which can extract local–global compound features as highly descriptive and distinctive representations of the vessels and other retinal structures.

As shown in Fig. 3, the SGA-Module consists of a light-weight ResNet structure (ResNet-Light), a Stimulus-Guided Adaptive Pooling Transformer (SGAP-Former), and a Stimulus-Guided Adaptive Feature Fusion (SGAFF) module. ResNet-Light generates the coarse-to-fine features $x^r \in \mathbb{R}^{C_r \times H_r \times W_r}$ by the multiple convolution blocks in the deep layers and the residual connection. Compared with ResEncoder, ResNet-Light has a similar structure with ResEncoder but the stride of each ResBlock in ResNet-Light is set to 1.

3.2.1. SGAP-Former

Parallel to the ResNet-Light branch, an additional branch is constructed to process the same features. The SGAP-Former analyses them by utilizing the long-range dependency of the transformer. The SGAP-Former is designed to establish the relationships between the nearby and remote pixels, and on top of that, to select the pooling ways in accordance with the feature property, emphasizing the relevant information of context and anatomy. In this way, it takes into account information from the global view rather than only within the local RF and suppresses the interference from irrelevant artefacts and blurred regions. As shown in Fig. 4, the SGAP-Former consists of two main modules: patch embedding and transformer encoder.

Patch embedding: the features x from the Downsampling module will be reshaped into features with two channels: the number of patches and the embedded dimension. Firstly, the input will be divided into several patches by the convolution with fixed strides:

$$x^{conv} = \sum_{n=0}^{C-1} \sum_{j,k=m \times ps}^{(m+1) \times ps} ((x)_n^{j,k} \star wg_n) + bs_n, \tag{1}$$

where $(x)_n^{j,k}$ denotes the n th channel's value at coordinates (j,k) in the input x, $\star$ is the 2D cross-correlation operator, wg_n and bs_n are the n th channel's weight and bias of the kernel. As the size of kernel and stride is the same as ps (ps is the patch size) and j, k are m times of ps (i.e. $j,k = m \times ps$, $m \in \mathbb{Z}$), the convolution can be regarded as the division. Then, transposing the channel of the flattened feature generates the projection features $x^{pj} \in \mathbb{R}^{P \times C_{pj}}$, where $P = (\frac{H}{ps} \times \frac{W}{ps})$ is the number of patches, and C_{pj} is the projection dimension. Since the projection on the distinct images can produce the same projection vector, it is necessary to insert the encoded positional information $x^{ep} \in \mathbb{R}^{P \times C_{em}}$, with C_{em} being the embedded dimension on the projection features:

$$(x^{pos})_0 = [(x^{pj})_0 \times W; (x^{pj})_1 \times W; ...; (x^{pj})_{p-1} \times W] + x^{ep}, \tag{2}$$

where $(x^{pj})_n$ is the n th patch in the embedding features x^{pj}, which omits the class token compared with the original ViT, $W \in \mathbb{R}^{C_{pj} \times C_{em}}$ is the weight matrix.

Transformer encoder: the encoded positional features x^{pos} are processed by the two residual structures. The Multi-Head Attention Module (MAM) is the combination of Layer Normalization (LayerNorm), Multi-Head Attention with Stimulus-Guided Adaptive Pooling (MHA-SGAP), and Dropout:

$$(x^{att})_{i+1} = \text{MHA-SGAP}(LN((x^{pos})_i)) + (x^{pos})_i, i = 0...N, \tag{3}$$

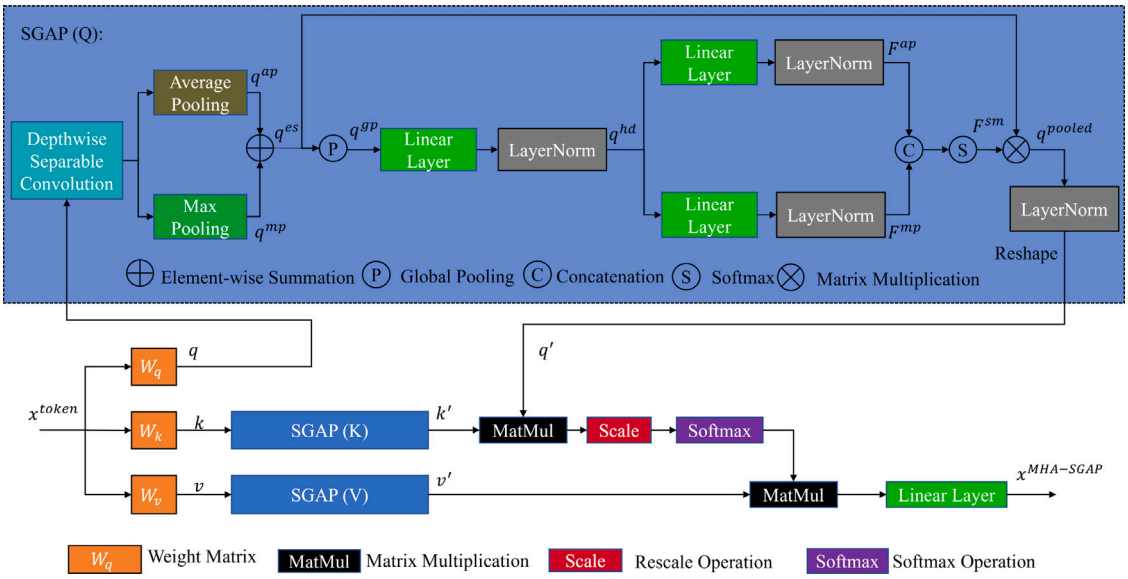


Fig. 5. Illustration of the MHA-SGAP module as the core component of the SGAP-Former. This module first generates the three embeddings query (q), key (k), and value (v) based on the token and then compacts these embeddings by selective pooling relying on the adaptive fusion features from average and maximum pooling. Despite the input, the SGAP(K) and SGAP(V) share the same structure with SGAP(Q).

where $(x^{pos})_i \in \mathbb{R}^{P\times C_{em}}$ is the i th input of MHA-SGAP modules, LN denotes the LayerNorm, N is the number of MAM. MHA-SGAP is based on the fundamental idea of multi-head attention (MHA) in traditional transformer, but added the design of stimulus-guided adaptive pooling, to make it more focused on the distinctive features. The Multi-Layer Perception Module (MLPM) consists of the LN, Multi-Layer Perception (MLP), and Dropout operation:

$$(x^{pos})_{i+1} = \mathrm{MLP}(LN((x^{att})_{i+1})) + (x^{att})_{i+1}, i = 0...N, \tag{4}$$

Notably, the MAM is directly connected with MLPM. In contrast, MAM focuses on the relationship between different blocks, namely, the long-term interactions, and MLPM discovers the relevance between the surrounding pixels, namely, the short-term connections. Followed by the three consecutive MAM and MLPM, the relationship between the flattened patches beyond the pixel level is built, which harnesses the contextual information to enhance the quality of predictions, especially for ambiguous or truncated areas.

MHA-SGAP: Stimulus-Guided Adaptive Pooling (SGAP) is proposed to work along with the traditional MHA so that the attention is able to focus on the most relevant details in morphology and anatomy, as well as to reduce the influence of redundant information.

According to the stimulus that will affect the visual cortical neurons' response namely, RF, SGAP is achieved by adaptively fusing average pooling and maximum pooling to simulate visual cortical neurons. That is, average pooling reduces the influence of the noisy or glitch pixels to summarize the implicit stimulus and segment blurred vessels from background, while maximum pooling pays attention to the regions with high contrast to respond to the explicit stimulus and decreases the influence of artefact. Notably, the selection of pooling heavily depends on the property of fed features. For example, the maximum pooling is more suitable for colour features rather than the average pooling.

Similar to the traditional MHA mechanism, the query (q), key (k), and value (v) in MHA-SGAP are derived based on the token $x^{token} \in \mathbb{R}^{P\times C_{em}}$:

$$q, k, v = \mathrm{Split}(x^{token} \times W^{q,k,v}), \tag{5}$$

where $W^{q,k,v} \in \mathbb{R}^{C_{em}\times 3C_{em}}$ is the weight matrix to increase the channel size by three times and the channel are equally split to obtain $q, k, v \in \mathbb{R}^{P\times C_{em}}$.

As depicted in Fig. 5, the SGAP module selectively pools the original input based on its q, k, and v. q is chosen as an example to illustrate the SGAP (Q) that is the same as SGAP (K), SGAP (V). First, q is reshaped and goes through the Depthwise Separable Convolution (Chollet, 2017) module including the two convolution blocks. The groups of the first block are set to C_{em} and the second one is set to 1, resulting in the depthwise convolution and pointwise convolution, respectively. Second, the features from averaged pooling $q^{ap} \in \mathbb{R}^{P\times C_p}$, where C_p is the channel of features, and maximum pooling $q^{mp} \in \mathbb{R}^{P\times C_p}$, undergo a concatenation and global average pooling operation. In the concatenation, the features are unsqueezed to enlarge the number of channels from 3 to 4 to facilitate the element-wise summation, leading to the fused features $q^{es} \in \mathbb{R}^{P\times C_p}$:

$$q^{es} = \sum_{c=0}^{2}((q^{ap})_c + (q^{mp})_c), \tag{6}$$

where $(q^{ap})_c$ and $(q^{mp})_c$ represent the c th channel of the unsqueezed q^{ap} and q^{mp}. Third, global average pooling extracts the descriptive feature $q^{gp} \in \mathbb{R}^{C_p}$ and filters out the interference information:

$$q^{gp} = \frac{1}{H\times W}\sum_{j=0}^{H}\sum_{k=0}^{W}(q^{es})^{j,k}, \tag{7}$$

where $(q^{es})^{j,k}$ denotes the coordinate (j, k) of the summation result q^{es}. Besides, the merged and pooled outputs are transferred to the latent space to facilitate the division:

$$q^{hd} = M(q^{gp}) = LN(q^{gp} \times W_G), \tag{8}$$

where M is the mapping function, $W_G \in \mathbb{R}^{C_p\times C_{hd}}$ is the Linear Function to rearrange the latent space, C_{hd} is the dimension of the latent space, LN adjusts the distribution of features obeying the Gaussian Distribution.

Then, the feature q^{hd} in the latent space goes through two parallel paths, both formed of a Linear Layer and LN to reduce the feature space, as shown in Fig. 5, the process is similar to Eq. (8). $F^{ap} = q^{hd} \times W_A$, $F^{mp} = q^{hd} \times W_M$ are the reduced feature space for the average pooling and maximum pooling path, respectively. The weight of Linear Layer in each path are $W_A \in \mathbb{R}^{C_{hd}\times\frac{C_{hd}}{2}}$ and $W_M \in \mathbb{R}^{C_{hd}\times\frac{C_{hd}}{2}}$. Dimension $\frac{C_{hd}}{2}$ means the two paths play an equally important role in feature selection. Moreover, the information flow from the two paths reconcatenated and softmax is operated to determine the attention in each path and fuse the

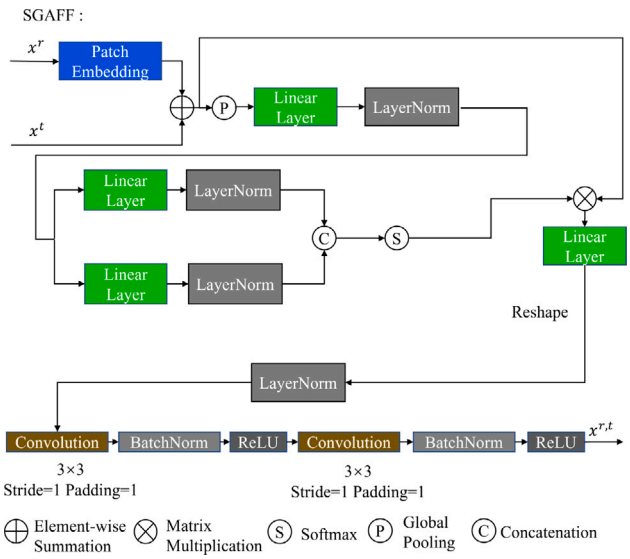


Fig. 6. Illustration of the SGAFF module. Although it is quite similar to SGAP, it processes the feature from two domains (CNN-based and transformer-based) and enlarges the RF. As a result, the feature representation capability is enhanced.

two distinct pooled features :

$$q_{wap} = \frac{e^{F^{ap}}}{e^{F^{ap}} + e^{F^{mp}}},$$
$$q_{wmp} = \frac{e^{F^{mp}}}{e^{F^{ap}} + e^{F^{mp}}}, \tag{9}$$
$$q_{wap} + q_{wmp} = 1,$$

where q_{wap} and $q_{wmp} \in \mathbb{R}^{\frac{C_{hd}}{2}}$ are the weights of the average pooled and maximum pooled feature, respectively, e is the exponential function and $q_{wap} + q_{wmp} = 1$ ensures the coefficient normalization.

Finally, the multiplication between the concatenated outputs and the adaptive weight through the skip path generates the adaptive pooled results $q^{pooled} \in \mathbb{R}^{P \times C_{em}}$:

$$q^{pooled} = q^{es} \times F^{sm}, \tag{10}$$

where the adaptive weight is F^{sm}, $F^{sm} = Concat(q_{wap}, q_{wmp})$, where $Concat$ denotes the concatenation operation. Then, an LN and reshape operation rearrange q^{pooled} to the original form of q.

After SGAP, the new query (q'), key (k'), and value (v') are connected together by multiplying, rescaling and softmax operations to obtain the attention value, and the following Linear Layer increases the dimension of the adaptive pooled feature by four times to compensate for the decreased dimension due to the pooling operation.

3.2.2. SGAFF

The locality allows CNN to clearly perceive the explicit stimulus, e.g. blood vessels' colour and shape easily, and the self-attention mechanism makes the transformer more sensitive to the implicit stimulus, e.g. the structure and the potential relationship between nearby and remote pixels. Thus, followed by the ResNet-Light and SGAP-Former, the SGAFF module is designed to adaptively allocate the weight of the features and embeddings, and rearrange and combine them in the latent space to generate a proper RF based on the task at hand, such as thin or thick vessel segmentation.

Compared with SGAP, the SGAFF processes the CNN-based features x^r and the pooled transformer-based embeddings x^t rather than only the traditional embeddings as shown in Eq. (5). As depicted in Fig. 6, the CNN features are first fed into the patch embedding layer which is the same as the first layer in the transformer to divide the feature into several patches and flatten the dimension from 4 to 3. This procedure ensures the features are rearranged into the same space where embeddings exist.

Then, similar to the SGAP module, the features integrate local and global information obtained by concatenation and averaging the flattened features. The multi-level features are transformed into high-dimension space by Linear Layer and LN. The features are passed through two paths including a Linear Layer, which rises to the latent space that double the embedded dimension, followed by LN for normalization. The respective weights for these two structures are obtained by the same operation as shown in Eq. (9). In addition, the calculated weights are multiplied with concatenated features, i.e., the weighted local-and-global features. Finally, the Linear Layer adjusts the channel of local-and-global features to match with the original x^r, reshapes them, and passes through the two convolution blocks (each block contains a 3×3 convolution kernel with stride 1 and padding 1, batch normalization and ReLU) to obtain the final output $x^{r,t}$.

In summary, the preprocessed retinal fundus images are downsampled and then fed into the SGA-Module. In the SGA-Module, ResNet-Light based on inductive bias encodes the coarse-to-fine information about the appearances of retinal vessels, and SGAP-Former based on the self-attention mechanism extracts the global context such as the structure of blood vessels. Following that, SGAFF adaptively fuses the local and global features to produce the proper RF for accurate vascular segmentation. The enhanced features are fed into the encoders and then go through the decoders. Finally, the upsampling module upsamples the decoded features to obtain segmentation of the retinal blood vessels.

Table 1
Summary of key information of the four datasets.

Dataset	Amount	Resolution	Disease	Observers	Year
STARE	20	605 × 700	10 pathology images	2	2000
DRIVE	40	565 × 584	7 pathology images	3	2004
CHASEDB1	28	999 × 960	28 normal images	2	2012
FIVES	800	2048 × 2048	600 pathology images	2	2021

4. Experimental results

To perform a thorough evaluation of the proposed method and a comparative analysis against the state of the art, the experiments are conducted on four public datasets: Digital Retinal Images for Vessel Extraction (DRIVE) (Staal et al., 2004), subset of Child Heart and Health Study in England (CHASEDB1) (Fraz et al., 2012) and Structured Analysis of the Retina (STARE) (Hoover et al., 2000) and Fundus Image Vessel Segmentations (FIVES) (Jin et al., 2022).

4.1. Datasets

This section introduces the four public datasets in detail and their key information is summarized in Table 1.

4.1.1. DRIVE

Staal et al. (2004) utilized the blood vessel ridges based on their elongated structures to classify the straight line into the blood vessel automatically on the Utrecht database. Hence, they generated the DRIVE dataset consisting of 40 images with the size of 565 × 584 in the Joint Photographic Experts Group (JPEG) format and being captured at the field of view (FOV) of 45°. Besides, the data is split into a training dataset and a test dataset with the same amount of 20 images. The training dataset is annotated by two annotators and the test dataset is segmented twice (set A and B) by them. As set A includes 12.7% more blood vessel annotation than set B, the researchers usually use the marks of set A as the ground truth.

4.1.2. STARE

In comparison with the DRIVE dataset, Hoover et al. (2000) proposed the region-based probing method to establish the STARE dataset with fewer false positives. The STARE dataset supplies 20 RGB images with 605 × 700 resolutions. For the clinical purpose, half of them have a few lesions on the blood structure and the remaining normal. Then, the whole dataset is manually marked using the tool in Hoover et al. (1994), which provides the proper magnification level and histogram transformation for visualization.

4.1.3. CHASEDB1

In contrast with DRIVE and STARE, the CHASEDB1 dataset is based on an ensemble system with the multiscale Gabor filter, morphological transformation, etc. A total of 28 images are collected from 14 school children of various ethnic groups regarding cardiovascular health from 46 primary schools, and the resolution of each image is 999 × 960 with FOV 30°. Besides, two individual annotators manually segmented each image and the segmentations of the first one are used as the standard annotations.

4.1.4. FIVES

Compared with the above retinal vessel image datasets, FIVES is a recently published dataset with remarkable advantages. First, it consists of 800 retinal images from 573 people in total, which is around 40 times the size of DRIVE and CHASEDB1, and 20 times that of STARE. The average age of patients is 48 and there are 469 and 331 images collected from women and men, respectively. Second, the image resolution is on a large scale 2048 × 2048 to benefit the image analysis. Third, the segmentation annotation has good intra- and inter-observer consistency. The labelling is individually performed by 3 ophthalmic practitioners and 24 medical staff. Finally, the image data is collected from 200 normal eyes, 200 AMD, 200 DR, and 200 GC, with a good balance between the retinal vessel and the ocular disease to help better facilitate the development and validation of computational algorithms.

In this paper, all the above datasets are employed for testing, while the FIVES dataset is viewed as the main dataset to evaluate the performance of the methods in the experiment, considering its high quality and quantity in both imagery data and annotation.

4.2. Experimental details

The designed model is trained and tested on the same hardware: Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50 GHz, four NVIDIA A100 40 GB GPUs, one RTX A5000 24 GB and the software environment is python 3.8, PyTorch1.8.1, CUDA 11.1, and cudnn 8.1.

Experimental configuration: the initial learning rate is 0.0002, the loss function is Cross Entropy + DICE, the batch size is 16, and the optimizer is Adam. Also, the number of epochs is 100 and the first 40 maintain the learning rate, and the rest decay the learning rate based on the current epoch.

Data augmentations: there are two kinds of prepared images: the RGB images and the CLAHED images. The original RGB images are converted into greyscale images and normalized, then Contrast Limited Adaptive Histogram Equalization (CLAHE) (Pizer et al., 1987) and gamma correction are applied to the normalized images to enhance the image quality, leading to the CLAHED images. On top of that, augmentation operations including flipping, rotating, translating, zooming, adding Gaussian noise, and randomly adjusting image brightness are applied to the prepared images to improve the generalization of the model.

Experimental metrics: There are six standard metrics for medical image segmentation: Recall (RC), Specificity (SP), Accuracy (ACC), IOU, F1 and Area Under Curve (AUC). These six metrics focus on different aspects of segmentation performance, but at the same time, they mutually restrain others, so all of them need to be examined, to give a thorough evaluation of the performance of methods in the retinal images.

Table 2
The ablation study of the proposed methods.

Approach	RC (%)	SP (%)	ACC (%)	IOU (%)	F1 (%)
ResNet-E	89.74	99.24	98.61	80.64	88.56
ResNet-N	91.35	99.32	98.84	83.12	90.28
ResVit-F	**92.80**	99.19	98.80	82.71	90.09
ResVit-L	91.70	99.31	**98.86**	83.37	90.47
ResVit-L-NC	92.53	99.23	98.85	83.00	90.11
SGA-L	92.71	99.19	98.79	82.21	89.65
SGA-2	89.98	99.30	98.70	81.88	89.39
SGA-F-NC	92.47	99.20	98.82	82.70	90.02
SGA-F (Ours)	91.62	**99.33**	**98.86**	**83.47**	**90.51**

4.3. Ablation studies

In this section, ablation studies carried out on the FIVES dataset are reported, with the key information given in Table 2. The different sizes of the embedded dimension of ResNet are compared. Then, based on these results, the effectiveness of SGAFF is shown and a few different combinations of CNN and traditional transformer are compared. Besides, under the condition that the MHA is replaced by the proposed MHA-SGAP and other modules are kept the same, the effect of the SGA-Module's position is also explored.

The embedded dimension. Taking the captured multi-scale information into consideration, the encoder of the original U-Net is replaced by the ResNet with the half expansion rate named ResNet-E. Comparing the results in Tables 2 and 3, it turns out that ResNet-E outperforms the standard U-Net. The embedded dimension increased from the original channel is the critical factor to determine the embedding space. In the first two rows of Table 2, ResNet-N with half the embedded dimension performs better than ResNet-E in all the five metrics, with the 2.48%, 1.72% and 1.61% gain of IOU, F1, and recall, respectively. For retinal fundus images, there is a high chance of performance overfitting using larger embedded dimensions.

The effectiveness of SGAFF. The embeddings provided by the transformer are reshaped into the same dimension and scale as the feature maps obtained by ResNet-N, and then they are added together to fuse the features from CNN and transformer. The results of this approach are denoted by the original methods' name followed by an 'NC' suffix. The fourth and fifth rows of Table 2 illustrates the effectiveness of SGAFF on high-level features provided by the ResVit-L, 0.37% and 0.36% higher in IOU and F1. As for SGA-F, the pooling operation removes irrelevant information and the last two rows of Table 2 depicts that the SGAFF operation outperforms the normal confusion operations in four metrics, especially 0.77% and 0.49% gains in IOU and F1. In general, the SGAFF module utilizes the features provided by CNN and embedding provided by transformer and effectively combines them to generate the local-and-global compound features for retinal vascular segmentation.

The introduction of transformer. Above the experiments, the SGAFF achieve superior performance than normal confusion, so the SGAFF is selected to fuse the features from CNN and transformer in the following experiment. As shown in Table 2, ResVit-L and SGA-F represent the introduction of traditional transformer in the last stage and SGAP-Former in the first stage respectively. The second and fourth rows of Table 2 illustrates that the introduction of the traditional transformer brings 0.35%, 0.25% and 0.19% gains in recall, IOU and F1 respectively. Similarly, the existence of SGAP-Former also gains 0.27%, 0.35% and 0.23% higher recall, IOU and F1 according to the second and last rows of Table 2. It proves that the introduction of different transformers at different stages has better performance due to the long-range dependency.

The location of ResVit. Unlike the SGA-Module, the ResVit only contains the combination of CNN and transformer with the original MHA rather than MHA-SGAP. According to the research on the embedded dimension, the different locations of ResVit are set the same

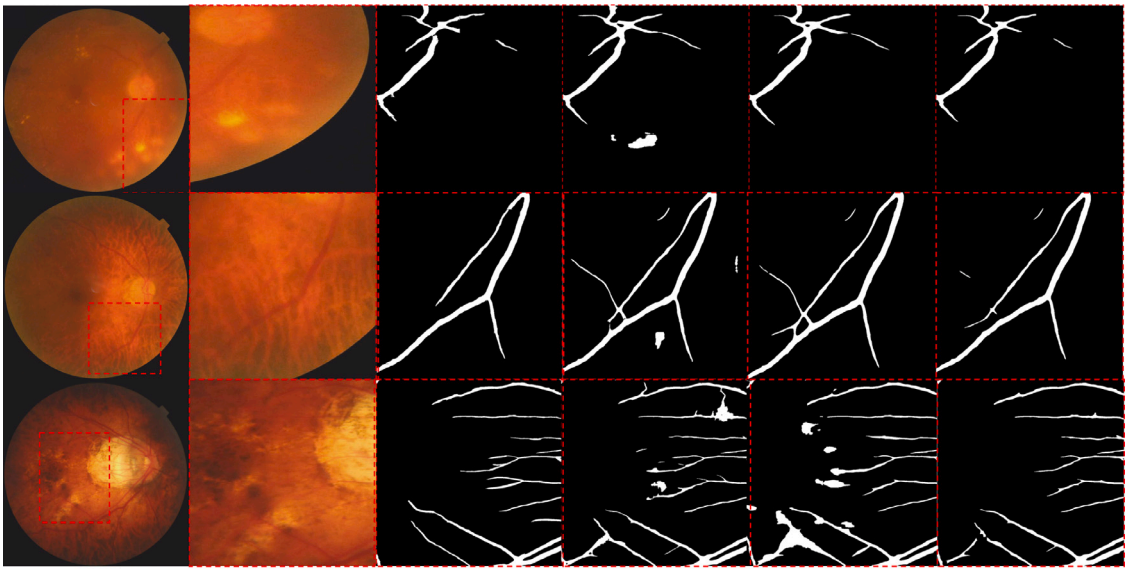


Fig. 7. A few examples to visually illustrate the result of ablation studies. The performance of ResNet-N, ResVit-F, and SGA-F is compared in the region of interest (ROI) of the retinal fundus image.

as the embedded dimension 32. The ResVit is placed at the first and the last stage of the encoder, namely ResVit-F and ResVit-L. The third and fourth rows of Table 2 show that ResVit-L achieves a minor gain of 0.66% and 0.38% in IOU and F1, but is 1.10% in recall lower than ResVit-F. It turns out that combining CNN and transformer at the last stage can handle abstract and compact information better while doing so in the first stage favours apparent and specific information.

The position of SGA-Module. Similar to the research on the location of ResVit, the SGA-Module is placed at the first and the last stage of the encoder, namely SGA-F and SGA-L. As it is shown in the sixth and last rows of Table 2, there is a 1.09% improvement in the recall whereas the rest four metrics decline to some extent such as a 1.26% decrease in IOU. On the contrary, although ResVit-F owns the highest recall which is 1.18% higher than SGA-F, it still has an inferior performance than SGA-F. There is remarkable progress between the SGA-F and ResVit-F, that is, 0.76% 0.42%, and 0.14% gains in IOU, F1, and specificity respectively. It proves that the SGAP operation outperforms the hierarchical convolution with different strides because SGAP operation occurs at an early stage that maintains the descriptive feature such as the colour and scale and provides intact features as much as possible.

The number of SGA-Module. Due to the high computational complexity of transformer, the combination of SGA-F and SGA-L including two SGA-Module named SGA-2 is only conducted during the ablation study to explore the influence of the number of SGA-Module. The sixth, seventh and last rows of Table 2 illustrates that the performance of SGA-2 achieves an inferior performance in almost all the metrics compared with SGA-F and SGA-L, because the two SGA-Module introduces multiple downsampling processes to compact the feature space and this may lead to removing some of the representative features. Besides, the number of parameters of SGA-F with the best performance is 14.41 M, which is far smaller than the number of parameters of SGA-2 — 408.35 M. Based on this analysis, we argue that the use of a single SGA-module is sufficient and more effective for retinal vascular segmentation.

Some visual results are given in Fig. 7. ResVit-F reduces the errors of ResNet-N in the first and second rows. However, the ResVit-F also brings more segmentation errors due to the redundant information and noise as shown in the third row and the fifth column in Fig. 7. To improve the performance of ResVit-F, the SGAP module in the SGAT-Net (i.e. SGA-F) emphasizes the relevant morphology information and suppresses the redundant information, resulting in the best prediction that matches well with the ground truth.

Table 3
Comparative results of relevant deep learning methods on the FIVES dataset.

Approach	RC (%)	SP (%)	ACC (%)	IOU (%)	F1 (%)	AUC (%)
U-Net	91.86	99.10	98.66	80.77	88.87	93.00
R2U-Net	84.52	98.99	98.09	74.65	84.92	92.38
Attention-Unet	92.73	99.07	98.68	80.73	88.81	92.72
GCN	91.91	99.22	98.79	82.60	90.02	93.99
Deeplab V3+	87.92	**99.33**	98.50	80.75	88.56	**94.85**
SK	89.77	99.12	98.58	79.94	88.35	93.34
CBAM	**93.30**	99.01	98.67	80.29	88.50	92.26
PSPNet	91.92	99.20	98.78	82.35	89.88	93.96
ENet	89.93	99.22	98.67	81.10	89.09	94.09
SegNet	84.84	98.99	98.13	74.98	85.09	92.44
Swin-Unet	92.27	99.22	98.82	82.76	90.13	94.02
TransU-Net	91.80	99.28	98.83	83.17	90.37	94.47
SGAT-Net (Ours)	91.62	**99.33**	**98.86**	**83.47**	**90.51**	94.67

4.4. Retinal vessel segmentation in the FIVES dataset

This section compares the results of the proposed SGAT-Net and the traditional networks, and further analyses their performance on different diseases and normal eyes.

4.4.1. Comparison with classical networks

The proposed SGAT-Net is compared with a few most relevant models including U-Net (Ronneberger et al., 2015), R2U-Net (Alom et al., 2018), Attention-Unet (Oktay et al., 2018), global convolutional network (GCN) (Peng et al., 2017), Deeplab V3+ (Chen et al., 2018), selective kernel (SK) (Li et al., 2019), CBAM (Woo et al., 2018), PSPNet (Zhao et al., 2017), ENet (Paszke et al., 2016), SegNet (Badrinarayanan et al., 2017), Swin-Unet (Cao et al., 2021), TransU-Net (Chen et al., 2021b) on the FIVES dataset. As shown in Table 3, the SGAT-Net outperforms the other methods in four out of the six metrics, with the exception where CBAM with the channel and spatial attention gains the highest recall 93.30% and the SGAT-Net also gains the second-highest AUC (94.67%) which is 0.18% lower than DeeplabV3. Besides, the SGAT-Net achieves the same specificity 99.33% as the DeeplabV3+ which encodes the multi-scale textual information better than any other model. Also, there is a slim margin of about 0.05% in accuracy between the SGAT-Net (98.86%) and the Swin-Unet (98.82%) which is purely based on transformer and TransU-Net (98.83%) that

Table 4
Class-wise results of the top methods on the FIVES dataset.

Method	Disease AMD						Disease DR					
	RC (%)	SP (%)	ACC (%)	IOU (%)	F1 (%)	AUC (%)	RC (%)	SP (%)	ACC (%)	IOU (%)	F1 (%)	AUC (%)
U-Net	94.56	99.09	98.78	84.40	91.36	94.08	91.93	99.09	98.68	80.42	88.84	92.90
Attention-Unet	94.58	99.07	98.75	83.95	91.08	93.83	**92.61**	99.03	98.67	80.10	88.68	92.49
Deeplab V3+	90.61	99.34	98.42	83.74	90.32	95.50	83.36	**99.47**	98.16	77.85	86.58	**95.56**
SK	93.02	99.09	98.68	83.52	90.85	94.24	89.49	99.17	98.61	79.75	88.47	93.59
TransU-Net	**94.80**	99.21	98.92	86.33	92.51	95.06	91.25	99.31	98.84	82.92	90.45	94.71
SGAT-Net (Ours)	93.67	**99.37**	**98.96**	**87.05**	**92.93**	**95.91**	91.40	99.33	**98.87**	**83.24**	**90.64**	94.78
Method	Disease GC						Normal eyes					
	RC (%)	SP (%)	ACC (%)	IOU (%)	F1 (%)	AUC (%)	RC (%)	SP (%)	ACC (%)	IOU (%)	F1 (%)	AUC (%)
U-Net	90.53	99.16	98.90	76.73	85.57	90.79	90.40	99.05	98.30	81.55	89.72	94.22
Attention-Unet	**92.56**	99.13	98.92	76.68	85.35	90.30	91.16	99.06	98.37	82.18	90.11	94.28
Deeplab V3+	85.59	**99.44**	98.90	77.95	86.45	**93.75**	**92.10**	99.09	98.50	83.45	90.89	94.58
SK	86.68	99.28	98.82	75.97	85.03	91.77	89.91	98.95	98.19	80.49	89.06	93.74
TransU-Net	89.77	99.41	**99.06**	**79.63**	**87.42**	93.04	91.36	**99.20**	98.52	83.79	91.09	**95.09**
SGAT-Net (Ours)	89.51	**99.44**	**99.06**	**79.63**	87.31	93.08	91.90	99.16	**98.55**	**83.93**	**91.16**	94.90

combines the CNN with transformer. Moreover, the SGAT-Net obtains 0.30% gains than the TransU-Net with the second largest IOU (83.17%) and is also 0.14% higher in terms of F1 than that of TransU-Net (90.37%).

4.4.2. Comparison on different diseases and normal eyes

The proposed method is also evaluated in qualitative and quantitative ways on cases with three different diseases and normal eyes to explore potential medical applications.

Qualitative results. Some samples of segmentation results are visually illustrated in Figs. 8 and 9. Each group includes the original image, ground truth, and the predictions from U-Net, Attention-Unet, Deeplab V3+, SK, TransU-Net, and the SGAT-Net. The red, yellow, and green dash bounding boxes represent the improved segmentation by the SGAT-Net, the existing shortcomings of the SGAT-Net, and the possible segmentations that are missed in the ground truth, respectively. When all the evaluated models generate similar predictions which are not present in the ground truth, a green bounding box is used to show that there may be a missing region in the ground truth mask.

1. **Disease AMD.** The SGAT-Net can distinguish the blood vessel from the background even in the artefact area according to the first red bounding box in Fig. 8(a). Besides, The second red bounding box with similar features to the blood vessel has the least error. However, the proposed model gains inferior performance on some blurred blood vessels as highlighted in the yellow bounding boxes.
2. **Disease DR.** Although the SGAT-Net is less able to detect the thin retinal vessels in the yellow bounding box in Fig. 8(b), it manages to infer the minor blood vessels in the vague area, as shown in the first and second red bounding boxes. Besides, the predictions of models as shown in the green bounding boxes agree that there may be a vessel that is missed in the ground-truth segmentation.
3. **Disease GC.** The SGAT-Net avoids making the common mistake by other models, as shown in the first red bounding box in Fig. 8(c) and its predicted mask is almost exactly the same as the annotation in the second red bounding box. Nevertheless, a few thin vessels in the yellow bounding box also causes wrong segmentation by the SGAT-Net.
4. **Normal eyes.** The first and second red bounding boxes in Fig. 9 reveal that the SGAT-Net is able to differentiate the retinal vessel from the retinal nerve fibres. Similarly, the SGAT-Net can clearly depict the arch blood vessel as shown in the third red bounding box. However, there is a few missed segmentation in the yellow bounding boxes.

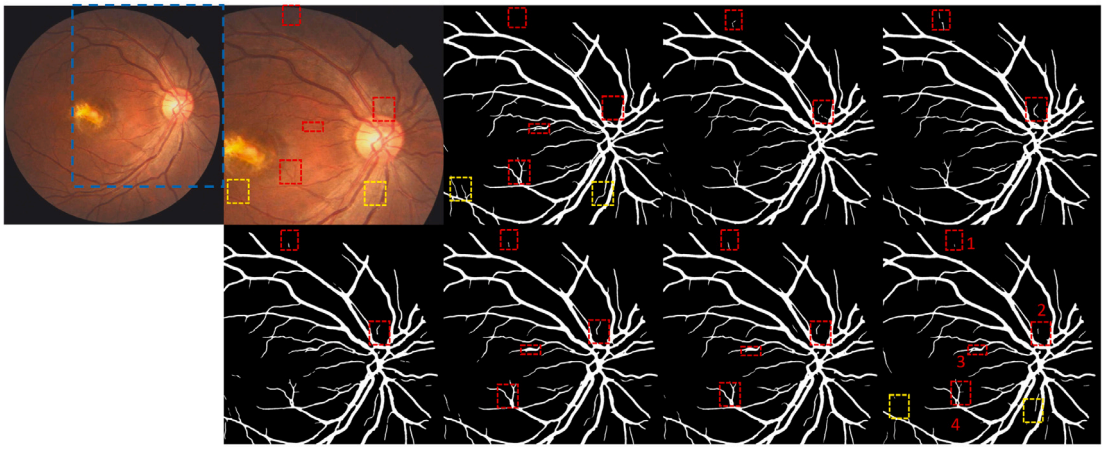
Quantitative results. In addition, Table 4 lists the performance of these classical methods on three diseases (AMD, DR, GC) and normal eyes. For disease AMD, the SGAT-Net, with specificity 99.37%, accuracy 98.96%, IOU 87.05%, F1 92.93% and AUC 95.91% surpasses the other model except for recall which is slightly lower (about 1.13%) than TransU-Net. For disease DR, although the recall of Attention-Unet and the specificity and AUC of Deeplab V3+ are 1.21%, 0.14% and 0.78% higher than SGAT-Net, the performance in other cases is inferior to the SGAT-Net. For disease GC, the SGAT-Net scores the highest, with the same accuracy (99.06%) and IOU (79.63%) with TransU-Net, and with the same specificity (99.44%) with Deeplab V3+. It also achieves comparative performance with TransU-Net in terms of F1 (87.31%) and the second-highest AUC 93.08%. For normal eye cases, SGAT-Net achieves remarkable performance, with the highest accuracy (98.55%), IOU (83.93%), and F1 (91.16%), and competitive recall and specificity of about 0.2% loss. Besides, the AUC of SGAT-Net is slightly lower than Deeplab V3+ (less than 0.2%).

In summary, the SGAT-Net outperforms similar models and makes a remarkable improvement in areas of common mistakes. Besides, its superiority in performance is reflected in the quantitative results such as the IOU and accuracy.

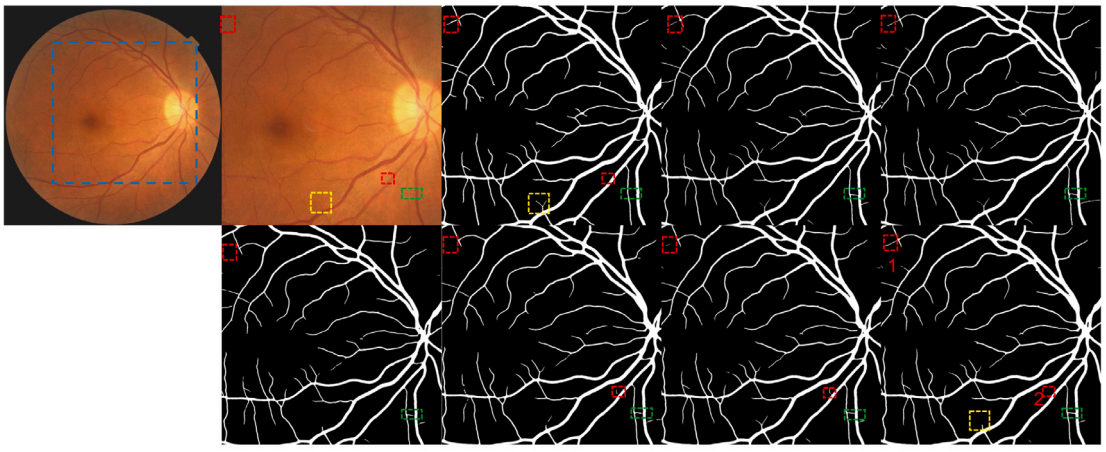
4.5. Retinal vessel segmentation in further selected dataset

On the other three public datasets, namely, DRIVE, STARE and CHASEDB1, the SGAT-Net is also compared with the traditional structure U-Net and R2U-Net, deformable U-Net (DU-Net) (Jin et al., 2019), bottom–top and top–bottom short connections in deeply supervised network (BTS-DSN) (Guo et al., 2019), dense dilated network (DD-Net) (Mou et al., 2019), multi-scale segmentation network (MS-Net) (Xia et al., 2021), hybrid deep segmentation network (HD-Net) (Yang et al., 2021), scale and context sensitive network (SCS-Net) (Wu et al., 2021), curvilinear structure segmentation network (CS2-Net) (Mou et al., 2021), feedback attention network (FA-Net) (Tomar et al., 2022), context-involved U-Net (Bridge-Net) (Zhang et al., 2022), pool-less residual segmentation network (PLRS-Net) (Arsalan et al., 2022), multi-scale context-aware network (CA-Net) (Wang et al., 2022), data-driven deep supervision (DDS) (Mishra et al., 2022), dilated convolutions U-Net (DilU-Net) (Hussain et al., 2022), and cascaded residual attention U-Net (CRAU-Net) (Dong et al., 2022).

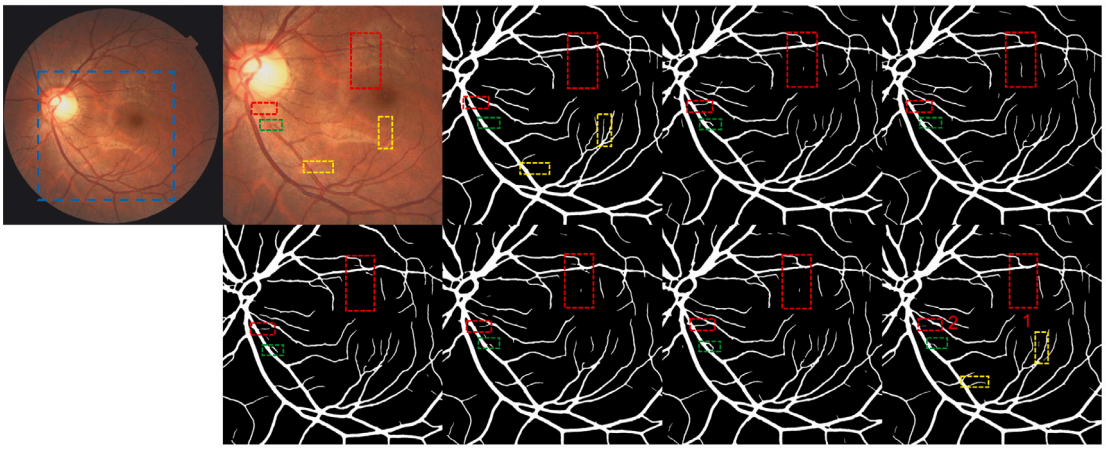
The quantitative evaluation results summarized in Tables 5 and 6 give an insight into the method's performance, generalizability, and robustness. Notably, the rows of SGAT-Net (RGB) and SGAT-Net (CLAHED) in Tables 5 and 6 represent experiments using the proposed method on RGB images and CLAHED images, respectively. Besides, the AUC of the methods is assessed on the DRIVE dataset because there is



(a) Disease AMD



(b) Disease DR



(c) Disease GC

Fig. 8. Some visual examples of the segmentation results of SGAT-Net on three diseases: (a) AMD, (b) DR, and (c) GC. In each disease's group, from left to right, the first row to the second row, it depicts the RGB image, the ground truth, the predicted segmentation masks of U-Net, Attention-Unet, Deeplab V3+, SK, TransU-Net, and SGAT-Net, respectively. For the stark comparison, the red, yellow, and green dash bounding boxes represent the improved segmentation by SGAT-Net, the existing shortcomings of it, and the possible annotations that are missed in the ground truth.

no official split between the training and test sets in the STARE and CHASEDB1 datasets.

For quantitative results, the SGAT-Net achieves the greatest overall performance and is the most robust model, with only a few exception cases where it scores lower than another method with a slim margin.

In the DRIVE dataset, the SGAT-Net obtains the highest F1 (83.32%) which is about 0.40% higher than the second-highest model HD-Net, the 86.32% recall outweighs the other models except for DDS. In other words, the SGAT-Net keeps the balance between precision and recall and is more stable. However, the performance in specificity and accuracy are similar to some models.

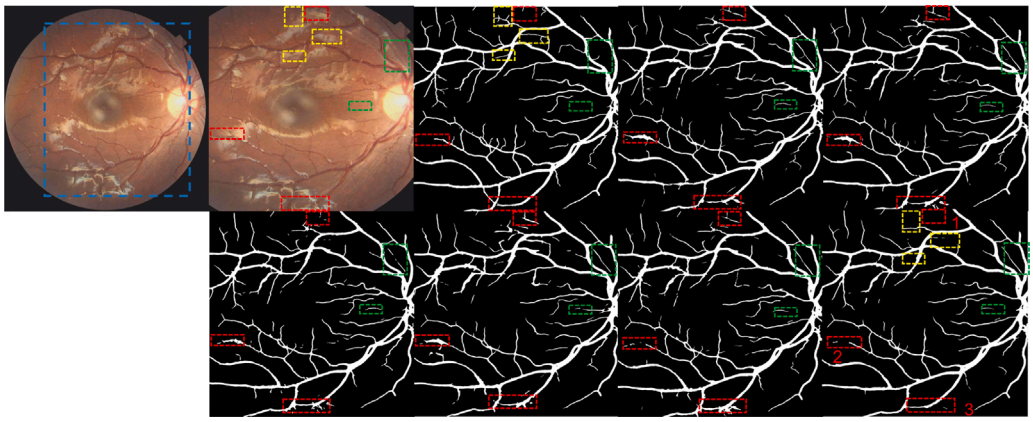


Fig. 9. Some visual examples of the segmentation results of SGAT-Net on the normal eyes. The display order and the legend of each image are the same as that of Fig. 8.

Table 5
The deep learning methods on the DRIVE dataset.

Dataset	Approach	F1 (%)	RC (%)	SP (%)	ACC (%)	AUC (%)
DRIVE	U-Net	81.42	75.37	98.20	95.31	97.55
	R2U-Net	81.71	77.92	98.13	95.56	97.84
	DU-Net	82.37	79.63	98.00	95.66	98.02
	BTS-DSN	82.08	78.00	98.06	95.51	97.96
	DD-Net	–	81.26	97.88	95.94	97.96
	MS-Net	82.10	81.20	98.00	95.40	–
	HD-Net	82.97	83.53	97.51	95.79	–
	SCS-Net	81.89	82.89	98.38	**96.97**	98.37
	CS2-Net	–	82.18	**98.90**	96.32	98.25
	FA-Net	81.83	82.15	98.26	–	–
	Bridge-Net	82.03	78.53	98.18	95.65	98.34
	PLRS-Net	–	82.69	98.17	96.82	98.35
	CA-Net	82.54	79.34	98.12	95.61	–
	DDS	–	**89.50**	96.30	95.68	98.14
	SGAT-Net (RGB)	83.09	87.26	97.62	96.62	90.35
	SGAT-Net (CLAHED)	**83.32**	86.32	97.74	96.62	91.05

Table 6
The deep learning methods on the two public dataset.

Dataset	Approach	F1 (%)	RC (%)	SP (%)	ACC (%)
STARE	U-Net	83.73	82.70	98.42	96.90
	R2U-Net	84.75	82.98	98.62	97.12
	DU-Net	81.43	75.95	**98.78**	96.41
	BTS-DSN	83.62	82.01	98.28	96.60
	DD-Net	–	83.91	97.69	96.85
	HD-Net	81.55	79.46	98.21	96.26
	SCS-Net	–	82.07	98.39	97.36
	CS2-Net	–	88.16	98.40	97.52
	Bridge-Net	82.89	80.02	98.64	96.68
	PLRS-Net	–	86.35	98.03	97.15
	DilU-Net	–	82.63	**98.78**	96.94
	SGAT-Net (RGB)	**85.12**	**89.28**	98.32	**97.64**
	SGAT-Net (CLAHED)	84.25	87.30	98.35	97.49
CHASEDB1	U-Net	77.83	82.88	97.01	95.78
	R2U-Net	79.28	77.56	98.20	96.34
	DU-Net	78.83	81.55	97.52	96.10
	BTS-DSN	79.83	78.88	98.01	96.27
	DD-Net	–	82.68	97.73	96.37
	MS-Net	81.90	82.00	98.30	97.00
	HD-Net	79.97	81.76	97.76	96.32
	SCS-Net	–	83.65	98.39	97.44
	PLRS-Net	–	83.01	98.39	97.31
	CRAU-Net	81.56	82.59	–	96.59
	FA-Net	81.08	85.44	98.30	–
	Bridge-Net	82.93	81.32	98.40	96.67
	DDS	–	**93.55**	96.45	96.25
	SGAT-Net (RGB)	**84.91**	87.01	**98.63**	**97.82**
	SGAT-Net (CLAHED)	83.98	88.00	98.42	97.72

Meanwhile, compared with the previous methods which aim at low-resolution and low-precision fundus images, the proposed method specializes in dealing with high-resolution images that are more suitable for clinical application, because they provide more details of retinal blood vessels and more relevant information for diagnosis. According to Table 1, the images in the DRIVES dataset are only one-fourth of resolution compared with that of the FIVES dataset, so it is necessary to upsample the original images to fit it with the proposed method during evaluation. However, upsampling leads to blurred regions of interest as well as the relatively large AUC gap between the proposed methods and the previous methods, as shown in Table 5.

On the contrary, the SGAT-Net accomplishes the best results in at least two metrics in the STARE and CHASEDB1 datasets. On STARE, SGAT-Net scores slightly lower (less than 0.50%) in specificity, while outperforming other methods in F1 (85.12%), recall (89.28%) and accuracy (97.64%). On CHASEDB1, the SGAT-Net outperforms models in F1 (84.91%) though its recall (87.01%) is a bit lower than DDS. Besides, the SGAT-Net gains a 0.23% advantage in specificity over than Bridge-Net and 0.38% benefits in accuracy over than SCS-Net.

For qualitative visualization, some samples of the segmentation results of the SGAT-Net are depicted in Fig. 10. The SGAT-Net is more sensitive to the thick blood vessels with descriptive features such as colour and preserves the anatomy structure, whereas being less able to segment the thin and blurred vascular with few pixels in area.

5. Discussion

Generally, the SGAT-Net achieves a better performance than the basic structure (U-Net), the attention gate in U-Net (Attention-Unet), the multi-scale feature extraction (Deeplab V3+), the adaptive kernel (SK), the fully transformer (Swin-Unet) and the combination of CNN and Transformer (TransU-Net) in the FIVES dataset. Besides, it also outperforms the existing models in the further common dataset (i.e. STARE and CHASEDB1 dataset) in most of the factors.

The SGA-Module overcomes the inherent limitations of both CNN and transformer because of the local–global compound features that are attributed to the inductive bias as well as the self-attention mechanism. For the disease AMD, the predicted segmentation of retinal vessels by SGAT-Net not heavily relies on distinctive features like colours but makes a reasonable inference in accordance with the blood vessel structure, e.g., it could distinguish the difference between the blood vessels and lesions in the second red bounding box in Fig. 8(a). It turns out that the global information provided by the transformer has a significant positive influence on retinal vessel segmentation.

Besides, the MHA-SGAP in the transformer adaptively determines the pooling approach rather than the direct use of ViT, obtaining better performance, especially for the vessels in blurry regions. For

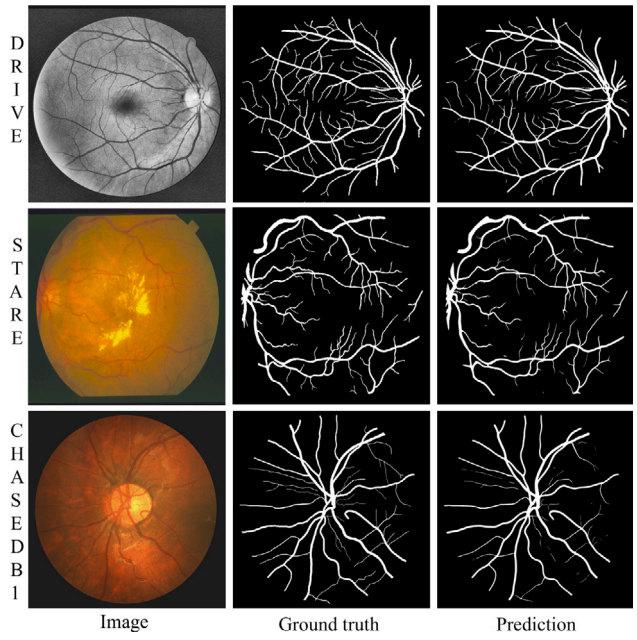


Fig. 10. Samples of segmentation masks by the proposed method in the three classical datasets (DRIVE, STARE, and CHASEDB1). The images are CLAHED images, RGB images, and RGB images for DRIVE, STARE, and CHASEDB1 respectively.

example, in the third and fourth red bounding boxes in Fig. 8(a), SGAT-Net produces a clear delineation of the blood vessel circle in disease AMD whereas TransU-Net depicts a filled circle. This is attributed to the SGAP operation in the self-attention mechanism that extracts the discriminative contextual embeddings and filters out the noise, leading to the strong feature representation ability.

Moreover, the SGAFF adaptively emphasizes on the importance of CNN and transformer to generate a proper RF, thus avoiding making the common mistakes by the existing method. For instance, for normal eyes, the SGAT-Net makes the most similar predictions with ground truth whereas the other evaluated methods misidentify the illumination as the retinal blood vessel in the third red bounding box in Fig. 9.

In addition, since the SGAT-Net is able to extract the local details and establish a long-range dependency simultaneously, and adaptive merge them based on tasks, the segmentation results by the SGAT-Net are more accurate and reasonable than the original annotation. For example, there are comparatively light and thin retinal vessels segmented by the SGAT-Net whereas the annotators ignore it or regard it as background as shown in green bounding box in Fig. 8(b).

However, we recognize that there are some limitations of the SGAT-Net leading to errors in the experiment. On the one hand, the method is unable to the serious class imbalance problem. For example, the SGAT-Net cannot distinguish the smaller retinal blood vessel from the background. On the other hand, the SGAT-Net is more sensitive to the minor change of light that human is unable to perceive, causing the missing object in the yellow bounding box in Fig. 9 where some other approaches perform better.

6. Conclusion

Retinal fundus images provide an efficient and non-invasive method for the early detection of prevalent ocular diseases. However, there still exist some challenges that impede clinical analysis. To address these challenges, an automated retinal vessel segmentation framework is proposed in this paper to assist ophthalmologist diagnosis. The proposed SGA-Module produces the local-and-global compound features based on inductive bias and self-attention mechanism to take into account the vascular details and anatomy structures. While ResEncoder provides the local details of retinal blood vessels, the SGAP-Former adaptively arranges the weighting of maximum and average pooling to refine the contextual embedding representation while filtering out redundant information. Also, the SGAFF module adaptively stresses the CNN-based features and transformer-based embedding, and aggregates them in the latent space to generate the appropriate RF based on the task. The qualitative and quantitative evaluation in the largest fundus image dataset (FIVES) and three classical retinal image datasets (DRIVE, STARE, and CHASEDB1) demonstrates outstanding performance over the other existing method, even in confusing areas where other methods incline to make mistakes. Without losing generalizability, the proposed methods can be easily adapted to address other medical image segmentation challenges where a high variety of appearances and anatomy is present.

CRediT authorship contribution statement

Ji Lin: Conceptualization, Formal analysis, Methodology, Software, Investigation, Writing – original draft. **Xingru Huang:** Software, Validation. **Huiyu Zhou:** Conceptualization, Writing – review & editing, Formal analysis. **Yaqi Wang:** Formal analysis, Funding acquisition, Writing – review & editing. **Qianni Zhang:** Conceptualization, Formal analysis, Supervision, Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This research was partially supported by the Royal Society International Exchanges Fund (IEC\NSFC\211269). Besides, the hardware was partially supported by NVIDIA Academic Grant.

References

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

Arsalan, M., Haider, A., Lee, Y.W., Park, K.R., 2022. Detecting retinal vasculature as a key biomarker for deep learning-based intelligent screening and analysis of diabetic and hypertensive retinopathy. Expert Syst. Appl. 200, 117009.

Ayhan, M.S., Kühlewein, L., Aliyeva, G., Inhoffen, W., Ziemssen, F., Berens, P., 2020. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. Med. Image Anal. 64, 101724.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Cai, B., Ma, L., 2022. A transformer-based cascade network with boundary enhancement loss for retinal vessel segmentation. In: 2022 26th International Conference on Pattern Recognition. (ICPR), IEEE, pp. 4292–4298.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2021. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537.

Chang, Y., Menghan, H., Guangtao, Z., Xiao-Ping, Z., 2021. Transclaw u-net: Claw u-net with transformers for medical image segmentation. arXiv preprint arXiv:2107.05188.

Chen, B., Liu, Y., Zhang, Z., Lu, G., Zhang, D., 2021a. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. arXiv preprint arXiv:2107.05274.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021b. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. (ECCV), pp. 801–818.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1251–1258.

Deng, K., Meng, Y., Gao, D., Bridge, J., Shen, Y., Lip, G., Zhao, Y., Zheng, Y., 2021. Transbridge: A lightweight transformer for left ventricle segmentation in echocardiography. In: International Workshop on Advances in Simplifying Medical Ultrasound. Springer, pp. 63–72.

Dong, F., Wu, D., Guo, C., Zhang, S., Yang, B., Gong, X., 2022. CRAUNet: A cascaded residual attention U-net for retinal vessel segmentation. Comput. Biol. Med. 105651.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Fraz, M.M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A.R., Owen, C.G., Barman, S.A., 2012. An ensemble classification-based approach applied to retinal blood vessel segmentation. IEEE Trans. Biomed. Eng. 59 (9), 2538–2548.

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

He, X., Tan, E.L., Bi, H., Zhang, X., Zhao, S., Lei, B., 2022. Fully transformer network for skin lesion analysis. Med. Image Anal. 77, 102357.

Hoover, A., Jean-Baptiste, G., Goldgof, D., Bowyer, K.W., 1994. A methodology for evaluating range image segmentation techniques. In: Proceedings of 1994 IEEE Workshop on Applications of Computer Vision. IEEE, pp. 264–271.

Hoover, A., Kouznetsova, V., Goldbaum, M., 2000. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Trans. Med. Imaging 19 (3), 203–210.

Huang, S., Li, J., Xiao, Y., Shen, N., Xu, T., 2022. RTNet: relation transformer network for diabetic retinopathy multi-lesion segmentation. IEEE Trans. Med. Imaging 41 (6), 1596–1607.

Hubel, D.H., Wiesel, T.N., 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. 160 (1), 106.

Hussain, S., Guo, F., Li, W., Shen, Z., 2022. DilUnet: A U-net based architecture for blood vessels segmentation. Comput. Methods Programs Biomed. 218, 106732.

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. Adv. Neural Inf. Process. Syst. 28.

Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., Ye, J., 2022. Fives: A fundus image dataset for artificial intelligence based vessel segmentation. Sci. Data 9 (1), 1–8.

Jin, Q., Meng, Z., Pham, T.D., Chen, Q., Wei, L., Su, R., 2019. DUNet: A deformable network for retinal vessel segmentation. Knowl.-Based Syst. 178, 149–162.

Kaur, J., Mittal, D., Singla, R., 2021. Diabetic retinopathy diagnosis through computer-aided fundus image analysis: A review. Arch. Comput. Methods Eng. 1–39.

Li, Y., Cai, W., Gao, Y., Hu, X., 2021a. More than encoder: Introducing transformer decoder to upsample. arXiv preprint arXiv:2106.10637.

Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., Goh, R., 2021b. Medical image segmentation using squeeze-and-expansion transformers. arXiv preprint arXiv:2105.09511.

Li, X., Wang, W., Hu, X., Yang, J., 2019. Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 510–519.

Li, Y., Wang, S., Wang, J., Zeng, G., Liu, W., Zhang, Q., Jin, Q., Wang, Y., 2021c. Gt u-net: A u-net like group transformer network for tooth root segmentation. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 386–395.

Li, Y., Yang, J., Ni, J., Elazab, A., Wu, J., 2021d. TA-net: triple attention network for medical image segmentation. Comput. Biol. Med. 137, 104836.

Li, Y., Zhang, Y., Liu, J.Y., Wang, K., Zhang, K., Zhang, G.S., Liao, X.F., Yang, G., 2022. Global transformer and dual local attention network via deep-shallow hierarchical feature fusion for retinal vessel segmentation. IEEE Trans. Cybern..

Mishra, S., Zhang, Y., Chen, D.Z., Hu, X.S., 2022. Data-driven deep supervision for medical image segmentation. IEEE Trans. Med. Imaging.

Mitani, A., Huang, A., Venugopalan, S., Corrado, G.S., Peng, L., Webster, D.R., Hammel, N., Liu, Y., Varadarajan, A.V., 2020. Detection of anaemia from retinal fundus images via deep learning. Nat. Biomed. Eng. 4 (1), 18–27.

Mou, L., Chen, L., Cheng, J., Gu, Z., Zhao, Y., Liu, J., 2019. Dense dilated network with probability regularized walk for vessel detection. IEEE Trans. Med. Imaging 39 (5), 1392–1403.

Mou, L., Zhao, Y., Fu, H., Liu, Y., Cheng, J., Zheng, Y., Su, P., Yang, J., Chen, L., Frangi, A.F., et al., 2021. CS2-net: Deep learning segmentation of curvilinear structures in medical imaging. Med. Image Anal. 67, 101874.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Pachade, S., Porwal, P., Kokare, M., Giancardo, L., Mériaudeau, F., 2021. NENet: Nested EfficientNet and adversarial learning for joint optic disc and cup segmentation. Med. Image Anal. 74, 102253.

Paszke, A., Chaurasia, A., Kim, S., Culurciello, E., 2016. Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147.

Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J., 2017. Large kernel matters–improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4353–4361.

Penno, G., Solini, A., Zoppini, G., Orsi, E., Zerbini, G., Trevisan, R., Gruden, G., Cavalot, F., Laviola, L., Morano, S., et al., 2012. Rate and determinants of association between advanced retinopathy and chronic kidney disease in patients with type 2 diabetes: the renal insufficiency and cardiovascular events (RIACE) Italian multicenter study. Diabetes care 35 (11), 2317–2323.

Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K., 1987. Adaptive histogram equalization and its variations. Comput. Vis., Graph. Image Process. 39 (3), 355–368.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Shen, X., Xu, J., Jia, H., Fan, P., Dong, F., Yu, B., Ren, S., 2022. Self-attentional microvessel segmentation via squeeze-excitation transformer unet. Comput. Med. Imaging Graph. 97, 102055.

Sivaprasad, S., Gupta, B., Crosby-Nwaobi, R., Evans, J., 2012. Prevalence of diabetic retinopathy in various ethnic groups: A worldwide perspective. Surv. Ophthalmol. 57 (4), 347–370.

Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B., 2004. Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging 23 (4), 501–509.

Szeskin, A., Yehuda, R., Shmueli, O., Levy, J., Joskowicz, L., 2021. A column-based deep learning method for the detection and quantification of atrophy associated with AMD in OCT scans. Med. Image Anal. 72, 102130.

Tham, Y.C., Li, X., Wong, T.Y., Quigley, H.A., Aung, T., Cheng, C.Y., 2014. Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis. Ophthalmology 121 (11), 2081–2090.

Thylefors, B., Negrel, A., 1994. The global impact of glaucoma. Bull. World Health Organ. 72 (3), 323.

Tomar, N.K., Jha, D., Riegler, M.A., Johansen, H.D., Johansen, D., Rittscher, J., Halvorsen, P., Ali, S., 2022. Fanet: A feedback attention network for improved biomedical image segmentation. IEEE Trans. Neural Netw. Learn. Syst..

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, X., Li, Z., Huang, Y., Jiao, Y., 2022. Multimodal medical image segmentation using multi-scale context-aware network. Neurocomputing 486, 135–146.

Wong, W.L., Su, X., Li, X., Cheung, C.M.G., Klein, R., Cheng, C.Y., Wong, T.Y., 2014. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: A systematic review and meta-analysis. Lancet Global Health 2 (2), e106–e116.

Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. (ECCV), pp. 3–19.

Wu, H., Wang, W., Zhong, J., Lei, B., Wen, Z., Qin, J., 2021. Scs-net: A scale and context sensitive network for retinal vessel segmentation. Med. Image Anal. 70, 102025.

Xia, H., Lan, Y., Song, S., Li, H., 2021. A multi-scale segmentation-to-classification network for tiny microaneurysm detection in fundus images. Knowl.-Based Syst. 226, 107140.

Xiao, W., Huang, X., Wang, J.H., Lin, D.R., Zhu, Y., Chen, C., Yang, Y.H., Xiao, J., Zhao, L.Q., Li, J.P.O., et al., 2021. Screening and identifying hepatobiliary diseases through deep learning using ocular images: A prospective, multicentre study. Lancet Digital Health 3 (2), e88–e97.

Yang, L., Wang, H., Zeng, Q., Liu, Y., Bian, G., 2021. A hybrid deep segmentation network for fundus vessels via deep-learning framework. Neurocomputing 448, 168–178.

Yu, H., Shim, J.-h., Kwak, J., Song, J.W., Kang, S.-J., 2022. Vision transformer-based retina vessel segmentation with deep adaptive Gamma correction. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. (ICASSP), IEEE, pp. 1456–1460.

Zhang, Y., He, M., Chen, Z., Hu, K., Li, X., Gao, X., 2022. Bridge-net: Context-involved U-net with patch-based loss weight mapping for retinal blood vessel segmentation. Expert Syst. Appl. 195, 116526.

Zhang, H., Ni, W., Luo, Y., Feng, Y., Song, R., Wang, X., 2023. Tunet-LBF: Retinal fundus image fine segmentation model based on transformer unet network and LBF. Comput. Biol. Med. 106937.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2881–2890.