Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research Article

LiViT-Net: A U-Net-like, lightweight Transformer network for retinal vessel segmentation

Le Tong [a], Tianjiu Li [a], Qian Zhang [a,*], Qin Zhang [b,*], Renchaoli Zhu [a], Wei Du [c], Pengwei Hu [d]

[a] The College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, No. 100 Haisi Road, Shanghai, 201418, China
[b] Ophthalmology Department, Jing'an District Central Hospital, No. 259, Xikang Road, Shanghai, 200040, China
[c] Laboratory of Smart Manufacturing in Energy Chemical Process, East China University of Science and Technology, No. 130 Meilong Road, Shanghai, 200237, China
[d] The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, 40-1 South Beijing Road, Urumqi, 830011, China

ARTICLE INFO

Keywords:
Retinal vessel segmentation
Transformer
Lightweight
Joint loss

ABSTRACT

The intricate task of precisely segmenting retinal vessels from images, which is critical for diagnosing various eye diseases, presents significant challenges for models due to factors such as scale variation, complex anatomical patterns, low contrast, and limitations in training data. Building on these challenges, we offer novel contributions spanning model architecture, loss function design, robustness, and real-time efficacy. To comprehensively address these challenges, a new U-Net-like, lightweight Transformer network for retinal vessel segmentation is presented. By integrating MobileViT+ and a novel local representation in the encoder, our design emphasizes lightweight processing while capturing intricate image structures, enhancing vessel edge precision. A novel joint loss is designed, leveraging the characteristics of weighted cross-entropy and Dice loss to effectively guide the model through the task's challenges, such as foreground-background imbalance and intricate vascular structures. Exhaustive experiments were performed on three prominent retinal image databases. The results underscore the robustness and generalizability of the proposed LiViT-Net, which outperforms other methods in complex scenarios, especially in intricate environments with fine vessels or vessel edges. Importantly, optimized for efficiency, LiViT-Net excels on devices with constrained computational power, as evidenced by its fast performance. To demonstrate the model proposed in this study, a freely accessible and interactive website was established (https://hz-t3.matpool.com:28765?token=aQjYR4hqMI), revealing real-time performance with no login requirements.

1. Introduction

Retinal imaging techniques, which enable noninvasive observation of the deep microvasculature in the human body, play a pivotal role in enabling doctors to intuitively detect early-stage changes in retinal vascular structures caused by diseases like diabetes and glaucoma. The urgent need for real-time analysis in scenarios such as emergency diagnostics, especially in remote areas with limited computational resources, underscores the necessity for efficient and lightweight models. This importance of retinal vascular analysis for detecting and diagnosing retinal diseases has spurred researchers to intensively explore ways to enhance the performance of segmentation algorithms, with a particular emphasis on models that can operate effectively in resource-constrained environments. However, due to challenges such as thin vessels, lesions, complex vascular structures, and low contrast in fundus images, achieving accurate segmentation remains a formidable task [1]. Lately, there has been a marked focus on the detection and segmentation of arteries and veins within retinal vessel images [2,3], where lightweight, efficient models can be particularly transformative.

In addition, the emergence of Vision Transformer (ViT) models in recent fields of computer vision, including object detection, image classification, and semantic segmentation. Unlike their convolutional neural network (CNN)-based predecessors, ViT models not only offer robust global context modeling but also display exceptional adaptability when extensively pretrained on downstream tasks. Consequently, several studies have leveraged Transformer blocks as the network's backbone for medical image analysis [4–6]. For instance, SwinUNet [7] employs hierarchical Transformer blocks to construct the encoder and decoder components of a U-Net-like architecture. Similarly, DS-TransUNet [5] introduces an encoder that accommodates inputs of

* Corresponding authors.
E-mail addresses: qianzhang@shnu.edu.cn (Q. Zhang), zqandzxf@163.com (Q. Zhang).

https://doi.org/10.1016/j.csbj.2024.03.003
Received 23 November 2023; Received in revised form 22 February 2024; Accepted 4 March 2024
Available online 19 March 2024

varying sizes. However, these models often falter when tasked with small datasets, such as those used in medical imaging. The considerable number of parameters in these models not only diminishes their efficacy but also accentuates their dependence on high-performance hardware, impeding practical algorithm implementation.

The success of fully convolutional networks (FCNs) has resulted in numerous FCN-based models, such as U-Net [8] and DeepLab [9] series, which are excellent establishing state-of-the-art models for semantic segmentation tasks. U-Net has emerged as the most widely utilized model in medical Several studies have integrated CNNs with ViTs to mitigate the limitations of ViT-based models [10–12]. Chen et al. proposed TransUNet [13], a framework that capitalizes on ViT's potential in medical image segmentation. The architecture of TransUNet comprises a convolutional component for feature extraction and a Transformer component to assist in encoding global context information. Zhang et al. introduced TransFuse, a parallel-in-branch architecture that combines Transformers and CNNs for efficient global and low-level spatial detail modeling in medical image segmentation; this architecture was further enhanced by a novel BiFusion module for effective feature fusion. However, those methods are not sufficiently effective at interweaving long-term dependencies with convolutional representations containing precise spatial information. Consequently, the advantages of combining both CNNs and ViT remain suboptimal.

In medical image analysis tasks, to minimize the overfitting problem caused by scarce samples, designing suitable loss functions to guide the learning process of models is crucial. Several loss functions, including cross-entropy, Dice loss, and focal loss, are often employed in this specialized domain of medical imaging. The cross-entropy loss is straightforward to understand and implement. Consequently, several methods for retinal vessel segmentation utilize binary cross-entropy to optimize the network [7,14,26]. Unfortunately, this approach tends to exhibit a bias toward the dominant class in imbalanced datasets. While Dice loss can effectively handle class imbalances, its noncontinuous and nonconvex nature makes it challenging to optimize [15]. In contrast, focal loss, an enhanced version of the cross-entropy loss, also addresses class imbalances effectively. Authors in [16–18,29] have employed Dice and focal loss functions, which prevent the network from inaccurately favoring well-performing samples (such as background objects) over samples of interest (such as vessels, especially micro-vessels). However, this kind of loss requires careful tuning due to the presence of an additional hyperparameter. It is essential to explore effective methods for integrating multiple loss functions to harness their respective advantages and make appropriate adjustments tailored to the tasks at hand.

In retinal image analysis, accurate, real-time, and generalizable models are essential for early ocular diseases diagnosis. Current models encounter challenges in vessel segmentation due to data dependencies and computational demands. This study addresses these key concerns:

- Establishing an efficient integration of the strengths of CNNs and Transformers to enhance feature extraction from retinal images, especially in accurately identifying minute vessels and vessel boundaries and improving computational efficiency.
- Devising a new loss function to guide the model more effectively in learning the intricate parts of retinal images, and its performance is enhanced on datasets with limited data.
- Exploring the model's generalizability across diverse datasets and evaluating its real-time performance in resource-constrained settings such as mobile devices or embedded systems are pivotal, especially in scenarios demanding instant analysis.

By addressing these objectives, our research contributes to retinal imaging techniques by developing a more efficient and effective segmentation algorithm. The remainder of the paper is organized as follows. Section 2 provides a review of related work in retinal image segmentation. Section 3 introduces the proposed LiViT-Net method, detailing the hybrid network architecture and the novel loss function. Section 4 describes the experimental setup, including dataset preparation, evaluation metrics, and implementation details. Section 5 presents and discusses the results, including comparisons with state-of-the-art methods. Finally, Section 6 concludes the paper.

2. Related work

In this section, deep learning-based methods for fundus vascular segmentation tasks and methods based on Transformer for medical image segmentation are reviewed.

2.1. Deep learning for retinal vessel segmentation

Deep learning methods can automatically learn image features from large quantities of data without manual intervention. Furthermore, they have become the leading methods in this domain. Gu et al. [19] developed the Context Encoder Network (CE-Net) to enhance high-level information capture and spatial information preservation for medical image segmentation. Jiang et al. [20] introduced a down-sampling coefficient and Joint Expansion Convolution to reduce information loss and address the "grid problem" in expansion convolution. Feng et al. [21] designed a cross-stitching network to integrate multi-scale features and increase approach robustness by eliminating preprocessing hyperparameters. Kamran et al. [22] proposed the multi-scale RV-Gan for improved retinal microvessel structure extraction. Wang et al. [23] devised a dual-channel encoder structure network with a feature fusion module, attention jump module, and Structure Loss for enhanced microvessel segmentation and boundary detection. Zhang et al. [24] presented Bridge-net, which combines recurrent neural networks (RNNs) with CNNs, and introduced a patch-based loss weight mapping to correct vessel morphology-related imbalances. The methodologies presented in that study [25] utilized Dense U-Nets to optimize the learning process. By mitigating the creation of redundant activation maps and conserving intricate information, these algorithms have accomplished improved predictions, while maintaining a minimum parameter usage and computational expense. These methods suffer from loss of spatial information, inability to use global information, and morphological differences between thick and thin vessels. Li et al. [26] developed a U-Net-based model with an attention module to capture global information and enhance features. Wang et al. [27] proposed HAnet, an end-to-end deep learning architecture with three decoder networks and an attention mechanism for retinal vessel segmentation. Similarly, the study [28] adopted a Weighted Attention Gate strategy in which extraneous background features were discarded to refine the segmentation process further. Several methods based on CNN were proposed for the task, such as Scs-Net [29], NFN+ [30], Sa-Nnet [31]. Several methods have contributed to optimizing model efficiency, such as Cc-net [32], which, by proposing a network compression strategy based on image complexity, significantly optimizes the model's parameter count and computational load, making an important contribution to the lightweight design of deep learning models. This demonstrates the utility of sophisticated techniques for enhancing the efficiency and effectiveness of retinal vessel segmentation. Researchers have proposed methods, such as attention modules and patch-based loss weight mapping, explicitly to address issues such as the loss of spatial information, and the inability to utilize global information and morphological disparities between thick and thin vessels. However, room for the development of more effective techniques to address additional problems and challenges remains.

2.2. Methods based on Transformer for medical image segmentation

In recent years, with the excellent performance of Vision Transformer (ViT) models in various fields of computer vision, ViT models have gradually been introduced into medical image processing. Chen et

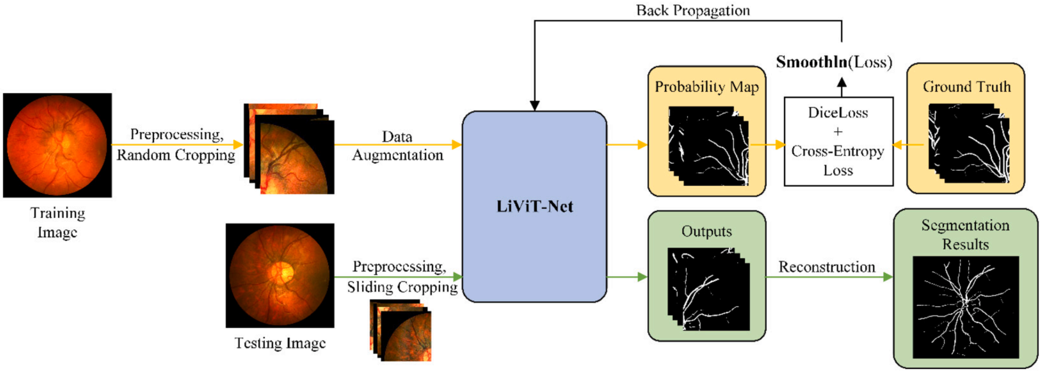


Fig. 1. Overview of the proposed method.

al. [12] introduce ViT modules into a model based on U-Net, which solved certain limitations of conventional convolution operations in modeling long-range dependencies. Cao et al. [6] proposed a U-Net structure composed purely of Transformer blocks for medical image segmentation, named Swin-UNet, which uses a layered Swin Transformer with shift windows as an encoder for extracting contextual features. DS-TransUNet [5] further added an encoder to accept inputs of different sizes. Wu et al. [33] proposed an adaptive Transformer network based on classic encoder-decoder architecture, called FAT-Net, which integrates a dual encoder consisting of CNN and Transformer branches to effectively capture long-range dependencies and global background information. However, current methods are still limited in their ability to accurately segment complex medical images, particularly when dealing with small objects or objects with irregular shapes. Additionally, these models require a large number of computational resources and may be computationally expensive.

In contrast, this paper presents a hybrid network architecture named LiViT-Net, which leverages the advantages of convolutions and Transformers. LiViT-Net utilizes a lightweight convolutional module (*inverted residual block*) to capture fine, high-resolution spatial information in shallow images prior to the *MobileViT+ block*. These convolutional modules encode precise spatial information at the pixel level, providing low-level but high-resolution features for subsequent processing. The proposed *MobileViT+ block*, which interweaves lightweight convolution and Transformer, enables distant dependencies and high-level object concepts to be fully integrated across different scales. Additionally, a parallel convolution module is incorporated to extract richer semantic information and better introduce inductive bias into the feature maps, thereby improving the model's generalizability and robustness.

The contributions of this paper are described in greater depth as follows:

- A new, U-Net-like, lightweight Transformer network is designed for retinal vessel segmentation. The *MobileViT+ block* is introduced, to amplify the model's sensitivity to vascular edges. Within this block, a local representation is employed by integrating parallel convolutions, which further enhances ViT's inductive bias and interpatch relations.
- A remapped, weighted joint loss mechanism is introduced to address the pronounced pixel imbalances and intricate vascular structures in retinal vessel segmentation. By synergizing weighted cross-entropy and Dice loss, our method emphasizes pixel-level accuracy while mitigating class disparities.
- Comprehensive tests were conducted on three renowned retinal image datasets: DRIVE, CHASEDB1, and HRF. The evaluations underscore the robustness of our proposed approach. Performance on edge devices also indicates its computational efficiency and promising potential for broader application in diverse scenarios.

3. Methodology

3.1. Preprocessing

To highlight the performance of the proposed model and to facilitate comparison with other methods, only the most commonly used preprocessing methods are applied in this paper. As shown in Fig. 1, the preprocessing step uses limited adaptive histogram equalization (CLAHE) [34] to enhance image quality. After that, random cropping is used to increase the number of images in each dataset.

3.2. LiViT-Net architecture

Overview. LiViT-Net, which evolved from the U-shaped model, consists of an encoder and a decoder. The segmentation procedure involved: (1) fundus image preprocessing and random cropping to obtain an image patch $X \in \mathbb{R}^{C\times H\times W}$, where C is the number of channels, H is the image height, and W is the image width; (2) feeding the cropped patches to LiViT-Net for prediction; and (3) reconstructing the output results. The overall flowcharts of the method and architecture are depicted in Fig. 1 and Fig. 2, respectively.

Encoder. The encoder, consists of CNN blocks and Transformer blocks. However, this model differs from traditional models, which typically integrate both components in the upper or lower layers of the network or in parallel encoding sections, a crossover arrangement of CNN blocks and Transformer blocks are employed to leverage the ability of CNNs to capture local spatial features and the capacity of Transformers to model long-range dependencies effectively.

Compared to CNNs, Transformers demonstrate a greater sensitivity to structural information in images, while CNNs are adept at capturing local details [35]. This insight informed the design of LiViT-Net, which utilizes an *inverted residual block* to extract local features from the input image before processing it with the *MobileViT+ block*. The *inverted residual block*, composed of CNNs, efficiently captures image details, while the *MobileViT+ block* excels at extracting structural information from the image. When employed in retinal vessel segmentation, these characteristics ensure enhanced edge detection of vessels and substantially mitigate instances of vessel segmentation fragmentation.

The existing Transformer-based networks mostly are too large in terms of the number of parameters and lack of inductive bias [36], increasing the difficulty of training pure Transformer networks on datasets with a small number of images. Inspired by the MobileViT [37], the *MobileViT+ block* which introduces spatial inductive bias into the Transformer blocks through unfolding and folding operations, is included in the Transformer portion of the LiViT-Net encoder. Consistent with previous scholarly endeavors, the computation method of self-attention within the *MobileViT+ block* can be described as follows:

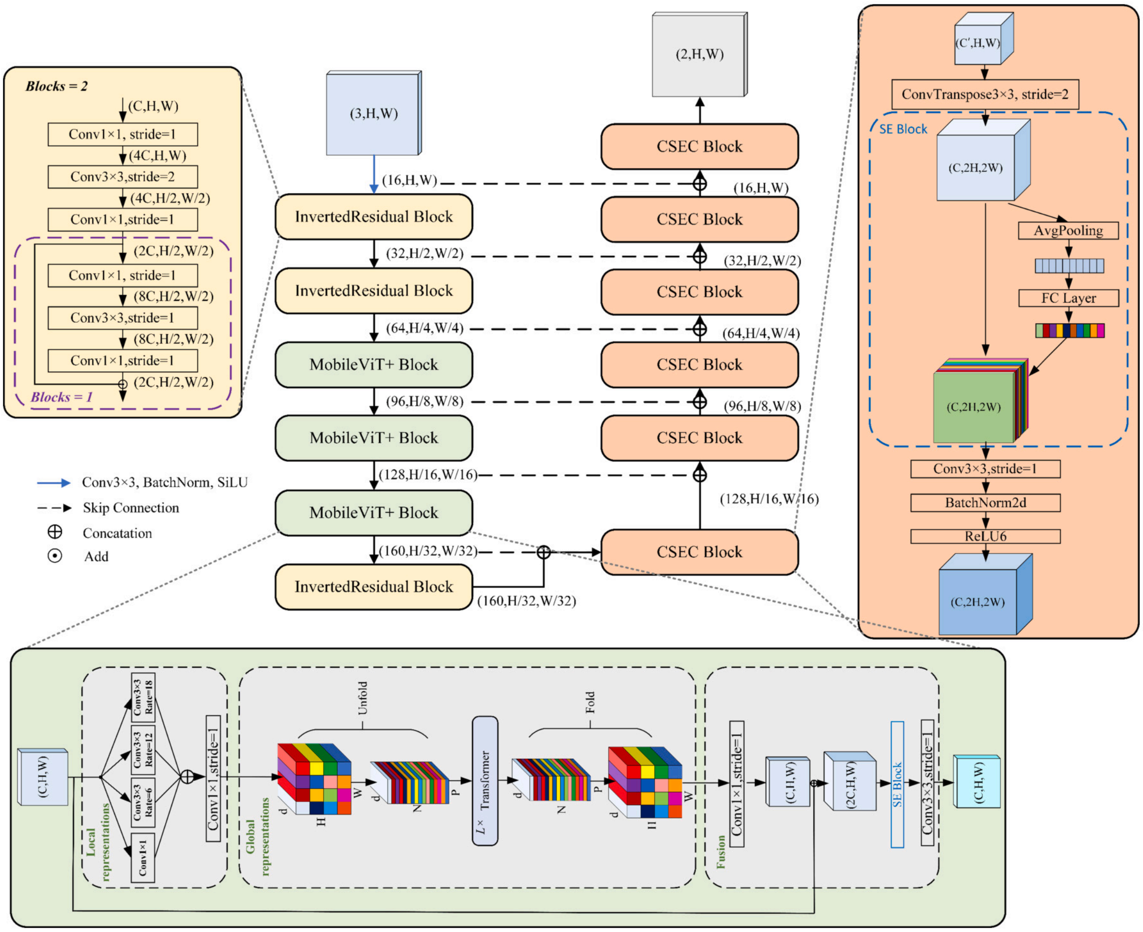


Fig. 2. Overall architecture of LiViT-Net.

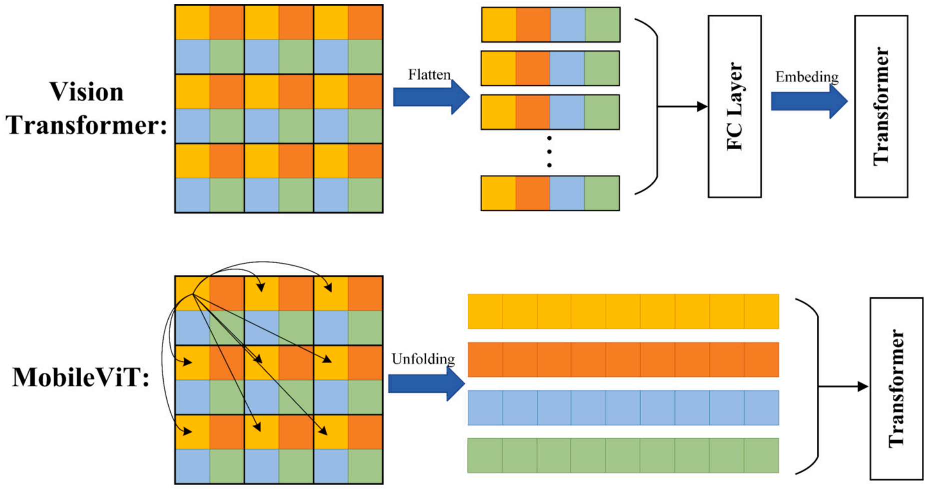


Fig. 3. Comparison between Vision Transformer and MobileViT.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{1}$$

Here, Q, K, and V denote the query, key, and value matrices, respectively. The variables $\sqrt{d}$ embody the dimensionality of the query or key. To mitigate the prospective issue of gradient vanishing in subsequent softmax procedures; the dot product results are divided by the square root of the feature dimension.

As illustrated in Fig. 3, in the *MobileViT+ block* workflow, the initial action on the input feature map is the innovative *local representation* step. This critical stage employs multiple parallel convolutions with varying kernel sizes. Specifically, *local representation* ensures detailed local modeling of the feature map while integrating essential spatial inductive bias, which is vital for capturing complex image structures. As shown in Fig. 2, the use of atrous convolutions in these parallel convolutions is a strategic choice. This technique significantly mitigates the computational burden while preserving the integral benefits of ViT, thus enhancing the structure capture capabilities of *MobileViT+ block* in a lightweight format. Mathematically, the computation of atrous convolutions is as follows:

$$X_L[i] = \bigoplus_{j=1}^{K} X[i+r_j]\cdot w_j \tag{2}$$

In this equation, $X[i]$ represents the input feature map, and $X_L[i]$ denotes the final feature map obtained by concatenating the feature maps. Specifically, for the parallel convolution operation, we have K different dilation rates denoted as $r_1, r_2, \ldots, r_K$. When calculating each element $X_L[i]$ of the merged feature map, the input feature map $X[i+r_j]$ is multiplied by the corresponding weight w_j for each dilation rate r_j. Finally, the results from all the parallel convolution operations are concatenated to obtain the final feature map $X_L[i]$. Integrating convolutions with different kernel sizes within the local representation is designed to extract a more nuanced semantic context. This enables each pixel to embody the surrounding feature information after this stage and proves beneficial in the context of the vessel segmentation task. Specifically, this approach allows for the extraction of more detailed information from images, thereby enhancing the accuracy when segmenting minute vessels. The feature map is then projected to the target dimension through a pointwise convolution step. This step compensates for the information loss that occurs during the unfolding process in the Transformer stage.

In the global representation step, LiViT-Net employs a novel approach with the MobileViT+ architecture, which is distinct from conventional Vision Transformers. Here, the input feature map $X_L \in \mathbb{R}^{d\times H\times W}$ undergoes an innovative patch division using the predefined patch size (w,h), resulting in $X_U \in \mathbb{R}^{d\times N\times P}$, where $P = wh$ represents the total number of pixels within a patch, and $N = \frac{HW}{P}$ signifies the patch count. The divided patches are then flattened and input into the Transformer for computation, as illustrated in equation (3), where p belongs to the set $\{1,\ldots,P\}$. The inter-patch relationships are encoded by leveraging Transformers, consequently yielding $X_G \in \mathbb{R}^{P\times N\times d}$. This unique patch processing, a core aspect of MobileViT+, significantly reduces the computational load. For instance, with a 2×2 patch size, complexity is reduced from $O(WHC)$ to a more manageable $O(\frac{WHC}{4})$.

$$X_G(P) = \text{Transformer}(X_U(p)), 1 \le p \le P \tag{3}$$

Furthermore, the convolution operations preceding the unfolding operations ensure that each pixel in X_L is enriched with information from adjacent pixels, as shown in Fig. 4. This means, for example, when red and blue pixels within a patch interact in the Transformer, the blue pixel has preencoded information from neighboring pixels, allowing comprehensive image encoding. Each cell in the illustrated grids corresponds to a patch and pixel, highlighting how MobileViT+ facilitates efficient information encoding across patches while maintaining lower computational demands.

As illustrated in Fig. 2, the *inverted residual block* processes the input feature map by first increasing and then decreasing its dimensions, facilitating the extraction of more information. Subsequently, a shortcut is employed to combine the *inverted residual block* with the original input feature map, generating a new, enriched feature map.

Decoder. In the decoder section of LiViT-Net, the implementation of the *CSEC block* plays a significant role, in facilitating enhanced assimilation of information from the encoder and the feature maps of the upstream layers. Such integration becomes particularly critical in the task of retinal vessel segmentation, where intricate features spanning various scales need to be seamlessly melded for accurate results.

As demonstrated in Fig. 2, this process begins with the concatenation of feature maps from the encoder and the upstream layers. A transposed Convolution is then employed to project the feature maps onto a specific dimension, followed by a squeeze-and-excitation (SE) block, colloquially known as the *SE block*, to enhance CNN models by explicitly modeling interchannel correlations. The primary basis of this block is the generation of distinctive weights, predicated upon the significance of each channel's feature map. This mechanism is mathematically defined as follows:

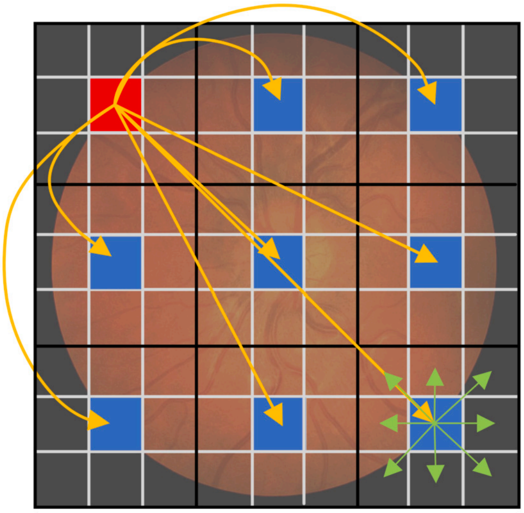


Fig. 4. After convolution, the pixels contain information from their surrounding pixels.

$$F_{\text{se}}(A) = F_{\text{scale}}(A, F_{\text{ex}}(F_{\text{sq}}(A))) \tag{4}$$

$$F_{\text{sq}}(A) = \frac{1}{HW} * \Sigma A(i,j,c) \tag{5}$$

$$s = F_{\text{ex}}(Z,W) = \sigma(g(Z,W)) = \sigma(W_2\delta(W_1 Z)) \tag{6}$$

$$\widetilde{X} = F_{\text{scale}}(A,s) = s * A \tag{7}$$

Herein, Eq. (5) (squeeze operation) condenses each channel's feature map into a scalar using global average pooling. Here, H and W denote the height and width of the feature map, respectively, while $\sum A(i,j,c)$ signifies the summation of each location and channel on the feature map. Eq. (6) (excitation operation) defines each channel's weights using a compressed-and-reconstructed fully connected layer structure. Z is the squeeze operation output, $W1$ and $W2$ are the two fully connected layer weights, σ denotes the sigmoid activation function, and δ is the ReLU activation function. Eq. (7) (the scale operation) adjusts the original feature map's channels based on the weights. Here, s represents the excitation operation output, and A is the SE block input.

The feature maps produced by the *SE block* are subsequently processed through a convolutional layer. This step extracts detailed information from the processed feature maps, providing a richer context for the minute vessels in retinal images. This enriched information is subsequently fed into the next layer, contributing to more accurate segmentation in subsequent steps.

3.3. Loss function

In medical imaging tasks, the region of interest (ROI) generally passes through a relatively small area within the image. This can cause the model to become trapped in a local minimum during training, resulting in significant prediction bias toward the background. To address this issue, the Dice coefficient is often used as a metric to assess the similarity between the model prediction and the target [38]. The calculation process of the Dice coefficient is shown in Eq. (8), where X represents the model output prediction and Y represents the target label:

$$\text{Dice coefficient} = \frac{2\cdot X\cap Y}{X+Y} \tag{8}$$

Milletari et al. [39] proposed the Dice loss to balance the relationship between the background and foreground based on the Dice coefficient. As shown in Eq. (9), where X represents the model output and Y represents the target label, the higher the Dice coefficient is, the lower the corresponding Dice loss is, indicating a greater degree of overlap between X and Y.

$$\text{Dice Loss} = 1 - \frac{2 \cdot X \cap Y}{X + Y} \tag{9}$$

The characteristic of weighted cross-entropy, which is suitable for binary classification tasks, particularly in imbalanced datasets, makes this approach highly applicable to such tasks. Here, weighted cross-entropy is referred to as the binary weighted cross-entropy loss function, which is represented as shown in formula (10), where X represents the prediction result, Y represents the target label, w_0 is the weight for the background class, and w_1 is the weight for the foreground class:

$$\begin{aligned}&\text{Weighted Cross-Entropy Loss}\\&= -\left(w_1 \cdot Y \cdot \log(X) + w_0 \cdot (1-Y) \cdot \log(1-X)\right)\end{aligned} \tag{10}$$

Joint loss, as formulated in Eq. (11), which integrates the strengths of Dice loss and weighted cross-entropy Loss are often applied to mitigate training instability with high gradients in weighted cross-entropy loss, especially when X and Y have minuscule values and the dataset is imbalanced. This combination ensures a holistic approach to segmentation cascades by balancing the contribution of each class to the loss function:

$$\text{Joint Loss} = \alpha \cdot \text{Dice Loss} + \beta \cdot \text{Weighted Cross-Entropy Loss} \tag{11}$$

Dice loss, known for its efficacy in balancing the representation of positive and negative samples, is crucial for enhancing the model's sensitivity, especially in detecting finer and sparser targets such as small blood vessels. However, its sole use might overlook the accuracy of individual pixels, which is where weighted cross-entropy loss plays a vital role. Weighted cross-entropy loss, obtained by focusing on pixel-level precision, improves the accuracy but could introduce class biases. Therefore, in the joint loss formulation, a higher weight is assigned to Dice loss (α) to ensure the model's robustness in capturing detailed features, while a smaller weight to weighted cross-entropy loss (β) sharpens pixel-level classification accuracy. The joint loss function effectively marries Dice loss and weighted cross-entropy loss, with a ratio of $\alpha/\beta = 0.8/0.2$, enhancing precision and class diversity in retinal vessel segmentation. This balanced formula improves the detection of complex structures while ensuring overall segmentation accuracy, crucial for accurate medical diagnoses.

Since fine vessels often pose a challenge in terms of accurate detection and delineation in the task of retinal vessel segmentation, we seek to remap joint loss, which is particularly crucial. The goal is to penalize regions where the Dice coefficient is small and the weighted cross-entropy loss is large, effectively improving the segmentation of such intricate areas. Conversely, we reward areas where the Dice coefficient is large and the weighted cross-entropy Loss is small, as these regions are typically easier to segment and often require less correction. This approach is inspired by the concept of sample weighting in loss functions, and we employ the Smoothln function [28] to remap the joint loss, as illustrated in Eq. (12):

$$\text{Smooth}_{ln}(x) = \begin{cases} 1 - \frac{1}{e^{2x}} & \text{if } x \le \sigma \\ (1-\sigma) \cdot (2x + \ln(1-\sigma)) + \sigma & \text{if } x > \sigma \end{cases} \tag{12}$$

Where $\sigma \in [0, 1)$ is set artificially to define linear or nonlinear reweighting. When $\sigma = 0$, it is a linear mapping, and when $\sigma \to 1$, it is a steeply smooth monotonically increasing concave curve. The variable x denotes the loss value to be remapped, derived from the joint loss in Eq. (11). More intuitively, the smooth joint loss amplifies the joint loss to varying extents, specifically penalizing samples with a low Dice coefficient and high weighted cross-entropy. This adjustment guides the learning process to concentrate more on the challenging regions of retinal images, such as the fine vessels and the edges of the blood vessels.

4. Experimental setup

4.1. Datasets

We use three public datasets—DRIVE [40], CHASEDB1 [41], and HRF [42]—for validating the LiViT-Net model's performance. These datasets, sourced from high-end medical optical instruments, offer an authentic evaluation environment for our method.

The DRIVE dataset consists of 40 fundus images captured by a Canon camera with a field-of-view (FOV) of 45° and a resolution of 584×565. The number of images for both the training and test sets is 20. For this dataset, the annotation of each fundus image was performed by two independent observers, and the annotation provided by the first observer was selected to evaluate the performance of the proposed method.

The CHASEDB1 dataset includes 14 images of 28 fundus photographs obtained from different children; these images were captured with an NM-200-D camera with a field of view (FOV) of 30° and a resolution of 999×960. For this dataset, each fundus image was annotated by two nonexpert observers. We selected the annotations of the first observer to evaluate the performance of the proposed method in this paper.

The HRF (High-Resolution Fundus) dataset consists of three types of images: healthy patients, patients with glaucoma, and diabetic retinopathy with 15 images of each type and a resolution of 3504×2336. For each image, a corresponding binary gold-standard vessel segmentation image was generated by a group of experts in retinal image analysis and the collaborating ophthalmologic clinic. In addition, masks for determining the FOV are provided for specific datasets.

4.2. Evaluation metrics

For the dense prediction task of retinal vessel segmentation, pixels considered as vessels (foreground) by the expert are defined as true positives (TPs) if it is correctly classified as vessel pixels; while the pixels erroneously detected as background are defined as false positives (FPs). For the dense prediction task of retinal vessel segmentation, pixels are considered to be background (nonvessels) by the expert are defined as true negatives (TNs) if they are correctly classified as background pixels; while the pixels erroneously detected as vessels are defined as false negatives (FNs).

To provide a more comprehensive assessment, a series of metrics are adopted, including accuracy (Acc), sensitivity (Sen), specificity (Spe), precision (Pre) and F1 score (F1). The definitions of the above metrics are as follows:

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

$$\text{Sen} = \frac{TP}{TP + FN} \tag{14}$$

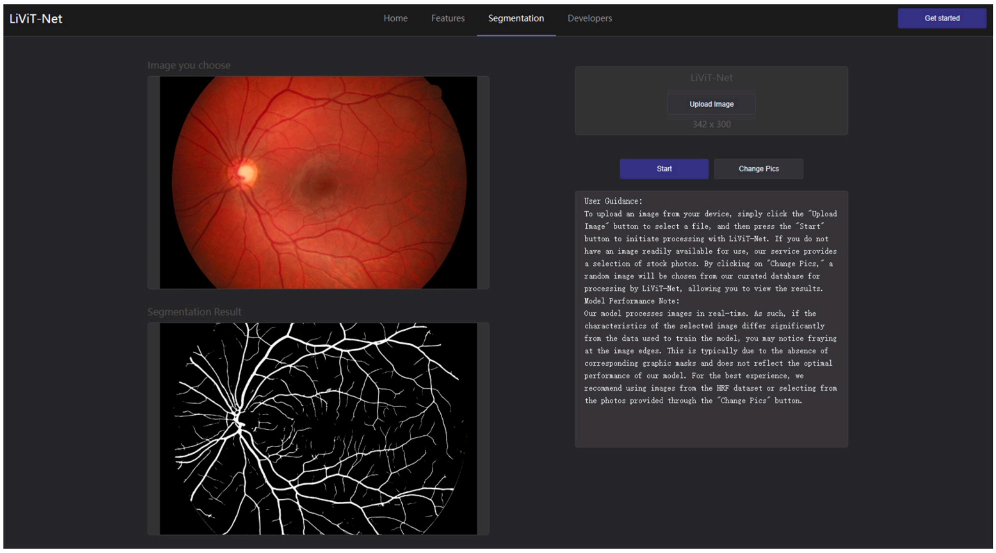
$$\text{Spe} = \frac{TN}{TN + FP} \tag{15}$$

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \tag{16}$$

In addition to these metrics, we utilize the structural similarity index measure (SSIM) because of its ability to measure perceptual quality in segmented images, which is vital in medical image analysis. The SSIM evaluates the similarity between segmented and original images in terms of structural integrity and luminance. Its formula is:

$$\text{SSIM}(x, y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{17}$$

Here, μ_x and μ_y represent average intensities, σ_x^2 and σ_y^2 are the variances of images x and y.



User Guidance:
To upload an image from your device, simply click the ″Upload Image″ button to select a file, and then press the ″Start″ button to initiate processing with LiViT-Net. If you do not have an image readily available for use, our service provides a selection of stock photos. By clicking on ″Change Pics,″ a random image will be chosen from our curated database for processing by LiViT-Net, allowing you to view the results.
Model Performance Note:
Our model processes images in real-time. As such, if the characteristics of the selected image differ significantly from the data used to train the model, you may notice fraying at the image edges. This is typically due to the absence of corresponding graphic masks and does not reflect the optimal performance of our model. For the best experience, we recommend using images from the HRF dataset or selecting from the photos provided through the ″Change Pics″ button.

Fig. 5. Screenshot of the website interface.

4.3. *Implementation details*

LiViT-Net's training was developed based on Pytorch, utilizing an NVIDIA T4 graphics card. Employing SGD with an initial learning rate of 0.02 and weight decay of 1e-4, we used a dynamic learning rate scheduler and increased the batch size to 8 for efficiency. To address the class imbalance inherent in retinal vessel segmentation, we utilized a weighted cross-entropy loss function during training. This choice was essential for enhancing the model's sensitivity, especially for recognizing smaller vessel structures. Kaiming initialization [43] was used to stabilize the gradients. The mixed precision training accelerated the process, achieving convergence within 250 epochs. Continuous performance monitoring and model checkpointing based on the Dice coefficient were integral to optimizing training effectiveness.

5. Results

In this section, the proposed model is deployed on a dedicated website (Fig. 5), allowing readers to gain a more intuitive understanding of its capabilities. The website, accessible at https://hz-t3.matpool.com:28765?token = aQjYR4hqMI, not only showcases the real-time performance of the model but also provides detailed insights into the application context and the methodology underpinning it. Users can interactively explore various features of the model, offering a hands-on experience that deepens the understanding of its practical application and effectiveness.

5.1. *Comparison with state-of-the-art methods*

5.1.1. *Segmentation performance*

We compare the proposed LiViT-Net with eleven commonly used and state-of-the-art methods, including U-Net [8], Cc-net [32], Ce-net [19], D-Net [20], Ccnet [21], CSU-Net [23], RV-Gan [22], Bridge-Net [24], U-Net with attention module [26], SCS-Net [29], and NFN+ [30].

Table 1 presents a comparison of the performances of our proposed method and other advanced methods on the RGB datasets (DRIVE, CHASE_DB1 and HRF). As demonstrated in our methodology, admirable results are achieved on two key metrics: F1 score and the SSIM score. A heightened F1 score indicates that our approach excels in terms of precision and recall, which is particularly relevant in the context of retinal medical image segmentation where correct identification (precision) and completeness of this identification (recall) are essential for accurate diagnosis.

As illustrated in Table 1, our method showcases a commendable SSIM score, emphasizing its ability to preserve the structural information of segmented images during retinal medical imaging. Such preservation of structural details, like variations in vessel width or branching patterns, is vital for detecting pathological changes, highlighting the effectiveness of our method in these contexts.

In real-world scenarios, particularly for devices with limited computational capabilities often found in resource-limited areas, achieving good performance across various metrics is more crucial than dominating a single metric. LiViT-Net provides the optimal balance in terms of performance, underscored by its remarkable F1 score and SSIM, which are essential for retinal imaging. This finding positions LiViT-Net as a reliable and effective choice for retinal vessel segmentation in practical applications.

5.1.2. *Efficiency*

In addition to better performance on the metrics, LiViT-Net also has superior requirements for hardware devices when showing similar performance. This is reflected in the number of model parameters and FLOPs (floating point operations per second), a standard measure of computational workload indicating the number of floating-point operations the model requires to process an image. Fig. 6 shows a comparison of the parameter count and computational load for various models. The area of each circle is indicative of the model's parameter count. A smaller circle area implies a lower number of model parameters, signifying a more compact and efficient model. The vertical axis of the figure represents the F1 score, a harmonic mean of precision and recall, providing a single metric for evaluating the model's accuracy in terms of both FPs and FNs. A higher F1 score indicates better overall model performance. The horizontal axis represents FLOPs for computation, with smaller values indicating a reduced dependency on hardware performance. As shown in Fig. 6, LiViT-Net clearly outperforms the other models in terms of both performance and computational efficiency. To more clearly showcase the features of the compared methods, the model parameter settings shown in the figure are based on the default settings from their respective papers. The computing FLOPs in the figure are calculated based on the input image size of 512 × 512 × 3.

Furthermore, we compared the real-time performance of several popular methods on a terminal device (Honor 30 pro Kirin 990 CPU). To vividly demonstrate the superiority of our LiViT-Net model, we observed the runtime, FLOPs, and the number of parameters of these models at the same resolution. As illustrated in Table 2, the LiViT-Net model not only has the shortest runtime but also possesses fewer FLOPs

Table 1
Comparison with other advanced models on three different datasets.

Datasets	Methods	Accuracy	Specificity	Sensitivity	F1 score	SSIM
DRIVE	Jiang [20], 2019	0.9709	0.9890	0.7839	0.8246	-
	Mishra [32], 2019	0.9789	0.9879	0.7655	0.7859	0.9655
	Feng [21], 2020	0.9528	0.9809	0.7625	-	-
	Wang [23], 2020	0.9565	0.9782	**0.8071**	0.8251	-
	Li [26], 2020	0.9568	0.9810	0.7921	-	-
	Wang [27], 2020	0.9581	0.9813	0.7991	0.8293	-
	Wu [30], 2020	0.9668	0.9790	0.8002	0.8295	-
	Kamran [22], 2021	0.9790	**0.9969**	0.7927	0.8690	0.9237
	Wu [29], 2021	0.9697	0.9838	0.8289	0.8189	-
	Zhang [24], 2022	0.9565	0.9818	0.7853	0.8203	-
	U-Net [8], 2015	0.9640	0.9808	0.7915	-	-
	LiViT-Net(Ours)	**0.9907**	0.9963	0.7657	**0.8772**	**0.9705**
CHASEDB1	Jiang [20], 2019	0.9721	0.9894	0.7839	0.8062	-
	Mishra [32], 2019	0.9551	**0.9915**	0.7175	0.7391	0.9033
	Feng [21], 2020	0.9635	0.9866	0.7760	0.7434	0.9136
	Wang [23], 2020	0.9706	0.9836	**0.8427**	0.8105	-
	Li [26], 2020	0.9635	0.9819	0.7818	-	-
	Wang [27], 2020	0.9670	0.9813	0.8239	0.8191	-
	Wu [30], 2020	0.9735	0.9855	0.7933	0.8369	-
	Kamran [22], 2021	0.9697	0.9806	0.8199	0.8957	**0.9266**
	Wu [29], 2021	**0.9744**	0.9839	0.8365	-	-
	Zhang [24], 2022	0.9667	0.9840	0.8132	0.8293	-
	U-Net [8], 2015	0.9716	0.9861	0.7617	-	-
	LiViT-Net(Ours)	0.9678	0.9833	0.7816	**0.8965**	0.9205
HRF	Jiang [20], 2019	0.9517	0.9789	0.7276	0.7685	0.9056
	Mishra [32], 2019	0.9660	**0.9904**	0.6755	0.7553	0.9184
	Feng [21], 2020	0.9635	0.9830	0.7318	0.8510	0.9233
	Wang [23], 2020	0.9673	0.9847	0.7604	0.7833	0.9188
	Li [26], 2020	0.9420	0.9491	**0.8578**	0.7765	0.8919
	Wang [27], 2020	0.9654	0.9843	0.7803	0.8074	-
	Wu [30], 2020	0.9656	0.9749	0.8114	-	-
	Kamran [22], 2021	0.9631	0.9846	0.8326	0.8737	0.9254
	Wu [29], 2021	**0.9687**	0.9823	0.8114	-	-
	Zhang [24], 2022	0.9590	0.9690	0.8570	-	-
	U-Net [8], 2015	0.9638	0.9776	0.8044	-	-
	LiViT-Net(Ours)	0.9686	0.9807	0.8284	**0.8740**	**0.9288**

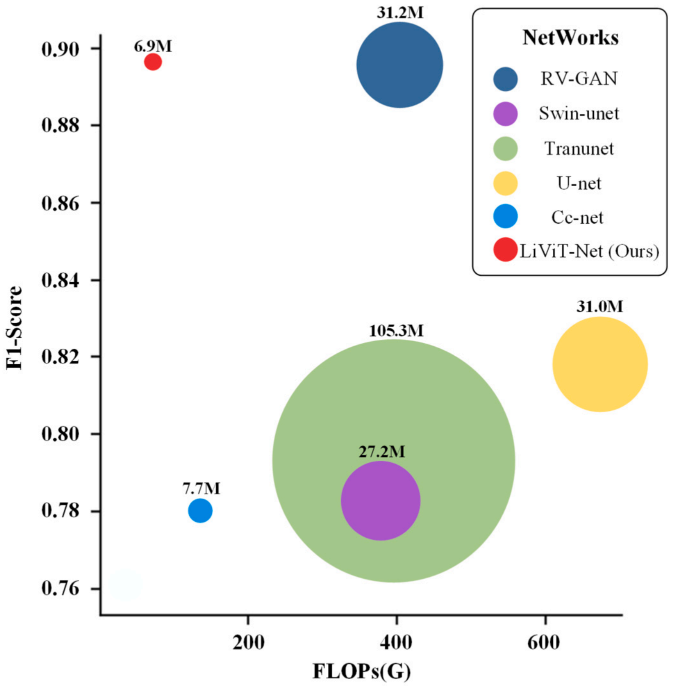


Fig. 6. LiViT-Net outperforms other state-of-the-art methods. The networks used were RV-GAN [22], Swin-UNet [6], TransUNet [13], Cc-Net [32], U-net [8], and LiViT-Net (Ours) on the CHASEDB1 dataset.

Table 2
Comparison of deep learning network models on a Honor 30 pro Kirin 990 CPU.

Network	Resolution	Run Time (ms)	FLOPs (G)	Params (M)
RV-GAN	512×512	7044	440.0	31.0
Swin-Unet	512×512	2944	94.7	27.2
TransUNet	512×512	6629	395.6	105.3
U-net	512×512	2575	670.6	31.2
Cc-net	512×512	2789	168	7.7
LiViT-Net (Ours)	512×512	**1195**	**71.1**	**6.9**

and fewer parameters, highlighting its friendliness and efficiency on terminal devices.

The LiViT-Net model demonstrates remarkable runtime efficiency, operating at 1195 ms, which is a significant reduction compared to TransUNet, Swin-Unet, and Cc-net. Specifically, LiViT-Net's runtime is about 78% faster than TransUNet's 6629 ms, 59% faster than Swin-Unet, and 57% faster than Cc-net's 2789 ms. This enhanced speed is critical for real-time applications, particularly in edge computing scenarios. In terms of computational complexity, measured in FLOPs, LiViT-Net requires only 71.1G, markedly less than U-net's 670.6G, Swin-Unet, TransUNet, and also lower than Cc-net's 168G. This reduction in computational demand makes LiViT-Net highly suitable for deployment on edge devices, where resources are limited.

LiViT-Net's design, with a mere 6.9 M parameters, showcases a remarkable compactness compared to TransUNet's 105.3 M and Cc-Net's 7.7 M, as revealed in our experiments. This significant reduction in parameters, coupled with its leading-edge runtime of only 1195 ms and

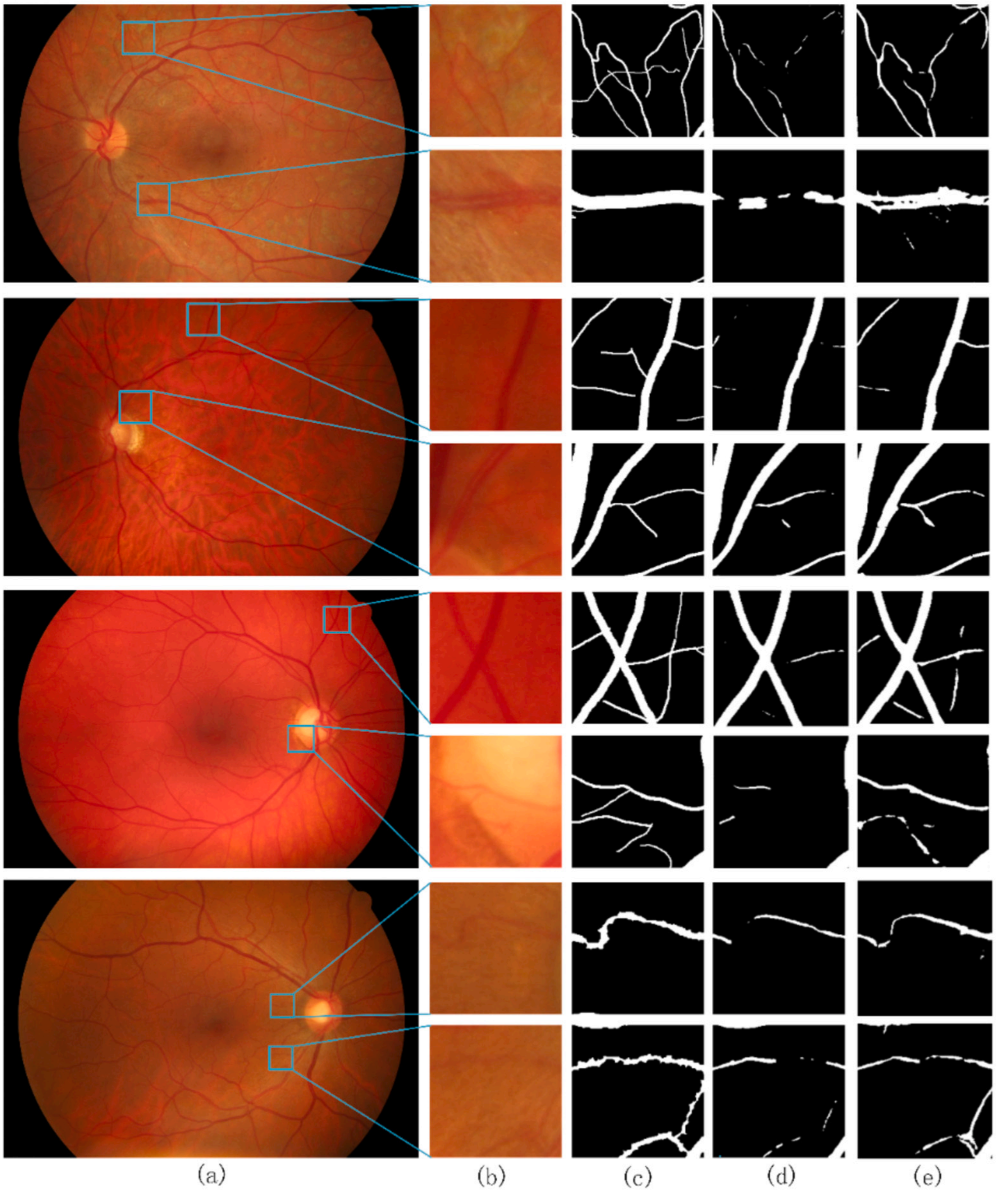


Fig. 7. Typical visual results for different methods in the ablation study on the HRF dataset. (a) Original image, (b) detailed view, (c) ground truth, (d) baseline, (e) LiViT-Net.

minimal computational demand of 71.1G FLOPs, underscores the model's tailored fit for edge devices. These results not only demonstrate LiViT-Net's ability to efficiently manage limited memory and processing power but also its superiority in operational speed and computational efficiency over models such as Cc-Net, making it an ideal solution for resource-constrained environments.

5.2. Ablation study

To evaluate the effectiveness of our proposed method, we conducted an ablation study. In this section, we meticulously validate the efficacy of both our model and the proposed loss function, with the results clearly illustrated through images and tables.

5.2.1. Effectiveness of the MobileViT+ block

To evaluate the effectiveness of the MobileViT+ block within our LiViT-Net, we established a baseline model. This variant of the LiViT-Net replaces the MobileViT+ block with an inverted residual block, serves as a point of comparison and is applied to the HRF dataset. Fig. 7 presents four distinctive instances of retinal vessel segmentation results, effectively demonstrating the ability of the proposed MobileViT+ module to segment vessels in varying scales. It is noteworthy that the module excels in accurately identifying minuscule vessels and discerning vessel edges more effectively, whereas the baseline network finds these tasks challenging. As shown in Table 3, compared with "baseline", "baseline + MobileViT+" improved the performance from 0.7345/0.8643 to 0.8284/0.8740 in terms of sensitivity/F1 score, which demonstrates that the long-range feature extraction is necessary to improve segmentation accuracy.

Table 3
Statistical comparison of ablation studies on HRF dataset.

Method	SE	SP	ACC	AUC	F1	SSIM
baseline	0.7345	**0.9889**	0.9666	**0.9836**	0.8643	0.9262
LiViT-Net	**0.8284**	0.9807	**0.9686**	0.9800	**0.8740**	**0.9288**

To quantify the generalization capacity and robustness of the proposed model, we conduct additional cross-dataset experiments and implement a cross-training evaluation on the DRIVE, CHASEDB1, and HRF

Table 4
Results of cross-training. The best values are marked by **bold**.

Methods	CHASEDB1 (train) DRIVE (test)				HRF (train) CHASEDB1 (test)			
	Acc	Sen	F1	Auc_Roc	Acc	Sen	F1	Auc_Roc
U-net	-0.0013	-0.1010	-0.0467	**-0.0166**	-0.0293	-0.4385	-0.3371	-0.0303
LiViT-Net	**-0.0013**	**-0.0642**	**-0.0392**	-0.0274	**-0.0082**	**-0.1027**	**-0.0677**	**-0.0026**

Table 5
Comparison of Proposed Loss (green) and CE+Dice loss.

Methods	DRIVE				HRF			
	Acc	Sen	F1	AUC_ROC	Acc	Sen	F1	AUC_ROC
CE Loss+Dice loss	**0.9907**	0.7412	0.8725	0.9556	0.9679	0.8255	0.8737	0.9795
Proposed Loss	**0.9907**	**0.7657**	**0.8772**	**0.9650**	**0.9686**	**0.8284**	**0.8740**	**0.9800**

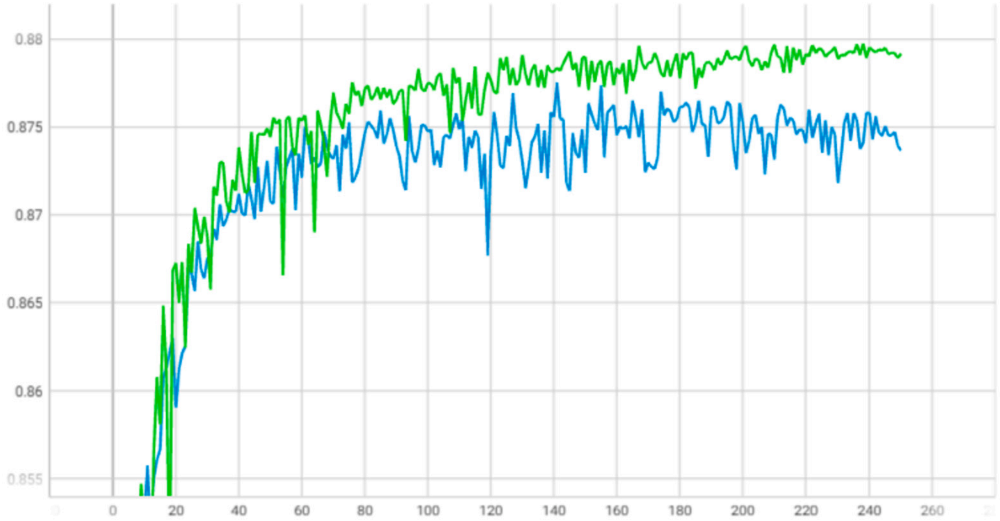


Fig. 8. Comparison of Proposed Loss (green) and CE+Dice loss on Dice coefficient.

datasets. We directly apply the well-trained proposed model to other datasets without retraining the model on the new dataset. Table 4 presents the statistical cross-training evaluation results of LiViT-Net, where a smaller value indicates a less significant fluctuation in the index post-cross-validation, indicating sufficient model robustness.

5.2.2. Effectiveness of the joint loss

To verify the effectiveness of the newly proposed joint loss, we compare it with the commonly used weighted cross-entropy loss and Dice loss by combining the Losses.

As shown in Table 5, the comparison results demonstrate that the model performs better when using the proposed joint loss. As illustrated in Fig. 8, the green line indicates the variations of metric Dice during the training process with the proposed joint loss, while the blue line represents the variations of the metric when using weighted cross-entropy Loss and Dice loss. It can be observed that the model trained with the newly proposed joint loss method demonstrates faster learning speed and superior performance under the same conditions.

5.2.3. Optimal configuration of loss ratios

To determine the optimal values for α and β in the joint loss function, defined as Eq. (11), we conduct ablation studies on the DRIVE dataset, varying their ratios to assess their influence on the model's performance (Table 6).

According to the experimental results, setting the α/β ratio to 0.2/0.8 yielded the highest F1 score (0.8772) among all the configurations. Moreover, the sensitivity, specificity, and global accuracy under these settings are comparable to those of other configurations, indicating that these combinations provide a good balance for the model, allowing it to effectively capture fine vessel regions while maintaining high classification accuracy. Additionally, the SSIM score also showed good predictive performance.

Table 6
Ablation study results with varying ratios of α and β in joint loss.

Ratio	Global Accuracy	Specificity	Sensitivity	F1 score	SSIM
0.0/1.0	**0.9907**	0.9968	0.7608	0.8709	**0.9707**
0.1/0.9	0.9906	0.9964	**0.7696**	0.8677	0.9703
0.2/0.8	**0.9907**	0.9963	**0.7657**	**0.8772**	**0.9705**
0.3/0.7	**0.9907**	0.9969	0.7614	**0.8759**	**0.9705**
0.5/0.5	0.9906	0.9966	0.7609	0.8671	0.9702
0.7/0.3	0.9906	0.9967	0.7567	0.8687	0.9703
0.9/0.1	0.9902	**0.9982**	0.7332	0.8445	0.9696
1.0/0.0	0.9905	**0.9974**	0.7422	0.8598	0.9702

Considering the characteristics of Dice loss and weighted cross-entropy loss, it is understandable why the α/β ratio of 0.2/0.8 achieved the best outcomes. Dice loss offers robustness against the sample imbalance, especially when target regions like vessels are sparse. On the other hand, weighted cross-entropy loss ensures accurate predictions for every pixel. Therefore, increasing the Dice loss weight helps individuals capture fine vessels and reduces misjudgments at boundary regions.

6. Conclusion

In this study, we propose a U-Net-like, lightweight Transformer network, LiViT-Net, for retinal vessel segmentation. By integrating the innovative local representation strategy and the MobileViT+ module, the model efficiently captures intricate image structures without excessive parameters, merging the benefits of CNNs and Transformers. Furthermore, a remapped, weighted joint loss is designed to address the challenges of retinal vessel segmentation, including pixel imbalances and complex vascular structures, ensuring a balance between accuracy and class equilibrium.

Transitioning to practical applications, LiViT-Net's architecture not only occupies the primary domain but is also remarkably adaptable to computational constraints. Specific tests conducted on edge devices demonstrate the ability of edge devices to work in environments with restricted computational capabilities, emphasizing their streamlined efficiency. Furthermore, comprehensive evaluations of esteemed databases such as DRIVE, CHASEDB1, and HRF reinforce our model's supremacy, as these evaluations consistently eclipse the performance metrics of its contemporaries.

In future endeavors, we intend to enhance the precision of LiViT-Net by exploring advanced optimization techniques. We aim to incorporate sophisticated regularization methods and structures, minimizing computational costs without compromising performance. We are also interested in adapting the model to various medical imaging fields to ensure its resilience and versatility. We are interested in introducing real-time data augmentation to bolster the model's adaptability. Engaging with medical experts is on our agenda, aligning LiViT-Net's capabilities with clinical needs for maximum impact.

Funding statement

This work is partially supported by the National Natural Science Foundation of China under Grant No. 62301320. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agency.

Declaration of competing interest

The authors affirm that there are no existing financial conflicts or personal affiliations which could be perceived as influencing the research presented in this manuscript.

References

[1] Wang S, Yin Y, Cao G, Wei B, Zheng Y, Yang G. Hierarchical retinal blood vessel segmentation based on feature and ensemble learning. Neurocomputing 2015;149:708–17.

[2] Ørskov M, Vorum H, Larsen TB, Lip GY, Bek T, Skjøth F. Similarities and differences in systemic risk factors for retinal artery occlusion and stroke: a nationwide case-control study. J Stroke Cerebrovasc Dis 2022;31(8):106610.

[3] Shi D, He S, Yang J, Zheng Y, He M. One-shot retinal artery and vein segmentation via cross-modality pretraining. Ophthalmol Sci 2023:100363.

[4] Karimi D, Vasylechko SD, Gholipour A. Convolution-free medical image segmentation using transformers. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24. Springer; 2021. p. 78–88.

[5] Lin A, Chen B, Xu J, Zhang Z, Lu G, Zhang D. DS-TransUNet: dual swin transformer U-Net for medical image segmentation. IEEE Trans Instrum Meas 2022;71:1–15.

[6] Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-UNet: unet-like pure transformer for medical image segmentation. In: European conference on computer vision. Springer; 2022. p. 205–18.

[7] Khan TM, Alhussein M, Aurangzeb K, Arsalan M, Naqvi SS, Nawaz SJ. Residual connection-based encoder decoder network (RCED-Net) for retinal vessel segmentation. IEEE Access 2020;8:131257–72.

[8] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer; 2015. p. 234–41.

[9] Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 2017;40(4):834–48.

[10] Zhang Y, Liu H, Hu Q. Transfuse: fusing transformers and CNNs for medical image segmentation. 2021. p. 14–24.

[11] Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: gated axial-attention for medical image segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24. Springer; 2021. p. 36–46.

[12] Chen B, Liu Y, Zhang Z, Lu G, Kong AWK. TransAttUnet: multi-level attention-guided U-Net with transformer for medical image segmentation. arXiv preprint. arXiv:2107.05274, 2021.

[13] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint. arXiv:2102.04306, 2021.

[14] Biswas R, Vasan A, Roy SS. Dilated deep neural network for segmentation of retinal blood vessels in fundus images. Iranian J Sci Technol Trans Electr Eng 2020;44:505–18.

[15] Chen C, Chuah JH, Ali R, Wang Y. Retinal vessel segmentation using deep learning: a review. IEEE Access 2021;9:111985–2004.

[16] He J, Jiang D. Fundus image segmentation based on improved generative adversarial network for retinal vessel analysis. In: 2020 3rd international conference on artificial intelligence and big data (ICAIBD). IEEE; 2020. p. 231–6.

[17] Wang D, Hu G, Lyu C. FRNet: an end-to-end feature refinement neural network for medical image segmentation. Vis Comput 2021;37:1101–12.

[18] Mou L, Chen L, Cheng J, Gu Z, Zhao Y, Liu J. Dense dilated network with probability regularized walk for vessel detection. IEEE Trans Med Imaging 2019;39(5):1392–403.

[19] Gu Z, Cheng J, Fu H, Zhou K, Hao H, Zhao Y, et al. CE-Net: context encoder network for 2D medical image segmentation. IEEE Trans Med Imaging 2019;38(10):2281–92.

[20] Jiang Y, Tan N, Peng T, Zhang H. Retinal vessels segmentation based on dilated multi-scale convolutional neural network. IEEE Access 2019;7:76342–52.

[21] Feng S, Zhuo Z, Pan D, Tian Q. Ccnet: a cross-connected convolutional network for segmenting retinal vessels using multi-scale features. Neurocomputing 2020;392:268–76.

[22] Kamran SA, Hossain KF, Tavakkoli A, Zuckerbrod SL, Sanders KM, Baker SA. RV-GAN: segmenting retinal vascular structure in fundus photographs using a novel multi-scale generative adversarial network. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part VIII 24. Springer; 2021. p. 34–44.

[23] Wang B, Wang S, Qiu S, Wei W, Wang H, He H. CSU-Net: a context spatial U-Net for accurate blood vessel segmentation in fundus images. IEEE J Biomed Health Inform 2020;25(4):1128–38.

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

[25] Lian S, Li L, Lian G, Xiao X, Luo Z, Li S. A global and local enhanced residual u-net for accurate retinal vessel segmentation. IEEE/ACM Trans Comput Biol Bioinform 2019;18(3):852–62.

[26] Li X, Jiang Y, Li M, Yin S. Lightweight attention convolutional neural network for retinal vessel image segmentation. IEEE Trans Ind Inform 2020;17(3):1958–67.

[27] Wang D, Haytham A, Pottenburgh J, Saeedi O, Tao Y. Hard attention net for automatic retinal vessel segmentation. IEEE J Biomed Health Inform 2020;24(12):3384–96.

[28] Cheng Y, Ma M, Zhang L, Jin C, Ma L, Zhou Y. Retinal blood vessel segmentation based on densely connected u-net. Math Biosci Eng 2020;17(4):3088–108.

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

[30] Wu Y, Xia Y, Song Y, Zhang Y, Cai W. NFN+: a novel network followed network for retinal vessel segmentation. Neural Netw 2020;126:153–62.

[31] Guo C, Szemenyei M, Yi Y, Wang W, Chen B, Fan C. SA-UNet: spatial attention U-Net for retinal vessel segmentation. In: 2020 25th international conference on pattern recognition (ICPR); 2021. p. 1236–42.

[32] Mishra S, Liang P, Czajka A, Chen DZ, Hu XS. Cc-net: image complexity guided network compression for biomedical image segmentation. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). IEEE; 2019. p. 57–60.

[33] Wu H, Chen S, Chen G, Wang W, Lei B, Wen Z. Fat-net: feature adaptive transformers for automated skin lesion segmentation. Med Image Anal 2022;76:102327.

[34] Pizer SM, Amburn EP, Austin JD, Cromartie R, Geselowitz A, Greer T, et al. Adaptive histogram equalization and its variations. Comput Vis Graph Image Process 1987;39(3):355–68.

[35] Naseer MM, Ranasinghe K, Khan SH, Hayat M, Khan F Shahbaz, Yang M-H. Intriguing properties of vision transformers. Adv Neural Inf Process Syst 2021;34:23296–308.

[36] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint. arXiv:2010.11929, 2020.

[37] Mehta S, Rastegari M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint. arXiv:2110.02178, 2021.

[38] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26(3):297–302.

[39] Milletari F, Navab N, Ahmadi S-A. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). IEEE; 2016. p. 565–71.

[40] Staal J, Abràmoff MD, Niemeijer M, Viergever MA, Van Ginneken B. Ridge-based vessel segmentation in color images of the retina. IEEE Trans Med Imaging 2004;23(4):501–9.

[41] Fraz MM, Remagnino P, Hoppe A, Uyyanonvara B, Rudnicka AR, Owen CG, et al. An ensemble classification-based approach applied to retinal blood vessel segmentation. IEEE Trans Biomed Eng 2012;59(9):2538–48.

[42] Köhler T, Budai A, Kraus MF, Odstrčilik J, Michelson G, Hornegger J. Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In: Proceedings of the 26th IEEE international symposium on computer-based medical systems. IEEE; 2013. p. 95–100.

[43] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 1026–34.