C^2AM-Unet: Coordinate and Channel Attention Mixing Flexible Architecture for Retinal Vessel Segmentation

Jingling Zhang[1], Shuting Zheng[2], Wen Zhang[3], Jia Gu[4*]

1 City University of Macau, China; D24092100270@cityu.edu.mo
2 City University of Macau, China; D23092110405@cityu.edu.mo
3 Florida Atlantic University, Boca Raton, FL 33431, USA; zhangw@fau.edu
4 City University of Macau, China; jiagu@cityu.edu.mo
* Correspondence: jiagu@cityu.edu.mo

Abstract: Recently, CNN-based and Transformer-based network have become the de-facto standard for vessel segmentation, due to their strong feature representation capabilities. However, CNNs fall short in explicitly capturing global dependencies between different spatial locations, while Transformers require quadratic computation by virtue of the long sequences and lack fine-grained local correlations. Seeking to model both local correlations and long-range dependencies, our research revisits this issue and introduces a novel retinal vessel segmentation network based on U-Net, the Coordinate and Channel Attention Mixing Flexible Architecture (C^2AM-Unet), a new lightweight network that combines the U-Net model with a multi-attention mechanism. Specifically, we present a core Dual Global Coordinate Attention (DGCA) module, which embeds two types of serial global spatial-specific property information: one with a direction-aware mechanism, responsible for encoding long-range dependencies and coordinate information in both horizontal and vertical directions, and another with a transformer-like coordinate module that focuses on the connections and interactions between each spatial location. Next, the C^2AM feature maps are aggregated with the direction-aware feature maps to enhance their ability to represent the objects of interest. To further strengthen the long-distance semantic dependencies, we utilize the transformer-like channel attention model to capture the cross-channel communication of global information. To verify the effectiveness of our method, we also construct other similar variant components that can be flexibly applied to other networks. The experimental results on three retinal vessel and two coronary angiography datasets, including DRIVE, STARE, CHASE_DB1, DCA1 and CHUAC, demonstrate that C^2AM-Unet achieves competitive scores compared to the state-of-the-art methods.

Keywords: retinal vessel segmentation; U-Net; coordinate attention; channel attention; DGCA

1. Introduction

Medical image segmentation is essential for automating the diagnostic process, as it isolates specific anatomical structures or abnormalities within images, facilitating more precise analyses. By segmenting different regions in medical images, clinicians can better localize and characterize pathologies, which is crucial for accurate diagnosis and treatment planning. Particularly in the case of retinal vascular segmentation, the task is challenging due to complex structures, such as low-contrast backgrounds, elongated tubular branches, tree-like structures, as well as variable and diverse morphologies, making

manual segmentation difficult. The points where vessels split or cross over each other are difficult to differentiate individual vessels. Therefore, the development of automated segmentation algorithms is highly imperative to significantly reduce the manual workload and potential for human error, leading to more consistent and reliable diagnostic outcomes. U-Net [1], introduced by Ronneberger in 2015, is a pioneering and widely recognized convolutional neural network (CNN) architecture, specifically designed for semantic segmentation in medical imaging, with its distinctive encoder-decoder architecture. U-Net's top-down encoder path generates a rich hierarchical feature representation that captures both local and contextual information, which is essential for understanding image content at multiple scales, while the bottom-up decoder path maps the learned representation back to the original image resolution, enabling precise pixel-by-pixel classification for accurate segmentation. These advantages have catapulted U-Net to widespread adoption and success across various medical image segmentation tasks. Building upon the U-Net architecture, several variants have been developed to enhance the performance of medical image segmentation, such as UNet++ [2], Att-UNet [3], and V-Net [4] by incorporating the attention mechanisms, or redesigning the skip connection path. Although these extensions largely enrich the benchmark approach, the inherent limitations of CNNs make it challenging to capture global and structural representations effectively.

Recently, the Visual Transformer (ViT) [5] has emerged as a pioneering architecture in the computer vision domain, drawing inspiration from the successes of Transformer NLP. Subsequently, to enhance the global representation capacity for medical image segmentation, some authors have begun to propose methods based on transformers, such as Swin-trans[6], MISSFormer[7], and DAE-Former[8], which leverage stacked transformer blocks to establish relationships between image patches at different locations, benefiting from multi-head self-attention or cross-attention. However, naive transformer-based methods have high computational and memory costs due to the sequence length when calculating the similarity of image patches[8]. Meanwhile, the basic ViT model underperforms in capturing fine-grained local representations compared to CNNs. In response, various hybrid architectures combining CNNs and Transformers have been introduced. TransUNet [9], which combines Transformers with a U-shaped architecture, demonstrates strong ability in extracting finer details. DE-DCGCN-EE [10] designs a dual encoder-based dynamic-channel network to enhance edge information using graph convolutional networks. GT-DLA-dsHFF [11], PLVS-Net [12] and DPF-Net[13] aim to reduce the semantic gap between encoder and decoder stages, while DPL-GFT-EFA [14] and MedSegDiff[15] utilize diffusion probabilistic to tackle the problem of lesions and image noise. However, these approaches have not yet achieved an optimal trade-off between computational efficiency and accuracy.

How to integrate local information and long-range interaction more effectively with low computation cost remains a key issue in the field of medical image segmentation. In this paper, we re-evaluate this issue: Beyond CNN and Transformer are there alternative innovative approaches? Undoubtedly, the answer is yes. By joining CNNs with novel attention mechanisms, we've established a more robust framework, called C^2AM-Unet. The core module is the Dual Global Coordinate Attention module (DGCA) component, which contains two types of attention layers: direction-sensitive coordinate attention and transformer-style attention. The direction-sensitive coordinate attention is applied to fuse coordinate information from different directions, which is responsible for the dynamic and variable orientation of the retinal vasculature. The transformer-style attention facilitates communication between pixel locations, simplifying the complexity to a single matrix multiplication process. To model remote

dependencies across multiple directions, the direction-sensitive features are transferred to the channel attention.

Additionally, we propose several alternative options that can effectively enhance performance. Overall, our approach adheres to the encoder-decoder U-shaped structure, featuring improved skip connections and an enhanced bottleneck layer. In the fourth layer, the original convolution block is replaced by the DGCA module and Channel Attention (stage 2 and 3). We validate the effectiveness of our method on five mainstream medical image segmentation datasets, achieving competitive performance. Meanwhile, the components proposed in this paper can be applied to other U-shaped networks to easily improve their accuracy and performance. In summary, we make following contributions: 1) We propose a new medical segmentation network, named C^2AM-Unet. 2) To encode long-range dependencies and local contextual information from multi-directions, we design a novel Dual Global Coordinate Attention module, DGCA. 3) We introduce multiple high-performance alternative methods, which can be flexibly integrated into other U-shaped methods. 4) Experiments on five benchmark datasets validate the promising effectiveness of C^2AM-Unet for accurately localizing thin blood vessel.

2. Related Work

CNN- and Transformer-based Methods for Medical Image Segmentation. The success of U-Net has spurred academics to introduce numerous enhanced variants based on its framework. U-Net++ [2] redesigns skip connections to enable flexible feature fusion for multi-scale feature maps. To automatically focus on task-relevant regions or targets, Attention U-Net [3] has achieved impressive results by applying attention gate technique to the standard U-Net without requiring complicated heuristics. Dai et al. [16] proposes a dual-path U-Net that efficiently bridges the encoder and the decoder semantic gap by re-using and re-exploring the interaction of historical information among the dual paths. The use of attention mechanisms and the reimagining of skip connections are all effective attempts to improve feature representation. Inspired by ViT [5], several researchers further improve the performance from other perspectives. For example, Transunet [9] firstly demonstrates us the pure Transformer-based U-shape net. MISSFormer [7] achieves excellent performance when employs the proposed Enhanced Transformer Block to strengthen long-range dependencies and complement local context, while DAE-Former[8] combines efficient spatial attention with transpose (channel) attention to comprehensively capture spatial and channel features in the image. Swin-trans [6], as a pioneer work in this direction, adapts the non-overlapping shifted windows which brought greater efficiency on various vision task. Other frameworks, such as GFT-EFA [14] and MedSegDiff [15] pay more attention to tackle image noise using diffusion probabilistic. Despite that integrating CNN and Transformer has largely emerged the local and global feature to improve the feature representation, how to create a strong connection between different spatial locations and generate a more accurate segmentation map remains a challenge.

Attention Mechanisms. Extensive works have been conducted on attention mechanisms across various vision tasks, yielding promising results [17], [18], [19], [20], [21], [22], [23], [24], [25]. These mechanisms can be broadly categorized into two major types: general attention mechanisms and transformer-based attention mechanisms. General attention mechanisms, while capable of capturing essential structural information for precise segmentation, fail to effectively model long-rang spatial relationships. The SE [18] attention mechanism utilizes 2D global average pooling to compress feature

maps efficiently, ensuring low computation complexity. However, by focusing primarily on inter-channel information, the SE attention mechanism neglects the critical role of positional information. To address this limitation, several subsequent methods, such as BAM [24] and CBAM [26], have been developed to partially recover positional information. Nevertheless, these approaches often fail to capture long-term dependencies due to their reliance on convolutional layers. On the other hand, transformer-based attention mechanisms have demonstrated significant success in modeling long-range interactions, as seen in methods like FTransCNN [27] and M-CnT [20]. However, these methods come at the cost of generating many parameters and consuming substantial memory resources. In the specialized domain of medical imaging, transformer-based methods may struggle to capture intricate details, particularly in representing thin tubular vessel structures. Inspired by these advancements, this paper proposes a novel approach that addresses these challenges by capturing fine-grained details and establishing long-distance dependencies without introducing an excessive number of parameters. By integrating global coordinate information, transformer-like attention, and inter-channel attention, our method achieves superior results compared to existing lightweight frameworks.

3. Methods

3.1 Architecture Overview

Figure 1 illustrates the proposed overall architecture of C^2AM-Unet, which builds upon the FR-Unet [28] design and consists of four main components: the encoder module, the DGCA, the channel attention module and the decoder module. Additionally, we propose two alternative frameworks that also demonstrate promising results. The encoder and decoder modules contain three stages of down-sampling and up-sampling layers, implemented using standard 2×2 convolution and a 2×2 fractionally-strided convolutions, respectively. Each stage of the encoder and decoder modules comprises multiple convolution blocks.

Unlike the FR-Unet, we substitute the modified residual block with a newly designed standard convolutional block, which consists of two 3×3 convolution (Conv) layers, two batch normalization (BN) layers and a LeakyReLU (LR) activation layer. To generate dense predictions, we retain the skip-connection structure, which fuses the feature maps from the previous block with those obtained from the down-sampling and up-sampling layers of adjacent stages.

The dual global coordinate attention module includes two types of coordinate attention mechanisms: direction-aware coordinate attention and transformer-like coordinate attention. The direction-aware mechanism embeds position-aware and direction-sensitive information from the horizontal and vertical axes, while the transformer-like mechanism models long-range spatial relationships between any two pixels. To further enhance the network's representational capacity, the direction-aware feature maps are fed into the channel attention module for inter-channel communication. Finally, we introduce two additional alternative frameworks, which also achieve impressive performance.

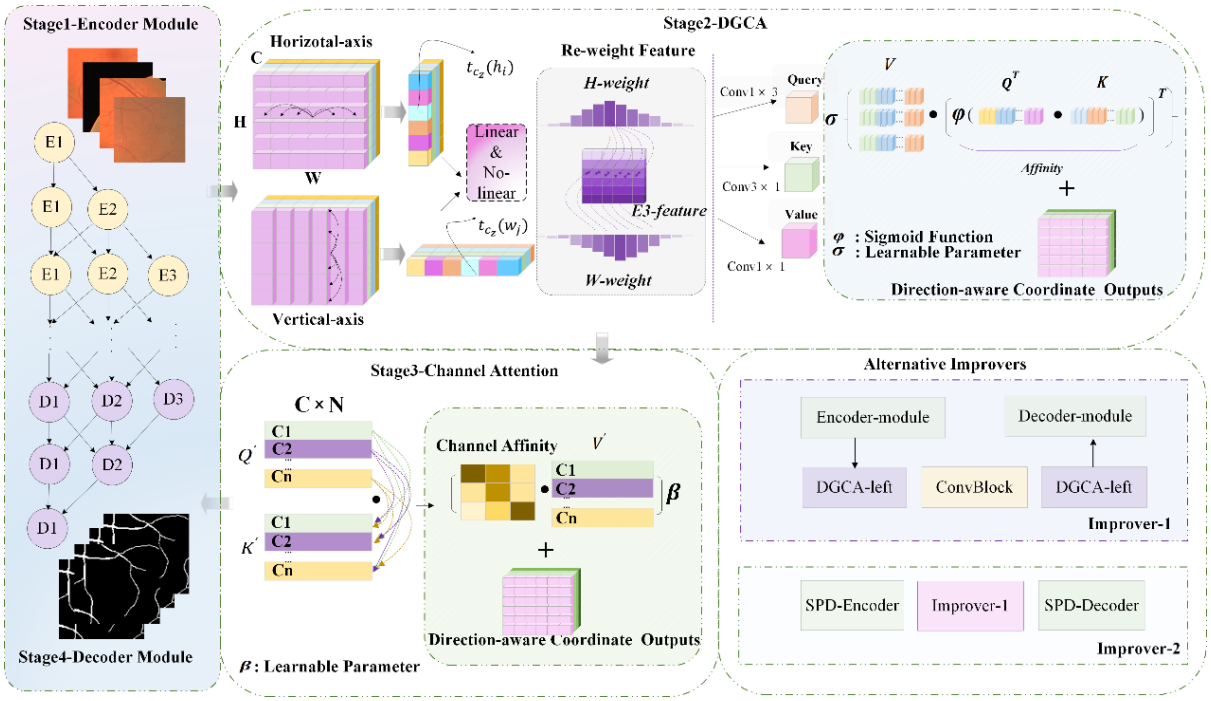

Figure 1. The overview of the proposed C²AM-Unet. It consists of four modules: the encoder module, the DGCA module, the channel attention module, and the decoder module. The components from stage2 and stage3 are flexibly combined to create alternative improver modules.

3.2 Dual Global Coordinate Attention Module (DGCA)

Considering the complex structure of the retinal vessel segmentation task, such as the thin, fragile local structures and the variable global tree-like morphology, it is essential to capture rich global contextual information from multiple perspectives. As demonstrated in many vision tasks[17], [18], [19], [25], spatial and channel information play crucial role to capture structure features and generate precise segmentation maps. Hence, we design a novel DGCA module that adheres to three key principles. First, it can effectively acquire rich spatial information, enabling the generation of accurate segmentation maps. Second, it must capture long-range dependencies to accurately identify regions of interest while maintaining low memory consumption. Finally, the design should accommodate the slender and variable characteristics of retinal blood vessels. Our proposed dual global coordinate attention module is divided into two steps: direction-aware coordinate attention and transformer-like coordinate attention generation. These steps will be described in detail in the following sections.

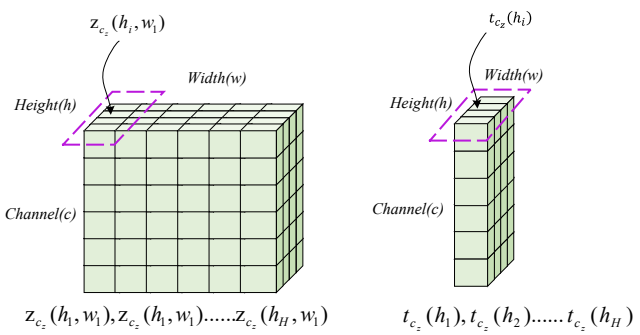

Figure 2. Direction-aware feature map generation in height direction.

In order to maximize the ability of capturing direction-sensitive spatial information and accommodating the slender tubular structure of blood vessels, we apply two parallel adaptive average pooling operations to compute the coordinate information from horizontal direction and vertical direction [25]. Given a feature map $Z \in R^{C\times H\times W}$, which is directly generated from the third layer of the encoder module. Here H, W and C denote the height, width and channel number of the feature map, respectively. We embed directional feature maps using two orthogonal average poling with kernel sizes of (H,1) and (W,1), dependently. As shown in Figure 2 (taking the horizontal direction as an example). The precise computation procedure is presented hereafter:

$$t_{c_z}(h_i) = \frac{1}{W}\sum_{j=1}^{W} z_c(h_i, w_j) \tag{1}$$

$$t_{c_z}(w_j) = \frac{1}{H}\sum_{i=1}^{H} z_c(h_i, w_j) \tag{2}$$

Here, $h_i(i \in [1, H])$ denotes the height index, $w_j(j \in [1, W])$ denotes the width index, $c_z(z \in [1, C])$ denotes the channel index. The transformed features of the c_z channel at each spatial location index (h_i, w_j) fuse the long-rang global direction-sensitive relationships and spatial location information, which is essential to model complex structure details. To fully leverage the embedded 1D features $t_c(h), t_c(w)$, we first reshape the width-transformed feature $t_c(w)$ to a size of $[H, 1]$, like the vertical direction. Then, we concatenate them and generate a new feature map $p_c \in R^{2H\times 1}$. A standard convolutional layer is then used as a local spatial filter to model the details. The calculation process is as follows:

$$g = \delta(BN(f^{1\times 1}(p_c))) \tag{3}$$

Where $f^{1\times 1}$ represents the convolutional operation with kernel size 1×1, BN represents batch normalization, and δ represents the non-linear hard-swish function. Then, the resulting features are split into two separate features g_h and g_w which are subsequently fed into another two distinct convolutional layers to generate the attention maps. Lastly, the final re-weighted feature can be written as:

$$Z^r = Z \odot S(C_h^{1\times 1}(g_h)) \odot \mathrm{S}(C_h^{1\times 1}(g_w)) \tag{4}$$

Where, $C_h^{1\times 1}$ and $C_w^{1\times 1}$ denote the different convolutional kernel size of 1×1, S denotes the sigmoid function, Z^r denotes the final re-weighted feature maps and $\odot$ represents element-wise multiplication.

To enhance segmentation accuracy, it is essential to integrate both local contextual details and broader semantic information in retinal vessel images. The direction-sensitive features computed earlier capture information from individual rows or columns at each spatial location but fail to account for contextual dependencies across different spatial locations. To address this limitation, we employ a coordinate transformer attention mechanism inspired by the transformer framework. By using features from both local and global perspectives, this approach significantly enhances the segmentation of fine vascular structures, particularly thin vessels.

First, we use two standard 3×3 convolutional layers to further encode rich local information and generate the local features. Let $X \in R^{C'\times H'\times W'}$ be the local feature. Then, we transformer it into three new different matrixes, separately, which are the query matric $Q \in R^{C'\times H'\times W'}$, key matrix $K \in R^{C'\times H'\times W'}$ and value matrix $V \in R^{C'\times H'\times W'}$ using a 1×3, a 3×1, and a 1×1 kernel convolutional layers. To amplify the internal features correlation, they are then flattened into a sequence of size $C' \times N'$, where $N' = H' \times W'$. Then, a matrix multiplication is used to calculate the affinity score between Q and K. Additionally, the softmax function is performed to normalize the affinity score and thus derive the

attention weight. Meanwhile, the value matric V is reshaped to $R^{C'\times N'}$ and is multiplied by attention weight using a matrix multiplication to calculate the affinity attention map. Finally, to avoid an overemphasis on specific elements, the affinity attention map is scaled by a trainable scalar parameter. This adjusted feature maps are then combined with the original input X through a residual connection, thereby boosting the model's capacity to generalization. The calculation method is shown in the following formula:

$$Coordinate_{output} = \sigma(V \cdot \varphi(Q^T \cdot K)) + X \tag{5}$$

Where, φ denotes the softmax function, which is used to compute the attention weight, σ denotes the learnable parameter and $\cdot$ is the matrix multiplication.

3.3 Channel Attention Module

Each dimension within a high-level feature representation corresponds to a distinct class activation. By leveraging channel attention, the model can accurately identify and prioritize channels containing critical global key features while allocating less attention to channels that lack significant information. This selective focus enhances the model's capability to effectively represent and emphasize essential features. Thus, in this section, we further explore the inter-relationships among channel maps using the computed coordinate attention maps. Unlike the SE attention mechanism [18], which relies on scalar operations, our approach employs matrix multiplication to explicitly model the interactions between different channels. We firstly reshape the local feature map $X \in R^{C'\times H'\times W'}$ to $C' \times N'$, where $N' = H' \times W'$, and then it is initialized as the Query matrix Q', the Key matrix K' and the Value matrix V'. To obtain the channel affinity score Aff, we perform a multiplication between Q' and the transposed K'. To enhance the contrast of the affinity matrix and make the relative differences of non-maximum values more pronounced, we calculate the maximum value of the affinity matrix in the last dimension (i.e., within each channel), and expand the maximum values to the same shape (C', C') as the affinity matrix using the expand_as function, so that element-wise subtraction can be performed by *Aff*.

$$Aff = Q' \cdot (K')^T \tag{6}$$

$$Aff_{new} = expand(\max(Aff)) - Aff \tag{7}$$

Where *expand* presents the expand_as function, max presents the max function and $Aff_{new} \in R^{(C',C')}$ presents the resulting new affinity matrix. Similar to above section, we then normalize the calculated affinity matrix with softmax function and then it is multiped with V' by a linear matrix multiplication to produce the attention-weighted Value matrix. In order to dynamically adjust the weights of the new attention feature map and the original input feature map X, we perform the following calculation formula:

$$Channel_{out} = \beta(softmax(Aff_{new}) \cdot V') + X \tag{8}$$

Where, $Channel_{out}$ represent the output of channel attention module, β represents the learnable parameter. Finally, the final attention output is obtained by adding the coordinate attention $Coordinate_{output}$ and the channel attention $Channel_{out}$, which is used to model the fine-grained interaction between the features from the two branches. To obtain the final precise segmentation, the final attention ouput is fed to the decoder module, with the deep supervision strategy[28].

3.4 Alternative Improvers

Considering the computational limitations inherent to some application scenarios, we also provide two alternative yet simple and flexible schemes that demonstrate superior performance, as illustrated in Figure 1. In the first improver, the network comprises an encoder-decoder module similar to C^2AM-Unet, two direction-aware coordinate attention modules (corresponding to the left part of DGCA), and one

ConvBlock. Notably, the embedded features from the encoder module are fed into the direction-aware coordinate attention module to enhance spatial feature representation. Following this, the ConvBlock, which contains two standard 3×3 convolutional layers, is used to compute local contextual features. Subsequently, the DGCA-left module is reapplied to improve the model's ability to restore detailed image information during the up-sampling process, which is followed by the decoder module. This reapplication operation enhances the model's focus on critical areas for image segmentation tasks during decoding.

In the second improver, to address the ambiguity present in the segmentation boundaries of retinal vessels, we design a new down-sampling module, called SPD-down, which utilizes a space-to-depth (SPD) layer initially introduced by Sunkara et al. [29]. Similarly, the corresponding encoder-decoder modules are referred to as the SPD_Encoder and SPD_Decoder modules. Compared to traditional *strided convolutional layer and pooling layer*, SPD-Conv effectively restores fine-grained information that is typically lost during conventional convolutional or pooling operations. As illustrated in Figure 3, this process begins with a convolutional layer with a stride of 1 to adjust the channel number from R to D. Assuming the resulting output feature map is $Y \in R^{(D,S,S)}$, it is then divided into four distinct sub-feature maps as follows:

$$M_0 = Y[\ldots, :: 2, :: 2]$$
$$M_1 = Y[\ldots, 1 :: 2, :: 2]$$
$$M_2 = Y[\ldots, :: 2, 1 :: 2]$$
$$M_3 = Y[\ldots, 1 :: 2, 1 :: 2]$$

By using the method of slicing with interval sampling, we obtain four sub feature maps $M_0, M_1, M_2, M_3 \in R^{(D,\frac{S}{2},\frac{S}{2})}$, each of which is reduced to half the size of the original. Then, we concatenate sub feature maps and hence obtain a new feature map that is one-half the original size but boasts a quadrupled channel count. Next, we execute ordinary convolutional operation to reduce the channel dimensions.

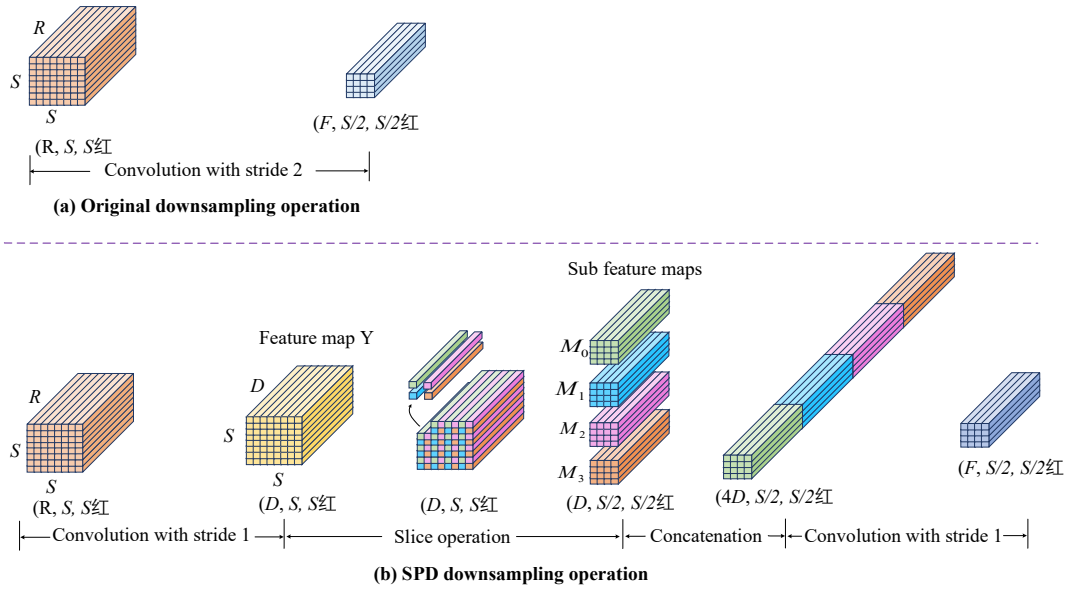

Figure 3. SPD-down module. (a) is our original operation with stride 2. (b) is the new designed SPD down-sampling operation using slicing strategy. We firstly use the convolution layer with stride 1 to change channel numbers, then utilize the slice operation to obtain four sub feature maps. Finally, the convolution layer is reused to decrease channel numbers.

4. Experiments

4.1 DATASETS

We evaluate our method on five public datasets with varying image resolutions and sizes, including DRIVE [30], STARE [31], CHASE_DB1 [32], DCA1 [33] and CHUAC [32]. The details are as follows:

1) DRIVE Dataset: The DRIVE dataset contains 40 fundus images, each measuring 584 ×565 pixels, evenly divided into 20 training images and 20 testing images. These images were captured using Canon CR5 non-mydriatic 3CCD camera, including 7 with abnormal pathology and 33 from health subjects. Each image includes two manual annotations provided by two different observers, and we use the first annotation as the ground truth.
2) STARE Dataset: The STARE dataset consists of 20 colored retinal images, each with dimension of 700 × 605 pixels. In alignment with SCS-Net[34], we adopt a 10-fold cross-validation method to split the training and testing sets.
3) CHASE_DB1 Dataset: The CHASE_DB1 dataset includes 28 fundus images collected from 14 schoolchildren, each with a resolution of 999×960 pixels. Consistent with recent studies [35], [36], we designate the first 20 images for training and the remaining 8 for testing.
4) DCA1 Dataset: The DCA1 dataset contains 130 X-ray coronary angiograms obtained from the Mexican Social Security Institute. Each image, manually annotated by an expert cardiologist, has a resolution of 300 × 300 pixels in PGM grayscale format. The database is divided into 100 images for training and 30 for testing.
5) CHUAC Dataset: The CHUAC dataset consists of 30 coronary angiography images, each originally measuring 189 × 189 pixels, with ground truth masks of 512 × 512 pixel. These masks were manually created by doctors using an image editing tool. As in previous work, the first 20 images are used for training and the remaining 10 for testing. To align the dimensions of the images with their ground truth, we resized the angiography images to 512 × 512 pixels using bilinear interpolation.

4.2 Implementation Details

The proposed framework is implemented on an NVIDIA A100 GPU using the Pytorch 2.0.1 deep learning framework. We employ the Adam optimizer to train the networks, with the learning rate and weight decay set to 0.0001 and 0.00001, respectively. Moreover, we utilize binary cross-entropy (BCE) as the loss function and adopt a cosine annealing learning rate scheduler. Specifically, to conserve computational resources and augment the data, we extract image patches with a 48 × 48 sliding window using the unfold function. In this study, we evaluate the model's segmentation performance by comparing the segmentation results to the ground truths using several widely used metrics, including the intersection over union (IoU), accuracy (Acc), sensitivity (Sen), specificity (Spe), F1 score (F1) and area under the curve (AUC). These criteria are defined as follows:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad Sen = \frac{TP}{TP+FN} \quad Spe = \frac{TN}{TN+FP} \quad F1 = \frac{2TP}{2TP+FP+FN} \quad IoU = \frac{TP}{TP+FP+FN} \quad (9)$$

For medical segmentation maps, pixels are divided into two categories: vascular pixels (those with blood flow) and nonvascular pixels (those without blood flow). The classification of each image results

in four potential outcomes: vascular pixels correctly identified as vessels (True Positive, TP), non-vascular pixels correctly identified as non-vessels (True Negative, TN), nonvascular pixels incorrectly predicted as vessels (False Positive, FP), vascular pixels incorrectly predicted as non-vessels (False Negative, FN).

4.3 Comparisons with the State-of-The-Art Methods

4.3.1 Quantitative Results

To evaluate the performance of our designed C^2AM-Unet, we conduct series experiments against state-of-the-art methods, including TransUNet [12], DE-DCGCN-EE [10], PLVS-Net [12], DPF-Net [12], MedSegDiff [15], GT-DLA-dsHFF [11] and DPL-GFT-EFA [14]. Also, we compare our framework with classic methods, such as U-Net [1], UNet++ [2] and Attention U-Net [3]. To demonstrate the advantages of our approach more effectively, we conduct a comprehensive evaluation incorporating both qualitative and quantitative analyses across five public benchmarks. Comparison results for the DRIVE and STARE datasets are presented in Table I, while results for the CHASE_DB1 dataset are shown in Table II. Additionally, results for DCA1 and CHUAC, two coronary angiography datasets, are provided in Table III.

TABLE I

COMPARISON WITH THE STATE-OF-THE-ART METHODS ON DRIVE AND STARE

Methods	DRIVE						STARE					
	ACC (%)	Sen (%)	Spe (%)	AUC (%)	F1 (%)	IoU (%)	ACC (%)	Sen (%)	Spe (%)	AUC (%)	F1 (%)	IoU (%)
U-Ne t[1]	96.78	80.57	98.33	98.25	81.41	68.64	97.30	78.50	98.84	98.75	81.18	68.56
UNet++ [2]	96.79	78.91	98.50	98.25	81.14	68.27	97.34	79.09	98.83	98.84	81.50	69.02
Attention U-Ne t[3]	96.62	79.06	98.31	97.74	80.39	67.21	97.30	78.04	98.87	98.78	81.06	68.39
TransUNet [9]	96.90	82.64	98.27	98.38	82.35	70.30	97.48	83.32	98.62	98.68	83.25	71.45
DE-DCGCN-EE [10]	97.05	83.59[3]	98.26	98.66	82.88	70.77	97.51	84.05	98.61	98.99	83.63	71.87
PLVS-Net [12]	96.79	81.95	98.19	98.32	81.56	68.89	97.28	83.08	98.43	98.71	82.20	69.87
DPF-Net [13]	96.96	83.05	98.30	98.53	82.70	70.54	97.44	83.60	98.56	98.76	83.04	71.17
MedSegDiff [15]	95.61	78.00	97.31	94.32	75.65	60.86	91.22	70.56	92.83	88.42	66.43	50.80
GT-DLA-dsHFF [11]	97.00	**83.77**	98.27	98.66	82.99	70.96	97.60	84.88	98.65	99.13	84.32	72.89
DPL-GFT-EFA [14]	97.05	83.61	98.34	98.69	**83.22**	71.03	97.70	85.31	98.69	99.17	84.74	73.64
Baseline*	97.05	83.13	98.40	98.87	83.07	71.06	97.65	87.39	98.67	99.42	86.75	76.70
Improver-1	97.06	82.77	98.46	98.88	83.09	71.10	97.68	80.61	98.85	99.43	86.73	76.66
Improver-2	97.04	83.30	98.38	98.87	83.06	71.06	97.71	85.55	**98.92**	99.44	86.83	76.80
C^2AM-Unet	**97.08**	82.37	**98.52**	**98.89**	83.11	**71.12**	**97.72**	**88.01**	98.68	**99.47**	**87.18**	**77.35**

The bold fonts represent the best results and the * denotes the retrained baseline results.

For the DRIVE dataset, the proposed C^2AM-Unet achieves ACC, Sen, Spe, AUC, F1 and IoU of 97.08%, 82.37%, 98.52%, 98.89%, 83.11% and 71.12%, respectively. Among these, ACC, Spe, AUC and IoU are the highest metrics compared to the methods listed in Table I. Although the GT-DLA-dsHFF [11] achieves the highest Sen among existing frameworks, its performance on the remaining metrics is significantly lower than that of our architecture. Specially, the Spe, which represents the segmentation accuracy of non-blood vessel pixels, is 0.25% lower than that of C^2AM-Unet. The GT-DLA-dsHFF method employs a deep-shallow hierarchical feature fusion strategy, comparing transformer and dual local attention network to extract features. This approach effectively reduces the sematic gap between deep and shallow features, leading to the highest Sen. However, its use of multiple dual global attention skip structures introduces redundant information, which limits its ability to segment thin vessels accurately. Although the F1 score of C^2AM-Unet is 0.11% lower than that of the DPL-GFT-EFA[14] approach, this trade-off is acceptable given the significant improvements in ACC, Spe, AUC and IoU metrics achieved by C^2AM-Unet. For the other two developed architectures, Improver-1 and Improver-2, they are slightly lower than our C^2AM-Unet, but still demonstrate quite commendable performance in

overall. It is worth highlighting that both methods have secured IoU score that rank just below our approach, surpassing all the other contenders.

Compared to the current methods outlined in TABLE I, our proposed frameworks exhibit superior feature representation capabilities, achieving improvements of 0.02%, 0.62%, 0.05%, 0.05%, 0.43% and 0.65% across all indicators on the STARE dataset. Notably, the Improver-2 secures a strong second place in overall performance, closely following the C^2AM-Unet method, particularly in the ACC, AUC, F1 and IoU metrics. Meanwhile, Improver-1 ranks third in terms of ACC, Spe and AUC indicators. Segmentation methods in [9] and [15] excel at capturing long-range information through transformer-based architectures, however, they overlook coordinate information and fail to adequately enhance blood vessel morphology, especially for thin vessels pixels. The recent DE-DCGCN-EE [10] prioritizes preserving vessel boundary information through a dual encoder designed for edge detection. This approach employs a dynamic channel graph convolutional network that maps image channels to a topological space, synthesizing channel features to tackle the challenge of underutilized channel information. However, its omission of coordinate and global information results in lower Sen and F1 scores, impacting the segmentation accuracy for distinguishing background pixels. The DPF-Net [13] explores abundant fine-grained details and broad contextual information using CNN and RNN network. Nevertheless, it struggles to capture long-range features effectively, limiting its impact on improving segmentation capabilities. Conversely, our proposed method achieves the best performance across nearly all metrics. It leverages dual coordinate attention and transformer-inspired channel attention modules to capture diverse feature expressions. Additionally, the progressive fusion strategy we propose facilitates the seamless integration of global and detailed adjacency relationships, delivering superior segmentation results.

TABLE II

COMPARISON WITH THE STATE-OF-THE-ART METHODS ON CHASEC_DB1 DATASETS

Methods	CHASE_DB1					
	ACC (%)	Sen (%)	Spe (%)	AUC (%)	F1 (%)	IoU (%)
U-Net [1]	97.43	76.50	98.84	98.36	78.98	65.26
UNet++ [2]	97.39	83.57	98.32	98.81	80.15	66.88
Attention U-Net [3]	97.30	83.84	98.20	98.48	79.64	66.17
TransUNet [9]	97.55	82.22	98.58	98.73	80.92	67.98
DE-DCGCN-EE [10]	**97.62**	84.00	98.56	98.98	**82.61**	**70.37**
PLVS-Net [12]	97.42	81.89	98.46	98.60	80.05	66.75
DPF-Net [13]	97.53	83.12	98.50	98.59	80.95	68.03
MedSegDiff [15]	95.98	78.57	97.15	93.60	70.93	55.01
GT-DLA-dsHFF [11]	97.57	83.39	98.52	98.83	81.20	68.35
DPL-GFT-EFA [14]	97.71	79.60	**98.93**	99.01	81.45	68.74
Baseline*	97.39	89.06	97.97	**99.12**	81.14	68.29
Improver-1	97.48	86.96	98.20	98.96	81.27	68.50
Improver-2	97.52	87.07	98.23	99.08	81.54	68.86
C^2AM-Unet	97.33	**89.29**	97.88	99.11	80.84	67.87

The bold fonts represent the best results and the * denotes the retrained baseline results.

TABLE II shows the competitive results on the CHASE_DB1 dataset. Sen of our proposed C^2AM-Unet, which denotes the segmentation performance of vessel pixels that have blood flowing through them, gets the highest metric with 89.29% compared with other approaches. Although the baseline gets the best AUC index with 99.12%, but it sacrifices the segmentation accuracy of foreground pixels, which

represents the robustness of the model for vessel pixels. For the comprehensive AUC indicator, which measures the network's classification effectiveness, out method falls short of the best performance by only 0.01%. Method in [10] archives the highest ACC, F1 and IoU metrics, with values of 97.62%, 82.61% and 70.37%, respectively. However, it suffers from lower Sen, Spe and AUC values, which are crucial to evaluate the network's overall capability, particularly in segmenting thin vessel pixels. Similarly, DPL-GFT-EFA [14] records the highest Spe indicator, but its lower Sen and AUC metrics are less favorable. While individual methods may excel in specific metrics, our proposed framework delivers the most balanced performance across all indicators.

To further illustrate the superior segmentation capabilities of our model, particularly in addressing challenging cases in medical image segmentation, we present a line chart in Figure 4 to evaluate its effectiveness. The x-axis represents the different methods listed in Table I and II, encoded numerically as 1, 2, etc., in top-to-bottom order. The y-axis corresponds to the AUC or IoU metric. This chart, which visually highlights the experimental results, demonstrates that our proposed model consistently delivers the most effective segmentation outcomes across three distinct datasets.

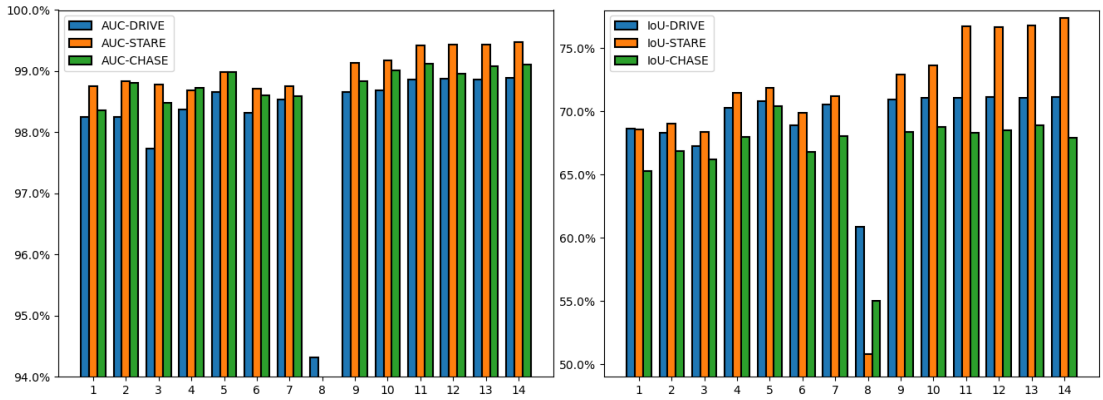

Figure 4. Bar chart comparison results on AUC and IoU metrics across DRIVE, STARE and CHASE_DB1 datasets. The x-axis represents the methods in Table I-II and they are encoded as 1, 2, 3, etc. from top to bottom. The y-axis represents the AUC or IoU performance on different datasets.

From TABLE III, almost all our proposed methods, including Improver-1, Improver-2 and C^2AM-Unet, achieve the best results on DCA1 dataset, where Improver-1 and Improver-2 obtain the best ACC with 97.81%, where Improver-2 gets the best AUC, F1 and IoU with 99.23%, 79.81% and 66.53%, respectively, and where C^2AM-Unet archives the best Spe with 98.91%. Compared with our approach, the baseline network has fewer parameters and faster training speed, due to the lack of coordinate and channel feature attention modules, but fail to meet the expected ACC, Spe, AUC, F1 and IoU metrics. U-Net [1] uses an encoder-decoder structure to capture contextual information and enables accuracy localization, yet, a considerable portion of intricate information is sacrificed, which results in diminished IoU. UNet++ [2] redesigns the skip connections to aggregate the multi-scale semantic scales features. Nevertheless, when it comes to segmenting thin vessels, the performance of this method is inferior to the C^2AM-Unet. Attention U-net [3] and CS-Net [27] considers attention mechanism to highlight salient features but leads to low AUC and F1 metrics. HRNet [49] focuses on high-resolution representation by fusing multi-stage networks but neglects the global information. These methods fail to construct an effective module to improve the segmentation ability of retinal vessels.

TABLE III

COMPARISON WITH THE STATE-OF-THE-ART METHODS ON DCA1 DATASET

Methods	DCA1					
	ACC (%)	Sen (%)	Spe (%)	AUC (%)	F1 (%)	IoU (%)
U-Net [1]	97.58	78.16	98.66	98.79	77.35	63.07
UNet++ [2]	97.61	79.54	98.62	98.84	77.86	63.75
Attention U-net [3]	97.55	79.86	98.53	98.55	77.48	63.24
CS-Ne t[21]	97.63	78.95	98.67	98.89	77.90	63.80
HRNet [37]	97.77	80.07	98.76	98.99	79.19	65.54
Baseline*	97.63	**86.19**	98.28	99.22	79.12	65.58
Improver-1	**97.81**	82.75	98.66	99.22	79.76	66.44
Improver-2	**97.81**	82.77	98.66	**99.23**	**79.81**	**66.53**
C^2AM-Unet	97.77	77.79	**98.91**	99.19	78.30	64.60

The bold fonts represent the best results.

We also explore the effectiveness of C^2AM-Unet on CHUAC dataset. Results in TABLE Ⅳ shows that the Spe index of our framework is the highest, with value of 99.60% and the Improver-2 archives the best Sen with 83.65%. The low contrast and significant variations in tubular structures make task more difficult for accurate targets segmentations. In particular, UNet++ [2] achieves the highest ACC at 98.12%, yet it underperforms in all other metrics compared to our approach. Similarly, while the baseline network attains the top AUC, its Sen and Spe scores are inferior to our results. Although the baseline model excels in F1 score and IoU, it falls short of our method in terms of ACC, Sen, Spe. These findings underscore the commendable generalization capabilities of our proposed method.

TABLE IV

COMPARISON WITH THE STATE-OF-THE-ART METHODS ON CHUAC DATASET

Methods	CHUAC					
	ACC (%)	Sen (%)	Spe (%)	AUC (%)	F1 (%)	IoU (%)
U-Net [1]	97.84	58.81	99.40	96.14	67.68	51.15
UNet++ [2]	**98.12**	66.87	99.37	97.07	73.23	57.77
Attention U-net [3]	98.00	65.26	99.13	95.06	71.54	55.69
HRNet [37]	98.11	74.56	99.06	98.70	75.26	60.33
CS-Net [21]	97.96	67.35	99.18	97.47	71.71	55.89
Baseline	98.03	81.71	98.68	**99.08**	**76.01**	**61.51**
Improver-1	98.00	81.76	98.65	99.06	75.73	61.16
Improver-2	97.80	**83.65**	98.37	99.00	74.47	59.55
C^2AM-Unet	97.48	82.12	**99.60**	98.04	74.30	61.20

The bold fonts represent the best results.

4.3.2 Qualitative Results

The complex architecture of blood vessels, poor contrast in retinal imagery, and localized lesions in pathological fundi, such as hemorrhagic spots and structural deformities, pose significant challenges for accurately segmenting retinal blood vessels. These artifacts are often misclassified as vessel pixels, adversely affecting segmentation performance. To visually demonstrate the effectiveness of our C^2AM-Unet, segmentation results are presented in Figure 5. Due to space limitations, we focus on comparing the baseline network results with those of our proposed methods.

In terms of visual performance, Figure 5 shows that the baseline method struggles to accurately detect the delicate structures of fine vessels. Additionally, the baseline frequently produces under-segmented regions, leading to false segmentations, as indicated by the green arrows. This limitation arises from an

inadequate feature combination strategy, which fails to effectively integrate deep and surface-level features within the network, reducing its ability to suppress external noise. When blood vessels are poorly distinguishable due to low contrast, even with preprocessing enhancements to the original vessel images, the final segmentations remain suboptimal. For example, in the third row of Figure 5, certain areas are incorrectly segmented due to similarities in texture between the vessels and the background. Furthermore, it is even difficult to segment thick blood vessels as the fourth row in Figure 5. Evidenced in the first, second and fifth rows that more thin vessels are misclassified as background pixels.

In contrast, by incorporating dual coordinate and channel features, our proposed methods demonstrate reliable performance in addressing class imbalance and consistently outperform other approaches across most datasets. The comprehensive visualizations highlight framework's exceptional ability to recognize intricate vessel details, while analysis across different datasets confirms that our method delivers consistently stable segmentation results, further validating its robustness and effectiveness.

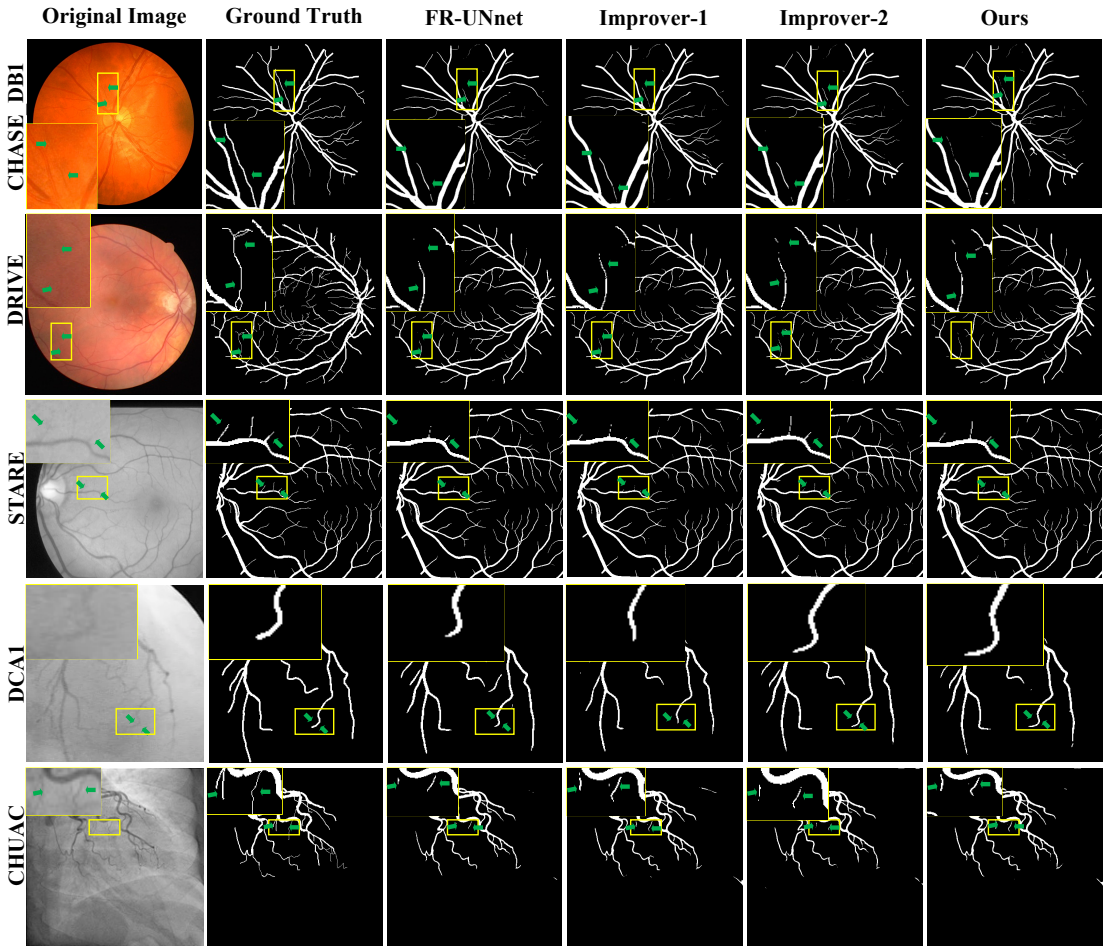

Figure 5. Visualization segmentation results on five datasets. The yellow-boxed areas identify instances where the proposed methods excel over FR-Unet method by making more efficient use of long-distance dependencies. Green arrows ……

4.4 Ablation Study

To substantiate the efficiency of each component of our proposed C^2AM-Unet, we conduct ablation

experiments on the widely recognized DRIVE and STARE datasets for verification. The symbols used in ablation studies are summarized in Table V and Table VI-VII show the quantitative results. We also visualize the detailed segmentation results to display the superiority of our proposed algorithm in Figure 6-7. The contribution of each component is discussed as follows.

4.4.1 Network Configurations

To ease the subsequent explorations, the network configurations which are part of the ablation experiment are firstly introduced. Experiments are all conducted in the same configuration environment and the FR-Unet [28] is utilized as our baseline framework. In this context, a total of ten distinct network setups have been created for the purpose of ablation studies as detailed below.

(a) Baseline: FR-Unet network. (b) Baseline + D_L + D_L: The architecture includes two D_L modules based on baseline network, namely, Improver-1 method. (c) Baseline + D_L+ CH_A: It means that the architecture is composed of one D_L module and one CH_A module based on baseline network. (d) Baseline + D_L + D_L + SPD_down: It means that the architecture is composed of two D_L modules and one SPD_ down module based on baseline network, namely, Improver-2 method. (e) Baseline + D_R / CH_A: It represents that the D_R and the CH_A modules are paralleled designed based on the baseline network. (f) Baseline+ SPD_down_D_L + D_R/ CH_A: It represents that the network consists of SPD_down_D_L, D_R and CH_A components based on the baseline network. (g) Baseline + Down _D_L + D_R/ CH_A: It presents that the network consists of Down _D_L, D_R and CH_A components based on the baseline network. (h) Baseline + Out_D_L+ D_R/ CH_A: It represents that the network consists of Out_D_L, D_R and CH_A components based on the baseline network. (i) Baseline + D_L + D_L + SPD_down + Out_weight: It represents that the network consists of two D_L, one SPD_ down and one Out_weight modules based on the baseline network. (j) Baseline + D_L + D_R / CH_A: It represents that the network consists of D_L, D_R and CH_A modules based on the baseline network , namely, C^2AM-Unet.

TABLE V. List of abbreviations that used in ablation studies

Notations	Explanation
/	Parallel connection
D_L	The left part of DGCA
D_R	The right part of DGCA
SPD_down	The strided convolution is replaced by non-strided convolution
CH_A	The channel attention of stage 3
Down_D_L	The D_L module is concatenated to the normal down-sampling operation.
Out_weight	Outputs are summed with weights
SPD_ down _D_L	The D_L module is concatenated to the SPD_down down-sampling operation.
Out_D_L	The D_L module is concatenated to the output layer.
Baseline+D_L+ D_L	The architecture includes two D_L modules based on baseline network.
Improver-1	
Baseline + D_L+ CH_A	It means that the architecture is composed of one D_L module and one CH_A module based on Baseline network.
Baseline + D_L + D_L+ SPD_down	It means that the architecture is composed of two D_L modules and one SPD_ down module based on baseline network.
Improver-2	
Baseline + D_R/ CH_A	It represents that the D_R and the CH_A modules are paralleled designed based on the baseline network based on the baseline network.
Baseline + SPD_down_D_L + D_R/ CH_A	It represents that the network consists of SPD_ down _D_L, D_R and CH_A components based on the baseline network.
Baseline + Out_D_L + D_R + CH_A	It represents that the network consists of Out_D_L, D_R and CH_A components based on the baseline network.
Baseline + D_L + D_L +	It represents that the network consists of two D_L, one

SPD_down + Out_weight	SPD_down and one Out_weight modules based on the baseline network.
Baseline + D_L+ D_R/ CH_A	It represents that the network consists of D_L, D_R and CH_A modules based on the baseline network , namely, C^2AM-Unet.

4.4.2 Effect of the D_L module

To verify the effect of direction-sensitive coordinate attention, the left part of DGCA module, we conduct experiments on DRIVE and STARE datasets in TABLE VI, comparing network (e) and (j). For DRIVE dataset, as can be shown, the Sen, AUC, F1, and IoU metrics are increased by 0.29%, 0.01%, 0.04% and 0.04%, respectively. For STARE dataset, the ACC, Sen, AUC, F1 and IoU indexes are increased to 97.72%, 88.01%, 99.47%, 87.18% and 77.35%, respectively. The network (j) outperformers network (e) on Sen index, owing to heightened sensitivity to thin blood targets. Network (e) excels at capturing global semantic information, but it struggles with the precise segmentation of thick and thin blood vessel edges. In contrast, the multi-feature fusion capability of network (j) leads to enhanced Sen.

To further explore the position influence of D_L module on overall performance, we additionally design the network (h), changing the location of D_L module to the output layer. It can be seen that the results of network (h) are far inferior to those of network (j) and demonstrates that mid-level features are more reliable for localizing blood vessel positions. Meanwhile, comparing (a) with (e), significant improvements are also generated on almost all the metrics for both two datasets when incorporate the transformer-like coordinate attention and channel attention for global contextual information. The use of DGCA module enhance the feature representation performance of the proposed framework.

TABLE VI

ABLATION EXPERIMENTS ON THE DRIVE AND STARE DATASETS.

Methods	DRIVE						STARE					
	ACC (%)	Sen (%)	Spe (%)	AUC (%)	F1 (%)	IoU (%)	ACC (%)	Sen (%)	Spe (%)	AUC (%)	F1 (%)	IoU (%)
(a) Baseline	97.05	83.13	98.40	98.87	83.07	71.06	97.65	87.39	98.67	99.42	86.75	76.70
(c) Baseline + D_L+ CH_A	96.67	88.57	97.46	98.86	82.25	69.87	96.77	74.93	98.90	98.32	80.15	66.94
(e) Baseline + D_R/ CH_A	97.09	82.08	98.55	98.87	83.07	71.08	97.70	85.97	98.87	99.43	86.80	76.77
(h) Baseline + Out_D_L+ D_R/ CH_A	96.99	83.19	98.33	98.53	82.78	70.65	97.60	85.38	98.83	99.23	86.25	75.94
(j) Baseline + D_L+ D_R/ CH_A	97.08	82.37	98.52	98.89	83.11	71.12	97.72	88.01	98.68	99.47	87.18	77.35

To intuitively observe the prediction results of different network configurations (a) to (j), we visualize the colored segmentation maps in Figure 6. Fluorescent blue pixels represent the correctly segmented vessel pixels, while red and yellow pixels indicate incorrectly segmented vessel pixel and background pixels, respectively. Comparing the red box in (a) with (e), fewer yellow pixels are produced. When the D_L module is added to network (e), inaccurately segmented vessel pixels are partially rectified, as shown in the lavender and blue boxes of Figure (j). Furthermore, most vessel pixels are corrected, however, some vessels at the edges of thick ones remain misidentified as yellow pixels.

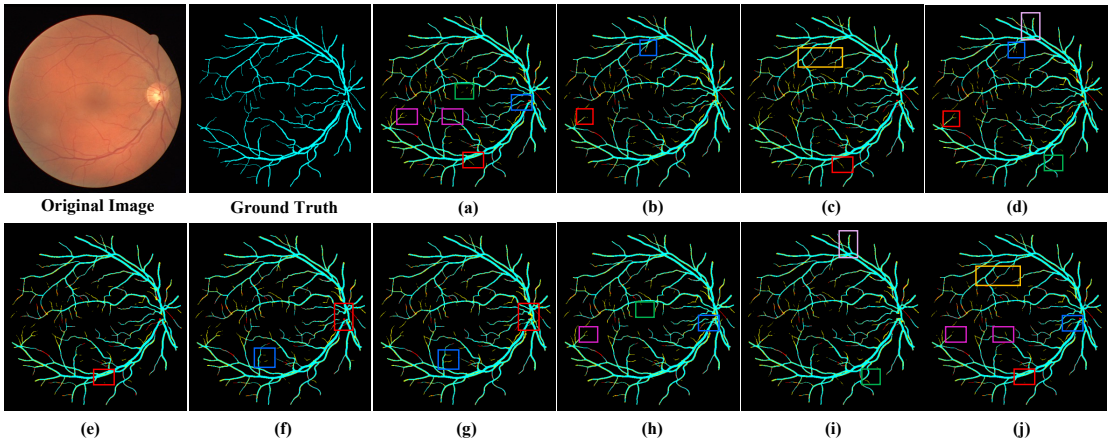

Figure 6. Visualized colored segmentation results of ablation study on DRIVE dataset of network (a)- (j). The fluorescent blue pixels present the correctly segmented vessel pixels, the red and yellow pixels present the incorrectly vessel pixels and background pixels, respectively.

Figure 7 presents the attention maps of network (a) to (j) generated using the Grad-CAM [38] technique. The visualization focuses on the last output layer to evaluate the segmentation performance on the DRIVE dataset. As shown in the fifth column of Figure 7, network (j) accurately pinpoints the locations of thick vessels and demonstrates a robust capacity for detecting them. Network (e) pays more attention to thick vessels and obtains accurately segmentation results with higher brightness than network (h). Obviously, network (e) possesses more detailed information by capturing remote dependency between adjacent pixels, which confirms our antecedent analysis. In summary, the incorporation of the proposed D_L component has led to the correct segmentation of almost all pixels, thereby validating the effectiveness of the global feature fusion strategy in enhancing the precision of thick vessel segmentation.

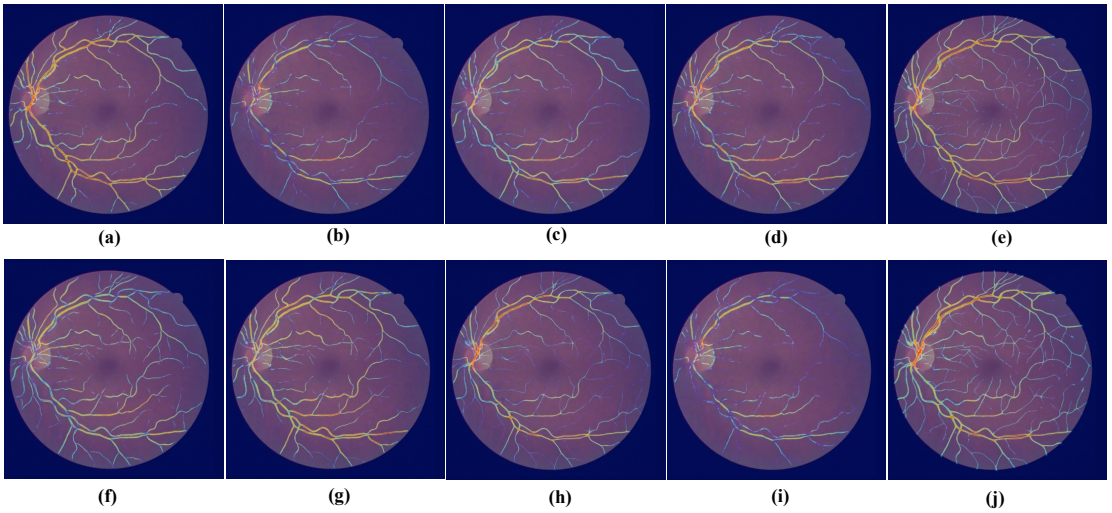

Figure 7. Visualization of attention maps using Grad-CAM of ablation study on DRIVE dataset of network (a)- (j). All attention maps are visualized from the last output layer.

4.4.3 Effect of the D_R module

As evidenced in TABLE VI, the introduction of D_R to the network (c) has contributed to a rise in five indicators both on DRIVE and STARE datasets. The most significant enhancements are observed in IoU and Sen, rising from 69.87% to 71.12% on DRIVE dataset and from 74.93% to 88.01% on STARE

dataset, respectively, demonstrating that our proposed D_R is effective in preserving fundus vessel edges and enhancing segmentation ability of the model. From Figure 6, it can be seen that more thin vessels are correctly predicted, and more vascular pixels are accurate localized because of the fused long-range contextual coordinate information. The instance presented in Figure 7 visually establishes that D_R strengthens the prioritization of attention on slender blood vessels.

4.4.4 Effect of the SPD_down module

Unlike traditional strided convolution layers, the SPD_Conv (non-strided convolution) retains significantly more information, reducing the model's dependence on high-resolution inputs. This enables the model to maintain strong performance even when handling lower-quality inputs and small objects. To evaluate the impact of SPD_Conv, we replace the strided convolution in network (g) with SPD_Conv, observing notable improvements across four metrics on both the DRIVE (Sen: 80.69% → 83.66%, AUC: 98.89%→98.90%, F1: 82.93%→83.15%, IoU: 70.89%→71.18%) and STARE datasets (ACC: 94.71% → 97.62%, Sen: 82.32% → 88.86%, AUC: 94.44% → 99.42%, IoU: 71.41% → 71.76%).

To further validate SPD_Conv performance, we conduct the same experiment on network (b), revealing significant improvements across all six metrics on the STARE dataset and comparable results on the DRIVE dataset. In terms of visual performance, as shown in Figure 6 (red and blue boxes in the second row) and Figure 7, the SPD_Conv-integrated method greatly enhances focus on the tiny pixels of blood vessels while effectively minimizing the misidentification of non-vessel pixels as blood vessels. In summary, SPD_Conv preserves more detailed information compared to the strided convolution down-sampling process. Incorporating the SPD_Conv strategy enhances the model's ability to capture thin targets and generate distinguish features.

TABLE VII

ABLATION EXPERIMENTS OF THE D_L MODULE ON THE DRIVE AND STARE DATASETS

Methods	DRIVE						STARE					
	ACC (%)	Sen (%)	Spe (%)	AUC (%)	F1 (%)	IoU (%)	ACC (%)	Sen (%)	Spe (%)	AUC (%)	F1 (%)	IoU (%)
(b) Baseline + D_L + D_L	97.06	82.77	98.46	98.88	83.09	71.10	97.68	80.61	98.85	99.43	86.73	76.66
(d) Baseline + D_L + D_L + SPD_ down	97.04	83.33	98.38	98.87	83.06	71.06	97.71	85.55	98.92	99.44	86.83	76.80
(f) Baseline + SPD_down _D_L + D_R/ CH_A	97.05	83.66	98.35	98.90	83.15	71.18	97.62	88.86	98.49	99.42	86.80	71.76
(g)Baseline + Down _D_L + D_R/ CH_A	97.11	80.69	98.71	98.89	82.93	70.89	94.71	82.32	99.82	94.44	86.93	71.41
(i)Baseline + D_L + D_L + SPD_ down + Out_weight	96.58	89.32	97.29	99.87	81.99	69.50	96.79	72.94	99.11	98.39	79.82	66.50

4.4.5 Effect of the output strategy module

To generate satisfactory segmentation results, we further explore the effect of different output strategies, deep supervision, and average summation. Here, the deep supervision method in network (i) performs a weighted summation of the last five output layers in the first decoder level, while average summation in network (d) performs an average summation of the outputs. Although deep supervision techniques can enhance segmentation accuracy to a certain degree, their efficacy may be suboptimal when dealing with retinal datasets characterized by class imbalance and limits the capacity of the framework.

In terms of visual performance in Figure 6, we observe that most of the blood vessels are now correctly segmented, with a small fraction mistakenly identified as red pixels, which, to some extent, could well verify the effectiveness of the average summation strategy on the blur vessel edge. The intensified

attention map in Figure 7, evidence that the average summation technique encompasses a dual focus, targeting both macrovascular and microvascular structures within the vasculature. In conclusion, the above complexity ablation studies and visualized segmentation results prove that our designed architectures are effective and outstanding.

4.4.6 Statistical Analysis

To further evaluate performance from a statistical perspective, a boxplot is created to compare the most representative method of (b), (d), (e), (g) and (j). Figure 8 presents a six-number summary of the DRIVE dataset, including the minimum, first quartile (Q1), median (Q2), third quartile (Q3), mean, and maximum values for three metrics. Due to the excessive number of negative samples in the fundus images and to present the results more intuitively, the TN metric is not displayed.

For the TP metric, C^2AM-Unet archives the highest mean, median, minimum, and Q3 values, demonstrating its strong ability to correctly segment targets thanks to the rich features extracted. Additionally, it obtains the lowest mean, median, minimum, maximum, Q1, and Q3 values for FN compared to all other methods. Compared to method (a), method (d) achieves lower FN and higher TP values across nearly all statistical metrics, further demonstrating that coordinate attention is better at accurately segmenting non-vascular pixels. Figure 8 also shows that methods (e), (g) and (b) achieve lower FP value across all six statistical metrics, indicating a stronger ability to segment vascular pixels accurately.

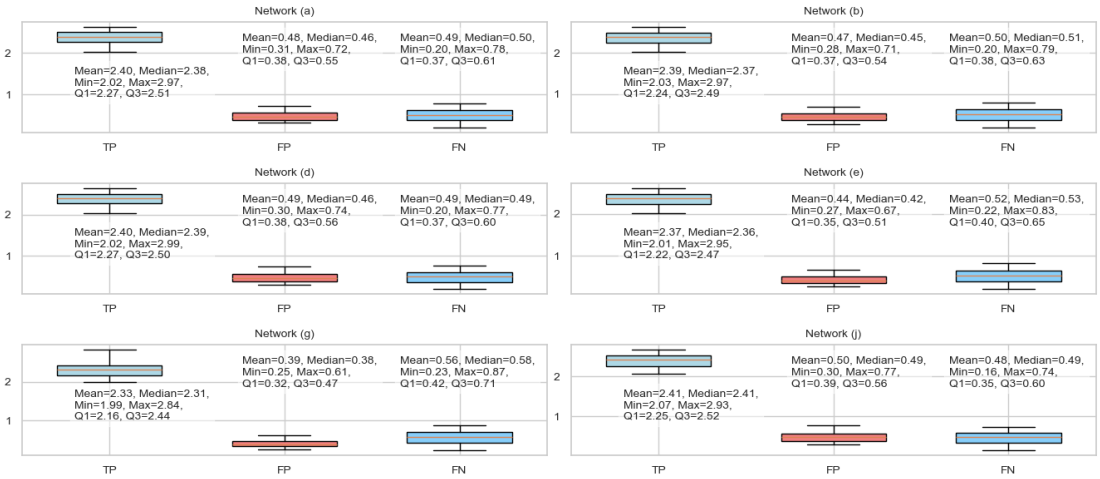

Figure 8. Static analysis on DRIVE dataset of TP, FP, FN metrics for different methods. The y-axis represents the pixel number of the image (in ten thousand) and the x-axis denotes the evaluation metrics.

CONCLUSION

In this paper, we propose a novel coordinate and channel attention mixing flexible architecture for retinal vessel segmentation, named C^2AM-Unet, to further fuse the local and global information of retinal vessels. Extensive qualitative and quantitative experiments on public datasets validate the effectiveness and advanced performance of C^2AM-Unet compared to state-of-the-art segmentation algorithms. The main contributions of this research are summarized as follows: (i) We introduce a groundbreaking deep segmentation network, C^2AM-Unet, specifically designed for the task of blood vessel segmentation. (ii) An innovative dual global coordinate attention (DGCA) fusion module is proposed for efficient feature representation. This module captures both long-range contextual information and local spatial

relationships between adjacent pixels.(iii) Two alternative methods are designed to efficiently detect small targets and accurately segment even the finest blood vessels. (iv) Experiments results on the DRIVE, CHASEDB1, STARE, DCA1 and CHASE_DB1 datasets demonstrate that C^2AM-Unet outperforms cutting-edge approaches by achieving precise segmentation of thick vessels and detecting a greater number of fine features in thin blood vessels.

ACKNOWLEDGMENT

This work was supported by the Science and Technology Development Fund of Macao under Grant No. 0002/2024/RIA1.

References

[1] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, *arXiv*: arXiv:1505.04597. http://arxiv.org/abs/1505.04597

[2] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation, *arXiv*: arXiv:1912.05074.

[3] O. Oktay, Attention U-Net: Learning Where to Look for the Pancreas, arXiv.org: https://arxiv.org/abs/1804.03999v3

[4] F. Milletari, N. Navab, S.-A. Ahmadi, V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, https://arxiv.org/abs/1606.04797v1

[5] A. Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv.org. https://arxiv.org/abs/2010.11929v2

[6] Z. Liu et al., Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, *arXiv*: arXiv:2103.14030. http://arxiv.org/abs/2103.14030

[7] X. Huang, Z. Deng, D. Li, X. Yuan, MISSFormer: An Effective Medical Image Segmentation Transformer, *arXiv*: arXiv:2109.07162. http://arxiv.org/abs/2109.07162

[8] R. Azad, R. Arimond, E. K. Aghdam, A. Kazerouni, D. Merhof, DAE-Former: Dual Attention-guided Efficient Transformer for Medical Image Segmentation, *arXiv*: arXiv:2212.13504. doi: 10.48550/arXiv.2212.13504.

[9] J. Chen et al., TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, *arXiv*: arXiv:2102.04306. doi: 10.48550/arXiv.2102.04306.

[10] Y. Li, Y. Zhang, W. Cui, B. Lei, X. Kuang, T. Zhang, Dual Encoder-Based Dynamic-Channel Graph Convolutional Network With Edge Enhancement for Retinal Vessel Segmentation, *IEEE Trans. Med. Imaging*, 41, 8, 1975–1989, doi: 10.1109/TMI.2022.3151666.

[11] Y. Li et al., Global Transformer and Dual Local Attention Network via Deep-Shallow Hierarchical Feature Fusion for Retinal Vessel Segmentation, *IEEE Trans. Cybern.*, 53, 9, 5826–5839, 9,2023, doi: 10.1109/TCYB.2022.3194099.

[12] M. Arsalan, T. M. Khan, S. S. Naqvi, M. Nawaz, I. Razzak, Prompt Deep Light-Weight Vessel Segmentation Network (PLVS-Net), *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 20, 2, 1363–1371, 3,2023, doi: 10.1109/TCBB.2022.3211936.

[13] J. Li, G. Gao, L. Yang, G. Bian, Y. Liu, DPF-Net: A Dual-Path Progressive Fusion Network for Retinal Vessel Segmentation, *IEEE Trans. Instrum. Meas.*, 72, 1–17, 2023, doi: 10.1109/TIM.2023.3277946.

[14] Y. Li et al., Diffusion Probabilistic Learning With Gate-Fusion Transformer and Edge-Frequency

Attention for Retinal Vessel Segmentation, *IEEE Trans. Instrum. Meas.*, 73, 1–13, 2024, doi: 10.1109/TIM.2024.3420264.

[15] J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin，Y. Xu, MedSegDiff-V2: Diffusion based Medical Image Segmentation with Transformer, *arXiv*: arXiv:2301.11798. doi: 10.48550/arXiv.2301.11798.

[16] D. Dai et al., I 2 U-Net: A dual-path U-Net with rich information interaction for medical image segmentation, *Med. Image Anal.*, 97, 103241, 10,2024, doi: 10.1016/j.media.2024.103241.

[17] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille，L.-C. Chen, Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation, *Computer Vision – ECCV 2020*, 12349, A. Vedaldi, H. Bischof, T. Brox，J.-M. Frahm，Lecture Notes in Computer Science, vol. 12349. , Cham: Springer International Publishing, 2020, 108–126. doi: 10.1007/978-3-030-58548-8_7.

[18] J. Hu, L. Shen，G. Sun, Squeeze-and-Excitation Networks, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, 6,2018, 7132–7141. doi: 10.1109/CVPR.2018.00745.

[19] S. Woo, J. Park, J.-Y. Lee，I. S. Kweon, CBAM: Convolutional Block Attention Module, *Computer Vision – ECCV 2018*, 11211, V. Ferrari, M. Hebert, C. Sminchisescu，Y. Weiss，Lecture Notes in Computer Science, vol. 11211, Cham: Springer International Publishing, 2018, 3–19. doi: 10.1007/978-3-030-01234-2_1.

[20] H. Huang et al., Combinatorial CNN-Transformer Learning with Manifold Constraints for Semi-supervised Medical Image Segmentation, *Proc. AAAI Conf. Artif. Intell.*, 38, 3, 2330–2338, 3,2024, doi: 10.1609/aaai.v38i3.28007.

[21] L. Mou et al., CS-Net: Channel and Spatial Attention Network for Curvilinear Structure Segmentation, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, 11764, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap，A. Khan，Lecture Notes in Computer Science, vol. 11764. , Cham: Springer International Publishing, 2019, 721–730. doi: 10.1007/978-3-030-32239-7_80.

[22] C. Chen, L. Yu, S. Min，S. Wang, MSVM-UNet: Multi-Scale Vision Mamba UNet for Medical Image Segmentation, *arXiv*: arXiv:2408.13735. http://arxiv.org/abs/2408.13735

[23] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell，S. Xie, A ConvNet for the 2020s, *arXiv*: arXiv:2201.03545. 2024 年 9 月 22 日. http://arxiv.org/abs/2201.03545

[24] J. Park, S. Woo, J.-Y. Lee，I. S. Kweon, BAM: Bottleneck Attention Module, *arXiv*: arXiv:1807.06514. doi: 10.48550/arXiv.1807.06514.

[25] Q. Hou, D. Zhou，J. Feng, Coordinate Attention for Efficient Mobile Network Design, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, 6,2021, 13708–13717. doi: 10.1109/CVPR46437.2021.01350.

[26] S. Woo, J. Park, J.-Y. Lee，I. S. Kweon, CBAM: Convolutional Block Attention Module, *arXiv*: arXiv:1807.06521. doi: 10.48550/arXiv.1807.06521.

[27] W. Ding et al., FTransCNN: Fusing Transformer and a CNN based on fuzzy logic for uncertain medical image segmentation, *Inf. Fusion*, 99, 101880, 11,2023, doi: 10.1016/j.inffus.2023.101880.

[28] W. Liu et al., Full-Resolution Network and Dual-Threshold Iteration for Retinal Vessel and Coronary Angiograph Segmentation, *IEEE J. Biomed. Health Inform.*, 26, 9, 4623–4634, 9,2022, doi: 10.1109/JBHI.2022.3188710.

[29] R. Sunkara，T. Luo, No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects, 2022 年 8 月 7 日, *arXiv*: arXiv:2208.03641. 2024 年 9 月 22 日. http://arxiv.org/abs/2208.03641

[30] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, B. Van Ginneken, Ridge-Based Vessel Segmentation in Color Images of the Retina, *IEEE Trans. Med. Imaging*, 23, 4, 501–509, 4,2004, doi: 10.1109/TMI.2004.825627.

[31] A. D. Hoover, V. Kouznetsova, M. Goldbaum, Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response, *IEEE Trans. Med. Imaging*, 19, 3, 203–210, 3,2000, doi: 10.1109/42.845178.

[32] A. Carballal et al., Automatic multiscale vascular image segmentation algorithm for coronary angiography, *Biomed. Signal Process. Control*, 46, 1–9, 9,2018, doi: 10.1016/j.bspc.2018.06.007.

[33] F. Cervantes-Sanchez, I. Cruz-Aceves, A. Hernandez-Aguirre, M. A. Hernandez-Gonzalez, S. E. Solorio-Meza, Automatic Segmentation of Coronary Arteries in X-ray Angiograms using Multiscale Analysis and Artificial Neural Networks, *Appl. Sci.*, 9, 24, 5507, 12,2019, doi: 10.3390/app9245507.

[34] H. Wu, W. Wang, J. Zhong, B. Lei, Z. Wen, J. Qin, SCS-Net: A Scale and Context Sensitive Network for Retinal Vessel Segmentation, *Med. Image Anal.*, 70, 102025, doi: 10.1016/j.media.2021.102025.

[35] S. Zhang et al., Attention Guided Network for Retinal Image Segmentation, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, 11764, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, A. Khan, Lecture Notes in Computer Science, vol. 11764. , Cham: Springer International Publishing, 2019, 797–805. doi: 10.1007/978-3-030-32239-7_88.

[36] W. Wang, J. Zhong, H. Wu, Z. Wen, J. Qin, RVSeg-Net: An Efficient Feature Pyramid Cascade Network for Retinal Vessel Segmentation, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 12265, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, L. Joskowicz, Lecture Notes in Computer Science, vol. 12265. , Cham: Springer International Publishing, 2020, 796–805. doi: 10.1007/978-3-030-59722-1_77.

[37] K. Sun, B. Xiao, D. Liu, J. Wang, Deep High-Resolution Representation Learning for Human Pose Estimation, *arXiv*: arXiv:1902.09212. doi: 10.48550/arXiv.1902.09212.

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, *arXiv*: arXiv:1610.02391. doi: 10.48550/arXiv.1610.02391.