SGEA-Net: Stimulus-Guided Gating with Multi-Scale Edge-Enhanced Aggregation for Retinal Vessel Segmentation

1st Zhiyuan Liu
School of Artificial Intelligence and Computer Science
Nantong University
Nantong, China
2744795165@qq.com

2nd Ziying Lu
School of Artificial Intelligence and Computer Science
Nantong University
Nantong, China
1004965779@qq.com

3rd Yifan Jiang*
State Key Laboratory of Internet of Things for Smart City
University of Macau
Macao, China
yc27495@umac.mo

4th Shu Jiang*
School of Artificial Intelligence and Computer Science
Nantong University
Nantong, China
Jshmjs45@ntu.edu.cn

***Abstract*—Retinal vessel segmentation plays a critical role in computer-aided screening and diagnosis of ocular diseases such as diabetic retinopathy. However, existing methods often suffer from discontinuous vessel boundaries, insufficient multi-scale semantic fusion, and weak channel-wise detail preservation, especially for thin capillaries. To address these challenges, we propose SGEA-Net, a novel network that integrates Multi-scale Edge-Enhanced Aggregation with Stimulus-Guided Gating for retinal vessel segmentation. Specifically, an Edge Attention Enhancement (EAE) block is designed to strengthen boundary continuity by exploiting dual-scale gradient cues while suppressing background noise. A Multi-scale Global Aggregation (MGA) block aligns hierarchical features into a unified resolution and enables adaptive global–local fusion across scales. In addition, a Stimulus-guided Adaptive Gated Bottleneck (SAGB) block dynamically recalibrates channel responses, improving the balance between global context and local vessel peaks. Experiments on three public benchmarks (DRIVE, CHASEDB1, and STARE) demonstrate that SGEA-Net consistently outperforms representative baseline methods, while identifying more continuous and detailed vessel structures. This work provides an effective and extensible solution for robust retinal vessel segmentation in fundus image analysis.**

***Index Terms*—retinal vessel segmentation, edge attention enhancement, dual-scale Sobel, gated bottleneck**

I. Introduction

The retina is one of the most information-dense tissues in the human eye, characterized by a highly branched and topologically complex vascular network [1]. Vessel diameters may gradually decrease from tens of micrometers to only a few micrometers. In practice, fundus images often suffer from low contrast, strong noise, blurred boundaries, and uneven illumination, due to imaging conditions and individual variations, making accurate retinal vessel segmentation particularly challenging. Retinal vascular diseases such as diabetic retinopathy, glaucoma, and macular degeneration frequently manifest early structural changes in vessel morphology and density. Therefore, reliable vessel segmentation is a critical prerequisite for early screening and clinical diagnosis [2], [3]. Extracting clear, continuous, and well-defined vessel structures from complex retinal backgrounds remains a fundamental yet difficult task.

Early retinal vessel segmentation methods relied primarily on traditional image processing and machine learning techniques, with handcrafted feature engineering. Representative approaches include vessel enhancement filtering (e.g., Frangi filter) [4], thresholding and edge-based operators (e.g., Sobel and Canny) [5], and region-based contour extraction using active contours and level sets [6]. With the development of machine learning, conventional supervised models, such as support vector machines (SVM) [7] and random forests (RF) [8], have also been explored with manually designed descriptors (e.g., texture, orientation, and intensity). Although effective for major vessel trunks, these methods are sensitive to parameter tuning, exhibit limited generalization, and often fail to preserve thin vessel continuity under noise, leading to fragmented boundaries.

In recent years, convolutional neural networks (CNNs) [9] have become dominant in medical image segmentation due to their strong ability to learn discriminative local representations. Among them, U-Net [10] serves as a standard backbone for retinal vessel segmentation. Meanwhile, Transformer-based architectures, originally developed in natural language processing [11], have demonstrated superior capability in modeling long-range dependencies through self-attention. Vision Transformers (ViT) [12] further extend this paradigm to visual

This work was supported by the National Natural Science Foundation of China (Grant No. 62406153).
*Corresponding author

tasks by partitioning an image into a sequence of tokens and capturing global contextual interactions. However, their high computational cost limits direct applicability in retinal segmentation scenarios.

To balance efficiency and global modeling, hybrid CNN-Transformer frameworks have been proposed. For example, TransU-Net [13] integrates a ViT module into a CNN encoder-decoder architecture, improving boundary continuity. Swin-UNet [14] adopts shifted-window attention to reduce complexity while maintaining translation invariance. Despite these advances, existing methods still struggle with (i) discontinuous and easily overlooked thin vessels, (ii) insufficient multi-scale semantic interaction, and (iii) unstable channel-wise feature fusion. These limitations highlight that boundary enhancement, effective scale aggregation, and adaptive channel regulation remain open challenges for robust retinal vessel segmentation.

To address these issues, we propose **SGEA-Net**, a novel framework that integrates *Stimulus-Guided Gating* with *Multi-scale Edge-Enhanced Aggregation*. By incorporating boundary-guided feature refinement, unified multi-scale semantic fusion, and adaptive channel-wise gating, SGEA-Net significantly improves thin-vessel continuity and segmentation robustness. The main contributions of this work are summarized as follows:

- We propose an *Edge Attention Enhancement (EAE) block* that explicitly incorporates dual-scale gradient stimuli into gated feature modulation, effectively strengthening thin-vessel boundaries while suppressing noise-induced fragmentation.
- We develop a *Multi-scale Global Aggregation (MGA) block* that unifies hierarchical features at a common resolution and enables scale-adaptive global–local fusion through lightweight attention-guided interaction, facilitating coherent vessel modeling across scales.
- We introduce a *Stimulus-guided Adaptive Gated Bottleneck (SAGB) block* that dynamically recalibrates channel responses via learnable gating, enhancing robustness and discriminability under challenging conditions such as low contrast and uneven illumination.

The source code, implementation details, and experimental data supporting the findings of this study are publicly available at https://github.com/laonanrenbangbangzhu/SGEA-Net.git.

II. Related Work

A. CNN-Based Medical Image Segmentation

Medical image segmentation is a fundamental component of computer-aided diagnosis and clinical decision-making. With the advent of fully convolutional networks [15], early deep learning-based segmentation methods relied primarily on convolutional neural networks (CNNs). Among these, U-Net has become a de facto backbone due to its encoder-decoder architecture and effective skip connections, and has been widely adopted in medical image analysis. Based on the U-Net, numerous variants have been proposed. For instance, Peng et al. [16] introduced U-Net v2, which progressively injects coarse-to-fine features via Hadamard-product fusion, mitigating information loss and improving segmentation accuracy. Pang et al. [17] proposed GA-UNet, which replaces standard convolutions with lightweight bottleneck layers and incorporates Convolutional Block Attention Module (CBAM) to emphasize informative responses across both channel and spatial dimensions, achieving strong performance with fewer parameters. Despite their effectiveness in capturing local structures, CNN-based methods remain limited in modeling long-range dependencies and global contextual interactions, and make insufficient use of pixel-wise positional information.

B. Transformer and CNN-Transformer Hybrid Segmentation

Recently, the Vision Transformer (ViT) paradigm has gained increasing prominence in computer vision, as self-attention enables effective modeling of global contextual relationships and long-range dependencies. However, the substantial computational overhead of vanilla Transformers often restricts their direct applicability in resource-constrained medical scenarios. To alleviate the limited global modeling capability of CNNs while maintaining efficiency, researchers have increasingly explored hybrid CNN-Transformer architectures. Chen et al. [18] proposed CTAUNet, which employs U-Net for pixel-level localization while embedding a Swin Transformer to capture global dependencies, thereby improving multi-scale consistency and structural continuity. Luo et al. [19] introduced PA-Net, integrating lightweight parallel Transformers with adaptive feature fusion: the CNN branch extracts multi-scale local representations, whereas the Transformer branch complements long-range interactions, improving recall while reducing computational complexity.

C. Vessel Modeling and Connectivity Enhancement

Retinal vessel segmentation remains highly challenging due to extremely thin capillaries, complex reticular topology, and strong sparsity. In addition, imaging artifacts such as uneven illumination, glare, macular reflexes, and hemorrhages often resemble vessel-like textures, making fine vessels prone to misclassification. Therefore, improving boundary clarity, preserving connectivity, and modeling multi-scale topology are essential for robust retinal vessel segmentation.

Several recent studies have explored vessel-aware strategies to enhance vessel continuity. Lin et al. [20] extended a CNN-based decoder with self-attention for global dependency modeling and introduced edge gating to modulate multi-scale features element-wise, helping the network focus on vessel patterns and boundaries while suppressing background interference. Li et al. [21] proposed TDCAU-Net, which uses a U-shaped CNN encoder with Transformer blocks and dilated convolutions to enlarge the receptive field and improve coarse-to-fine vessel consistency across scales. Zhang et al. [22] presented TU-Net in a two-stage coarse-to-fine framework, where a CNN–Transformer module generates a coarse segmentation and an LBF block refines boundaries and thin-branch details. Huang et al. [23] designed HiDiffSeg, which combines a lightweight CNN with Sobel-based boundary attention to

sharpen boundaries and reduce fragmentation under heavy noise. Liu et al. [24] introduced HRD-Net and proposed a global aggregation block at the output stage for multi-branch fusion and channel recalibration, further improving vessel continuity through multi-scale information.

Although these methods have made progress in topology modeling, boundary refinement, and multi-scale adaptation, they still struggle to stably preserve fine vessels, efficiently fuse hierarchical features at unified resolutions, adaptively regulate channels under challenging imaging conditions, and maintain a dynamic equilibrium between global and local dimensions across channels.

III. Methods

A. Overview

This section presents the proposed **SGEA-Net** in detail. As shown in Fig. 1(a), SGEA-Net follows a U-Net-style encoder-decoder architecture and consists of three key components: the *Edge Attention Enhancement (EAE) block*, the *Multi-scale Global Aggregation (MGA) block*, and the *Stimulus-guided Adaptive Gated Bottleneck (SAGB) block*.

Given an input fundus image, the EAE block is first applied at the shallow stage to generate an edge-sensitive attention map, providing boundary-enhanced cues for subsequent feature learning to improve segmentation continuity and stability. In the encoder, SAGB blocks are embedded into each of the four downsampling stages to adaptively recalibrate channel-wise responses. The resulting edge attention cues are injected into multi-scale encoder features via element-wise (Hadamard) multiplication, thereby strengthening thin-vessel continuity under noise interference.

In the decoder, hierarchical representations are progressively restored through four upsampling stages. At each stage, skip connections are employed to fuse low-level spatial details with high-level semantic features, while a CBAM-based attention module [17] further refines salient vessel responses. Finally, the MGA block is incorporated at the output stage to align multi-scale decoded features to a unified resolution and perform adaptive global–local feature aggregation.

The detailed formulations of the EAE, SAGB, and MGA blocks are presented in the following subsections.

B. Edge Attention Enhancement (EAE) Block

Retinal vessel images exhibit substantial appearance variations between major trunks and extremely thin branches, especially near vessel boundaries, where foreground pixels occupy only a small proportion. This often leads to fragmented contours and discontinuities at junction regions. Existing boundary-enhancement approaches mainly rely on either explicit edge attention amplification or uncertainty modeling via fuzzy membership functions. However, these strategies remain insufficient for robustly preserving thin-vessel continuity in noisy backgrounds. Motivated by the boundary-attention design in [23], we propose the *Edge Attention Enhancement (EAE) block*, which constructs a dual-scale edge-magnitude stimulus and applies gated modulation to emphasize weak vessel boundaries while suppressing noise. The overall procedure is illustrated in Fig. 1(b).

1) Learnable grayscale projection: Given an input fundus image $x \in \mathbb{R}^{N\times C\times H\times W}$, we first employ a learnable 1×1 convolution to project the multi-channel input into a single-channel grayscale representation $g \in \mathbb{R}^{N\times 1\times H\times W}$:

$$g = \mathrm{Conv}_{1\times 1}(x) = \sum_{c=1}^{C} w_c\, x_c, \tag{1}$$

where w_c denotes the learnable projection weight for the c-th channel, and $C = 3$ is the number of input channels. This operation allows the model to adaptively learn the most discriminative channel combination for edge extraction.

2) Dual-scale Sobel edge magnitude: Thin capillaries are often weak at the native resolution and can be easily overwhelmed by background textures. However, at lower resolutions, noise and high-frequency textures are averaged out, resulting in more robust structural details. Therefore, we compute the Sobel gradient magnitudes at both the original and downsampled scales to obtain complementary edge stimuli:

$$M^{(1)} = \sqrt{\left(G_x^{(1)}\right)^2 + \left(G_y^{(1)}\right)^2 + \varepsilon}, \tag{2}$$

$$M^{(2)} = \sqrt{\left(G_x^{(2)}\right)^2 + \left(G_y^{(2)}\right)^2 + \varepsilon}, \tag{3}$$

where $G_x^{(1)}$ and $G_y^{(1)}$ denote horizontal and vertical Sobel gradients of g, and $G_x^{(2)}$ and $G_y^{(2)}$ are computed from the downsampled map $\tilde{g}$. ε is a small constant for numerical stability.

3) Scale fusion and gated edge attention: The two magnitude maps are concatenated and fused along the scale dimension using a pixel-wise softmax weighting:

$$M = \sum_{k\in\{1,2\}} \pi^{(k)} \cdot M^{(k)}, \qquad \pi^{(k)} = \frac{\exp(M^{(k)})}{\sum_j \exp(M^{(j)})}, \tag{4}$$

yielding a unified edge-magnitude response M. To ensure stable contrast, M is further normalized into $[0, 1]$, denoted as M_{norm}.

Then we apply a parameterized sigmoid gating function to generate the edge attention map:

$$A = \beta + \alpha \cdot \sigma(s \cdot (t - M_{\mathrm{norm}})), \tag{5}$$

where α and β control the amplitude and bias, t is the threshold center, and s adjusts the slope of the soft gate. $\sigma(\cdot)$ denotes the Sigmoid activation. The resulting attention map $A \in [0, 1]^{N\times 1\times H\times W}$ explicitly enhances weak vessel boundaries while suppressing background noise.

The generated attention map is subsequently injected into encoder features via element-wise multiplication, providing boundary-sensitive guidance for downstream segmentation.

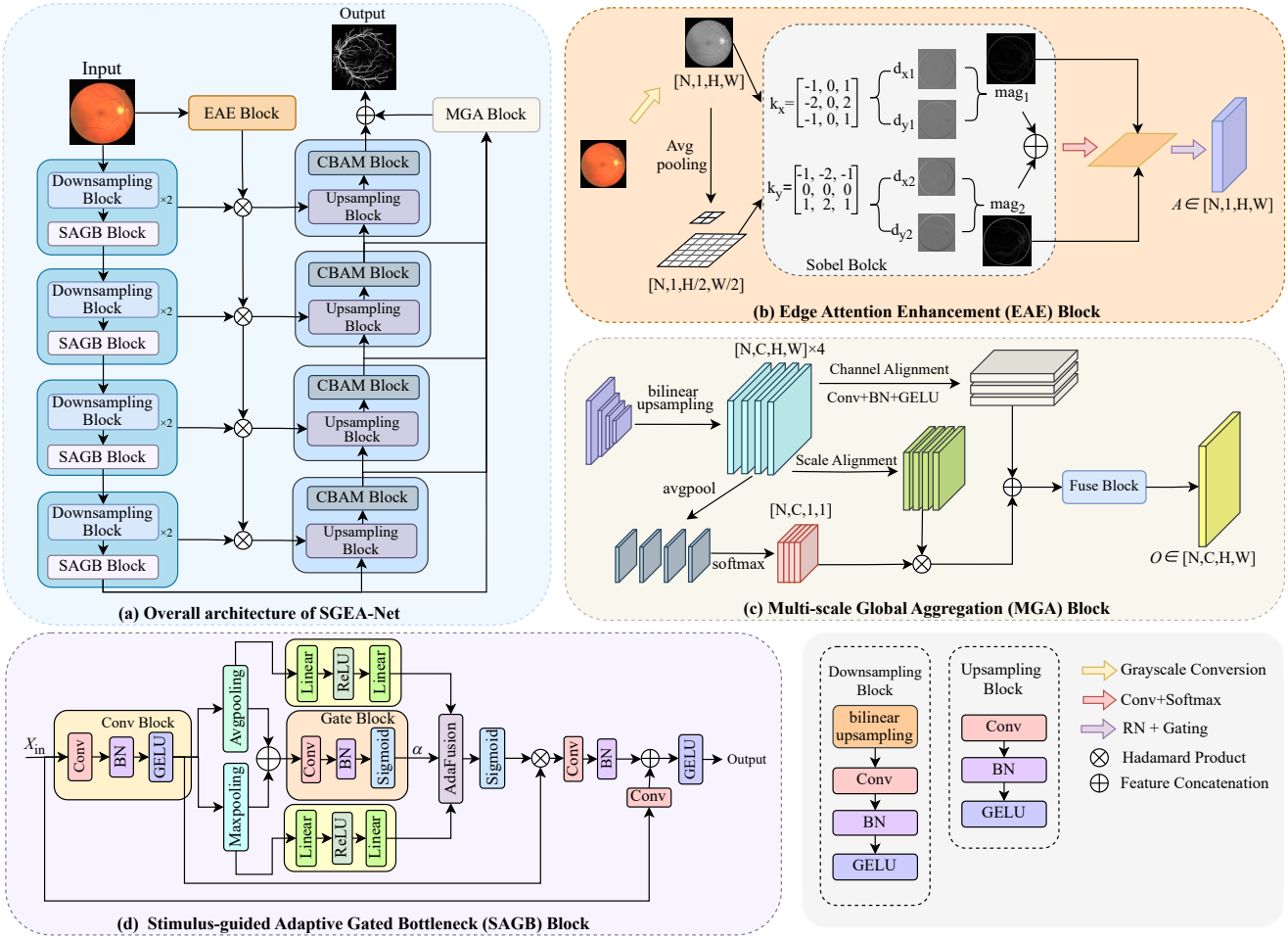

Fig. 1. Overview of the proposed SGEA-Net and its main components: (a) overall architecture of SGEA-Net; (b) structure of the EAE Block; (c) structure of the MGA Block; (d) structure of the SAGB Block. Symbols and operators are illustrated in the legend.

C. Multi-scale Global Aggregation (MGA) Block

Although downsampling operations in the U-Net encoder enlarge the receptive field, they inevitably discard fine spatial details and boundary cues, which are crucial for preserving the continuity of thin vessels. While skip connections partially compensate for this degradation, decoder features from a single scale may still lack sufficient global topological awareness. Therefore, we design the *Multi-scale Global Aggregation (MGA) block* to perform adaptive, pixel-level aggregation across hierarchical decoder features, as illustrated in Fig. 1(c).

Let $\{F^{(l)} \in \mathbb{R}^{N\times C_l\times H_l\times W_l}\}_{l=1}^{L}$ denote the multi-level feature maps extracted from L decoder stages. Each feature is first projected into a unified channel dimension C via a 1×1 convolution, followed by normalization. Then, all features are bilinearly upsampled to the resolution of the last decoder stage, yielding aligned representations $\{\widetilde{F}^{(l)} \in \mathbb{R}^{N\times C\times H\times W}\}_{l=1}^{L}$.

1) Scale-adaptive global aggregation: To enable cross-scale interaction, global average pooling is applied to each aligned feature map to extract channel-wise statistics. These descriptors are passed through a lightweight MLP with a bottleneck of $\max(\lceil C/r\rceil, 8)$ (where r is the reduction ratio), producing per-scale channel scoring vectors $s_{n,l,c}$. Softmax normalization is then applied along the scale dimension, yielding channel-specific attention weights across scales. The aggregated feature map is formulated as:

$$Y_{n,c,i,j} = \sum_{l=1}^{L} \mathrm{Softmax}_l(s_{n,l,c})\ \widetilde{F}^{(l)}_{n,c,i,j}, \tag{6}$$

where $\widetilde{F}^{(l)}$ denotes the aligned feature map at scale l, and $Y_{n,c,i,j}$ represents the scale-adaptive fused backbone feature.

2) Complementary refinement branch: Besides weighted fusion, MGA further introduces a refinement pathway by concatenating aligned multi-scale features and compressing them back to C channels:

$$R = \mathrm{GELU}\Big(\mathrm{BN}\left(w_{\mathrm{cat}} * [\,\widetilde{F}^{(1)}; \cdots ; \widetilde{F}^{(L)}\,]\right)\Big), \tag{7}$$

$$Y = Y + \gamma R, \tag{8}$$

where w_{cat} denotes the fusion weight, γ is a learnable residual scaling coefficient, Y is the output backbone feature map, and the residual injection improves feature completeness and boundary preservation.

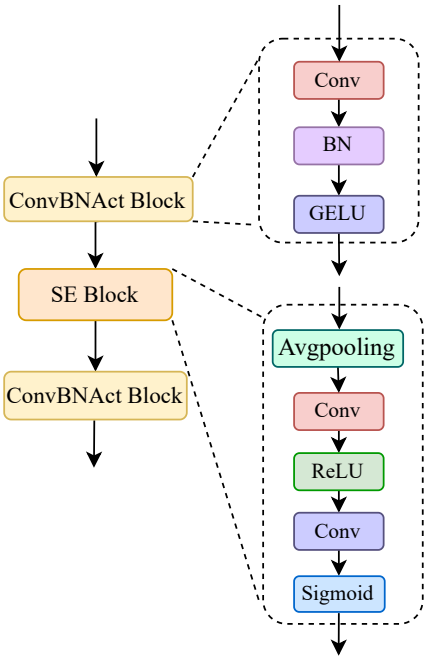

Fig. 2. Structure of the fusion block in the proposed MGA Block.

3) Fusion head with channel attention: Finally, the aggregated feature is refined using a lightweight fusion block, as shown in Fig. 2. As illustrated, the fusion head follows a *ConvBNAct–SE–ConvBNAct* design, where ConvBNAct denotes a 3×3 convolution followed by batch normalization and GELU activation, and SE represents a squeeze-and-excitation bottleneck for channel recalibration. The refinement operation is defined as:

$$\mathcal{F}_{\text{refine}}(Y) = \text{ConvBNAct}\Big(\text{SE}\big(\text{ConvBNAct}(Y)\big)\Big), \tag{9}$$

$$O = w_{\text{head}} * \mathcal{F}_{\text{refine}}(Y), \tag{10}$$

where w_{head} is a 1×1 output projection, and $O \in \mathbb{R}^{N\times C\times H\times W}$ is the final MGA output feature.

D. Stimulus-guided Adaptive Gated Bottleneck (SAGB) Block

Thin retinal capillaries are often fragile and easily corrupted by noise and illumination variations, making channel representations unstable across scales. To address this issue, we propose the *Stimulus-guided Adaptive Gated Bottleneck (SAGB) block*, which dynamically balances global contextual trends and salient local peaks through adaptive channel-wise gating, as illustrated in Fig. 1(d).

1) Local feature embedding: Given an input feature map x at an arbitrary scale, SAGB first applies a local context embedding via a 3×3 convolution, batch normalization, and GELU activation:

$$Y = \text{GELU}(\text{BN}(w_{3\times3} * x)), \tag{11}$$

where $w_{3\times3}$ denotes the convolution kernel and Y is the intermediate feature representation.

2) Dual-stimulus channel descriptors: Next, we use global average pooling and global max pooling to generate two complementary stimulus vectors that capture each channel's response from a global perspective: $v_{\text{avg}} = \text{GAP}(Y)$ and $v_{\max} = \text{GMP}(Y)$. These descriptors capture the mean trend and salient peaks of each channel, respectively, and are fed into a dual-branch MLP to generate channel-importance cues that reflect both aspects:

$$a = f_{\text{avg}}(v_{\text{avg}}) = w_2^{\text{avg}}\,\delta(w_1^{\text{avg}}\,v_{\text{avg}}), \tag{12}$$

$$m = f_{\max}(v_{\max}) = w_2^{\max}\,\delta(w_1^{\max}\,v_{\max}). \tag{13}$$

Here, w_1 and w_2 are the learnable parameters of the dual-branch perceptrons, and $\delta(\cdot)$ represents the ReLU activation function. The outputs a and m characterize the mean response trend and peak saliency of each channel. For subsequent gating fusion, they are mapped into channel-wise descriptors in $\mathbb{R}^{N\times C\times1\times1}$, denoted as $\tilde{a}$ and $\tilde{m}$, respectively.

3) Adaptive gating fusion: Instead of directly averaging the two stimuli, SAGB learns a dynamic gating coefficient to adaptively balance them. Specifically, the gate is computed by concatenating the two pooled descriptors, applying a 1×1 convolution, normalizing, and applying a Sigmoid activation as follows:

$$\alpha = \sigma\Big(\text{BN}\left(w_{1\times1}^{g} * [\,v_{\text{avg}} \,\|\, v_{\max}\,]\right)\Big), \tag{14}$$

where $\alpha \in [0,1]^{N\times C\times1\times1}$ controls the relative contribution of the average-trend and peak-saliency stimuli.

The fused channel attention mask is then obtained as:

$$A = \sigma(\alpha \odot \tilde{a} + (1-\alpha) \odot \tilde{m}), \tag{15}$$

where $\odot$ denotes element-wise multiplication, and the resulting A serves as a stimulus-guided channel-wise gating map that selectively enhances vessel-related activations.

4) Channel recalibration and residual output: Finally, the input feature is reweighted by A, projected through a 1×1 convolution, and injected back via a residual shortcut:

$$Y_{\text{out}} = \text{GELU}\big(\text{BN}\big(w_{1\times1}^{o} * (Y \odot A)\big) + x\big), \tag{16}$$

where Y_{out} denotes the output feature after adaptive channel gating.

In this way, SAGB provides stimulus-guided adaptive channel regulation, improving robustness and thin-vessel continuity across hierarchical encoder stages.

IV. Experiments

A. Dataset

1) DRIVE dataset: The DRIVE dataset, constructed by Staal et al. [25], contains 40 color fundus images with a resolution of approximately 584×565 pixels and a 45° field of view (FOV). Following the standard protocol, 20 images are used for training and the remaining 20 for testing. The training set is annotated by one expert, while the test set is annotated by two experts. In our evaluation, the first expert annotation is adopted as the ground truth reference.

2) CHASEDB1 dataset: CHASEDB1 consists of 28 fundus images acquired from both eyes of 14 school-aged children [26], with a resolution of 999×960 pixels and a 30° FOV. Each image is manually delineated by two trained graders, with a corresponding vessel segmentation annotation. As in prior work, we use the first grader's annotation as the reference standard and split the dataset into 20 training images and 8 test images.

3) STARE dataset: The STARE dataset includes 20 fundus images with a resolution of 605×700 pixels, covering various retinal pathologies, including macular degeneration, hypertensive retinopathy, and diabetic retinopathy [27]. Images are captured with a TopCon camera under a 35° FOV. Manual vessel annotations are provided by two experts, and 15 images are used for training, while the remaining 5 are used for testing in this study.

B. Implementation Details and Metrics

All models were implemented in PyTorch and trained using the Adam optimizer for 300 epochs with a batch size of 16. The initial learning rate was set to 5×10^{-3}. Standard data augmentation, including random rotation, flipping, and intensity perturbations, was applied to improve generalization. During inference, we employed a sliding-window strategy with a patch size of 64×64 and a stride of 32 pixels to ensure consistent evaluation across datasets with different resolutions. All experiments were conducted on a server equipped with an Intel(R) Core(TM) i5-13500H CPU and an NVIDIA RTX 4050 GPU.

To assess the similarity between the predicted segmentation and ground-truth annotations, we adopted five evaluation metrics: accuracy (ACC), sensitivity (SE), specificity (SP), F1-score (F1), and area under the curve (AUC). These metrics characterize complementary aspects of segmentation performance and exhibit inherent trade-offs; therefore, they were jointly considered to provide a comprehensive evaluation of the proposed method on retinal images.

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

To demonstrate the superiority of the proposed method for retinal vessel segmentation, we comprehensively evaluated **SGEA-Net** on three publicly available benchmarks: DRIVE, CHASEDB1, and STARE. We compare our approach with seven representative segmentation models, including U-Net [10], TransU-Net [13], SGAT-Net [20], TDCAU-Net [21], TUnet-LBF [22], HiDiffSeg [23], and HRD-Net [24].

Retinal vessel segmentation is particularly challenging because of the severe class imbalance between the major trunks and extremely thin branches. Thus, preserving the continuity of thin vessels is critical for high-quality segmentation. Fig. 3 presents qualitative comparisons among U-Net, TransU-Net, and our SGEA-Net across the three datasets. U-Net tends to miss subtle capillaries, while TransU-Net often produces discontinuities at thin-vessel junctions. In contrast, SGEA-Net better captures fine vessels and maintains structural connectivity, resulting in more complete segmentation.

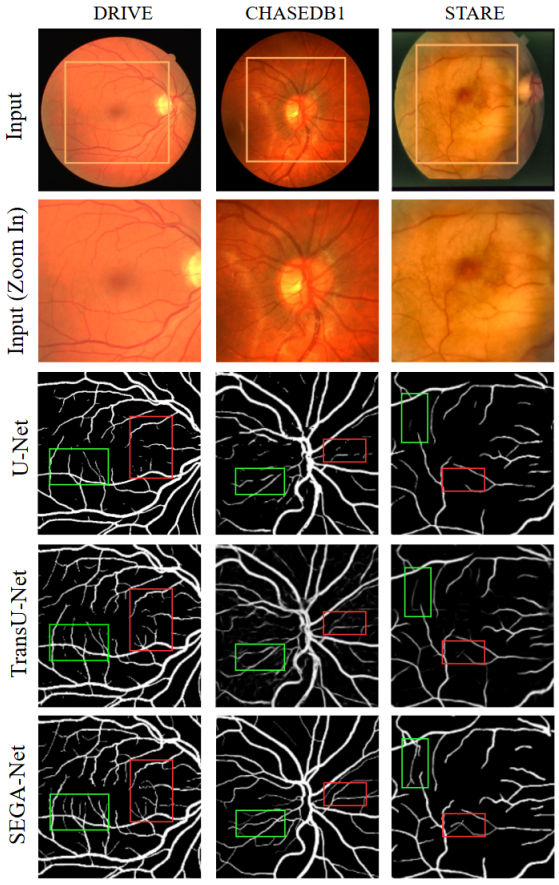

Fig. 3. Qualitative comparison of retinal vessel segmentation results on the DRIVE, CHASEDB1, and STARE datasets produced by U-Net, TransU-Net, and the proposed SGEA-Net.

Table I summarizes the quantitative comparison between our method and seven competing models across five evaluation metrics. Overall, results on the three datasets show that SGEA-Net achieves superior performance and strong generalization. On DRIVE, SGEA-Net obtained the best ACC (0.9773), SP (0.9875), and AUC (0.9896), while achieving a competitive F1 score of 0.8467, only 0.0012 lower than the best baseline. On CHASEDB1, it achieved the highest ACC (0.9816), SE (0.8471), F1 (0.8463), and AUC (0.9869), with F1 improving by 0.0051 over the second-best result. On STARE, SGEA-Net further achieved the best ACC (0.9885), SP (0.9834), F1 (0.8392), and AUC (0.9889), showing robust performance under more challenging conditions. Overall, SGEA-Net delivers the most consistent and competitive segmentation results, highlighting its potential for computer-aided retinal disease screening.

D. Ablation Studies

In this subsection, ablation experiments were conducted on DRIVE, CHASEDB1, and STARE to evaluate the effectiveness of the three proposed components: EAE, MGA, and SAGB. Using U-Net as the baseline, these modules were progressively added following an incremental strategy.

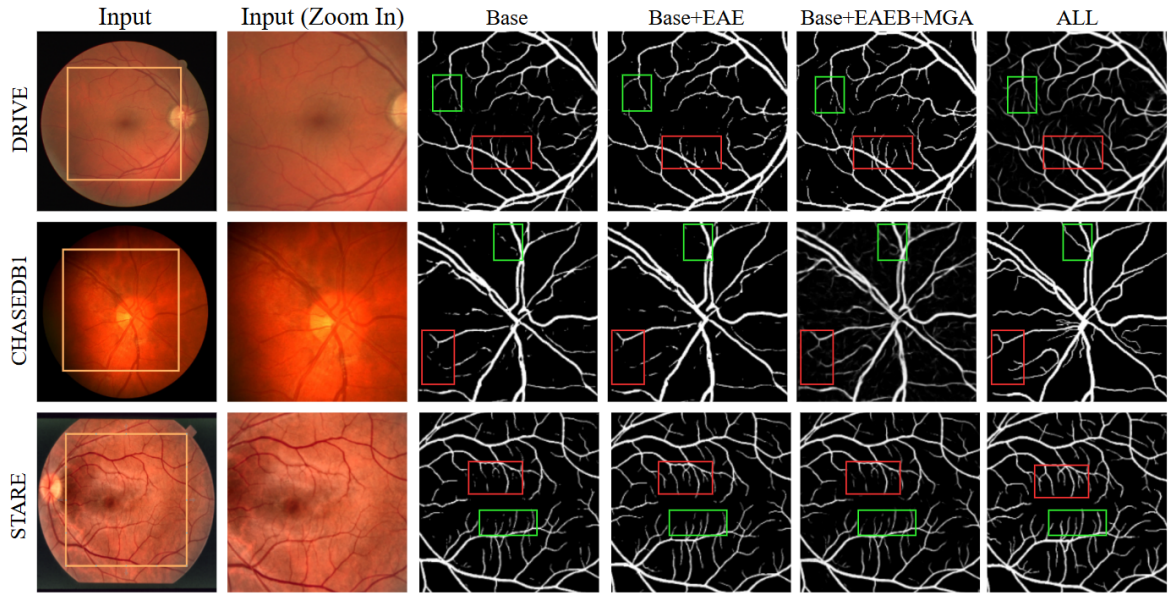

Fig. 4. Qualitative ablation results on the DRIVE, CHASEDB1, and STARE datasets. From left to right: input image, zoomed region, Base (U-Net), Base+EAE, Base+EAE+MGA, and the full model (Base+EAE+MGA+SAGB).

TABLE I
COMPARISON OF U-NET, TRANSU-NET, SGAT-NET, TDCAU-NET, TUNET-LBF, HIDIFFSEG, HRD-NET, AND OUR METHOD ON DRIVE, CHASEDB1, AND STARE DATASETS.

Model Name	ACC	SE	SP	F1	AUC
DRIVE					
U-Net	0.9561	0.7721	0.9841	0.8162	0.9763
TransU-Net	0.9656	0.7963	0.9795	0.8301	0.9743
SGAT-Net	0.9663	0.8019	0.9773	0.8313	0.9115
TDCAU-Net	0.9558	0.8178	0.9796	0.8353	0.9794
TUnet-LBF	0.9663	0.8138	0.9801	0.8289	0.9650
HiDiffSeg	0.9712	0.8317	0.9862	**0.8479**	0.9892
HRD-Net	0.9724	0.8309	0.9813	0.8314	0.9875
SGEA-Net (ours)	**0.9773**	**0.8318**	**0.9875**	0.8467	**0.9896**
CHASEDB1					
U-Net	0.9579	0.7993	0.9801	0.7801	0.9773
TransU-Net	0.9638	0.7931	0.9793	0.8019	0.9734
SGAT-Net	0.9742	0.8109	0.9814	0.8412	0.9203
TDCAU-Net	0.9735	0.8239	0.9816	0.8226	0.9807
TUnet-LBF	0.9711	0.8349	0.9809	0.8196	0.9680
HiDiffSeg	0.9804	0.8468	**0.9837**	0.8217	0.9863
HRD-Net	0.9783	0.8374	0.9794	0.8316	0.9826
SGEA-Net (ours)	**0.9816**	**0.8471**	0.9825	**0.8463**	**0.9869**
STARE					
U-Net	0.9581	0.7943	0.9715	0.8342	0.8826
TransU-Net	0.9653	0.8286	0.9768	0.8127	0.9763
SGAT-Net	0.9752	0.8365	0.9827	0.8389	0.9130
TDCAU-Net	0.9742	0.8297	0.9762	0.8133	0.9826
TUnet-LBF	0.9761	0.8364	0.9806	0.8196	0.9721
HiDiffSeg	0.9806	**0.8433**	0.9657	0.8117	0.9867
HRD-Net	0.9851	0.8379	0.9831	0.8273	0.9824
SGEA-Net (ours)	**0.9885**	0.8398	**0.9834**	**0.8392**	**0.9889**

TABLE II
ABLATION STUDIES ON DRIVE, CHASEDB1, AND STARE DATASETS.

Case	ACC	SE	SP	F1	AUC
DRIVE					
Base	0.9561	0.7921	0.9840	0.8242	0.9855
Base+EAE	0.9717	0.8146	0.9869	0.8297	0.9831
Base+EAE+MGA	0.9723	0.8302	**0.9881**	0.8369	0.9876
SGEA-Net	**0.9773**	**0.8318**	0.9875	**0.8467**	**0.9896**
CHASEDB1					
Base	0.9578	0.7993	0.9701	0.7983	0.9673
Base+EAE	0.9648	0.8239	0.9769	0.8268	0.9762
Base+EAE+MGA	0.9807	0.8310	**0.9836**	0.8371	0.9838
SGEA-Net	**0.9816**	**0.8471**	0.9825	**0.8463**	**0.9869**
STARE					
Base	0.9531	0.7943	0.9745	0.8142	0.9755
Base+EAE	0.9746	0.8324	0.9803	0.8338	0.9743
Base+EAE+MGA	0.9755	0.8298	**0.9856**	0.8384	0.9833
SGEA-Net	**0.9885**	**0.8398**	0.9834	**0.8392**	**0.9889**

Fig. 4 visualizes the qualitative evolution from the baseline to the full SGEA-Net. Introducing the EAE block enhances the detection of thin vessels by strengthening boundary cues. Adding the MGA block further improves multi-scale feature fusion and boundary delineation, although minor discontinuities may still occur in extremely thin branches. With all components enabled, the complete SGEA-Net effectively alleviates fragmentation and produces the most continuous vessel structures.

Table II reports the quantitative ablation results. On DRIVE, EAE improves sensitivity and accuracy, while MGA further enhances vessel connectivity and increases F1 and AUC. With

all modules combined, SGEA-Net achieves the best overall performance in ACC, SE, F1, and AUC. On CHASEDB1, EAE improves recall and F1, whereas MGA increases specificity by reducing false positives, and the full model shows the most balanced results. On STARE, MGA improves specificity and robustness, while the complete SGEA-Net achieves the highest ACC, AUC, and F1.

These ablation studies confirm that each module contributes complementary benefits, and their combination yields robust and accurate retinal vessel segmentation.

V. Conclusion

In this paper, we propose SGEA-Net to address key challenges in retinal vessel segmentation, including boundary discontinuity, limited multi-scale semantic fusion, and unstable channel-wise detail preservation. By integrating edge-aware enhancement, unified multi-scale global aggregation, and stimulus-guided adaptive channel gating, SGEA-Net effectively improves thin-vessel continuity and segmentation robustness. Experiments on three public benchmarks, DRIVE, CHASEDB1, and STARE, show that SGEA-Net outperforms both classical and recent state-of-the-art methods in ACC, SE, F1, and AUC. Qualitative results further demonstrate more continuous and detailed capillary segmentation, especially in challenging junctions and low-contrast regions. Ablation studies confirm that Edge Attention Enhancement improves boundary visibility and thin-vessel detection, while Multi-scale Global Aggregation enhances cross-scale consistency and vascular connectivity. Together with the Stimulus-guided Adaptive Gated Bottleneck, these modules yield the best overall performance.

Overall, this work presents an effective and extensible framework that jointly models edge information, multi-scale feature aggregation, and adaptive channel regulation for retinal vessel segmentation, offering practical value for fundus image analysis and computer-aided clinical screening. Future work will focus on lightweight deployment and cross-domain generalization under broader retinal imaging conditions.

References

[1] P. K. Verma and J. Kaur, "Systematic review of retinal blood vessels segmentation based on AI-driven technique," *Journal of Imaging Informatics in Medicine*, vol. 37, pp. 1783–1799, 2024.

[2] G. Ayoub, "Disease diagnosis using retinal vasculature: Insights from flammer syndrome and ai," *Brain Sci.*, vol. 15, no. 9, p. 919, 2025.

[3] M. R. K. Mookiah, S. Hogg, T. J. MacGillivray, V. Prathiba, R. Pradeepa, V. Mohan, R. M. Anjana, A. S. Doney, C. N. A. Palmer, and E. Trucco, "A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification," *Medical Image Analysis*, vol. 68, p. 101905, 2021.

[4] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI'98)*, vol. 1496, pp. 130–137, 1998.

[5] B. S. Tchinda, D. Tchiotsop, M. Noubom, V. Louis-Dorr, and D. Wolf, "Retinal blood vessels segmentation using classical edge detection filters and the neural network," *Informatics in Medicine Unlocked*, vol. 23, p. 100521, 2021.

[6] Y. Zhao, X. Wang, X. Wang, and Y. Shih, "Retinal vessels segmentation based on level set and region growing," *Pattern Recognition*, vol. 47, pp. 2437–2446, 2014.

[7] A. Osareh and B. Shadgar, "Retinal vessel extraction using gabor filters and support vector machines," in *Advances in Computer Science and Engineering, Communications in Computer and Information Science*, vol. 6, Springer, 2009, pp. 356–363.

[8] B. Hesko, "Pixel-wise segmentation of the blood vessels using random forests," in *Proceedings of the 24th Conference STUDENT EEICT 2018*, 2018, pp. 605–609.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proceedings of Medical image computing and computer-assisted intervention (MICCAI)*, pp. 234–241, 2015.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)2021*, 2021.

[13] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

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

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Computer-Assisted Intervention (CVPR)*, 2015, pp. 3431–3440.

[16] Y. Peng, M. Sonka, and D. Z. Chen, "U-Net v2: Rethinking the skip connections of U-Net for medical image segmentation," *arXiv preprint arXiv:2311.17791*, 2023.

[17] B. Pang, L. Chen, Q. Tao, E. Wang, and Y. Yu, "GA-UNet: A lightweight Ghost and Attention U-Net for medical image segmentation," *Journal of Imaging Informatics in Medicine*, vol. 37, pp. 1874–1888, 2024.

[18] Z. Chen, S. Chen, and F. Hu, "CTA-UNet: CNN-transformer architecture UNet for dental CBCT images segmentation," *Physics in Medicine & Biology*, vol. 68, p. 175042, 2023.

[19] X. Luo, L. Peng, Z. Ke, J. Lin, and Z. Yu, "PA-Net: A hybrid architecture for retinal vessel segmentation," *Pattern Recognition*, vol. 161, p. 111254, 2025.

[20] J. Lin, X. Huang, H. Zhou, Y. Wang, and Q. Zhang, "Stimulus-guided adaptive transformer network for retinal blood vessel segmentation in fundus images," *Medical Image Analysis*, vol. 89, p. 102929, 2023.

[21] C. Li, Z. Li, and W. Liu, "TDCAU-Net: retinal vessel segmentation using transformer dilated convolutional attention-based U-Net method," *Physics in Medicine & Biology*, vol. 69, no. 1, p. 015003, 2024.

[22] H. Zhang, W. Ni, Y. Luo, Y. Feng, R. Song, and X. Wang, "TUnet-LBF: retinal fundus image fine segmentation model based on transformer Unet network and LBF," *Computer in Biology and Medicine*, vol. 159, p. 106937, 2023.

[23] W. Huang and F. Liu, "HiDiffSeg: a hierarchical diffusion model for blood vessel segmentation in retinal fundus images," *Expert Systems Applications*, vol. 253, p. 124249, 2024.

[24] J. Liu, D. Zhao, J. Shen, P. Geng, Y. Zhang, J. Yang, and Z. Zhang, "HRD-Net: high resolution segmentation network with adaptive learning ability of retinal vessel features," *Computers in Biology and Medicine*, vol. 173, p. 108295, 2024.

[25] J. J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE transactions on medical imaging*, vol. 23, no. 4, pp. 501–509, 2004.

[26] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman, "An ensemble classification-based approach applied to retinal blood vessel segmentation," *IEEE transactions on biomedical engineering*, vol. 59, no. 9, pp. 2538–2548, 2012.

[27] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE transactions on medical imaging*, vol. 19, no. 3, pp. 203–210, 2000.