DOI: 10.19884/j.1672-5220.202506008

# SGFNet: An Attention-Enhanced Network for Colonic Polyp Segmentation in Endoscopic Images

PANG Bo, LIU Guohua*
College of Computer Science and Technology, Donghua University, Shanghai 201620, China

**Abstract: Accurate segmentation of colonic polyps in endoscopic images is vital for early diagnosis and prevention of colorectal cancer (CRC). However, this task remains challenging due to the high variability in polyp morphology, low contrast with surrounding mucosa, and frequent boundary ambiguity. To address these issues, we propose a selective-gated fusion network (SGFNet), a lightweight semantic segmentation network based on the U-Net architecture. SGFNet incorporates two targeted modules: the SSE-Encoder, which integrates selective kernel convolution (SKConv) and squeeze-and-excitation (SE) attention to enhance multi-scale representation and channel recalibration; the PGF-Unit, which employs a gated parallel polarized self-attention (PPSA) mechanism to improve boundary-sensitive feature fusion. The model is evaluated on a unified dataset of 1 612 annotated images from Kvasir-SEG and CVC-ClinicDB. Experimental results demonstrate that SGFNet achieves superior performance in mean intersection over union (mIoU), pixel accuracy (PA), and F1 score, outperforming several representative baseline models. With its balanced design and strong generalization, SGFNet offers a practical and interpretable solution for real-world medical image segmentation and provides insights into attention-based module integration for future computer-aided diagnosis systems.**

**Keywords**: colonic polyp segmentation; U-Net; multi-scale feature encoding; attention mechanism; SGFNet; medical image analysis
**CLC number**: TP391.41 **Document code**: A
**Article ID**: 1672-5220(2026)03-0101-12

Open Science Identity (OSID)

## 0 Introduction

Colorectal cancer (CRC) is one of the most common and deadly malignancies worldwide. In 2022, CRC accounted for approximately 9.6% of all new cancer cases and 9.3% of global cancer-related deaths[1]. As the primary precancerous lesion of CRC, colonic polyps play a critical role in early diagnosis and prevention. Clinical studies have shown that timely detection and removal of colonic polyps can significantly reduce the incidence of CRC[2]. Given this clinical importance, accurate segmentation of colonic polyps in endoscopic images is essential for computer-aided diagnosis and clinical decision support systems. However, this task remains challenging due to the high variability in polyp morphology, low contrast with surrounding mucosa, and frequent boundary ambiguity. These factors often lead to missed detections, especially for small-sized polyps[3]. To address these challenges, deep learning-based segmentation methods, particularly those built upon the U-Net architecture, have achieved notable success in medical imaging tasks. Several variants have been proposed to improve feature extraction and boundary localization, such as UNet++[4], ResUNet++[5], and SANet[6]. While these models achieve promising results, practical limitations remain, including class imbalance, inadequate detail preservation, and suboptimal multi-scale fusion, which hinder their reliability in clinical deployment.

Motivated by clinical findings and dataset analyses, this paper proposes a lightweight and practical network, the selective-gated fusion network (SGFNet). SGFNet targets two persistent challenges in colonic polyp segmentation: 1) wide variation in polyp morphology and size, which limits the capacity of fixed-receptive-field convolutions to capture multi-scale patterns; 2) the prevalence of low-contrast polyps with ambiguous boundaries, which often results in feature confusion and segmentation inaccuracy.

To address these issues, we propose a dual-strategy design. First, a compound SSE-Encoder is introduced in the encoder path by integrating selective kernel convolution (SKConv) and squeeze-and-excitation (SE) attention. This combination enables scale-adaptive receptive fields while reinforcing semantically salient channels. Second, a polarized gated fusion self-attention (PGF-Unit) is incorporated at the skip connections to selectively enhance edge-sensitive spatial information and suppress irrelevant background features.

Unlike previous methods that stack or parallelize spatial attention (SA) and channel attention (CA)

Received date: 2025-06-27
* Correspondence should be addressed to LIU Guohua, email: ghliu@dhu.edu.cn
Citation: Pang B, Liu G H. SGFNet: an attention-enhanced network for colonic polyp segmentation in endoscopic images[J]. *Journal of Donghua University (English Edition)*, 2026, 43(3): 101-112.

mechanisms without spatial-scale awareness, SGFNet emphasizes structurally lightweight yet context-adaptive encoding and boundary-preserving fusion. This tailored architecture addresses real-world segmentation bottlenecks while maintaining interpretability and computational efficiency.

To validate the effectiveness of our model, we conduct experiments on a merged dataset comprising 1 612 annotated images from Kvasir-SEG[7] and CVC-ClinicDB[8]. Comparative experiments and ablation studies demonstrate that SGFNet delivers robust segmentation performance with low computational cost, making it a promising solution for clinical polyp segmentation.

# 1 Methods

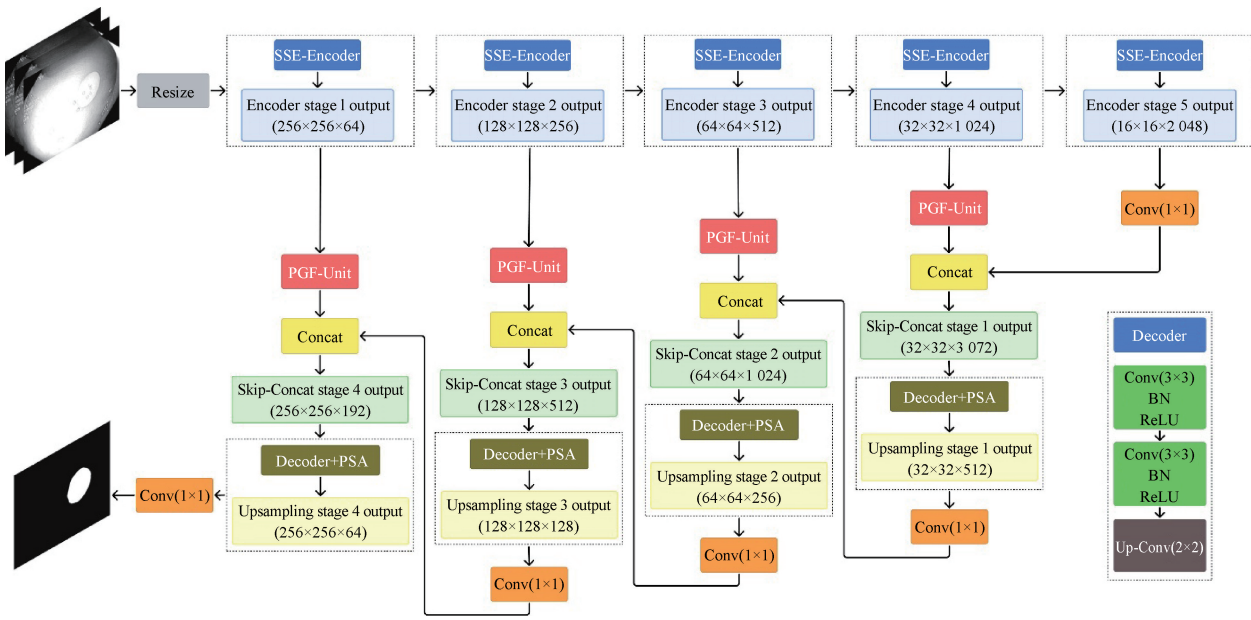
## 1.1 Overview of SGFNet architecture

SGFNet is a semantic segmentation framework designed to accurately capture targets with complex shapes and ambiguous boundaries. By employing a selective-gated fusion mechanism, the network effectively integrates multi-level feature information, enhancing segmentation accuracy and robustness across various challenging segmentation tasks. It builds upon the encoder-decoder architecture of U-Net, incorporating two targeted modules at distinct stages.

1) The SSE-Encoder, integrated into the encoder path, enhances multi-scale feature extraction while dynamically suppressing irrelevant background responses.

2) The PGF-Unit, placed at skip connections, improves the selective fusion of contextual and spatial features, with particular focus on boundary refinement.

The overall architecture of SGFNet is illustrated in Fig. 1.

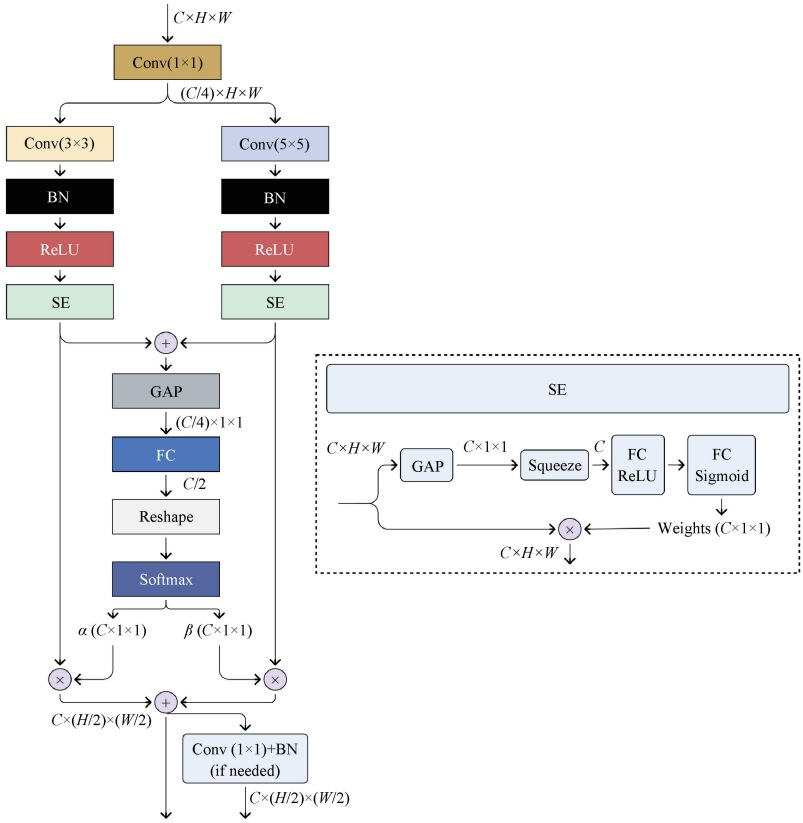



Conv—convolution; Concat—concatenation; BN—batch normalization; ReLU—rectified linear unit; PPSA—parallel polarized self-attention.

Fig. 1 Overall architecture of SGFNet

## 1.2 SSE-Encoder: selective feature extraction with dynamic recalibration

To enhance feature representation for colonic polyp segmentation, we propose the SSE-Encoder, a compound module that combines SKConv with SE attention in a unified design. This design is motivated by two challenges: 1) wide variation in polyp size and shape necessitating adaptive receptive fields; 2) boundary ambiguity that requires improved semantic discrimination across channels. The detailed architecture of the SSE-Encoder is illustrated in Fig. 2.




GAP—global average pooling; FC—fully connected; $\alpha$, $\beta$—Softmax-derived CA weights for 3 × 3 and 5 × 5 branch feature fusions, respectively; $C$—number of channels; $H$—height; $W$—width.

Fig. 2 Structure of SSE-Encoder module

SKConv enables adaptive spatial receptive fields but lacks sensitivity to channel semantics, while SE attention enhances channel-wise feature discrimination but ignores spatial scale diversity. In this work, we combine the two to create a hybrid encoder block that jointly addresses multi-scale spatial adaptation and semantic feature enhancement. This fusion allows SSE-Encoder to adaptively respond to both inter-scale and inter-channel variations, which is critical for segmenting small and low-contrast polyps.

We also customize the original modules to better suit the characteristics of medical image segmentation, which poses different constraints compared to their original use in natural image classification. For example, we reduce kernel and channel complexity to improve computational efficiency and modify activation placement to better capture small-target features.

The resulting SSE-Encoder is lightweight, fully convolutional, and capable of robustly encoding polyp regions with diverse sizes and low contrast. These properties make SSE-Encoder particularly effective for capturing fine structures under noisy backgrounds.

Compared to existing attention integrations such as CBAM[9] or dual-branch spatial-channel attention modules[10], our SSE-Encoder differs in both placement and design philosophy. Recent work, such as MSCMNet[11], proposed in the context of visible-infrared person re-identification, has also demonstrated the effectiveness of multi-scale semantic feature mining. MSCMNet adopts a semantic correlation strategy across multiple scales to enhance cross-modal discriminability. In contrast, our approach targets scale variation and semantic imbalance in medical segmentation by directly combining SKConv's adaptive receptive field learning with SE attention's channel-wise recalibration, forming a compact and encoder-focused design.

### 1.2.1 SKConv

SKConv was originally introduced in the field of image classification to address the rigidity of fixed receptive fields in standard convolutional neural networks[12]. It allows the network to dynamically select among multiple convolution branches with different kernel sizes, typically 3 × 3 and 5 × 5, based on the context of the input features. The selection is guided by a learned attention mechanism, enabling the model to focus on the most informative scale at each spatial location.

Although originally developed for classification, SKConv's scale-aware adaptation makes it particularly suitable for medical segmentation tasks like colonic polyp detection. Polyp sizes in endoscopic images vary widely,

from large protrusions to flat lesions under 10 pixels, posing challenges for fixed-kernel CNNs in capturing such structural diversity.

The detailed pipeline of SKConv, including its multi-branch processing and attention-guided feature fusion, is visualized in Fig. 3.

SKConv consists of three core operations: split, fuse, and select. Given an input feature map $\boldsymbol{X} \in \mathbf{R}^{C\times H\times W}$, in the split stage, the input feature map is processed through $K$ parallel convolutional branches with different kernel sizes (e.g., $3\times 3$ and $5\times 5$) to extract the multi-scale feature representations from $\boldsymbol{X}$. For each branch $k$, the output is denoted as $\widetilde{\boldsymbol{U}}$ (e.g., $\widetilde{\boldsymbol{U}}$ for the $3\times 3$ branch and $\hat{\boldsymbol{U}}$ for the $5\times 5$ branch in Fig. 3), which is defined as

$$\widetilde{\boldsymbol{U}}(k) = f_k(\boldsymbol{X}),\ k = 1, \cdots, K, \tag{1}$$

where $f_k$ denotes the convolutional operation of the $k$th branch.

In the fuse stage, the output feature maps from all branches are integrated into a unified feature representation $\boldsymbol{U}$ (the aggregated feature) through element-wise summation. A global average pooling operation $F_{\mathrm{gp}}$ is then applied to generate channel-wise statistics:

$$\boldsymbol{U} = \sum_{k=1}^{K} \widetilde{\boldsymbol{U}}^{(k)},\ \boldsymbol{S} = F_{\mathrm{gp}}(\boldsymbol{U}), \tag{2}$$

where $F_{\mathrm{gp}}$ denotes global average pooling, which compresses the spatial dimensions $H\times W$ into a single value for each channel; $\boldsymbol{S}$ represents the global channel-wise descriptor that summarizes the global information of each channel.

In the select stage, the aggregated statistics are fed into a fully connected bottleneck $F_{\mathrm{fc}}$ to compute the compact feature vector $\boldsymbol{Z}$:

$$\boldsymbol{Z} = F_{\mathrm{fc}}(\boldsymbol{S}) = \boldsymbol{W}_2\,\delta(\boldsymbol{W}_1\boldsymbol{S}), \tag{3}$$

where $\boldsymbol{W}_1 \in \mathbf{R}^{(C/r)\times C}$ and $\boldsymbol{W}_2 \in \mathbf{R}^{C\times(C/r)}$ denote the weight matrices; $\delta(\cdot)$ represents the ReLU activation function; $r$ is the channel reduction ratio.

The attention weights for each channel and branch are then obtained by applying a softmax function over $\boldsymbol{Z}$. In our two-branch case, these weights are denoted as $a$ (for the $3\times 3$ branch) and $b$ (for the $5\times 5$ branch):

$$a_c^{(k)} = \frac{\exp(z_c^{(k)})}{\sum_{j=1}^{K}\exp(z_c^{(l)})},\ \sum_{k=1}^{K} a_c^{(k)} = 1, \tag{4}$$

where $a_c^{(k)}$ denotes the attention weight assigned to the $k$th branch at the $c$th channel; $z_c^{(k)}$ and $z_c^{(l)}$ represent the feature values corresponding to the $c$th channel and the $k$th branch, and the $c$th channel and the $l$th branch in the compact feature vector $\boldsymbol{Z}$, respectively.

The final output $\boldsymbol{V}$ is obtained through weighted fusion of the branch features using the computed attention weights:

$$\boldsymbol{V}_c(i,j) = \sum_{k=1}^{K} a_c^{(k)}\widetilde{\boldsymbol{U}}_c^{(k)}(i,j),\ \boldsymbol{V} \in \mathbf{R}^{C\times H\times W}, \tag{5}$$

where $\boldsymbol{V}_c(i,j)$ denotes the feature value of the final output feature map $\boldsymbol{V}$ at the $c$th channel, $i$th row, and $j$th column; $\widetilde{\boldsymbol{U}}_c^{(k)}(i,j)$ represents the feature value of the $k$th branch's output feature map $\widetilde{\boldsymbol{U}}^{(k)}$ at the $c$th channel, $i$th row, and $j$th column.

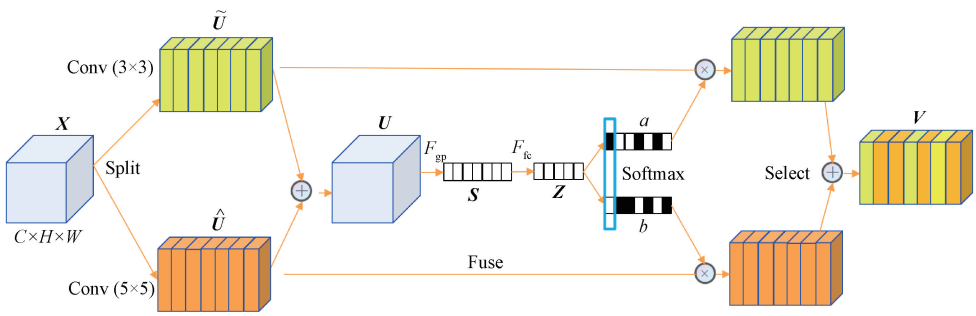



Fig. 3 Structure of SKConv module

In our encoder, SKConv is adapted to enable per-region receptive field modulation, enhancing the network's ability to represent both large and small polyps. To better fit the U-Net architecture and medical data constraints, we simplify the original design by reducing branch complexity and adopting lightweight convolutional operations.

In our SKConv implementation, we adopt two parallel convolution branches ($M=2$) with kernel sizes of $3\times 3$ and $5\times 5$, respectively. Each branch outputs the same number of channels and uses grouped convolution with group number $G=32$, followed by BN and ReLU activation. The branch outputs are then aggregated via global average pooling $F_{\mathrm{gp}}$ and passed through a two-layer fully connected bottleneck $F_{\mathrm{fc}}$ to produce soft attention weights $a$ and $b$.

This attention mechanism reduces the feature dimension from $C$ to $C/16$ and then expands it to $2C$, producing soft weights that correspond to each branch. These weights are used to perform an attention-

weighted fusion of the parallel features. The entire structure is lightweight, receptive-field adaptive, and designed to enhance multi-scale feature extraction efficiency in encoder-side representation learning, which is crucial in resource-limited medical segmentation tasks.

### 1.2.2 SE attention mechanism

SE attention was first introduced in the context of image classification as part of the SENet architecture[13]. It enhances network performance by learning channel-wise dependencies, enabling emphasis on informative channels while suppressing irrelevant ones. The SE block compresses global spatial information via pooling, processes it through a bottleneck, and produces channel-wise weights for recalibrating the input features.

The SE block follows a three-stage process—squeeze, excitation, and recalibration—for adaptive channel-wise feature refinement. It first models channel dependencies via global pooling (squeeze), then generates attention weights (excitation), and finally scales the feature maps accordingly (recalibration). This mechanism enhances informative channels while suppressing irrelevant or redundant ones, thereby significantly improving the discriminative power of feature representations.

In the squeeze phase, spatial information is compressed into channel descriptors through global average pooling:

$$\boldsymbol{s}_c = F_{\mathrm{sq}}(\boldsymbol{X}_c) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{X}_{c,i,j}, \tag{6}$$

where $\boldsymbol{X}_{c,i,j}$ denotes the pixel value at the $c$th channel, $i$th row, and $j$th column; $\boldsymbol{s}_c$ represents the global average pooling result for the $c$th channel, indicating its global response intensity; $F_{\mathrm{sq}}(\cdot)$ denotes the squeeze operation, which compresses spatial information into a compact channel-wise descriptor.

The overall pipeline of the SE block with its three-stage operation is illustrated in Fig. 4.

In the excitation phase, the compressed vector is processed through a two-layer fully connected network with activation functions to learn inter-channel dependencies and generate the attention weight $e$:

$$e = F_{\mathrm{ex}}(\boldsymbol{S},\boldsymbol{W}) = \sigma(\boldsymbol{W}_2\mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{S})),\ e \in \mathbf{R}^C, \tag{7}$$

where $F_{\mathrm{ex}}(\cdot, \boldsymbol{W})$ denotes the excitation operation parameterized by the weight matrix $\boldsymbol{W} = \{\boldsymbol{W}_1, \boldsymbol{W}_2\}$; $\sigma(\cdot)$ denotes the sigmoid activation function.

In the final recalibration stage, the original feature maps are channel-wise scaled using the generated attention weights:

$$\widetilde{\boldsymbol{X}}_{c,i,j} = F_{\mathrm{scale}}(\boldsymbol{X},e) = e_c \cdot \boldsymbol{X}_{c,i,j},\ \widetilde{\boldsymbol{X}} \in \mathbf{R}^{C\times H\times W}, \tag{8}$$

where $\widetilde{\boldsymbol{X}}_{c,i,j}$ denotes the feature value at the $c$th channel, $i$th row, and $j$th column of the output feature map; $e_c$ denotes the attention weight corresponding to the $c$th channel; $\widetilde{\boldsymbol{X}}$ is the recalibrated output feature map; $F_{\mathrm{scale}}(\cdot, e)$ denotes the channel-wise scaling operation that applies the learned attention weight $e$ to the original features.

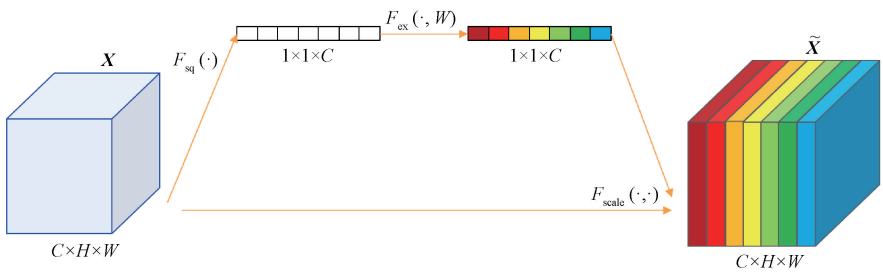



Fig. 4 Structure of SE block module

Despite its origin in classification, the SE mechanism aligns well with polyp segmentation, where subtle and low-contrast features are often masked by background noise. In particular, channel-level semantic cues, such as texture, color intensity, or shape abstraction, can play a key role in distinguishing polyps from mucosal folds or other artifacts.

We integrate SE attention into the SSE-Encoder to complement SKConv, reinforcing the model's focus on informative channels, especially under class imbalance. Compared to SKConv alone, SE offers a complementary path for channel recalibration, helping retain critical signals after spatial abstraction. We further adapt SE by reducing its bottleneck ratio and introducing residual links, improving compatibility with deep segmentation backbones.

In our implementation of the SE block, we set the channel reduction ratio $r = 16$. The excitation module consists of two sequential $1 \times 1$ convolutional layers (i.e., a bottleneck multi-layer perceptron), with ReLU and Sigmoid activations applied respectively. This structure reduces the channel dimensionality from $C$ to $C/16$ and then restores it to $C$. The resulting attention map is used to perform channel-wise recalibration on the original feature map via multiplication. This lightweight design enhances semantic focus while preserving computational efficiency, making it well-suited for real-time segmentation scenarios.

### 1.3 PGF-Unit

To improve semantic consistency and boundary localization during decoding, we propose the PGF-Unit, a lightweight attention-based fusion module placed at the skip connections of SGFNet. It addresses a common U-Net limitation: direct skip connections may introduce spatial noise or weaken semantic precision near boundaries. PGF-Unit provides an adaptive and selective mechanism for feature reuse, focusing on boundary-aware refinement. The detailed architecture of the PGF-Unit is illustrated in Fig. 5.

At the core of PGF-Unit is the PPSA mechanism, which independently models spatial and channel-wise dependencies in two parallel branches, as shown in the left and middle parts of Fig. 5. Originally proposed for pixel-wise regression tasks such as saliency prediction, PPSA provides a flexible framework for capturing fine-grained spatial structure and cross-channel semantic relevance. We deploy a simplified PPSA at skip connections, where high-resolution details and mid-level semantics co-exist and contribute most to boundary recovery. Specifically, the CA branch (left) learns to emphasize semantically relevant channels, while the SA branch (middle) refines spatial context, both working together to improve boundary delineation—a key challenge in polyp segmentation.

To further control the fusion process, a lightweight gating branch is introduced in parallel with the attention pathway, as shown on the right side of Fig. 5. The gate computes a global modulation vector from the input features, which modulates the outputs of the two attention branches, reducing over-activation and suppressing background noise. Finally, we apply residual fusion to combine the gated attention features with the original input, as indicated by the element-wise addition operations in Fig. 5, ensuring information preservation and stable gradient flow.

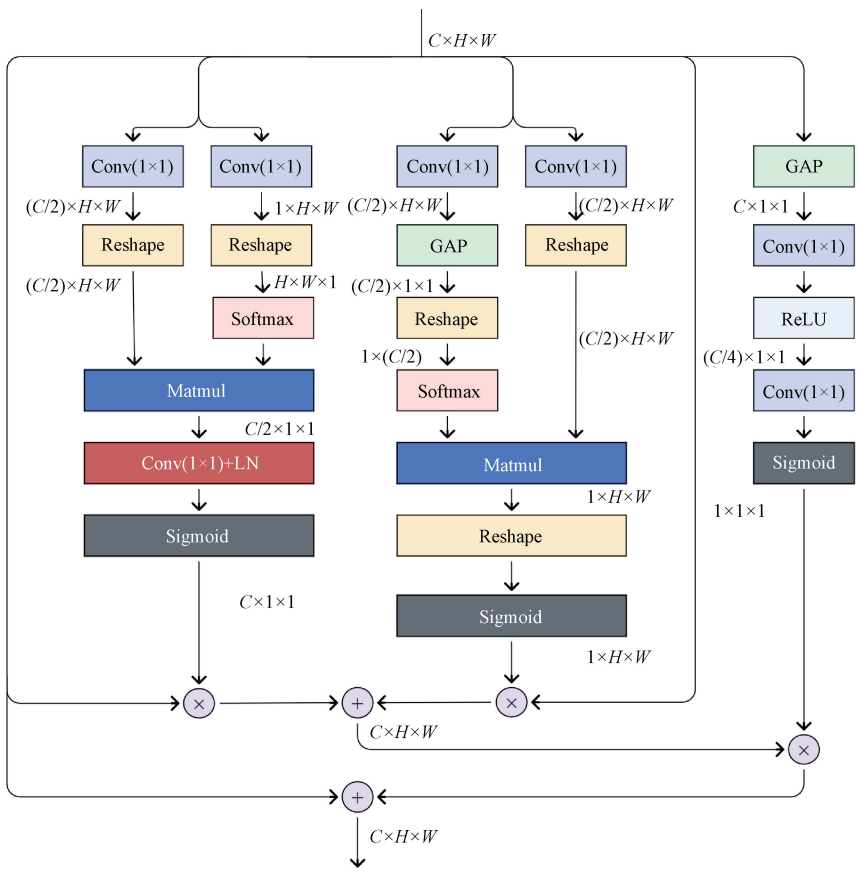



Fig. 5 Structure of PGF-Unit module

We position the PGF-Unit at the skip connections rather than in the encoder or decoder paths based on both theoretical considerations and empirical results. In U-Net-like architectures, skip connections deliver high-resolution but semantically weak feature maps from the encoder to the decoder. These features are essential for accurate boundary reconstruction but may contain noise or lack discriminative context.

By applying the gated attention mechanism at this junction, the PGF-Unit selectively enhances edge-relevant spatial cues while suppressing irrelevant activations, thus optimizing feature fusion. Our ablation study confirms this placement choice: moving PGF-Unit to the encoder or decoder stages results in a slight drop in segmentation performance, while skip connection placement consistently yields the highest mean intersection over

union (mIoU) and F1 score. This evidence highlights the skip connection as the most effective location for boundary-sensitive refinement.

Originally introduced for pixel-wise regression tasks such as saliency detection[14], PPSA separates the attention mechanism into two orthogonal branches, CA and SA, which are processed in parallel rather than in sequence. This decoupling enables more focused modeling of fine-grained image details while maintaining computational efficiency.

The PPSA module is based on the polarized self-attention (PSA) mechanism and adopts a parallel architecture to separately model CA and SA. Its structure is illustrated in Fig. 6.

In the CA branch, the input feature map undergoes global average pooling along the spatial dimensions to generate the CA map $\boldsymbol{Y}_c$, computed as

$$\boldsymbol{Y}_c = \sigma(\mathrm{LN}(\boldsymbol{W}_z(\boldsymbol{W}_v^c(\boldsymbol{X}) \otimes \mathrm{softmax}(\mathrm{reshape}(\boldsymbol{W}_q^c(\boldsymbol{X}))))) \otimes \boldsymbol{X}, \quad (9)$$

where $\boldsymbol{W}_q^c$ is the query projection in the CA branch; $\boldsymbol{W}_v^c$ is the value projection in the CA branch; $\boldsymbol{W}_z$ is the final projection layer after attention aggregation.

In the SA branch, the input feature map undergoes global average pooling along the channel dimension to generate the SA map $\boldsymbol{Y}_s$, computed as

$$\boldsymbol{Y}_s = \sigma(\mathrm{reshape}(\mathrm{softmax}(\mathrm{reshape}(\mathrm{GAP}(\boldsymbol{W}_q^s(\boldsymbol{X})))) \otimes \mathrm{reshape}(\boldsymbol{W}_v^s(\boldsymbol{X})))) \otimes \boldsymbol{X}, \quad (10)$$

where $\boldsymbol{W}_q^s$ is the query projection in the SA branch; $\boldsymbol{W}_v^s$ is the value projection in the SA branch; GAP denotes the global average pooling operation.

Finally, the outputs from the channel and spatial branches are summed to achieve parallel fusion in PPSA.

$$\boldsymbol{Y} = \boldsymbol{Y}_c + \boldsymbol{Y}_s. \quad (11)$$

In SGFNet, a lightweight variant of PPSA is applied at skip connections and decoder stages, which are locations rich in spatial detail and semantic context. This setup helps refine feature fusion and improves the localization of small or low-contrast polyps with ambiguous edges. This dual-branch attention improves boundary precision while preserving computational efficiency, making it a practical fit for clinical segmentation tasks.

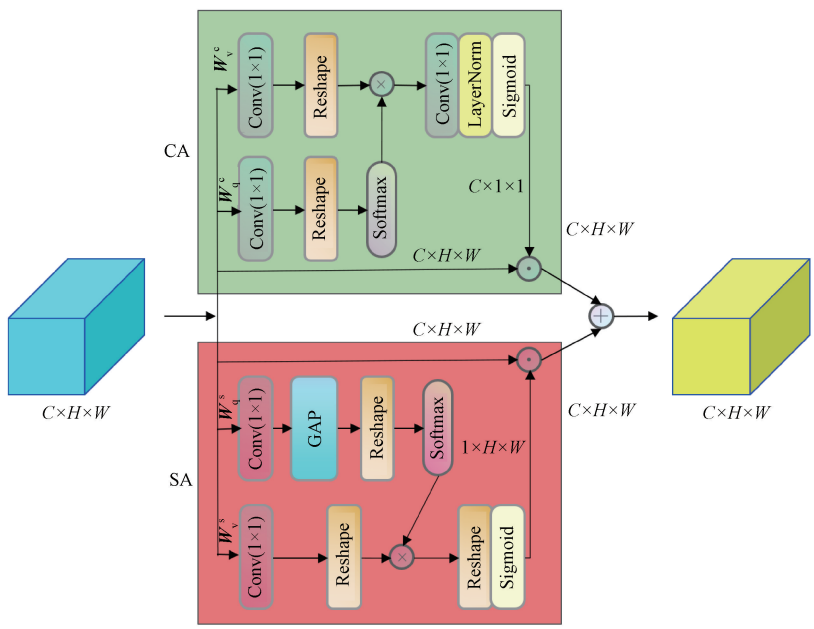



Fig. 6 Structure of PPSA module

## 1.4 Loss function

Class imbalance, often manifested as a long-tailed distribution between foreground and background pixels, is a common challenge in many semantic segmentation datasets. This imbalance tends to bias models toward the dominant background class, making it difficult to accurately segment smaller or less distinct objects. In colonic polyp segmentation tasks, label images typically exhibit extreme class imbalance between foreground (polyp) and background pixels. Using standard cross-entropy loss tends to bias learning toward the background class, leading to under-segmentation of small or low-contrast polyps.

To mitigate this issue, we adopt a weighted cross-entropy (WCE) loss $L_{\mathrm{WCE}}$, inspired by class-sensitive weighting strategies[15], and tailor it to emphasize polyp regions during training. By assigning higher weights to polyp pixels, the loss encourages the network to focus on small, underrepresented regions and improves discrimination. The mathematical formulation of this loss

function is

$$L_{\mathrm{WCE}} = -\frac{1}{N}\sum_{g=1}^{N}\sum_{l=1}^{L} w_l \cdot y_{g,l} \cdot \log(p_{g,l}), \quad (12)$$

where $N$ is the total number of pixels in the input image; $L$ is the number of classes (here $L=2$, corresponding to polyp and background); $w_l$ denotes the weight assigned to class $l$; $y_{g,l}$ is the one-hot label of the $g$th pixel for class $l$ (i.e., $y_{g,l}=1$ if the $g$th pixel belongs to class $l$, otherwise $y_{g,l}=0$); $p_{g,l}$ represents the predicted probability that the $g$th pixel belongs to class $l$; $\log(\cdot)$ denotes the natural logarithm function.

The weight $w_l$ is determined by the ratio of samples in class $l$ to the total number of samples across all classes, expressed as

$$w_l = \frac{L \cdot N_l^{-1}}{\sum_{g=1}^{L} N_g^{-1}}, \quad (13)$$

where $N_l$ denotes the number of pixels belonging to class $l$; $N_g^{-1}$ represents the reciprocal of the number of pixels belonging to the $g$th class.

### 1.5 Model evaluation metrics

To validate the effectiveness of the SGFNet model, this study employs three key performance metrics: mIoU[16], pixel accuracy (PA), and F1 score.

The mIoU metric, calculated as the mean of class-wise intersection over union (IoU) values, quantifies the region overlap between predicted segmentation and ground truth. Higher mIoU values indicate superior segmentation performance, computed as

$$m_{\mathrm{IoU}} = \frac{1}{L}\sum_{l=1}^{L}\mathrm{IoU}_l = \frac{1}{L}\sum_{l=1}^{L}\frac{|\hat{Y}_l \cap Y_l|}{|\hat{Y}_l \cup Y_l|}, \quad (14)$$

where $m_{\mathrm{IoU}}$ represents the mIoU value; $\hat{Y}_l$ and $Y_l$ denote the predicted and ground-truth pixel sets for class $l$, respectively.

PA quantifies the proportion of correctly classified pixels, defined as the ratio of true positives plus true negatives to the total number of pixels. The calculation formula for PA $P_{\mathrm{a}}$ is

$$P_{\mathrm{a}} = \frac{T_{\mathrm{p}} + T_{\mathrm{n}}}{T_{\mathrm{p}} + T_{\mathrm{n}} + F_{\mathrm{p}} + F_{\mathrm{n}}}, \quad (15)$$

where $T_{\mathrm{p}}$ (true positive) refers to the number of polyp pixels that are correctly identified and classified as polyps; $T_{\mathrm{n}}$ (true negative) represents the number of background pixels that are accurately recognized and classified as background; $F_{\mathrm{p}}$ (false positive) denotes the background pixels that are incorrectly misclassified as polyps; $F_{\mathrm{n}}$ (false negative) stands for the polyp pixels that are mistakenly classified as background.

The F1 score $F_1$ represents the harmonic mean of precision and recall, providing a balanced assessment of the model's accuracy and completeness in detection. Mathematically equivalent to the Dice coefficient in binary semantic segmentation tasks, it is calculated as

$$F_1 = \frac{2T_{\mathrm{p}}}{2T_{\mathrm{p}} + F_{\mathrm{p}} + F_{\mathrm{n}}}. \quad (16)$$

## 2 Experimental Results and Analyses

The proposed model was implemented in PyTorch and trained on a Linux system (Ubuntu 22.04 LTS). The training was performed on a machine equipped with an NVIDIA RTX 3090 GPU and 64 GB RAM. We used a batch size of 4 and an initial learning rate of 0.001. The model was trained for 300 epochs using the Adam optimizer.

### 2.1 Dataset integration and preprocessing

In this study, we constructed a unified polyp segmentation dataset by integrating two publicly available datasets: Kvasir-SEG and CVC-ClinicDB. Although both datasets offer pixel-level annotations, they differ in file formats, resolutions, and encoding schemes, which necessitate careful preprocessing to ensure compatibility.

The original Kvasir-SEG dataset consists of 1 000 color images in JPEG format and their corresponding JPEG-encoded binary masks. The images vary in resolution, ranging from 332 × 487 to 1 920 × 1 072 pixels. On the other hand, CVC-ClinicDB contains 612 images and masks stored in TIFF format, all with a fixed resolution of 384 × 288 pixels. Additionally, CVC-ClinicDB masks are stored as grayscale TIFFs, which require binarization before training.

We perform the following preprocessing steps to unify the datasets:

1) convert all images to JPEG format and all masks to PNG format for consistent encoding and compression;

2) ensure all ground-truth masks are binarized to 0 and 255 pixel values, regardless of the original encoding;

3) verify image-mask correspondence and filename consistency across both datasets.

Note that image resizing (e.g., to 512 × 512) is not part of the preprocessing pipeline but is handled dynamically during training.

After preprocessing, the unified dataset contains 1 612 image-mask pairs, which are randomly split into training (80%), validation (10%), and testing (10%) sets. The merged dataset combines the high-resolution diversity of Kvasir-SEG with the structural consistency of CVC-ClinicDB. This helps improve training stability and enhances cross-domain robustness.

Although our pipeline does not introduce novel transformations, it provides a systematic solution to a common challenge in medical image segmentation: the inconsistency in annotation formats, encoding standards,

and resolution variability across different public datasets.

By unifying the image and mask formats, enforcing binary consistency in annotations, and preserving resolution information for downstream resizing, we create a reliable, reproducible, and extensible dataset foundation. This integrated dataset not only facilitates our own experiments but also lays the groundwork for future polyp segmentation studies that require consistent cross-dataset training.

To our knowledge, few prior studies have formally documented the data integration process for Kvasir-SEG and CVC-ClinicDB, despite their frequent joint usage. Our work provides a reproducible protocol that supports consistent cross-dataset training.

While both Kvasir-SEG and CVC-ClinicDB are widely used in polyp segmentation research, they differ substantially in image resolution ($720 \times 576$ vs. $384 \times 288$), annotation granularity (polygonal vs. pixel-level labeling), and acquisition conditions (e. g., endoscope models and lighting setups). By integrating them into a unified training set, we introduce realistic domain shifts that promote robustness to diverse clinical scenarios. This heterogeneous composition serves as a practical proxy for generalization assessment in the absence of external unseen datasets, simulating variability encountered in real-world deployment.

## 2.2 Comparative experiments

We conducted comparative experiments to evaluate SGFNet against six representative segmentation models for colonic polyp segmentation: the classical U-Net, its enhanced variant UNet++, the lightweight attention-based SANet, the multi-scale residual model ResUNet++, the edge-refined EnhancedNet, and the transformer-based TransUNet, which integrates global self-attention with convolutional backbones for medical segmentation. The results are summarized in Table 1.

**Table 1** Performance comparison of different semantic segmentation models on polyp segmentation task

| Model | mIoU/% | PA/% | F1 score |
|---|---|---|---|
| U-Net | 81.75 | 88.57 | 0.903 |
| UNet++ | 83.56 | 88.15 | 0.916 |
| SANet | 88.15 | 92.43 | 0.940 |
| ResUNet++ | 87.68 | 92.53 | 0.933 |
| EnhancedNet | 86.52 | 91.75 | 0.934 |
| TransUNet | 88.09 | 92.66 | 0.935 |
| SGFNet (ours) | **89.53** | **93.98** | **0.948** |

SGFNet achieves the highest performance across all metrics, outperforming the best convolutional baseline SANet by 1.38% (note that all percentage increases mentioned in this paper are absolute improvements) in mIoU, and further surpassing TransUNet by 1.44% in mIoU and 1.32% in PA. This highlights SGFNet's robustness and effectiveness over both classical and transformer-based segmentation models.

These performance gains reflect the effectiveness of SGFNet's coordinated architectural design, which explicitly targets the challenges of polyp segmentation. In contrast to models with isolated enhancements (e. g., deeper backbones or standard attention modules), SGFNet integrates two complementary modules in a unified framework, leading to consistent performance gains across all evaluation metrics.

## 2.3 Segmentation performance

To qualitatively evaluate SGFNet's segmentation performance, we conducted a comparative visualization study against four representative models: EnhancedNet, ResUNet++, SANet, and TransUNet, using representative validation samples. The results are presented in Fig. 7, where each row corresponds to a test case displaying the original image, ground truth mask, and predicted segmentation maps from these models. In all images, white regions denote predicted polyp areas, and black regions represent the background.

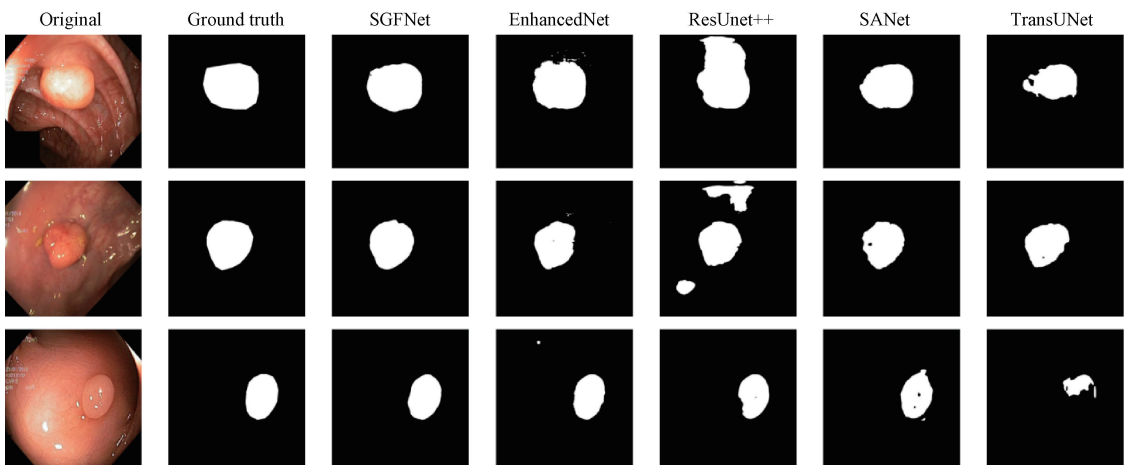



Fig. 7 Visual comparison of segmentation results from different models on colonic polyp images

SGFNet consistently generates segmentation masks with sharper boundaries, more complete regions, and better background suppression, closely matching the ground truth. Its predictions maintain smooth contours and exhibit minimal fragmentation or false positives, preserving both global structure and fine-grained details.

In comparison, EnhancedNet captures the general polyp shape but tends to blur boundaries and over-segment the foreground, particularly in rows 1 and 2 of Fig. 7, where obvious edge misclassifications occur. ResUNet++ exhibits foreground fragmentation and insufficient background suppression, especially in the upper region of row 2. SANet generates smooth predictions but often misses small or low-contrast polyps, leading to incomplete masks or hollow regions (e. g., rows 2 and 3). TransUNet, while benefiting from global self-attention, exhibits internal holes or boundary gaps in some cases (e. g., rows 1 and 3), likely due to the difficulty of integrating global context with fine edge cues.

These observations highlight the effectiveness of SGFNet's multi-scale encoding (via SSE-Encoder) and boundary-aware fusion (via PGF-Unit), particularly for capturing fine boundaries and detecting small lesions. These results confirm SGFNet's effectiveness in diverse endoscopic scenarios, reinforcing its value in clinical applications.

Although explicit boundary-specific metrics such as boundary IoU or Hausdorff distance are not included in this study, the reported mIoU and F1 score provide effective proxies for evaluating segmentation quality near object boundaries. Both metrics are sensitive to discrepancies between predicted and ground-truth masks, particularly at the edges. A higher mIoU indicates better spatial overlap, while the F1 score reflects a balance between precision and recall, both of which are directly influenced by accurate contour localization.

Visual results in Fig. 7 show that SGFNet consistently produces smooth and compact contours with minimal false positives or leakage. While these examples do not include extremely challenging cases such as villous or flat polyps, they still demonstrate SGFNet's robustness in maintaining boundary integrity under varied lighting and background conditions.

### 2.4 Ablation study

To assess the individual and combined contributions of SSE-Encoder and PGF-Unit, we conducted a series of ablation experiments. All variants were trained under identical settings on the same validation set and evaluated using mIoU, PA, and F1 score. The results are shown in Table 2.

SGFNet outperforms all other variants across metrics, achieving an mIoU improvement of 2.62% and a PA increase of 2.02% over the baseline ResUNet. Among single-module variants, the SSE-Encoder yields the largest gains, indicating its strong effect on multi-scale feature encoding. The combination of both modules in SGFNet leads to further gains, confirming their complementary effects.

**Table 2** Performance comparison of different improvement strategies on colonic polyp segmentation task

| Model | mIoU/% | PA/% | F1 score |
|---|---|---|---|
| ResUNet | 86.91 | 91.96 | 0.934 |
| ResUNet+SSE-Encoder | 88.64 | 93.30 | 0.945 |
| ResUNet+PGF-Unit(encoder) | 86.51 | 91.05 | 0.929 |
| ResUNet+PGF-Unit (skip connection) | 87.25 | 92.30 | 0.937 |
| ResUNet+PGF-Unit (decoder) | 86.70 | 91.20 | 0.931 |
| SGFNet (ours) | **89.53** | **93.98** | **0.948** |

These ablation results highlight the distinct contributions of each module: the SSE-Encoder primarily enhances semantic encoding, while the PGF-Unit refines boundary precision. Their integration delivers balanced improvements across metrics, offering practical insights for future segmentation model design.

### 2.5 Model complexity analysis

To evaluate the computational efficiency of SGFNet, we compared its number of parameters and floating-point operations (FLOPs) with several representative polyp segmentation models under identical input conditions. These two metrics serve as standard, framework-independent indicators of model size and inference complexity. The comparison is summarized in Table 3.

**Table 3** Complexity comparison of SGFNet and baseline models

| Model | Number of parameters | FLOP |
|---|---|---|
| TransUNet | $45.95 \times 10^6$ | $29.64 \times 10^9$ |
| EnhancedNet | $49.51 \times 10^6$ | $96.93 \times 10^9$ |
| UNet++ | $220.33 \times 10^6$ | $91.05 \times 10^9$ |
| SGFNet (ours) | $\mathbf{38.35 \times 10^6}$ | $\mathbf{93.98 \times 10^9}$ |

As shown in Table 3, SGFNet contains $38.35 \times 10^6$ parameters, about 82.59% fewer than UNet++ and 16.54% fewer than TransUNet, while maintaining comparable FLOPs to much larger models. This indicates that SGFNet is significantly more compact while retaining sufficient representational capacity for high-accuracy segmentation. The relatively balanced FLOPs ($93.98 \times 10^9$) suggest a moderate inference cost, which remains acceptable for real-time deployment scenarios. Overall, SGFNet achieves a favorable trade-off between segmentation performance and computational overhead, making it a strong candidate for resource-constrained clinical environments such as portable endoscopy systems or edge devices.

## 3 Conclusions

In this work, we propose SGFNet, a lightweight and effective semantic segmentation framework based on a

U-Net backbone. SGFNet achieves strong performance on colonic polyp segmentation in endoscopic images, demonstrating its capability in handling challenging medical segmentation tasks. Rather than pursuing highly complex architectures, SGFNet emphasizes a balance between structural simplicity and segmentation performance. Experiments on a unified dataset combining Kvasir-SEG and CVC-ClinicDB demonstrate consistent improvements across mIoU, PA, and F1 score. Ablation studies further confirm the complementary roles of SSE-Encoder and PGF-Unit. Overall, SGFNet provides a practical and interpretable solution for medical image segmentation, particularly suitable for resource-constrained clinical scenarios. Future work will explore more adaptive attention mechanisms and the deployment of SGFNet in real-time endoscopic diagnostic systems.

## References


[1] Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. *CA: A Cancer Journal for Clinicians*, 2024, 74(3): 229-263.

[2] Yang X, Zhu C, Li Y, et al. Incidence and mortality of post-polypectomy colorectal cancer in patients with low-risk adenomas: a systematic review and meta-analysis of observational studies [J]. *Digestive Diseases*, 2023, 41(2): 206-216.

[3] Liu J Q, Zhang W W, Liu Y, et al. Polyp segmentation based on implicit edge-guided cross-layer fusion networks [J]. *Scientific Reports*, 2024, 14: 11678.

[4] Zhou Z W, Rahman Siddiquee M M, Tajbakhsh N, et al. UNet++: a nested U-Net architecture for medical image segmentation [C]//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Cham: Springer, 2018: 3-11.

[5] Jha D, Smedsrud P H, Riegler M A, et al. ResUNet++: an advanced architecture for medical image segmentation [C]//2019 IEEE International Symposium on Multimedia (ISM). Piscataway, NJ: IEEE, 2020: 225-2255.

[6] Wei J, Hu Y W, Zhang R M, et al. Shallow attention network for polyp segmentation [C]// Medical Image Computing and Computer Assisted Intervention-MICCAI 2021. Cham: Springer, 2021: 699-708.

[7] Jha D, Smedsrud P H, Riegler M A, et al. Kvasir-SEG: a segmented polyp dataset [C]// MultiMedia Modeling. Cham: Springer, 2020: 451-462.

[8] Bernal J, Sánchez F J, Fernández-Esparrach G, et al. WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians[J]. *Computerized Medical Imaging and Graphics*, 2015, 43: 99-111.

[9] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module [M]// Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.

[10] Song W H, Gao M L, Chehri A, et al. Dual-branch and triple-attention network for pan-sharpening[J]. *Applied Intelligence*, 2024, 54(17): 8041-8058.

[11] Hua X C, Cheng K, Lu H, et al. MSCMNet: multi-scale semantic correlation mining for visible-infrared person re-identification [J]. *Pattern Recognition*, 2025, 159: 111090.

[12] Li X, Wang W H, Hu X L, et al. Selective kernel networks [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 510-519.

[13] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7132-7141.

[14] Liu H J, Liu F Q, Fan X Y, et al. Polarized self-attention: towards high-quality pixel-wise regression [PP/OL]. arXiv (2021-07-08) [2025-04-02]. https://arxiv.org/abs/2107.00782.

[15] Dai L, Wang J L, Chen Z Y, et al. Non-equivalent point cloud segmentation method based on convolutional neural network[J]. *Journal of Donghua University (Natural Science)*, 2019, 45(6): 862-868. (in Chinese).

[16] Rahman M A, Wang Y. Optimizing intersection-over-union in deep neural networks for image segmentation [C]//Advances in Visual Computing. Cham: Springer, 2016: 234-244.

# SGFNet：一种用于内窥镜图像结肠息肉分割的注意力增强网络

庞　博，刘国华*
东华大学 计算机科学与技术学院，上海 201620

**摘　要：**准确分割内窥镜图像中的结肠息肉对于早期诊断和预防结直肠癌至关重要。然而，由于息肉形态差异大、与周围黏膜对比度低以及边界模糊等问题，这项任务仍然具有挑战性。为了解决这些问题，我们提出了一种选择性门控融合网络（selective-gated fusion network，SGFNet），这是一种基于 U-Net 架构的轻量级语义分割网络。SGFNet 包含两个目标模块：SSE-Encoder，它集成了选择性核卷积和挤压与激励注意力机制，以增强多尺度表征和通道重校准；PGF-Unit，它采用门控并行极化自注意力机制来改进边界敏感的特征融合。SGFNet 在 Kvasir-SEG 和 CVC-ClinicDB 的 1 612 张带注释图像的统一数据集上进行了评估。实验结果表明，SGFNet 在平均交并比（mean intersection over union，mIoU）、像素精度（pixel accuracy，PA）和 F1 值方面均表现出色，超越了多个代表性基线模型。凭借其均衡的设计和强大的泛化能力，SGFNet 为现实世界的医学图像分割提供了实用且可解释的解决方案，并为未来计算机辅助诊断系统中基于注意力机制的模块集成提供了借鉴。

**关键词：**结肠息肉分割；U-Net；多尺度特征编码；注意力机制；SGFNet；医学图像分析