中图法分类号:TP18;TP391.41　　文献标识码:A　　文章编号:1006-8961(2026)11-4133-14

论文引用格式:Liu S J and Tong M J. 2026. Swin-EdgeNet: medical image segmentation with multi-scale edge enhancement based on Transformer. Journal of Image and Graphics, 31(11):4133-4146(刘善洁, 童孟军. 2026. Swin-EdgeNet:基于Transformer的多尺度边缘增强医学图像分割模型. 中国图象图形学报, 31(11):4133-4146)[DOI:10.11834/jig.250592]

Swin-EdgeNet:基于Transformer的多尺度边缘增强医学图像分割模型

刘善洁,童孟军*

浙江农林大学人工智能学院,杭州　310000

摘　要:目的　传统医学图像分割方法依赖手工特征提取,难以应对复杂多样的医学图像,而现有深度学习方法在边缘细节保留和多尺度特征提取方面仍存在不足。本文旨在提出一种新型深度学习框架,以提升医学图像分割的精度与鲁棒性。**方法**　提出融合Swin Transformer与U-Net的混合架构,其中编码器采用Swin Transformer V2-B以捕获分层特征及长范围依赖关系,解码器引入多尺度边缘密集模块(multi-scale edge dense module,MS-EDM)以增强边缘保留能力,并结合双重注意力机制(包括细节增强双重注意力(detail-enhanced dual attention,DEVA)与双重注意力(dual attention,DA))以动态聚焦关键区域。为进一步优化模型效率,将传统Conv2d层替换为KanConv,在保持性能的同时减少参数数量。**结果**　在ISIC2018(International Skin Imaging Collaboration)皮肤病变、肺部分割、眼底血管分割和息肉分割4个公开数据集上的实验表明,本文方法在多项评价指标上均优于当前主流方法。具体而言,在ISIC2018数据集中,Dice和IoU(intersection over union)分别达到89.87%与83.97%;在肺部分割任务中Dice系数达97.62%;在眼底血管分割中灵敏度达80.19%,显著优于对比方法;在息肉分割任务中Dice系数达93.80%,召回率提升6.58%。可视化结果进一步验证了本文方法在边缘保留与小目标识别方面的优势。**结论**　本文提出的融合Swin Transformer与多尺度边缘增强机制的分割模型,能够有效综合Transformer的全局建模能力与CNN(convolutional neural network)的局部特征提取优势,在多种医学图像分割任务中均表现出更优的分割精度与泛化能力。

关键词:医学图像分割;神经网络架构;Swin Transformer;U-Net;多尺度特征提取;注意力机制

Swin-EdgeNet: medical image segmentation with multi-scale edge enhancement based on Transformer

Liu Shanjie, Tong Mengjun*

College of Artificial Intelligence, Zhejiang A&F University, Hangzhou　310000, China

Abstract: Objective Medical image segmentation is crucial in computer-aided diagnosis, treatment planning, and clinical research. Traditional segmentation methods often rely on handcrafted features and classical image processing algorithms, which often struggle with the complexity, heterogeneity, and low-contrast characteristics commonly found in modern medical imaging data. Although deep learning methods, particularly convolutional neural networks (CNNs) and, more recently, Transformer-based architectures, have advanced the field through automatic feature extraction and state-of-the-art performance, considerable challenges remain. Existing deep learning models often have limited capability to preserve fine boundary details and effectively leverage multiscale contextual information, leading to suboptimal segmentation of irregular

收稿日期:2025-12-05;修回日期:2026-03-19;预印本日期:2026-03-26

*通信作者:童孟军　tmj@zafu.edu.cn

or small anatomical structures and pathological regions. Aiming to address these limitations, this study aims to design and validate a novel deep learning framework that synergistically combines the advantages of hierarchical feature learning, long-range dependency modeling, and precise edge delineation to enhance the accuracy, robustness, and generalizability of medical image segmentation across diverse tasks and datasets. **Method** This paper introduces Swin-EdgeNet, a hybrid encoder-decoder architecture for medical image segmentation that innovatively integrates a Swin Transformer backbone with a U-Net style decoder enhanced with specialized modules. The encoder is built upon Swin Transformer V2-B, which serves as a powerful feature extractor. Its shifted-window self-attention mechanism efficiently captures hierarchical feature representations and establishes long-range spatial dependencies across the image, providing a rich global context that is often missing in pure CNN architectures. The corresponding decoder is designed to progressively recover spatial resolution and generate precise segmentation maps. Aiming to specifically address the challenge of boundary ambiguity, a multiscale edge dense module (MS-EDM) is introduced within the decoder. This module employs parallel convolutional pathways with different receptive fields (dilated convolutions) to simultaneously capture and fuse features at multiple scales. Additionally, the dense connections within MS-EDM facilitate feature reuse and strengthen gradient flow, preserving and enhancing fine edge details and local textures throughout the upsampling process. Furthermore, a dual-attention mechanism is incorporated to dynamically focus the model's capacity on semantically critical and structurally salient regions. This mechanism comprises two components: a detail-enhanced visual attention (detail-enhanced dual attention, DEVA) module and a dual attention (dual attention, DA) module. The DEVA module refines feature maps along the channel dimension by emphasizing interdependent channel maps, while the DA module combines spatial and channel attention to selectively aggregate contextual information from relevant positions and feature channels. This dual strategy helps suppress irrelevant background noise while highlighting subtle structures and pathological areas. Aiming to optimize computational efficiency without compromising performance, standard Conv2d layers in specific parts of the decoder are replaced with KanConv layers, a parameter-efficient alternative that reduces the overall model parameters. The entire network is trained end-to-end using a combined loss function, typically involving Dice loss and cross-entropy loss, to address class imbalance. **Result** Extensive experiments and quantitative evaluations were conducted on four public benchmark datasets encompassing distinct medical image segmentation tasks to validate the effectiveness and generalization capability of the proposed Swin-EdgeNet. On the International Skin Imaging Collaboration (ISIC2018) skin lesion segmentation dataset, the proposed method achieved a Dice similarity coefficient (DSC) of 89.87% and an intersection over union (IoU) of 83.97%, outperforming existing state-of-the-art methods such as Rolling-Unet. The high recall rate of 93.44% indicates the model's strong capability to capture complete lesion areas while minimizing false negatives. For the lung segmentation task, the model also demonstrated exceptional performance, attaining a DSC of 97.62% and an IoU of 95.44%, surpassing strong baselines such as DCSAU-Net (DSC: 97.56%, IoU: 95.35%) and HTC-Net (DSC: 97.48%, IoU: 95.19%). Additionally, the model maintained high precision (97.12%), recall (95.94%), and F1-score (96.44%), indicating a robust balance between accuracy and coverage. In the challenging domain of retinal fundus image vessel segmentation, where detecting thin capillary structures is difficult, the proposed method achieved a notably high sensitivity of 80.19%, substantially exceeding that of models such as U-Net++ (61.56%) and M2SNet (43.46%). This finding highlights its enhanced capability for segmenting low-contrast, fine-grained structures. The model also attained an F1-score of 77.77% and an AUC of 88.63%. For the polyp segmentation task, Swin-EdgeNet achieved a DSC of 93.80% and a recall of 91.93%, representing improvements of 2.12% and 6.58%, respectively, over models such as Trans-UNet. The F1-score for polyp segmentation reached 89.81%. Qualitative visual analysis further corroborated these quantitative findings. As illustrated in the figures, the proposed method produces segmentation masks with noticeably sharper boundaries and more accurate contours for skin lesions. Additionally, the method retains more complete small vessel branches in retinal images and provides more precise coverage for flat and sessile polyps compared with other models, which often suffer from under-segmentation or over-segmentation issues. **Conclusion** This paper presents Swin-EdgeNet, a robust and effective deep-learning framework for medical image segmentation. By integrating the global contextual understanding of the Swin Transformer encoder with the detailed, localized feature recovery of a progressively enhanced U-Net decoder, the proposed model effectively bridges the gap between high-level semantics and low-level spatial precision. The introduced MS-EDM and dual-attention mechanism

collectively address key challenges in medical imaging, including the preservation of critical edge information, the effective utilization of multiscale contextual information, and the dynamic focus on diagnostically relevant regions. Additionally, the replacement of standard convolutions with KanConv contributes to a more parameter-efficient design. Comprehensive experimental results across four diverse and challenging clinical tasks—skin lesion segmentation, lung segmentation, retinal vessel segmentation, and polyp segmentation—consistently demonstrate that the proposed method achieves superior or highly competitive performance compared with current state-of-the-art approaches. Overall, these results confirm the strong generalization capability of Swin-EdgeNet and highlights its potential for improving the accuracy and reliability of computer-aided diagnosis systems.

Key words: medical image segmentation; neural network architecture; Swin Transformer; U-Net; multi-scale feature extraction; attention mechanism

0 引　言

医学图像分割是临床诊断和手术规划的关键环节（Litjens 等，2017）。早期的阈值分割（Otsu，1979）、区域生长（Adams 和 Bischof，1994）和边缘检测（Canny，1986）等方法依赖手工特征，在处理复杂病理结构时往往因抗噪性和泛化力不足而受限（Litjens 等，2017）。

深度学习的兴起为医学图像分割带来了革命性突破。全卷积网络（Lecun 等，2015）首次实现了端到端的像素级预测，为后续研究奠定了基础。U-Net（Ronneberger 等，2015）凭借其独特的对称编码器—解码器结构和跳跃连接，在医学图像分割领域取得了显著成功。国内学者也对从卷积神经网络到视觉 Transformer 的演进历程进行了全面的系统梳理（石军 等，2025）。在此基础之上，研究者提出了一系列改进方案：UNet++（Zhou 等，2018）通过嵌套密集连接增强了特征传播效率；Attention U-Net（Oktay 等，2019）引入了注意力机制，使模型能够聚焦于重要区域；3D U-Net（Çiçek 等，2016）扩展了三维医学图像的处理能力。

然而，卷积神经网络（convolutional neural network，CNN）固有的局部感受野限制了其对长程依赖关系的建模能力，在处理具有复杂空间结构的解剖目标时表现不佳（Wang 等，2021）。为此，研究者引入了具有全局建模能力的 Transformer（Vaswani 等，2017）。

为了平衡局部细节与全局语义，CNN 与 Transformer 的混合架构应运而生：Trans-UNet（Chen 等，2024）率先将 Transformer 嵌入 U-Net 瓶颈层；UNETR（UNEt Transformers）（Hatamizadeh 等，2022）则利用纯 Transformer 编码器结合 CNN 解码器处理 3D 容积数据；nnFormer（Zhou 等，2023）进一步将 Transformer 集成至 nnU-Net 框架以提升特征表达。此外，针对医学图像普遍存在的边界模糊问题，XBound-Former（Wang 等，2023）等研究开始探索基于 Transformer 的多尺度边界增强机制。虽然 Swin Transformer（Liu 等，2021）通过窗口机制降低了计算复杂度，但如何在保持高效推理的同时，进一步解决跨域泛化难（Ghafoorian 等，2017；Oktay 等，2019）及人体器官结构复杂、微小病灶边缘模糊导致分割精度不足（蒋婷和李晓宁，2024）的问题，仍是当前亟待解决的挑战。

针对上述问题，本研究提出了一种新型医学图像分割框架 Swin-EdgeNet。该框架以 Swin Transformer V2（Liu 等，2022）作为编码器，充分利用其强大的全局上下文建模能力；在解码器中引入多尺度边缘密集模块和双重注意力机制，增强对边缘细节和小目标的感知能力；同时采用 KanConv 替代传统卷积，在保持性能的前提下降低模型复杂度。通过在多个公开数据集上的系统验证，证明了该方法在保持优异分割性能的同时，具有更好的泛化能力和边缘保持特性。

1 方　法

1.1 整体架构

本文提出的 Swin-EdgeNet 模型是一种专为医学图像分割设计的深度神经网络架构，其核心思想是通过层次化的特征提取与精细化的特征融合实现高精度的像素级分割。如图 1 所示，模型整体采用编码器—解码器结构，但与传统 U-Net 不同，其编码器采用先进的 Swin Transformer V2-B 作为主干网络，而解码器部分则创新性地结合了多尺度边缘密集模

块(multi-scale edge dense module, MS-EDM)和双重注意力(detail-enhanced dual attention, DEVA)机制,逐步恢复空间细节并抑制无关噪声,最终实现病灶区域的精准定位。

在编码阶段,模型首先通过一个动态调整的前置卷积层将输入图像适配到Swin Transformer的输入维度。当处理医学影像常见的单通道数据时,该层会自动调整卷积核参数,避免传统RGB三通道预训练模型的特征错位问题。在编码器末端,模型引入了MS-EDM模块作为桥梁。通过并行执行不同膨胀率的空洞卷积捕获多尺度上下文信息,并利用门控融合机制自适应地加权整合各分支特征,显著提升了模型对病灶大小和形状的鲁棒性。解码器部分采用对称的U型结构,为了解决深层特征在传递过程中细节丢失的问题,在解码器的上采样层级嵌入了细节增强双重注意力(DEVA)机制,利用跳跃连接传递的浅层特征动态校准注意力权重,从而有效抑制背景噪声并聚焦微小病灶。特别地,为了在引入上述复杂注意力模块的同时防止模型参数量过度膨胀,在解码器的上采样阶段策略性地引入了KanConv替代部分标准卷积层(Conv2d)。这一设计作为一种高效的参数控制机制,在保持模型非线性表达能力的同时,实现了分割精度与计算负载的有效平衡。下面将详细描述各个核心模块的设计细节。

1.2 编码器结构

医学图像通常具有高分辨率且包含多尺度的解剖结构,这对骨干网络的显存效率与训练稳定性提出了极高要求。为此,本文选用Swin Transformer V2作为编码器,旨在利用其强大的全局建模能力捕获长程依赖,同时克服传统Transformer在处理高分辨影像时的容量瓶颈。如图1所示,编码器首先通过Patch Partition模块对输入图像进行分块,随后经

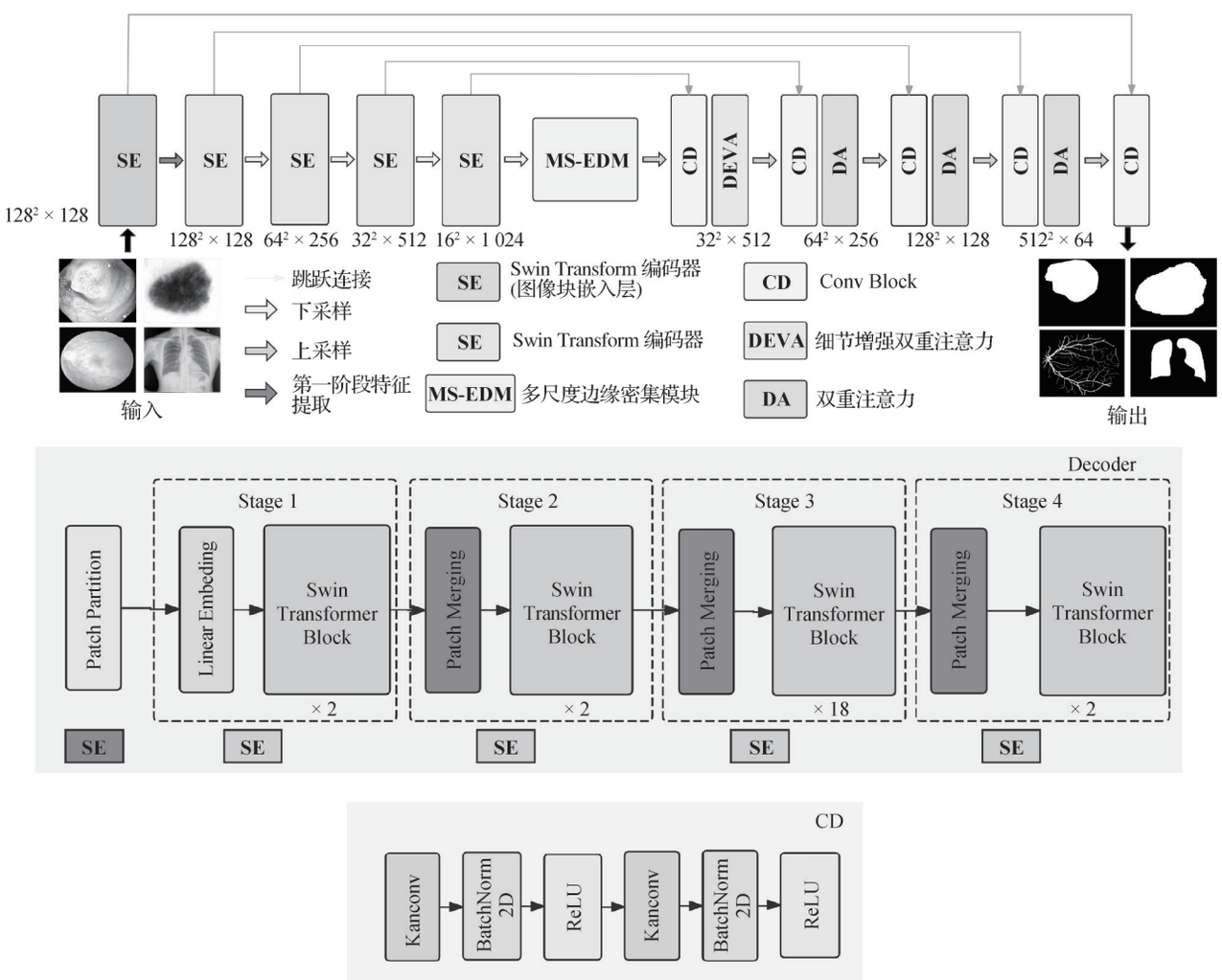


图1 本文提出的Swin-EdgeNet架构

Fig. 1 The architecture of our proposed Swin-EdgeNet

4 个 Stage 构建多尺度特征。除 Stage 1 外，其余 Stage 均通过 Patch Merging 层进行下采样。

相较于 Swin Transformer V1，Swin Transformer V2 在 3 个关键技术上实现突破，使其更适配医学分割任务。

1）为解决部分注意力头被少数像素对主导的问题，V2 引入缩放余弦注意力函数，其像素对（i，j）的相似度计算为

$$Sim\left(q_i,k_j\right)=\cos\left(q_i,k_j\right)/\tau+B_{ij} \tag{1}$$

式中，B_{ij}是像素 i 和 j 的相对位置偏差，τ 是一个可学习的标量，在不同的头和层之间不共享，且值设置为大于 0.01。余弦函数固有的归一化特性确保了注意力权值的平稳分布。

2）V2 摒弃了直接学习参数化偏置的方法，采用小规模元网络生成连续位置偏置。具体为

$$B\left(\Delta x,\Delta y\right)=G\left(\Delta x,\Delta y\right) \tag{2}$$

式中，G 为默认两层的多层感知机（multilayer perceptron，MLP）（含 ReLU（rectified linear unit）激活）。该设计通过元网络为任意相对坐标生成偏置值，使其能够灵活适应不同窗口大小的任务。为增强大范围坐标的外推能力，模型将对数间隔坐标转换引入。具体为

$$\widehat{\Delta x}=sign\left(\Delta x\right)\cdot\log\left(1+\left|\Delta x\right|\right) \tag{3}$$

$$\widehat{\Delta y}=sign\left(\Delta y\right)\cdot\log\left(1+\left|\Delta y\right|\right) \tag{4}$$

式中，Δx，Δy 与 $\widehat{\Delta x}$，$\widehat{\Delta y}$ 分别为线性间隔与对数间隔坐标。

3）V2 将预归一化调整为后归一化方案，即在残差连接前进行归一化操作，此举有效提升了训练稳定性。

本文采用 SwinV2-B 作为编码器，其配置参数为：初始通道数 $C=128$，各 Stage 块数配置 block =（2，2，18，2）。该配置在保证强大的特征提取能力的同时，与后续解码器中的 KanConv 参数控制策略相配合，实现了模型性能与计算开销的良好平衡。

1.3 多尺度边缘密集模块

为了解决医学图像中“强语义、弱边界”的矛盾，本文提出了多尺度边缘密集模块（MS-EDM）。如图 1 所示，被部署在编码器与解码器的跳跃连接处，用于承接并增强多尺度特征。该模块在结构上由两个核心子组件构成：1）级联边缘注意力模块（cascaded edge attention，CEA）：负责利用 Sobel 算子从多尺度特征中提取边缘先验，解决边界模糊问题；2）动态门控融合机制：负责自适应地整合上下文信息与边缘特征，替代传统的线性相加。与传统的密集空洞卷积（dense atrous convolution，DAC）（Gu 等，2019）或特征金字塔网络（feature pyramid network，FPN）不同，MS-EDM 不仅利用不同膨胀率捕获多尺度上下文，更通过上述两个组件的协同作用，实现了基于边缘清晰度的特征重加权。

1.3.1 级联边缘注意力模块

级联边缘注意力模块通过改进的特征金字塔网络（FPN）架构，创新性地融入了级联边缘注意力机制。为了精准捕捉病灶轮廓，每个金字塔层级均引入了可学习的 Sobel 算子生成边缘响应图，如图 2 所示。具体而言，水平与垂直方向的梯度特征通过平方和开方运算进行融合，并引入微小常数（1×10^{-6}）以确保数值稳定性。边缘响应经 1×1 卷积调整后，通过 sigmoid 激活生成注意力权重，与原始特征进行点乘实现边缘增强。这种设计使网络能够自适应强化病灶边界特征，特别适用于医学图像中常见的弱边缘组织分割。

在多尺度特征融合过程中，模块采用级联式信息传递策略：高层级（如 $\boldsymbol{P}5$）的边缘信息通过双线性插值下采样后与低层级（如 $\boldsymbol{P}4$）特征相加，再输入下一级边缘注意力模块。这种设计保持了边缘特征在不同尺度间的连续性。各层级特征在融合前均经过 3×3 卷积平滑处理，以减轻上采样引入的混叠效应。最终输出通过 1×1 卷积将特征维度映射至目标类别数，确保网络适应不同分割任务的需求。

1.3.2 动态门控融合机制

在特征融合阶段，本文设计了动态门控机制以实现自适应特征融合。如图 3 所示，该机制摒弃了传统的串联或简单相加方式，转而采用基于注意力权重的智能融合策略。

具体而言，模块首先将 DAC 与 CEA 两个支路的输出特征在通道维度进行拼接，随后通过包含 1×1 卷积与 sigmoid 激活的轻量级网络生成空间—通道自适应的融合权重（取值范围 0～1）。最终输出采用门控加权求和公式，具体为

$$\boldsymbol{F}^{\text{out}}=gate\otimes\boldsymbol{A}+\left(1-gate\right)\otimes\boldsymbol{B} \tag{5}$$

式中，$\otimes$ 表示逐元素乘法，$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 分别代表 DAC 与 CEA 支路的特征图，$gate$ 表示由特征融合网络生成

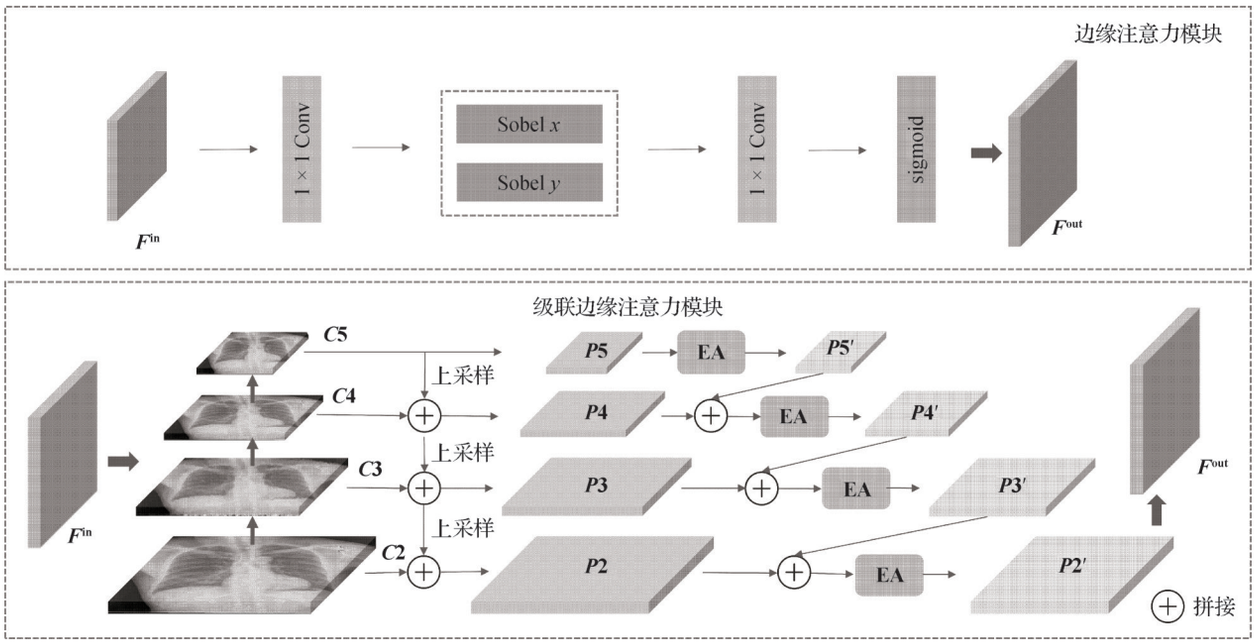


图2　Swin-EdgeNet的级联边缘注意力结构

Fig. 2　Cascaded edge attention structure of Swin-EdgeNet

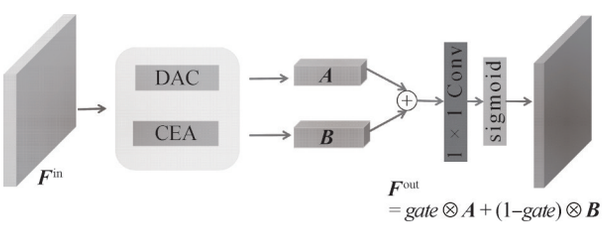


图3　Swin-EdgeNet 的多尺度边缘密集模块

Fig. 3　Multi-scale edge dense module of Swin-EdgeNet

的动态门控权重。这种融合设计具有三重优势：1）门控机制使网络能够根据局部特征特性动态调整各支路的贡献比例，如在组织边界区域增强边缘支路权重，而在同质区域侧重上下文特征；2）1×1卷积的引入实现了跨通道特征交互，避免了简单加权导致的通道间独立性损失；3）保持原始通道维度的设计确保模块可灵活嵌入各类分割网络，不会引起特征维度膨胀。

1.4　细节增强双重注意力

在医学图像分割任务中，细节信息的保留和关键区域的聚焦对于提高分割精度至关重要。经典的双重注意力模块（dual attention，DA）以卷积块注意力模块（convolutional block attention module，CBAM）（Woo等，2018）为代表，其通过串联通道与空间注意力子模块，从输入特征中自适应地提取关键信息。尽管CBAM在自然图像中表现优异，但其注意力的计算仅依赖于当前层级的深层特征。在医学分割中，经过多次下采样后的深层特征往往丢失了微小病灶的边界纹理，导致CBAM难以在这些模糊区域生成准确的注意力响应。

为了进一步恢复在深层网络中丢失的细节，本文提出了细节增强双重注意力（DEVA）机制，嵌入到解码器的上采样层级之后。与CBAM仅利用单输入特征不同，DEVA的核心创新在于引入了“双输入”交互策略，旨在通过动态调整特征权重，增强模型对目标区域的聚焦能力，同时保留更多的细节信息。DEVA机制如图4所示，是在保留传统的CBAM基础上，引入了浅层特征的细节信息，以增强模型对目标区域的细节捕捉能力。这种设计不仅增强了模型对复杂结构的聚焦能力，更从结构上解决了传统注意力机制在深层网络中细节丢失的问题。具体而言，对于输入的特征图 $\boldsymbol{F} \in \mathbf{R}^{C \times H \times W}$ 和细节特征 $\boldsymbol{F}_d$，通过全局平均池化（*GAP*）和全局最大池化（*GMP*）分别获取每个通道的全局统计信息。具体为

$$\boldsymbol{F}_d = Conv_{3\times 3}\left(\boldsymbol{F}_{sh}\right) \tag{6}$$

$$\begin{aligned} \boldsymbol{F}_{\text{gap}} &= GAP\left(\boldsymbol{F} + \boldsymbol{F}_d\right) \in \mathbf{R}^C \\ \boldsymbol{F}_{\text{gmp}} &= GMP\left(\boldsymbol{F} + \boldsymbol{F}_d\right) \in \mathbf{R}^C \end{aligned} \tag{7}$$

式中，$\boldsymbol{F}_{sh}$ 是细节特征，保留了浅层特征中的细节信息。空间注意力模块在DEVA中同样引入了细节特征 $\boldsymbol{F}_d$ 以增强空间注意力的计算。对于通道增强后的特征图 $\boldsymbol{F}_c$ 和细节特征 $\boldsymbol{F}_d$，通过卷积层提取其空间

特征。具体为

$$\boldsymbol{F}_{sh} = Conv\left(\boldsymbol{F}_c + \boldsymbol{F}_d\right) \tag{8}$$

后续的操作与空间注意力相同。通过细节卷积模块提取浅层特征的细节信息，并将其融入通道注意力和空间注意力的计算中，可以增强模型对目标区域的细节捕捉能力，尤其是在处理具有复杂结构和模糊边界的医学图像时，能够显著提高分割精度。

2 实　验

2.1 数据集

在本文的实验中，采用4个公开的医学图像数据集验证所提出方法的泛化能力，包括皮肤病变ISIC2018（International Skin Imaging Collaboration）、肺部X光CHNCXR（Shenzhen chest X-ray set）、眼底血管DRIVE（digital retinal images for vessel extraction）和消化道息肉Kvasir数据集。由于医学图像具有显著的领域差异性，这种多数据集组合能够全面评估模型在不同解剖结构和病理特征下的表现。ISIC2018、CHNCXR和Kvasir数据集均调整为512×512像素的标准分辨率图像，分别包含3 664幅皮肤镜图像、708幅胸部X光片和1 000幅内窥镜图像，这3个数据集的图像尺寸统一性有助于控制预处理阶段的复杂度（Codella等，2018）。特别地，DRIVE眼底数据集原始图像分辨率为960×960像素，考虑到该数据集仅有40幅总图像的实际情况，将其裁剪为128×128像素的小块，这一策略在不引入人工合成数据的前提下，有效地将样本数量扩充至原始规模的16倍（640幅），既缓解了数据稀缺问题，又保持了血管分割任务所需的局部细节特征。皮肤病变和眼底血管数据集均使用官方划分的训练/验证/测试集，而肺部和息肉数据集则按照7∶1.5∶1.5的比例进行随机划分以适配研究需求。

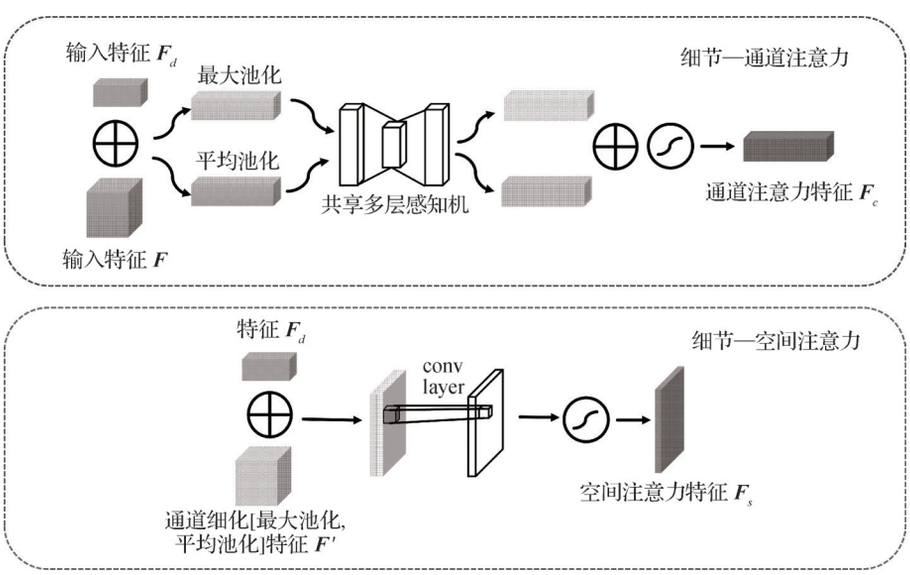


图4　Swin-EdgeNet的细节增强双重注意力模块图

Fig. 4　Detail-enhanced dual attention module of Swin-EdgeNet

2.2 实验细节

本研究的实验基于统一的计算环境开展，硬件配置采用单块NVIDIA RTX 4090（24 GB显存）显卡，搭载Intel Xeon Platinum 8352V处理器，系统内存90 GB，系统盘容量30 GB，所有实验均在该环境下执行以确保结果的可复现性。训练策略上，针对不同的医学图像分割任务采用差异化的参数配置。对于皮肤病变（ISIC2018）、肺部（CHNCXR）和息肉（Kvasir）分割任务，初始学习率设置为0.000 1，批次大小（batch size）为4；而眼底血管（DRIVE）分割任务因数据特性采用更高的初始学习率0.001，批次大小为8。所有任务均使用二元交叉熵损失（binary cross-entropy loss，BCE）和Dice损失的加权组合作为优化目标。数据预处理与增强策略根据任务特点进行定制化设计。皮肤病变、肺部、息肉数据集采用随机水平/垂直翻转、±10°旋转以及色彩抖动作为训练集增强手段，并统一应用ImageNet标准的归一化参数，验证集仅做归一化；而眼底数

据集训练和验证集仅图像进行归一化处理。所有输入图像在进入网络前均经过标准化处理，确保数据分布的一致性。训练过程中采用Adam（adaptive moment estimation）优化器进行参数更新，配合动态学习率调整策略以加速收敛并避免局部最优。测试阶段通过计算所有样本的指标均值作为最终性能评价基准，所有实验均重复3次取平均以消除随机性影响。

2.3 与先进方法的对比分析

本研究在多个医学图像分割任务上进行了全面的性能评估，与近年来最先进的方法（state-of-the-art，SOTA）进行了对比，包括U-Net、U-Net++、Swin-UNet（Cao等，2023）、Trans-UNet、M^{2}SNet（Zhao等，2026）、DCSAU-Net（Xu等，2023）、HTC-Net（Tang等，2024）和Rolling-Unet（Liu等，2024）等。其中，Rolling-Unet模型下的不同配置S/L，根据本文算法使用模型进行选择。实验结果以定量（表1—表4）和定性（图5—图8）两种方式呈现。在定量指标方面，表1（皮肤病变）、表2（肺部CT）、表3（眼底血管）和表4（内窥镜息肉）详细对比了本文方法与最新方法的核心指标；在图5—图8中，进一步通过可视化结果展示不同方法在典型样本上的分割效果差异，验证了方法的先进性。以下分别对不同数据集的实验结果进行分析。

如表1所示，在ISIC2018皮肤病变数据集中本文方法在多项关键指标上（包括Dice、IoU、精确率（Precision，Pre）、召回率（recall，Rec）、准确率（accuracy，Acc）、灵敏度（sentivity，Sen）、特异性（specificity，Spe）、F1和AUC（area under curve））优于主流分割算法。与Rolling-Unet相比，在Dice和交并比（intersection over union，IoU）上分别提升了0.23%和0.27%。在与XBound-Former的对比结果可以发现尽管XBound-Former取得了最高的召回率，但本文方法在IoU（83.97%）和准确率（93.02%）上均优于XBound-Former，且F1分数（87.24%）同样保持领先。这表明本文提出的模型在处理复杂病灶时，有效兼顾了病灶的完整检出与边界的精确分割，避免了XBound-Former存在的过分割问题，从而获得更优的区域重叠度（IoU）。可视化结果如图5所示。

表1 不同方法在ISIC2018数据集上的分割性能

Table 1 Segmentation performance of different methods on the ISIC2018 dataset

/%

方法	Dice ↑	IoU ↑	Pre ↑	F1 ↑	Rec ↑	Acc ↑
U-Net（2015）	85.07	77.64	80.92	79.59	86.14	89.63
U-Net++（2018）	86.90	79.85	82.02	82.66	89.28	90.87
Swin-UNet（2021）	88.16	81.02	81.94	84.76	92.15	91.78
Trans-UNet（2024）	89.14	81.73	85.71	86.00	93.34	92.89
XBound-Former（2022）	**89.89**	83.51	83.65	87.12	**94.61**	92.96
M^{2}SNet（2026）	88.97	82.50	83.64	86.86	92.74	92.58
HTC-Net（2024）	89.52	83.61	86.34	87.06	91.41	92.87
Rolling-Unet(S)（2024）	89.64	83.70	**86.62**	**87.45**	92.33	92.98
本文	89.87	**83.97**	86.11	87.24	93.44	**93.02**

注：加粗字体表示各列最优结果，“↑”表示值越高越好。

在表2肺部分割任务中，本文方法表现出卓越的分割能力。与当前先进的Rolling-Unet相比，本文方法将Dice系数和IoU分别提升至97.62%和95.44%，达到表中的最佳水平。在衡量轮廓重合度的IoU指标上，本文方法优于XBound-Former（95.44% vs 95.31%）及Rolling-Unet（95.44% vs 95.35%）。这表明本文提出的网络结构能更有效地提取边缘特征，从而在肺尖和肋膈角等细节区域生成更贴合真实标签的分割掩码。可视化结果如图6所示。

如表3所示，针对微细血管密集的DRIVE眼底数据集，结果显示，本文方法在捕捉血管细节方面具有显著优势。本文方法在灵敏度（Sen）上取得了80.19%的结果，显著优于当前SOTA方法Rolling-

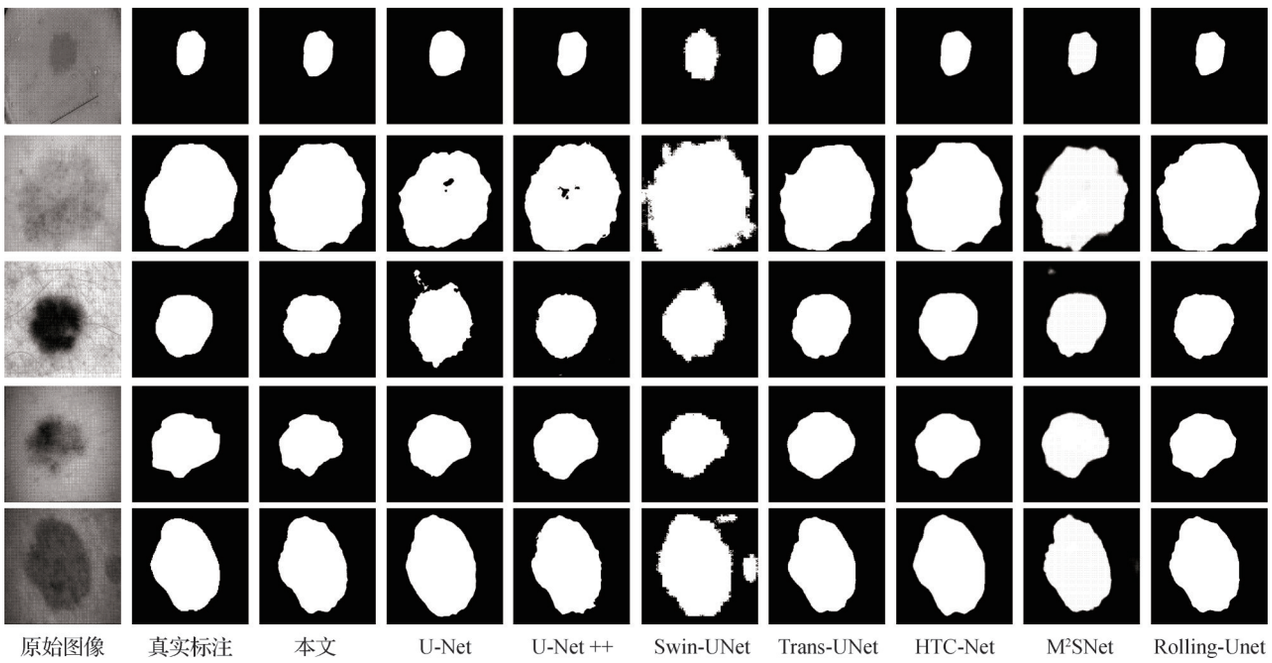


图5　皮肤病变分割结果的可视化对比

Fig. 5　Visualization comparison of skin lesion segmentation results

表2　不同方法在CHNCXR数据集上的分割性能

Table 2　Segmentation performance of different methods on the CHNCXR dataset

/%

方法	Dice ↑	IoU ↑	Pre ↑	Rec ↑	F1 ↑	Acc ↑
U-Net(2015)	96.71	93.75	96.28	94.11	95.07	97.57
U-Net++(2018)	97.17	94.60	96.72	94.99	95.75	97.91
Swin-UNet(2021)	95.94	92.33	94.76	93.27	93.91	97.00
Trans-UNet(2024)	97.04	94.36	96.79	94.54	95.56	97.83
XBound-Former(2022)	97.55	95.31	97.07	95.80	96.34	98.18
DCSAU-Net(2023)	97.56	95.35	97.02	95.58	96.35	98.21
HTC-Net(2024)	97.48	95.19	96.71	95.95	96.24	98.13
Rolling-Unet(S)(2024)	97.57	95.35	96.77	**96.15**	96.37	98.19
本文	**97.62**	**95.44**	**97.12**	95.94	**96.44**	**98.23**

注：加粗字体表示各列最优结果。“↑”表示值越高越好。

Unet(77.21%)和DCSAU-Net(77.03%)。这一指标的大幅提升(约3%)表明，本文模型极大地减少了微小血管的漏检率，如图7所示。尽管F1分数略低于Rolling-Unet，但由于灵敏度的大幅提升，使得本文方法在临床应用中能提供更完整的血管拓扑结构信息。

如表4所示，本文方法在Dice(93.80%)和F1分数(91.93%)上分别超越Trans-UNet 2.12%和5.89%。图8的息肉分割结果进一步表明：传统U-Net系列(如U-Net++)对扁平状息肉(图8第1行)的识别能力较弱(F1仅49.68%)，而Swin-EdgeNet凭借双重注意力机制的聚焦能力，实现了对各类形态息肉的精准覆盖。

2.4　消融实验

为系统验证各模块的贡献，本研究选择肺部分割(CHNCXR)和视网膜血管分割(DRIVE)两个具有显著解剖结构差异的数据集进行消融实验。

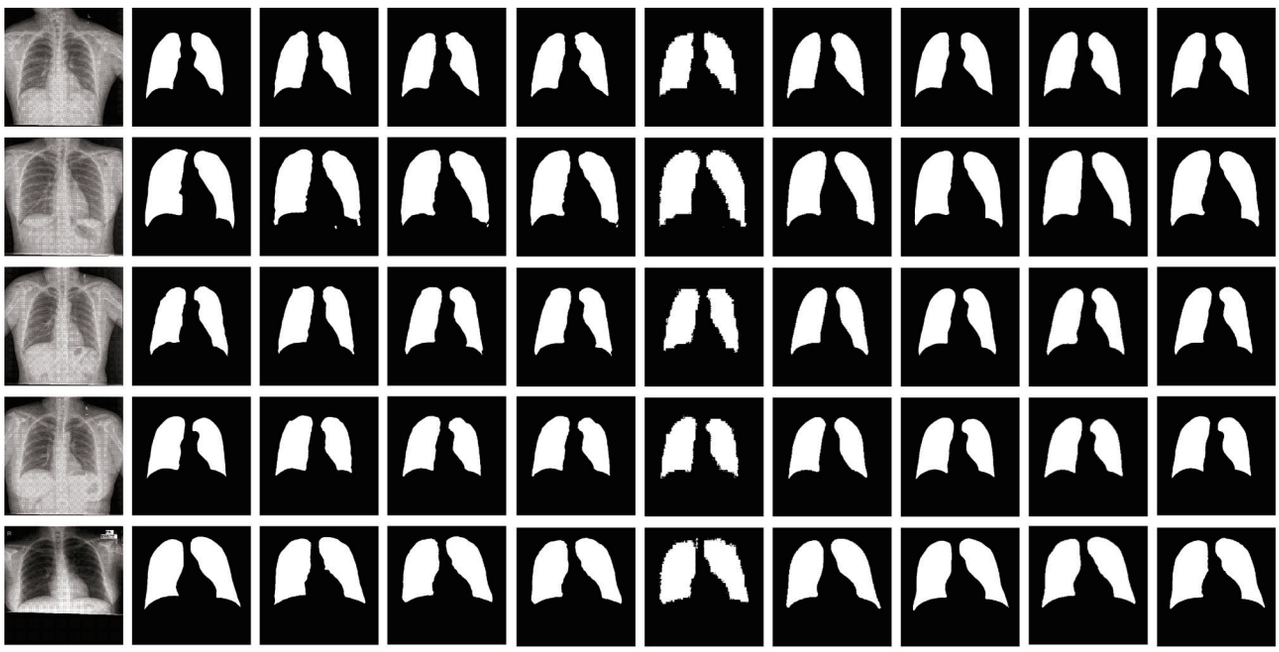


图6　肺部分割结果的可视化对比

Fig. 6　Visualization comparison of lung lesion segmentation results

表3　不同方法在DRIVE数据集上的分割性能

Table 3　Segmentation performance of different methods on the DRIVE dataset

/%

方法	Sen ↑	Spe ↑	F1 ↑	AUC ↑
U-Net(2015)	64.16	96.12	66.23	80.64
U-Net++(2018)	61.56	96.85	71.00	80.20
MSNet(2021)	69.49	96.18	66.57	82.83
M^2SNet(2026)	43.46	97.49	54.66	70.97
DCSAU-Net(2023)	77.03	97.67	78.36	87.58
Rolling-Unet(L)(2024)	77.21	97.69	**78.60**	87.72
本文	**80.19**	**97.76**	77.77	**88.63**

注:加粗字体表示各列最优结果。“↑”表示值越高越好。

2.4.1　不同组件的消融实验

表5进一步探讨了在CHNCXR数据集上,不同组件对肺部分割性能的影响。从表中可以看出,随着各个组件的逐步加入,分割性能逐渐提升。例如加入密集空洞卷积后,Dice系数提升至97.54%,IoU提升至95.29%,这表明密集空洞卷积能够有效增强特征的表达能力。此外,边缘增强注意力(edge enhancement attention,EEA)与门控融合机制的加入进一步优化了边界细节,Dice系数进一步提升至97.59%,IoU提升至95.41%。最终,当所有组件全部集成时,Dice系数达到了97.62%,IoU为95.44%,精确率为97.12%。这表明各个组件之间具有良好的协同作用,能够显著提升分割性能。在DRIVE数据集上(表6)也观察到了相同的趋势:各组件的协同作

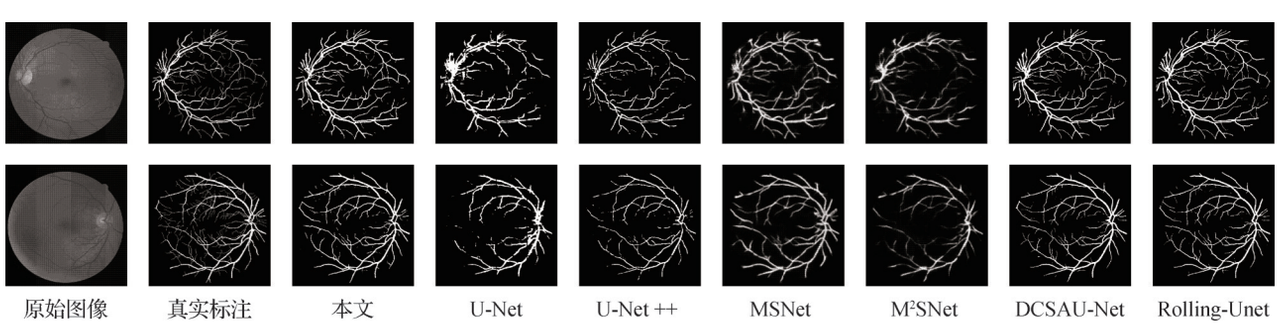


图7　眼底图像分割结果的可视化对比

Fig. 7　Visualization comparison of segmentation results for fundus images

表 4　不同方法在 Kvasir 数据集上的分割性能

Table 4　Segmentation performance of different methods on the Kvasir dataset

/%

方法	Dice ↑	IoU ↑	Pre ↑	Rec ↑	F1 ↑	Acc ↑
U-Net(2015)	73.70	64.71	71.91	53.50	54.46	88.67
U-Net++(2018)	72.06	63.12	70.08	49.68	51.44	88.18
Trans-UNet(2024)	91.68	86.32	90.48	86.04	86.11	95.84
M^2SNet(2026)	87.04	81.25	**93.45**	73.57	78.22	93.57
HTC-Net(2024)	91.27	85.69	90.52	85.35	85.60	95.51
Rolling-Unet(L)(2024)	91.25	86.18	87.97	88.60	85.49	95.56
本文	**93.80**	**86.79**	91.02	**91.93**	**89.81**	**96.72**

注：加粗字体表示各列最优结果。“↑”表示值越高越好。

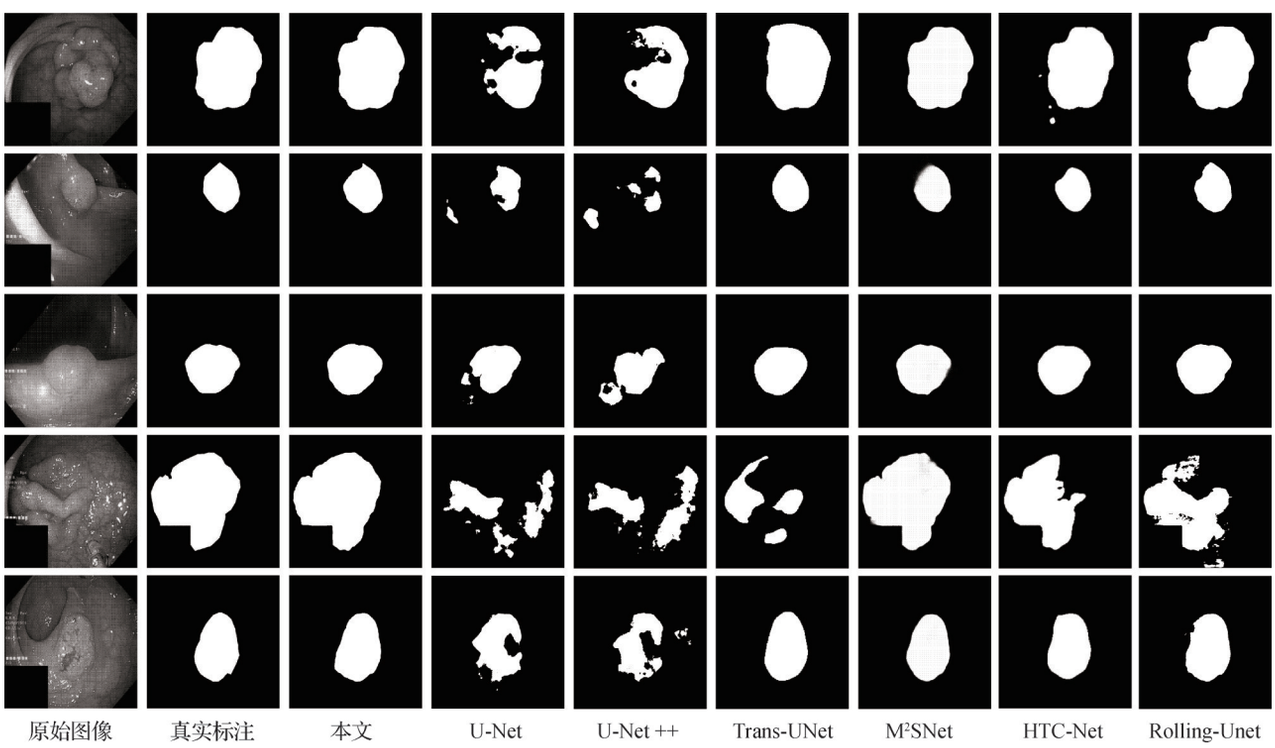


图 8　息肉分割结果的可视化对比

Fig. 8　Visualization comparison of segmentation results for polyp images

用使灵敏度从基线的 96.71% 稳步提升至 97.62%，验证了本文架构在处理细微结构时的有效性。

2.4.2　边缘注意力消融实验

边缘注意力机制在医学图像分割中具有重要作用，能够有效突出目标边缘特征，从而提高分割精度。本研究中，设计了 4 种不同的边缘注意力模块（EEA、EEA1、EEA2 和 EEA3），并对其性能进行了评估，EEA 是 FPN 结构中的 p2 输出结合 EA（edge attention），EEA1 是 FPN 特征融合后加入 EA，EEA2 在 FPN 每一个阶段中加入 EA，EEA3 是级联边缘注意力。

表 7 展示了在 CHNCXR 数据集上，不同边缘注意力模块对肺部分割性能的影响。从表中可以看出，不同模块在不同指标上各有侧重。其中，EEA2 在 Dice 系数（97.59%）、IoU（95.42%）、精确率（97.26%）和准确率（98.23%）上取得了并列最优表现；而 EEA3 则在召回率（95.98%）、F1 分数（96.44%）以及准确率（98.23%，与 EEA2 并列）上表现最佳。这表明，EEA2 更侧重于提升分割的整体重

表5　在CHNCXR数据集上不同组件的影响

Tabel 5　Effects of different components on the CHNCXR dataset

/%

方法	Dice ↑	IoU ↑	Pre ↑	Rec ↑	F1 ↑	Acc ↑
Backbone(U-Net)	96.71	93.75	96.28	94.11	95.07	97.57
UNet_Swin + DA	97.54	95.29	96.96	95.87	96.32	98.17
UNet_Swin + EEA	97.59	95.41	97.06	**95.98**	**96.44**	**98.23**
UNet_Swin + EB	97.60	95.40	**97.16**	95.84	96.42	98.22
UNet_Swin + MEFF	97.61	95.42	97.06	95.97	96.42	98.21
完整模型	**97.62**	**95.44**	97.12	95.94	**96.44**	**98.23**

注：加粗字体表示各列最优结果。“↑”表示值越高越好。

表6　在DRIVE数据集上不同组件的影响

Table 6　Effects of different components on the DRIVE dataset

/%

方法	Sen ↑	Spe ↑	F1 ↑	AUC ↑
Backbone (U-Net)	96.71	93.75	96.28	94.11
UNet_Swin + DA	97.54	95.29	96.96	95.87
UNet_Swin + EEA	97.59	95.41	97.06	**95.98**
UNet_Swin + EB	97.60	95.40	**97.16**	95.84
UNet_Swin + MEFF	97.61	95.42	97.06	95.97
完整模型	**97.62**	**95.44**	97.12	95.94

注：加粗字体表示各列最优结果。“↑”表示值越高越好。

表7　在CHNCXR数据集上不同边缘注意力的影响

Table 7　Effects of different edge attention on the CHNCXR dataset

/%

方法	Dice ↑	IoU ↑	Pre ↑	Rec ↑	F1 ↑	Acc ↑
EEA	97.56	95.41	97.04	95.97	96.42	98.22
EEA1	97.48	95.15	96.67	95.97	96.24	98.13
EEA2	**97.59**	**95.42**	**97.26**	95.78	96.43	**98.23**
EEA3	**97.59**	95.41	97.06	**95.98**	**96.44**	**98.23**

注：加粗字体表示各列最优结果。“↑”表示值越高越好。

合度与精确度，而级联边缘注意力模块（EEA3）通过在特征金字塔的每一层级进行边缘强化，有效降低了漏检率，提升了召回率和综合F1分数，保证了边缘信息在多尺度特征融合过程中的连续性。

表8进一步探讨了在DRIVE数据集上，不同边缘注意力模块对眼底血管分割性能的影响。结果表明，各模块的表现差异更为显著。基础EEA模块在

表8　在DRIVE数据集上不同边缘注意力的影响

Table 8　Effects of different edge attention on the DRIVE dataset

/%

方法	Sen ↑	Spe ↑	F1 ↑	AUC ↑
EEA	**81.48**	97.17	77.49	**89.33**
EEA1	78.76	97.62	77.52	88.19
EEA2	78.60	97.66	77.59	88.13
EEA3	78.59	**97.81**	**77.76**	88.14

注：加粗字体表示各列最优结果。“↑”表示值越高越好。

敏感性（81.48%）和AUC（89.33%）上取得了最优结果；而EEA3则在特异性（97.81%）和F1分数（77.76%）上表现最佳。这充分说明，单一的EEA模块在捕捉血管网络整体的敏感度方面更具优势，而级联设计（EEA3）则更擅长精确分割背景与血管的边界，从而在F1分数上取得了综合最优。再加一段：综合表7、表8，EEA3在两个数据集上F1均最优（CHNCXR：96.44%；DRIVE：77.76%），并在CHNCXR取得最高召回率及并列最优的Dice与准确率，在DRIVE取得最高特异性。其级联式边缘注意力设计能有效突出目标边缘特征，兼顾分割精度与召回能力，故最终选择EEA3。

3　结　论

本研究提出了一种融合Swin Transformer与U-Net架构的新型医学图像分割框架，通过引入多尺度边缘稠密模块与双重注意力机制，有效应对了不同医学影像中解剖结构多样性与病理特征差异所带

来的挑战。系统实验在4个公共数据集(ISIC2018、CHNCXR、DRIVE和Kvasir)上展开，验证了本文方法在分割精度与模型鲁棒性方面相较于现有先进技术的优势，表明该框架具备良好的任务适应性与结构泛化能力。从方法学角度总结，本文工作的主要特点包括：多尺度边缘信息与深层语义特征的协同提取机制，以及结合空间与通道维度的双重注意力融合策略，从而在复杂医学图像中实现更准确的结构边界识别与内部一致性表达。这些设计在多项分割任务中均体现出稳定且一致的性能提升。

然而，本研究仍存在若干局限：网络整体参数量较大，在计算资源受限场景下的部署效率尚需优化；此外，当前验证数据集的模态与病理类型仍有一定局限，对极端罕见形态的泛化能力有待进一步考察。后续研究将着重于轻量化结构设计，扩展多中心、多设备采集的实际临床数据验证，并探索结合自监督学习以降低标注依赖。同时，将该框架拓展至动态影像实时分割与多模态融合诊断，亦是未来值得探索的方向。

参考文献(References)

Adams R and Bischof L. 1994. Seeded region growing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(6): 641-647[DOI: 10.1109/34.295913]

Canny J. 1986. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8(6): 679-698[DOI: 10.1109/TPAMI.1986.4767851]

Cao H, Wang Y Y, Chen J, Jiang D S, Zhang X P, Tian Q, et al. 2023. Swin-UNet: unet-like pure transformer for medical image segmentation//Proceedings of ECCV 2022. Lecture Notes in Computer Science: Springer: #13803[DOI: 10.1007/978-3-031-25066-8_9]

Chen J N, Mei J R, Li X H, Lu Y, Yu Q, Wei Q, et al. 2024. TransUNet: rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis, 97: #103280[DOI: 10.1016/j.media.2024.103280]

Çiçek Ö, Abdulkadir A, Lienkamp S S, Brox T and Ronneberger O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation//Proceedings of Medical Image Computing and Computer-Assisted Intervention. Lecture Notes in Computer Science: Springer: 424-432[DOI: 10.1007/978-3-319-46723-8_49]

Codella N C F, Gutman D, Celebi M E, Helba B, Marchetti M A, Dusza S W, et al. 2018. Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)//2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). Washington, USA: IEEE: 168-172[DOI: 10.1109/ISBI.2018.8363547]

Ghafoorian M, Mehrtash A, Kapur T, Karssemeijer N, Marchiori E, Pesteie M, et al. 2017. Transfer learning for domain adaptation in MRI: application in brain lesion segmentation//Medical Image Computing and Computer Assisted Intervention. Quebec City, Canada: Springer: 516-524[DOI: 10.1007/978-3-319-66179-7_59]

Gu Z W, Cheng J, Fu H Z, Zhou K, Hao H Y, Zhao Y T, et al. 2019. CE-Net: context encoder network for 2D medical image segmentation. IEEE Transactions on Medical Imaging, 38(10): 2281-2292[DOI: 10.1109/TMI.2019.2903562]

Hatamizadeh A, Tang Y C, Nath V, Yang D, Myronenko A, Landman B, et al. 2022. UNETR: transformers for 3D medical image segmentation//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 1748-1758[DOI: 10.1109/WACV51458.2022.00181]

Jiang T and Li X N. 2024. Segmentation of abdominal CT and cardiac MR images with multi scale visual attention. Journal of Image and Graphics, 29(1): 268-279 (蒋婷, 李晓宁. 2024. 采用多尺度视觉注意力分割腹部CT和心脏MR图像. 中国图象图形学报, 29(1): 268-279)[DOI: 10.11834/jig.221032]

Lecun Y, Bengio Y and Hinton G. 2015. Deep learning. Nature, 521(7553): 436-444[DOI: 10.1038/nature14539]

Litjens G, Kooi T, Bejnordi B E, Setio A A A, Ciompi F, Ghafoorian M, et al. 2017. A survey on deep learning in medical image analysis. Medical Image Analysis, 42: 60-88[DOI: 10.1016/j.media.2017.07.005]

Liu Y T, Zhu H J, Liu M T, Yu H Y, Chen Z H and Gao J. 2024. Rolling-Unet: revitalizing MLP's ability to efficiently extract long-distance dependencies for medical image segmentation//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 3819-3827[DOI: 10.1609/aaai.v38i4.28173]

Liu Z, Hu H, Lin Y T, Yao Z L, Xie Z D, Wei Y X, et al. 2022. Swin transformer V2: scaling up capacity and resolution//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 11999-12009[DOI: 10.1109/CVPR52688.2022.01170]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, et al. 2021. Swin transformer: hierarchical vision transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 9992-10002[DOI: 10.1109/ICCV48922.2021.00986]

Otsu N. 1979. A threshold selection method from Gray-Level histograms. IEEE Transactions on Systems, Man, and Cybernetics, 9(1): 62-66[DOI: 10.1109/TSMC.1979.4310076]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional net-

works for biomedical image segmentation//Proceedings of Medical Image Computing and Computer-Assisted Intervention. Lecture Notes in Computer Science: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, et al. 2019. Attention gated networks: learning to leverage salient regions in medical images. Medical Image Analysis, 2019, 53: 197-207 [DOI:10.1016/j.media.2019.01.012]

Shi J, Wang T T, Zhu Z Q, Zhao M F, Wang B X and An H. 2025. Deep learning-based medical image segmentation methods. Journal of Image and Graphics, 30(6): 2161-2186 (石军, 王天同, 朱子琦, 赵敏帆, 王炳勋, 安虹. 2025. 基于深度学习的医学图像分割方法综述. 中国图象图形学报, 30(6): 2161-2186) [DOI: 10.11834/jig.240467]

Tang H, Chen Y B, Wang T, Zhou Y B, Zhao L X, Gao Q Q, et al. 2024. HTC-Net: a hybrid CNN-transformer framework for medical image segmentation. Biomedical Signal Processing and Control, 88: #105605 [DOI: 10.1016/j.bspc.2023.105605]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang J C, Chen F, Ma Y X, Wang L S, Fei Z D, Shuai J W, et al. 2023. XBound-Former: toward cross-scale boundary modeling in transformers. IEEE Transactions on Medical Imaging, 42(6): 1735-1745 [DOI: 10.1109/TMI.2023.3236037]

Wang W H, Xie E Z, Li X, Fan D P, Song K T, Liang D, et al. 2021. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 548-558 [DOI: 10.1109/ICCV48922.2021.00061]

Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: convolutional block attention module//Proceedings of Computer Vision — ECCV 2018. Lecture Notes in Computer Science: Springer: #11211 [DOI: 10.1007/978-3-030-01234-2_1]

Xu Q, Ma Z C, He N and Duan W T. 2023. DCSAU-Net: a deeper and more compact split-attention U-Net for medical image segmentation. Computers in Biology and Medicine, 154: #106626 [DOI: 10.1016/j.compbiomed.2023.106626]

Zhao X Q, Jia H P, Pang Y W, Lyu L, Tian F, Zhang L H, et al. 2026. M^2SNet: multi-scale in multi-scale subtraction network for medical image segmentation. Machine Intelligence Research: #1122-1144 [DOI: 10.1007/s11633-026-1662-9]

Zhou H Y, Guo J S, Zhang Y H, Han X G, Yu L Q, Wang L S, et al. 2023. nnFormer: volumetric medical image segmentation via a 3D transformer. IEEE Transactions on Image Processing, 32: 4036-4045 [DOI: 10.1109/TIP.2023.3293771]

Zhou Z W, Siddiquee M M R, Tajbakhsh N and Liang J M. 2018. UNet++: a nested U-Net architecture for medical image segmentation//Proceedings of DLMIA ML-CDS 2018. Lecture Notes in Computer Science: Springer: #11045 [DOI: 10.1007/978-3-030-00889-5_1]

作者简介

刘善洁,女,硕士研究生,主要研究方向为医学图像分割和深度学习。E-mail: 1084585774@qq.com

童孟军,通信作者,男,教授,主要研究方向为大数据分析和人工智能。E-mail:tmj@zafu.edu.cn