中图法分类号:TP18;TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2026)07-2569-18

论文引用格式:Wu Q Q, Deng X, Shao H J and Wang F. 2026. ViBound-Net: a hybrid attention network for colorectal polyp segmentation integrating visual and boundary-focused mechanisms. Journal of Image and Graphics, 31(7):2569-2586(吴琪琪, 邓星, 邵海见, 王飞. 2026. 视觉与边界融合的混合注意力结直肠息肉分割. 中国图象图形学报, 31(7):2569-2586)[DOI:10.11834/jig.250488]

视觉与边界融合的混合注意力结直肠息肉分割

吴琪琪,邓星*,邵海见,王飞

江苏科技大学计算机学院,镇江　212003

摘　要:**目的**　为提升结直肠息肉在复杂内镜图像中的分割准确性与鲁棒性,针对边界模糊、形态多样与光照不均等难点,构建一种在临床部署约束下兼顾精度与效率的轻量化方法。**方法**　提出混合注意力网络ViBound-Net,以EfficientNetV2S为编码骨干,设计局部—全局上下文融合注意力(local-global contextual attention fusion,LGCAF)模块以统一建模细粒度与全局语义;在解码端引入多尺度残差注意融合块(multi-scale residual attention fusion block,MRAFB)与边界引导的尺度感知特征增强(boundary-aware multi-scale feature refinement,BSFE)模块,强化跨尺度交互与边界细化;并配合深层监督与Dice+Focal复合损失缓解类别不平衡、稳定训练。**结果**　为保证可比性,采用统一的数据预处理与增强策略,在Kvasir-SEG、CVC-ClinicDB和CVC-300等5个公开基准上开展系统评估,并与多种代表性方法进行对比;同时报告复杂度指标(参数量与FLOPs)与推理时间的相对表现,以体现轻量化优势。ViBound-Net在Kvasir-SEG、CVC-ClinicDB和CVC-300上,mDice分别为0.908 1、0.936 8和0.871 0,mIoU(mean intersection over union)分别为0.852 7、0.887 8和0.798 9;在ETIS-LaribPolypDB中的平均绝对误差(mean absolute error,MAE)为0.011 7,分割区域在形状保持、轮廓完整性与区域一致性方面优于多数对比方法。可视化结果显示,本文方法在小目标与贴边病灶上边界更为贴合,能显著抑制背景噪声。复杂度与效率对比表明,ViBound-Net在较低参数量与FLOPs的前提下保持竞争性精度,推理延迟满足实时应用需求。**结论**　通过融合局部—全局语义与边界敏感建模,ViBound-Net在结构紧凑的条件下能实现高精度与良好跨域泛化,可适配复杂临床内镜下的任务需求,具备集成至结直肠癌早筛工作流的应用潜力。后续将进一步面向极端低对比与跨设备域偏移场景,加强边界正则与自适应学习,以持续提升稳健性。

关键词:医学图像分割;结直肠息肉分割;混合注意力网络;边界引导;多尺度特征增强;轻量化模型

ViBound-Net: a hybrid attention network for colorectal polyp segmentation integrating visual and boundary-focused mechanisms

Wu Qiqi, Deng Xing*, Shao Haijian, Wang Fei

School of Computer, Jiangsu University of Science and Technology, Zhenjiang　212003, China

Abstract: **Objective** Accurate delineation of colorectal polyps in colonoscopy images is a key prerequisite for a computer-aided diagnosis (CAD) in colorectal cancer (CRC) screening, as adenomatous polyps are common precursors of CRC, and their timely detection and removal can substantially reduce incidence and mortality. However, in routine endoscopy, automated polyp segmentation remains challenging for three main reasons. First, polyp boundaries are frequently indistinct:

收稿日期:2025-10-17;**修回日期**:2026-01-03;**预印本日期**:2026-01-10

*通信作者:邓星　xdeng@just.edu.cn

基金项目:国家自然科学基金项目(62202210)

Supported by: National Natural Science Foundation of China (62202210)

subtle intensity transitions, mucus coverage, and folds of surrounding mucosa can obscure lesion contours, making boundary localization error-prone. Second, polyp appearance markedly varies across size, shape, texture, and attachment morphology. This diversity complicates stable feature representation and can trigger over-segmentation (including mucosal folds) or under-segmentation (missing flat or small lesions). Third, colonoscopy imaging often exhibits non-uniform illumination, specular highlights, and lens-related artifacts. These factors distort local contrast and degrade model robustness. Meanwhile, practical deployment introduces additional constraints: endoscopy systems typically process high-resolution streams at high frame rates and must respond with low latency on resource-limited hardware. Accordingly, the objective of this study is to develop a lightweight yet accurate polyp segmentation model that 1) balances global semantic understanding with fine-grained boundary detail, 2) achieves efficiency suitable for routine clinical workflows under deployment constraints, and 3) generalizes robustly across heterogeneous datasets collected with different devices and acquisition conditions. **Method** We propose ViBound-Net, a lightweight hybrid-attention segmentation network designed to jointly emphasize semantic context and boundary fidelity while maintaining a compact computational footprint. The network adopts EfficientNetV2S as a lightweight encoder to provide strong feature extraction with favorable parameter-compute efficiency. We introduce a local-global contextual attention fusion (LGCAF) module that integrates fine structural cues (effectively captured by local receptive fields) with global contextual dependencies (captured via attention) to address the local-global trade-off that commonly limits purely convolutional or purely attention-based approaches. This fusion aims to improve semantic consistency in challenging backgrounds while preserving lesion-specific structural patterns that are essential for small or camouflaged polyps. We further design a boundary-guided scale-aware feature enhancement (BSFE) module to explicitly strengthen boundary modeling. BSFE leverages multi-scale representations and boundary-aware guidance to refine edge-related features, targeting failure modes, such as contour breaks, boundary blurring, and partial omission around low-contrast margins. In decoding, we utilized a multi-scale residual attention fusion block to aggregate cross-scale information while stabilizing optimization through residual pathways. This design encourages coherent reconstruction of polyp regions across scales and reduces sensitivity to size variability. In addition, compact channel-spatial attention is incorporated to promote feature consistency and adaptively emphasize salient lesion cues without introducing heavy computation. Deep supervision is used to guide intermediate predictions and improve convergence stability, particularly for boundary-sensitive learning. Model training adopts a compound Dice + focal loss to mitigate foreground-background imbalance and enhance learning stability under diverse lesion sizes and appearance variations. Images are resized to 352×352, normalized, and augmented using random scaling, flipping, and rotation to improve robustness to viewpoint and scale changes commonly observed in endoscopy. Optimization follows a unified protocol using AdamW (learning rate 1×10^{-4}, batch size 4) to ensure a fair comparison across methods. Experiments are conducted on three widely used public colonoscopy benchmarks——Kvasir-SEG, CVC-ClinicDB, and CVC-300——which collectively cover diverse resolutions, polyp sizes and shapes, and background complexity, for evaluation. We report segmentation accuracy using Dice similarity coefficient (Dice) and mean intersection over union (mIoU). Moreover, we report the mean absolute error (MAE) to effectively reflect boundary sensitivity and contour adherence. Given that deployability is an explicit goal, we assess the model efficiency through parameter count, FLOPs, and per-image inference time under the same experimental protocol and input scale. Finally, we provide qualitative visual comparisons to analyze contour adherence, over-/under-segmentation behaviors, and robustness under low contrast, specular highlights, and lesions that resemble surrounding mucosa. **Result** Under the unified evaluation protocol, ViBound-Net achieves strong performance across all three benchmarks. Specifically, this model attains mDice/mIoU of 0.908 1/0.852 7 on Kvasir-SEG, 0.936 8/0.887 8 on CVC-ClinicDB, and 0.871 0/0.798 9 on CVC-300. In boundary-sensitive evaluation on the ETIS-LaribPolypDB dataset, ViBound-Net reports an MAE of 0.011 7, indicating improved contour sharpness and faithful delineation. These results suggest that the proposed local-global fusion and boundary-guided enhancement effectively reduce common segmentation artifacts, such as contour leakage into mucosal folds, fragmented edges around flat polyps, and missing regions under low contrast. ViBound-Net delivers competitive or superior accuracy on most metrics while maintaining a compact design compared with strong convolutional and transformer-based baselines under the same protocol. For example, on Kvasir-SEG, this model improves over M2SNet in mDice and mIoU by 2.6 and 3.5 percentage points, respectively, and it reduces boundary error relative to DUCK-Net,

reflecting a reliable edge localization. Across datasets, qualitative comparisons show that ViBound-Net produces smooth, better-adhered contours with few over-/under-segmentation artifacts, particularly for small lesions, low-contrast polyps, and cases affected by specular highlights. From an efficiency perspective, ViBound-Net consistently demonstrates lower computational burden than many transformer-heavy counterparts (as reflected by fewer parameters and reduced FLOPs), supporting deployment-oriented constraints where latency and memory are critical. Overall, these results indicate that the proposed architecture provides a favorable accuracy-efficiency trade-off, aligning with the practical requirements of routine endoscopy systems. **Conclusion** This study presents ViBound-Net, a lightweight hybrid-attention network for colorectal polyp segmentation that explicitly targets the simultaneous demands of semantic reliability, boundary precision, and deployment efficiency. ViBound-Net achieves a high segmentation accuracy and a low boundary error on three public benchmarks while remaining computationally efficient by combining a compact EfficientNetV2S encoder with a LGCAF module and a boundary-guided scale-aware feature enhancement module and reinforcing multi-scale decoding via residual attention fusion and compact channel-spatial attention. These findings support the feasibility of integrating ViBound-Net into CAD pipelines for CRC screening, where stable boundary delineation can assist downstream clinical tasks, such as polyp characterization, measurement, and resection guidance. Future work will extend the method in several directions: 1) incorporating temporal cues to exploit video continuity and minimize frame-to-frame inconsistency in real endoscopy streams; 2) improving cross-device and cross-center robustness through explicit domain adaptation strategies; 3) enhancing interpretability to effectively communicate model decisions to clinicians; and 4) conducting prospective clinical validation to quantify the workflow-level effect, including time savings, detection consistency, and potential reduction of missed lesions in routine practice.

Key words: medical image segmentation; colorectal polyp segmentation; hybrid attention network; boundary guidance; multi-scale feature enhancement; lightweight model

0 引　言

结直肠癌(colorectal cancer, CRC)是消化系统中发病率与致死率均较高的恶性肿瘤,多由腺瘤性息肉恶变发展而来。以美国为例,2024年新增CRC病例约152 810例、相关死亡53 010例(Siegel等,2024)。研究表明,规范化早期筛查与及时息肉切除可显著降低CRC发病率与病死率(Jha等,2023);当病灶处于可切除阶段干预时,5年生存率可超过90%,而晚期可降至15%以下(Wu等,2024)。结肠镜可直接定位并切除息肉,被视为CRC筛查金标准,但其诊断准确性高度依赖操作者经验,对体积较小、边界模糊或具“伪装”特征的息肉仍存在漏检风险(Brenner等,2014)。因此,构建可靠的自动分割模型以辅助临床识别具有重要意义。

真实结肠镜场景下,图像背景复杂且非结构化、类间对比度低,边界易与黏膜褶皱混淆;息肉在形状、颜色与附着形态上差异显著,并叠加不同设备与医疗中心带来的成像风格差异,使模型表征与跨域泛化面临挑战(Jha等,2020)。围绕这些难点,相关研究大体经历了卷积网络主导、注意力/Transformer(唐霖峰 等,2023)引入以及混合结构与弱/半监督策略融合的演进脉络。

在卷积网络(convolutional neural network, CNN)方向,U-Net及其变体凭借编码—解码结构与跳跃连接得以广泛应用(Ronneberger等,2015;周涛 等,2021),并出现以PraNet(parallel reverse attention network)为代表的边界增强方法以改善轮廓刻画(Fan等,2020)。然而,卷积算子受限于局部感受野,对长程依赖与全局语义建模不足,复杂背景与低对比度场景下仍易产生分割缺失与轮廓模糊。

随着全局建模需求的提升,Transformer类架构被引入息肉分割并取得进展,例如Polyp-PVT(polyp segmentation with pyramid vision Transformers)的多尺度语义建模与CTNet(contrastive Transformer network)的低对比度辨识增强(Dong等,2023;Xiao等,2024)。但在高分辨率输入下,全局注意往往带来较高算力与显存开销,限制近实时部署;同时,小目标与边界细节的稳定刻画仍是影响临床可用性的关键瓶颈。

兼顾细节与语义的需求推动了混合结构的发展,如FCBFormer(fully convolutional branch-Transformer)实现局部—全局特征融合(Sanderson 和 Matusze-

wski, 2022), VANet(viewpoint-aware network)通过视角相关机制提升跨姿态泛化能力(Cai等, 2024)。在标注成本较高或域移显著的条件下,弱/半监督策略也被用于提升泛化性能,例如PolypMixNet利用伪标签一致性改善小样本分割(Jia等, 2024)。总体而言,现有方法虽然持续提升精度,但在边界保持、计算效率与跨域稳定性之间实现统一平衡方面仍具挑战。

临床结肠镜应用通常具备高帧率、高分辨率特点,系统需在资源受限平台上实现低时延推理;模型过大将带来显存占用与响应速度问题,影响医生操作节奏。因此,面向部署的分割模型应在紧凑结构下强化边界与小目标刻画,并将参数量、每秒浮点运算次数(floating point operations per second, FLOPs)与推理时延纳入统一权衡。

基于上述动机,本文提出轻量化混合注意力网络ViBound-Net。模型以EfficientNetV2S为骨干(Tan和Le, 2021),通过局部—全局上下文融合注意力(local-global contextual attention fusion, LGCAF)模块协同建模局部细节与全局语义;设计边界引导的尺度感知特征增强(boundary-guided scale-aware feature enhancement, BSFE)模块以强化结构与边界细节;解码阶段引入多尺度残差注意融合块(multi-scale residual attention fusion block, MRAFB)与紧凑通道—空间注意单元(compact channel-spatial attention unit, CCSAU)促进层级特征融合与边界连贯性,并采用Dice与Focal复合损失提升对小目标与模糊边界的敏感性与训练稳定性。本文在统一硬件和输入尺度下量化评估参数量、FLOPs、单帧时延与显存占用,并与代表性方法对比,结果如图1所示。可以看出,ViBound-Net在降低资源消耗的同时保持竞争性精度,体现出更优的"精度—效率"折中与部署潜力。

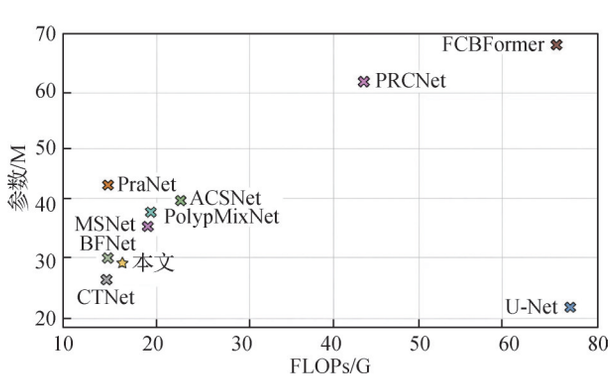


图1 模型散点图

Fig. 1 Scatter plot of models

本文主要贡献如下:1)提出基于EfficientNetV2S编码器的轻量化息肉分割框架ViBound-Net,在边界敏感任务中兼顾计算效率与语义表达能力;2)提出LGCAF与BSFE的联合集成,增强对复杂结构与模糊边界的表征能力;3)通过MRAFB与CCSAU提升层级特征融合与边界连贯性。在5个公开数据集上取得领先性能,展现出良好的泛化性与鲁棒性。

1 方法与模型设计

ViBound-Net的整体结构如图2所示,旨在高效应对息肉分割中的边界模糊、尺度差异与形态多样等挑战。模型采用EfficientNetV2S作为编码主干,并集成局部—全局上下文融合注意力(LGCAF)模块与边界引导的尺度感知特征增强(BSFE)模块,分别提升语义建模能力与边界刻画精度。网络先由主干提取多尺度特征,随后LGCAF融合细节与上下文信息以优化语义一致性;解码阶段借助BSFE强化边界特征表达,结合跳跃连接与多尺度残差注意融合块(MRAFB)实现跨层级交互,并辅以紧凑通道—空间注意单元(CCSAU)提升边界连贯性。训练过程中引入深层监督机制,使模型更敏感于细粒度结构。

1.1 编码器:基于EfficientNetV2S的多尺度特征提取

在ViBound-Net的编码器设计中,采用EfficientNetV2S作为主干,以在保证计算效率的同时获得更具判别性的分割特征。相较ResNet50(residual network), EfficientNetV2S通过Fused-MBConv将标准卷积与倒残差结构结合,兼顾多尺度感知与高效计算,并配合渐进式学习与基于AutoML的架构搜索(automated machine learning, NAS)获得更优的网络配置。模型自5个阶段提取特征(E-Stage 1—E-Stage 5):E-Stage 1(block1b_add, 1/2)与E-Stage 2(block2d_add, 1/4)侧重浅层纹理与边界线索,增强对小息肉的敏感性;E-Stage 3(block3d_add, 1/8)在空间结构与语义表征间取得平衡;E-Stage 4(block5f_add, 1/16)与E-Stage 5(block6l_add, 1/32)强化全局语义与前景—背景判别能力,从而在边界模糊与尺度不均衡场景下保持结构完整性。在此基础上,进一步引入

LGCAF 以整合局部细节与上下文信息，提升编码特征的一致性与表达质量。

尽管 EfficientNetV2S 具备较强的多尺度表征能力，医学图像分割仍易受局部细节缺失、边界模糊与复杂背景干扰影响；仅依赖深层语义也易造成语义抽象与细粒度信息的不平衡。为此，本文在编码阶段引入多尺度的 LGCAF 机制（见图 3），以在多层级特征间进行高效上下文融合：在兼顾全局结构与局部细节的同时提升特征一致性与判别性，从而增强对息肉边界与小尺度目标的精确识别。

LGCAF 模块旨在在统一框架下同时保留细粒度空间信息与高层次语义表征，其核心思想是将局部注意力计算与全局上下文建模有机结合。对于输入特征图 $\boldsymbol{X} \in \mathbf{R}^{H \times W \times C}$，首先通过 1 × 1 卷积获得中间表示 $\boldsymbol{X}'$。为建模局部空间依赖关系，第 2 个 1 × 1 卷积结合 softmax 归一化生成局部注意力图 $\boldsymbol{A}_{\text{local}}$，经逐元素乘法得到局部增强特征 $\boldsymbol{X}_{\text{local}}$。同时，对 $\boldsymbol{X}'$ 进行全局平均池化（global average pooling，GAP）以提取上下文语义，获得全局语义向量 $\boldsymbol{C}_{\text{global}}$，并经全连接层与非线性激活函数变换。该全局描述子被重塑并广播至与 $\boldsymbol{X}'$ 相同的空间尺寸，通过逐元素乘法融合形成全局增强特征 $\boldsymbol{X}_{\text{global}}$。最终，将两分支输出以逐元素加法整合，得到融合表征 $\boldsymbol{X}_{\text{enhanced}}$，并通过残差连接与原始输入 $\boldsymbol{X}$ 结合，生成输出特征 $\boldsymbol{X}_{\text{output}}$。

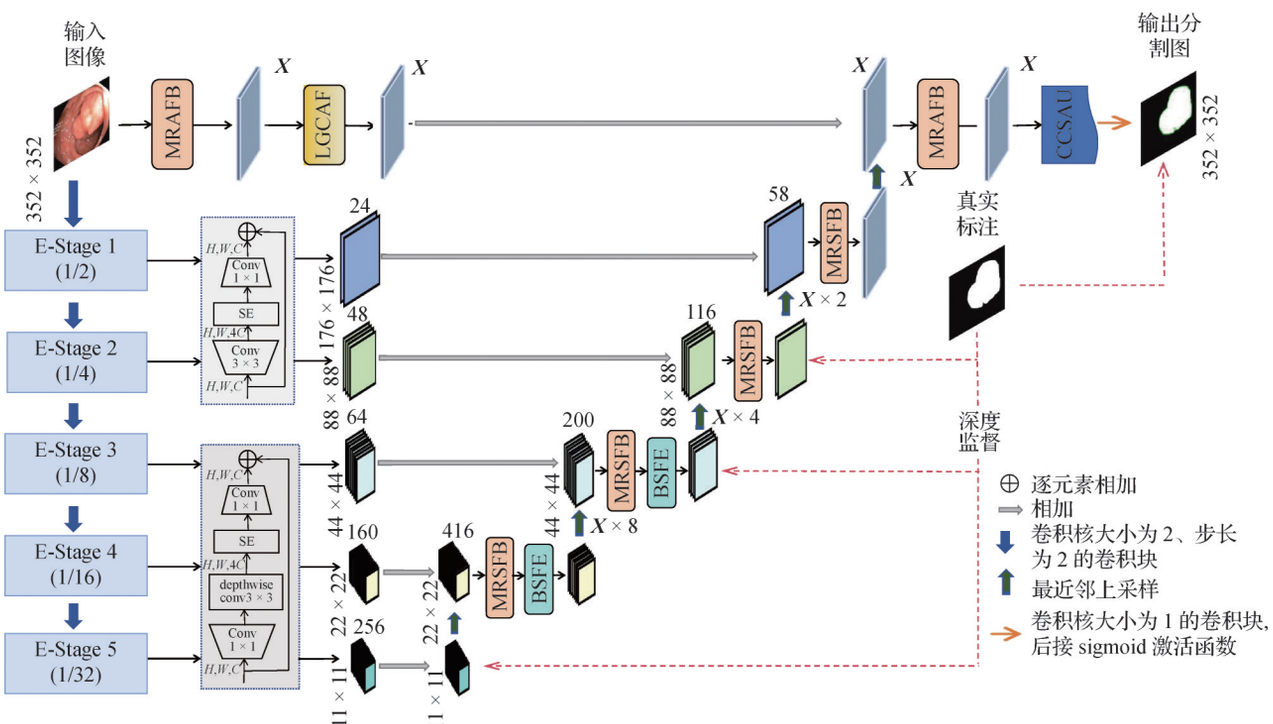


图 2　ViBound-Net 网络总体结构

Fig. 2　Overall architecture of ViBound-Net

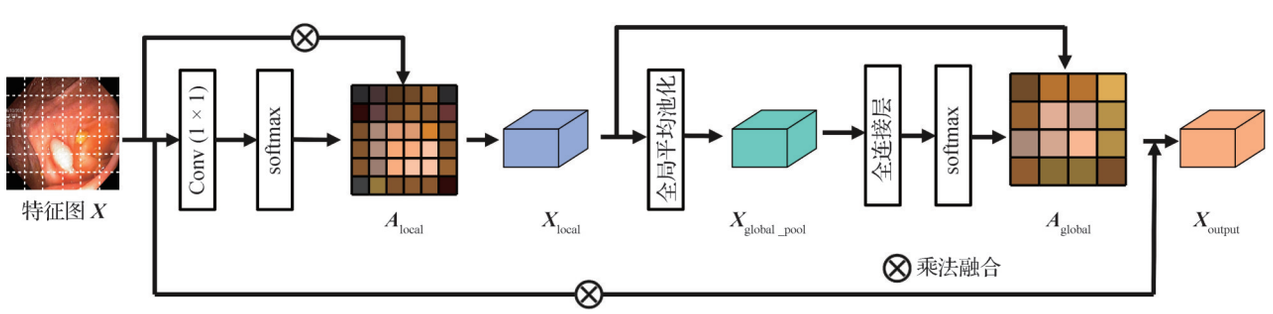


图 3　局部—全局上下文注意力融合模块结构图

Fig. 3　Architecture of the local-global contextual attention fusion module

1.2 解码器：多尺度特征融合与边界感知优化

在ViBound-Net的解码阶段，模型通过逐步上采样与多尺度特征融合逐层重建高分辨率分割图，从而恢复更丰富的空间细节。医学图像分割常面临目标尺度差异大、边界模糊及背景干扰等挑战，传统解码器在特征重建过程中易丢失细粒度信息，尤其在小目标或复杂结构分割时表现欠佳。为克服这些问题，ViBound-Net引入多尺度残差注意融合MRAFB，以强化跨尺度语义建模；同时设计BSFE（见图4），用于细化边界表征并提升结构一致性。两者协同作用，使模型在复杂临床成像条件下兼顾全局一致性与局部精细性，实现目标区域的精确刻画。

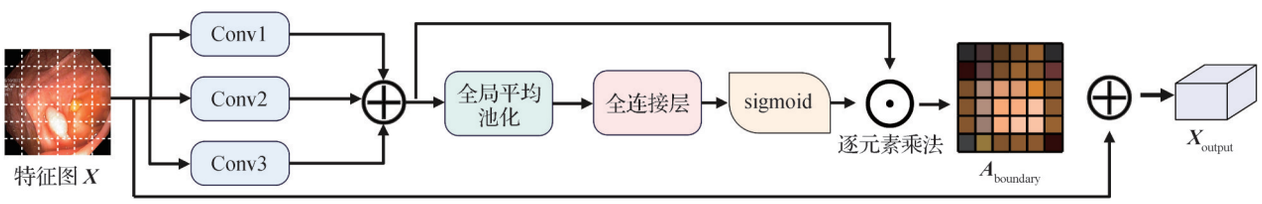


图4 边界感知多尺度特征精炼结构图

Fig. 4 Architecture of boundary-aware multi-scale feature refinement

1.2.1 多尺度特征融合

ViBound-Net通过跳跃连接将编码器的多层特征与解码器对应层融合，从而在重建过程中保持空间细节与语义一致性。为增强解码阶段的特征表征，每次融合操作后均引入多尺度残差注意融合块（MRAFB）。解码器共包含5个层级，在第i个解码层中，其特征图$\boldsymbol{F}_{\mathrm{d},i}$由当前解码输入与对应编码器输出拼接并经上采样（*UpSampling2D*）操作得到，具体为

$$\boldsymbol{F}_{\mathrm{d},i}=UpSampling2D\left(\boldsymbol{F}_{\mathrm{d},i+1}\oplus\boldsymbol{F}_{\mathrm{e},i}\right) \tag{1}$$

式中，$\boldsymbol{F}_{\mathrm{d},i+1}$表示来自更深层解码器的特征，$\boldsymbol{F}_{\mathrm{e},i}$表示同层级的编码器特征，$\oplus$为特征拼接操作。

为进一步改善边界刻画，特别是在低对比度或复杂形态的医学图像中，ViBound-Net在解码阶段引入边界引导的尺度感知特征增强（BSFE）模块。该模块结合多尺度卷积与边界感知注意机制，以提升结构边界敏感性。其包含两个关键阶段：多尺度边界特征提取与全局边界信息建模。首先，对输入特征$\boldsymbol{X}$施加不同感受野的卷积以捕获多尺度边界线索，得到

$$\boldsymbol{B}_{\mathrm{multi}}=Conv_{3\times3}(\boldsymbol{X})+Conv_{5\times5}(\boldsymbol{X})+Conv_{7\times7}(\boldsymbol{X}) \tag{2}$$

随后，对$\boldsymbol{B}_{\mathrm{multi}}$进行全局平均池化（GAP），经可学习变换与sigmoid激活（σ）生成边界注意力权重，具体为

$$\boldsymbol{A}_{\mathrm{boundary}}=\sigma\left(\boldsymbol{W}_{\mathrm{b}}\cdot GAP(\boldsymbol{B}_{\mathrm{multi}})\right) \tag{3}$$

该组权重与边界特征逐元素相乘，实现对边界特征的注意力加权，$\boldsymbol{W}_{\mathrm{b}}$为边界注意力权重，进而得到注意力增强特征。具体为

$$\boldsymbol{X}_{\mathrm{boundary}}=\boldsymbol{A}_{\mathrm{boundary}}\odot\boldsymbol{B}_{\mathrm{multi}} \tag{4}$$

最后，残差连接将输入$\boldsymbol{X}$加回以恢复上下文信息，形成最终边界细化表示，具体为

$$\boldsymbol{X}_{\mathrm{refined}}=\boldsymbol{X}_{\mathrm{boundary}}+\boldsymbol{X} \tag{5}$$

该边界感知增强机制在复杂解剖结构与模糊边界条件下显著提升了解码器输出的精度与鲁棒性。

1.2.2 多尺度残差注意力融合模块

在许多卷积结构中，同时捕获细粒度局部特征与全局上下文仍是一项关键挑战，尤其在医学图像分割等密集预测任务中。为此，本文提出多尺度残差注意融合块（MRAFB），结合多尺度卷积特征提取与注意力机制优化，并借鉴卷积块注意模块（convolutional block attention module，CBAM）（Woo等，2018）的思想。如图5所示，MRAFB由3部分组成：多尺度卷积处理、通道注意力和空间注意力。

在特征提取阶段，模块采用3条并行分支：广域卷积、中域卷积和可分离卷积，以捕获不同感受野下的上下文信息。同时引入来自残差分支的深层特征$\boldsymbol{X}_{\mathrm{res1}}$和$\boldsymbol{X}_{\mathrm{res2}}$，增强表征深度，而可分离卷积分支$\boldsymbol{X}_{\mathrm{step}}$在保持效率的同时降低计算复杂度。各分支输出聚合为统一特征图。具体为

$$\boldsymbol{X}_{\mathrm{multi}}=\boldsymbol{X}_{\mathrm{ws}}+\boldsymbol{X}_{\mathrm{ms}}+\boldsymbol{X}_{\mathrm{res1}}+\boldsymbol{X}_{\mathrm{res2}}+\boldsymbol{X}_{\mathrm{step}} \tag{6}$$

式中，$\boldsymbol{X}_{\mathrm{ws}}$与$\boldsymbol{X}_{\mathrm{ms}}$分别代表广域与中域卷积分支输出。

在特征聚合后，MRAFB采用两阶段注意机制进一步优化表征。在通道注意阶段，对$\boldsymbol{X}_{\mathrm{multi}}$分别执行全局平均池化（GAP）与全局最大池化（global max pooling，GMP），并通过共享变换计算注意权重，具

体为

$$\boldsymbol{A}_{\text{channel}} = \sigma\left(\boldsymbol{W}_{\text{c}} \cdot GAP(\boldsymbol{X}_{\text{multi}}) + GMP(\boldsymbol{X}_{\text{multi}})\right) \quad (7)$$

式中，$\boldsymbol{W}_{\text{c}}$为可学习权重矩阵。所得通道注意权重通过逐元素乘法作用于输入特征，具体为

$$\boldsymbol{X}_{\text{channel}} = \boldsymbol{A}_{\text{channel}} \odot \boldsymbol{X}_{\text{multi}} \quad (8)$$

随后，执行空间注意力计算，对特征沿通道维度分别进行平均池化与最大池化，将所得两种池化结果拼接（Concat），再经 7 × 7 卷积层与 sigmoid 激活函数处理，生成空间注意图，具体为

$$\boldsymbol{S} = Concat\left(GAP(\boldsymbol{X}_{\text{channel}}), GMP(\boldsymbol{X}_{\text{channel}})\right)$$

$$\boldsymbol{A}_{\text{spatial}} = \sigma\left(Conv_{7\times7}(\boldsymbol{S})\right) \quad (9)$$

最终，空间注意图作用于通道优化特征，得到模块输出，具体为

$$\boldsymbol{X}_{\text{MRAFB}} = \boldsymbol{A}_{\text{spatial}} \odot \boldsymbol{X}_{\text{channel}} \quad (10)$$

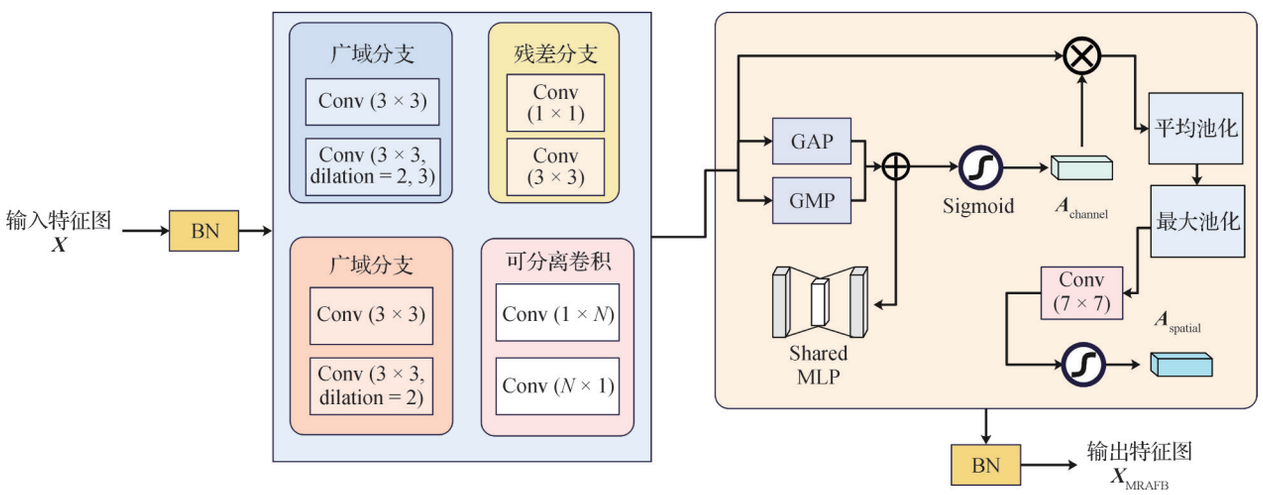


图 5　多尺度残差注意力融合块结构图

Fig. 5　Architecture of the multi-scale residual attention fusion block

1.2.3　紧凑型通道—空间注意力单元

为进一步提升特征判别性与分割稳定性，ViBound-Net 在解码末端引入 CCSAU，结构如图 6 所示。

该模块结合通道与空间注意机制，使模型能在空间与语义层面自适应调整特征权重，从而在保留边界细节的同时强化全局一致性。通道注意通过全局平均池化（GAP）从输入特征 $\boldsymbol{X}$ 提取通道描述子，生成注意权重，具体为

$$\boldsymbol{A}_{\text{channel}} = \sigma\left(\boldsymbol{W}_{\text{c}} \cdot GAP(\boldsymbol{X})\right) \quad (11)$$

式中，$\boldsymbol{W}_{\text{c}}$为可学习矩阵。该权重与输入特征逐元素相乘，得到通道增强特征，具体为

$$\boldsymbol{X}_{\text{channel}} = \boldsymbol{A}_{\text{channel}} \odot \boldsymbol{X} \quad (12)$$

随后，进行空间注意优化，对 GAP 输出施加 5 × 5 卷积并经 sigmoid 激活得到空间权重，具体为

$$\boldsymbol{A}_{\text{spatial}} = \sigma\left(Conv_{5\times5}\left(GAP(\boldsymbol{X})\right)\right) \quad (13)$$

再与输入特征逐元素相乘，得到空间增强特征，即

$$\boldsymbol{X}_{\text{spatial}} = \boldsymbol{A}_{\text{spatial}} \odot \boldsymbol{X} \quad (14)$$

CCSAU 模块使网络在保持边界细节的同时提升全局一致性与分割稳定性。

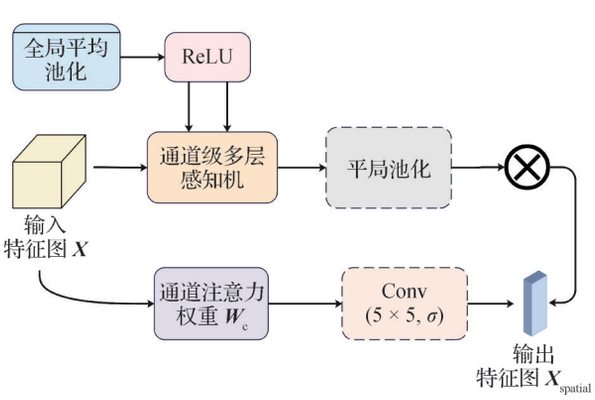


图 6　紧凑型通道—空间注意力单元结构图

Fig. 6　Architecture of the compact channel-spatial attention unit

1.2.4　用于息肉分割的复合损失函数

医学图像分割，尤其是在息肉分割任务中，常常面临诸多挑战，例如目标区域体积较小、边界模糊、形态复杂，以及背景像素占比极大。这些因素往往导致严重的类别不平衡，并使得模型难以在难分区域中有效学习。为提升小目标的检测精度并增强模型对复杂边界结构的敏感性，在 ViBound-Net 中设计

了一种复合损失函数，将Dice Loss与Focal Loss相结合（Lin等，2017）。通过自适应调整两者之间的权重，该损失函数在优化不平衡数据集分割性能的同时，还能强化模型在训练过程中对难分区域的关注。

1）Dice Loss：优化不平衡数据下的分割精度。医学图像分割任务中，目标区域通常仅占图像整体的极小部分，背景区域却占据绝对主导地位。此类严重的类别不平衡问题会严重削弱传统损失函数，如二元交叉熵（binary cross entropy，BCE）的优化效果，易导致模型倾向于预测背景区域，进而忽略息肉这类小目标结构。为解决该问题并提升分割精度，本文引入Dice Loss，该损失函数在处理不平衡数据分布时具备显著优势，其定义为

$$L_{\mathrm{Dice}} = 1 - \frac{2\sum\left(y_{\mathrm{true}} \cdot y_{\mathrm{pred}}\right) + \varepsilon}{\sum y_{\mathrm{true}} + \sum y_{\mathrm{pred}} + \varepsilon} \tag{15}$$

式中，y_{true}与y_{pred}分别表示真实标签与预测概率图，ε为一个小常数（通常取1.0），用于避免分母为零并稳定训练过程。

2）Focal Loss：促进难分区域的鲁棒学习。Dice Loss虽然能有效缓解类别不平衡问题，通过直接最大化预测结果与真实标签的重叠度以提升模型性能，但在优化难分类样本时仍存在局限性。尤其在息肉分割任务中，息肉边界模糊、形态复杂及背景噪声干扰等问题，易导致模型偏向关注易分类的背景区域，而忽略模糊的目标结构。针对这一问题，本文引入Focal Loss，该损失函数在传统交叉熵损失的基础上引入动态调制因子，可强化模型对困难样本的学习权重。其计算式为

$$L_{\mathrm{Focal}} = -\alpha\left(1 - p_t\right)^{\gamma}\log\left(p_t\right) \tag{16}$$

式中，p_t表示真实类别的预测概率，α为类别平衡因子，用于应对类别分布不均，γ为聚焦参数，用以降低易分类样本的贡献。在本文实验中，γ经验设定为2.0，以有效引导模型在训练过程中更加关注难分区域。

3）ViBound-Net的综合损失函数。为了将Dice Loss与Focal Loss的互补优势整合到统一优化目标中，构建ViBound-Net的复合损失函数，既能提升整体分割精度，又能强化模型对难分区域的关注。其定义为

$$L_{\mathrm{total}} = \lambda_{\mathrm{Dice}} L_{\mathrm{Dice}} + \lambda_{\mathrm{Focal}} L_{\mathrm{Focal}} \tag{17}$$

式中，λ_{Dice}和λ_{Focal}为权重系数，用于调节Dice Loss和Focal Loss的相对贡献。通过实验验证，设定$\lambda_{\mathrm{Dice}} = 0.85$、$\lambda_{\mathrm{Focal}} = 0.15$，保证模型在整体分割指标上保持较强性能的同时，对稀缺与模糊区域也具备足够敏感性。

2 实验与结果

2.1 实验设置

2.1.1 数据集

选用5个公开且权威的肠息肉分割数据集作为基准，覆盖多分辨率与多场景，均来源于真实结肠镜图像并由医学专家标注、胃肠病学专家审核，包括Kvasir-SEG（Jha等，2020）、CVC-ClinicDB（Bernal等，2015）、CVC-ColonDB（Bernal等，2012）、ETIS-LaribPolypDB（Silva等，2014）和CVC-300。Kvasir-SEG含1 000幅多分辨率图像（332×487像素至1 920×1 072像素），是肠息肉分割的常用基准；CVC-ClinicDB含612幅标准分辨率图像（384×288像素），适于评估低分辨率场景性能；CVC-ColonDB含300幅图像（574×500像素），体现息肉形态多样性；ETIS-LaribPolypDB含196幅高分辨率图像（1 255×966像素），用于检验复杂场景下的鲁棒性；CVC-300含300幅图像（500×574像素），作为新测试集评估未知场景的适应性。为便于同行复现与后续研究，本文所使用的结直肠息肉分割数据及对应标注已存储于ScienceDB（science data bank）（DOI：10.57760/sciencedb.j00240.00089），访问链接：https://www.scidb.cn/c/j00240。

为保证评估的全面性与可比性，本文在Kvasir-SEG和CVC-ClinicDB数据集上分别选取800幅和490幅图像用于模型训练，剩余图像用于可见域测试，以验证模型在已观测数据分布下的学习充分性与拟合性能；同时，在未参与训练的CVC-ColonDB、ETIS-LaribPolypDB和CVC-300数据集上开展不可见域测试，以此模拟临床场景下模型的跨域泛化能力。除整体性能验证外，还设计系统性消融实验，通过逐步剔除或替换关键模块来量化各模块的独立贡献度。所有对比实验均采用统一的数据划分方案与评估流程，确保实验的一致性、公平性和可复现性。

2.1.2 实现细节

本文模型基于Keras框架并结合TensorFlow实现，在配备NVIDIA GeForce RTX 4060 Ti GPU（显存

16 GB)的计算环境下训练。为增强泛化并减轻过拟合,所有输入图像经标准化后统一调整至 352 × 352 像素;训练阶段采用随机缩放、水平/垂直翻转与随机旋转等数据增强。优化器选用 AdamW,学习率与权重衰减均为 1×10^{-4},批量大小为 4;在 Kvasir-SEG 的 800 幅和 CVC-ClinicDB 的 490 幅图像上训练 100 个 epoch。

为保证后续复杂度与实时性评估的可比性与可复现性,在与训练相同的硬件/软件环境下采用统一评测口径:以 352 × 352 像素的输入、Batch = 1 测量单帧推理时延(ms),并据此按帧速率 = 1 000/ms 计算帧率(frame per second,FPS);计时使用多轮预热后多次平均。所有对比均在上述口径下完成。

2.1.3　评价指标

采用平均骰子相似系数(mean Dice coefficient,mDice)、平均交并比(mean intersection over union,mIoU)、加权 F 测度(F_1)(Margolin 等,2014)、结构测度(S_α)(Fan 等,2017)、均值 E 测度(mE_ξ)(Fan 等,2018)和平均绝对误差(mean absolute error,MAE)6 种指标,从区域重叠、结构相似性和边缘精准度等多个层面综合评估模型的分割性能。mDice 和 mIoU 主要衡量预测分割结果与真实掩膜的区域一致性,F_1 结合精确度与召回率权衡模型对目标区域的敏感性,S_α 反映分割结果在对象级和区域级别上的结构完整性,mE_ξ 评估分割掩膜与地面真实值的视觉对齐程度,MAE 计算预测掩膜与真实掩膜之间的平均像素误差,数值越低表明分割质量越高。

2.1.4　对比方法

将 ViBound-Net 与 13 种当前先进(state-of-the-art, SOTA)的息肉分割方法开展系统性对比实验,对比模型覆盖经典卷积范式、注意力融合架构、Transformer 架构及混合架构等主流技术路线,具体包括 U-Net、ResNet(Targ 等,2016),以及代表性 SOTA 方法(ACSNet(adaptive context selection network)(Zhang 等,2020)、PraNet、MSNet(multi-scale subtraction network)(Zhao 等,2021)、M^2SNet(multi-scale in multi-scale subtraction network)(Zhao 等,2023)、DUCK-Net、PolypMixNet、FCBFormer(fully convolutional branch Transformer)、BFNet(boundary-guided feature-aligned network)(Yue 等,2025)、CTNet(contrastive Transformer network)、PRCNet(parallel reverse convolutional attention network)、SAM-L(segment anything model-large)(Zhou 等,2023)),以保障评测的广度与权威性。为确保实验公平性,U-Net、ResNet、ACSNet、PraNet、MSNet、M^2SNet、DUCK-Net 及 PolypMixNet 均在统一训练配置(含损失函数、优化器、数据增强策略等)下完成复现;受显存资源限制,复现实验统一设置 batch size = 4(多数相关文献设定为 16 或 32)。该配置虽然可能对部分对比模型的最终性能产生轻微影响,但同步凸显了 ViBound-Net 在低资源环境下的高效适配能力。

对于 FCBFormer、BFNet、CTNet、PRCNet、SAM-L 等公开 SOTA 方法,直接引用原论文在相同数据集与评价指标的结果,以保证对比的可比性。在 CVC-300 数据集上,因 BFNet 与 CTNet 的原论文未提供该数据集的实验数据,故选取 C^2ENet(Mao 等,2025)作为补充对比基线。实验结果表明,ViBound-Net 在多个关键评价指标上持续展现优势,且在部分数据集上取得最优性能,充分验证了其在息肉分割任务中的有效性与鲁棒性。

2.2　有效性对比

2.2.1　定量结果

由不同数据集的定量比较结果可以看出,ViBound-Net 相较近年来多种高性能分割方法,在多个权威基准数据集上均表现出稳定且显著的优势,并在部分任务中达到当前最优水平,充分验证了方法的有效性与适应性。

如表 1 所示,在 Kvasir-SEG 数据集上,ViBound-Net 在 mDice 与 mIoU 指标上分别较 M^2SNet 提升 2.6% 与 3.5%,S_α 指标亦较 M^2SNet(0.934 7)提高 2.4%,表明该模型在病变区域结构保持与整体一致性方面具备更优性能。在细节刻画与边界精度方面,MAE 降至 0.027 2,较 DUCK-Net(0.041 6)下降 1.4%,较 PRCNet 下降 4.1%,体现了模型在抑制背景噪声与增强边界清晰度方面的显著改进。综合来看,ViBound-Net 在 mDice、mIoU、S_α 和 MAE 等核心指标上均优于大多数对比方法,进一步验证了其在病变结构完整性与边界精确度保持方面的优势。

表 2 显示,在 CVC-ClinicDB 数据集中,本文方法在 mDice、mIoU 和 F_1 方面均取得领先,相较于 PRCNet,分别提升 1.2%、1.9% 和 2.5%,而相比 PolypMixNet,提升幅度更为明显,分别达到 3.8%、6.1% 和 3.7%。在结构感知能力方面,S_α 达到 0.952 5,比 MSNet 提升 1.6%,同时 mE_ξ 在所有对比方法中保持

表1 Kvasir-SEG数据集上的对比结果

Table 1 The comparative results on the Kvasir-SEG dataset

方法	年份	mDice	mIoU	F_1	S_α	mE_ξ	MAE
U-Net(Ronneberger等,2015)	2015	0.797 1	0.700 4	0.807 8	0.878 3	0.921 1	0.058 5
ResNet(Targ等,2016)	2016	0.787 5	0.686 7	0.803 6	0.874 6	0.902 6	0.055 7
ACSNet(Zhang等,2020)	2020	0.887 8	0.825 5	0.889 3	0.936 3	0.960 6	0.030 4
PraNet(Fan等,2020)	2020	0.872 0	0.797 9	0.884 8	0.855 9	0.873 2	0.037 7
MSNet(Zhao等,2021)	2022	0.890 4	0.828 7	0.899 4	0.952 1	0.937 8	0.034 1
FCBFormer(Sanderson和Matuszewski,2022)	2022	0.905 0	0.851 0	0.898 0	0.924 0	**0.961 0**	0.027 0
M^2SNet(Zhao等,2023)	2023	0.882 5	0.817 9	0.888 0	0.934 7	0.955 9	0.033 8
SAM-L(Zhou等,2023)	2023	0.782 0	0.710 0	0.773 0	0.832 0	–	0.061 0
DUCK-Net(Dumitru等,2023)	2023	0.862 4	0.790 1	0.874 4	0.916 1	0.937 9	0.041 6
PolypMixNet(Jia等,2024)	2024	0.882 3	0.817 1	0.886 1	0.931 0	0.954 7	0.035 3
PRCNet(Li等,2024)	2024	0.799 0	0.716 0	0.763 0	0.837 0	0.886 0	0.068 0
CTNet(Xiao等,2024)	2024	0.917 0	0.863 0	0.910 0	0.928 0	0.959 0	0.023 0
BFNet(Yue等,2025)	2025	**0.933 0**	**0.887 0**	**0.929 0**	–	–	**0.019 0**
ViBound-Net(本文)	2026	0.908 1	0.852 7	0.911 4	**0.958 3**	0.945 3	0.027 2

注:加粗字体表示各列最优结果。“–”表示该项数据未提供、未报告,或无法统计。

表2 CVC-ClinicDB数据集上的对比结果

Table 2 The comparative results on the CVC-ClinicDB dataset

方法	年份	mDice	mIoU	F_1	S_α	mE_ξ	MAE
U-Net(Ronneberger等,2015)	2015	0.829 1	0.760 4	0.812 2	0.886 5	0.977 1	0.022 7
ResNet(Targ等,2016)	2016	0.802 1	0.623 5	0.803 7	0.854 5	0.975 5	0.024 5
ACSNet(Zhang等,2020)	2020	0.881 2	0.793 2	0.870 3	0.916 7	0.975 2	0.017 5
PraNet(Fan等,2020)	2020	0.900 6	0.848 9	0.895 8	0.926 2	0.981 0	0.016 2
MSNet(Zhao等,2021)	2022	0.923 6	0.877 3	0.905 8	0.936 2	0.985 0	0.016 3
FCBFormer(Sanderson和Matuszewski,2022)	2022	0.915 0	0.861 0	0.924 0	–	–	0.010 0
M^2SNet(Zhao等,2023)	2023	0.923 4	0.876 1	0.903 8	0.916 2	0.978 1	0.018 3
SAM-L(Zhou等,2023)	2023	0.578 0	0.526 0	0.563 0	0.744 0	–	0.057 0
DUCK-Net(Dumitru等,2023)	2023	0.849 5	0.802 2	0.879 7	0.900 4	0.977 0	0.023 0
PolypMixNet(Jia等,2024)	2024	0.899 2	0.827 3	0.904 5	0.876 9	**0.986 3**	0.013 7
PRCNet(Li等,2024)	2024	0.925 0	0.869 0	0.916 0	0.943 0	–	0.009 0
CTNet(Xiao等,2024)	2024	0.936 0	0.887 0	0.934 0	0.952 0	0.983 0	**0.006 0**
BFNet(Yue等,2025)	2025	**0.942 0**	**0.898 0**	**0.949 0**	–	–	**0.006 0**
ViBound-Net(本文)	2026	0.936 8	0.887 8	0.941 0	**0.952 5**	0.984 8	0.012 6

注:加粗字体表示各列最优结果。“–”表示该项数据未提供、未报告,或无法统计。

最优。值得注意的是,本文方法在MAE方面同样具备较强优势,相较于PolypMixNet下降0.11%,相比DUCK-Net下降1.04%,充分说明其在病变区域分割的精细度和边界保持能力方面的稳定性和优越性。

面对更具挑战性的CVC-ColonDB和ETIS-LaribPolypDB数据集，本文方法依然展现出较强的竞争力。表3显示，在CVC-ColonDB数据集中，mDice和mIoU相较于PRCNet，分别提升5.3%和6.3%，相比PolypMixNet，提升6.9%和6.4%，进一步验证了其在复杂背景下的分割能力。在ETIS-LaribPolypDB数据集中，如表4所示，本文方法在mE_ξ上较CTNet提升4.8%，同时MAE也比PRCNet下降1.0%，在边界保持能力和局部区域感知能力方面均达到了较优水平。如表4所示，本文方法在小目标病变检测任务中展现出更高的准确性和鲁棒性。

此外，在CVC-300数据集中，本文方法在各个评价指标上保持领先，尽管C^2E-Net在mDice和mIoU上略高，但本文方法整体性能更具稳定性与泛化能力。表5进一步验证了其在全局结构保持、边界刻画及小目标检测方面的优势，表明该方法在肠道病变自动分割任务中具备更强的竞争力和实用价值。

表3　CVC-ColonDB数据集上的对比结果

Table 3　The comparative results on the CVC-ColonDB dataset

方法	年份	mDice	mIoU	F_1	S_α	mE_ξ	MAE
U-Net(Ronneberger等，2015)	2015	0.607 3	0.499 9	0.620 5	0.780 7	0.818 0	0.063 9
ResNet(Targ等，2016)	2016	0.530 5	0.436 2	0.548 9	0.717 0	0.738 4	0.063 3
ACSNet(Zhang等，2020)	2020	0.709 1	0.634 6	0.711 3	0.826 1	0.863 8	0.039 3
PraNet(Fan等，2020)	2020	0.713 7	0.628 1	0.745 3	0.812 7	0.872 4	0.037 6
MSNet(Zhao等，2021)	2022	0.679 5	0.602 2	0.702 6	0.816 7	0.864 0	0.054 8
FCBFormer(Sanderson和Matuszewski，2022)	2022	0.786 0	0.707 0	0.793 0	–	–	0.033 0
M^2SNet(Zhao等，2023)	2023	0.708 6	0.625 3	0.720 8	0.840 0	0.848 2	0.039 4
SAM-L(Zhou等，2023)	2023	0.468 0	0.422 0	0.463 0	0.690 0	–	0.054 0
DUCK-Net(Dumitru等，2023)	2023	0.691 3	0.616 9	0.707 8	0.834 5	0.857 7	0.043 5
PolypMixNet(Jia等，2024)	2024	0.672 1	0.598 2	0.678 7	0.800 4	0.889 1	0.040 9
PRCNet(Li等，2024)	2024	0.688 0	0.599 0	0.660 0	0.800 0	–	0.051 0
CTNet(Xiao等，2024)	2024	0.813 0	0.734 0	0.801 0	**0.874 0**	0.919 0	**0.027 0**
BFNet(Yue等，2025)	2025	**0.820 0**	**0.742 0**	**0.819 0**	–	–	0.028 0
ViBound-Net(本文)	2026	0.741 1	0.662 1	0.749 5	0.845 8	**0.937 2**	0.037 1

注：加粗字体表示各列最优结果。“–”表示该项数据未提供、未报告，或无法统计。

2.2.2　定性结果

如图7所示，在5个公开数据集上ViBound-Net与多种主流分割方法进行了视觉对比，包括PraNet、M^2SNet、DUCK-Net和PolypMixNet等。结果显示，ViBound-Net在不同类型的图像中均展现出更强的结构感知与边界贴合能力。尤其在CVC-ClinicDB和Kvasir-SEG等高质量数据集中，其分割结果在形状保持、轮廓完整性及区域一致性方面均显著优于其他方法。与DUCK-Net和MSNet存在的锯齿状边界不同，ViBound-Net生成的边缘更为平滑且紧贴真实结构；同时，其结果也避免了PolypMixNet常出现的过度包围现象。得益于边界与主体区域的协同建模机制，ViBound-Net能够在保证全局结构完整性的同时有效控制局部冗余，体现出优越的几何约束性与区域内聚性。

在CVC-ColonDB与ETIS-LaribPolypDB等结构模糊、噪声干扰明显以及小目标密集的复杂场景中，ViBound-Net同样保持了稳定的性能与较强的鲁棒性。相比之下，M^2SNet与PraNet在低对比度或贴边息肉的分割任务中易出现目标缩减或轮廓偏移现象，而ViBound- Net能在模糊边界条件下实现轮廓的有效推进与形态恢复，生成结果在拓扑结构上与真实标注保持高度一致。这种优势不仅源于边缘增强策略，更得益于其全局语义、细粒度边缘与空间上

表4 ETIS-LaribPolypDB数据集上的对比结果

Table 4 The comparative results on the ETIS-LaribPolypDB dataset

方法	年份	mDice	mIoU	F_1	S_α	mE_ξ	MAE
U-Net(Ronneberger等,2015)	2015	0.386 5	0.304 6	0.403 7	0.624 9	0.679 4	0.074 6
ResNet(Targ等,2016)	2016	0.308 2	0.231 4	0.328 8	0.601 8	0.676 6	0.079 1
ACSNet(Zhang等,2020)	2020	0.692 3	0.622 9	0.696 0	0.831 4	0.812 9	0.017 8
PraNet(Fan等,2020)	2020	0.582 6	0.515 1	0.614 6	0.731 7	0.794 0	0.018 2
MSNet(Zhao等,2021)	2022	0.571 2	0.501 2	0.602 4	0.747 5	0.812 4	0.034 3
FCBFormer(Sanderson和Matuszewski,2022)	2022	0.786 0	0.686 0	0.814 0	–	–	0.015 0
M^2SNet(Zhao等,2023)	2023	0.585 7	0.516 7	0.601 5	0.765 6	0.846 3	0.027 4
SAM-L(Zhou等,2023)	2023	0.551 0	0.507 0	0.544 0	0.751 0	–	0.030 0
DUCK-Net(Dumitru等,2023)	2023	0.516 4	0.440 9	0.554 0	0.734 3	0.749 0	0.056 9
PolypMixNet(Jia等,2024)	2024	0.598 8	0.527 5	0.608 1	0.767 6	0.886 1	0.019 0
PRCNet(Li等,2024)	2024	0.685 0	0.598 0	0.639 0	0.824 0	–	0.022 0
CTNet(Xiao等,2024)	2024	**0.813 0**	**0.734 0**	0.801 0	**0.874 0**	0.915 0	0.014 0
BFNet(Yue等,2025)	2025	0.800 0	0.723 0	**0.825 0**	–	–	0.017 0
ViBound-Net(本文)	2026	0.745 4	0.662 1	0.753 5	0.845 6	**0.962 7**	**0.011 7**

注:加粗字体表示各列最优结果。“–”表示该项数据未提供、未报告,或无法统计。

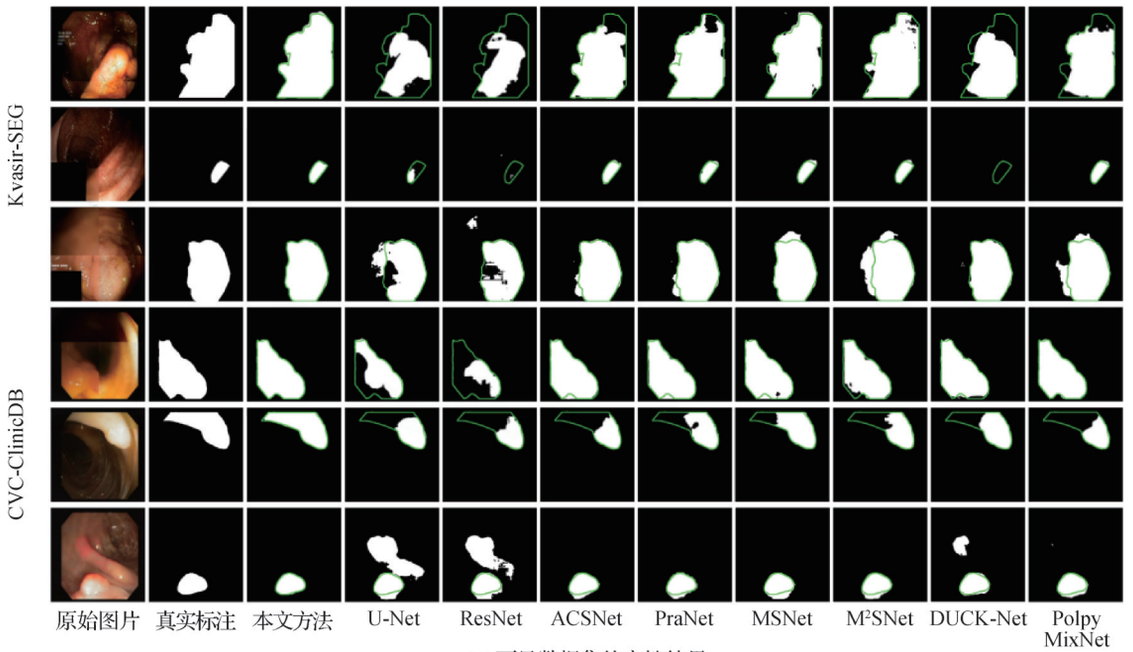
表5 CVC-300数据集上的对比结果

Table 5 The comparative results on the CVC-300 dataset

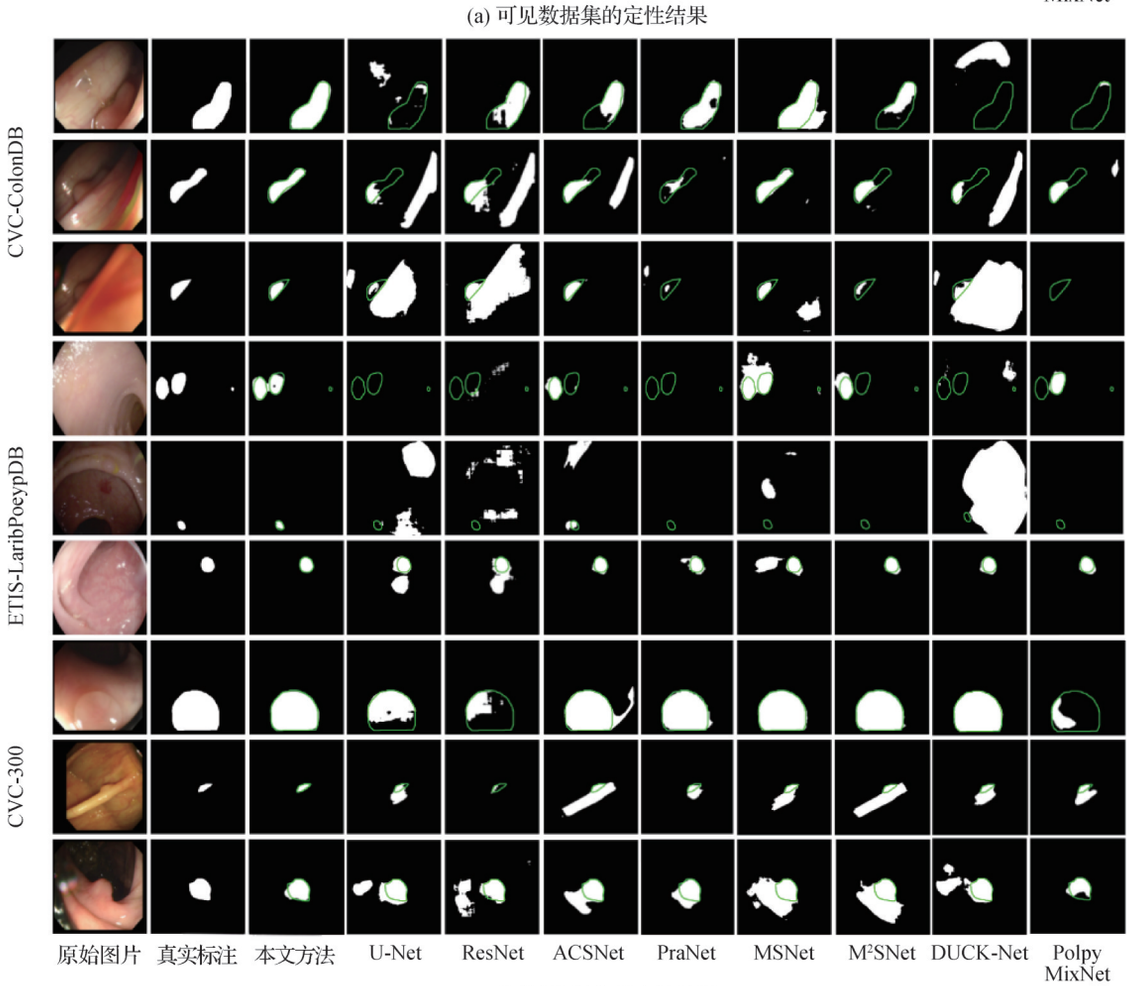
方法	年份	mDice	mIoU	F_1	S_α	mE_ξ	MAE
U-Net(Ronneberger等,2015)	2015	0.733 3	0.617 5	0.743 2	0.860 8	0.896 9	0.023 6
ResNet(Targ等,2016)	2016	0.636 3	0.531 2	0.668 3	0.786 6	0.958 1	0.020 9
ACSNet(Zhang等,2020)	2020	0.861 9	0.772 8	0.823 2	0.920 9	0.840 4	0.009 5
PraNet(Fan等,2020)	2020	0.860 1	0.784 9	0.867 5	0.918 9	0.920 9	0.009 1
MSNet(Zhao等,2021)	2022	0.849 2	0.774 8	0.863 8	0.919 6	0.911 9	0.013 9
FCBFormer(Sanderson和Matuszewski,2022)	2022	–	–	–	–	–	–
M^2SNet(Zhao等,2023)	2023	0.828 8	0.756 3	0.837 0	0.906 4	0.960 5	0.016 4
SAM-L(Zhou等,2023)	2023	0.726 0	0.676 0	0.729 0	0.849 0	–	0.020 0
DUCK-Net(Dumitru等,2023)	2023	0.824 9	0.742 9	0.839 7	0.916 8	0.974 9	0.015 5
PolypMixNet(Jia等,2024)	2024	0.810 3	0.733 3	0.817 3	0.901 3	0.979 0	0.013 9
PRCNet(Li等,2024)	2024	0.819 0	0.736 0	0.777 0	0.891 0	–	0.018 0
CTNet(Xiao等,2024)	2024	–	–	–	–	–	–
C^2ENet(Mao等,2025)	2025	**0.898 0**	**0.829 0**	0.863 0	0.932 0	0.957 0	**0.007 0**
ViBound-Net(本文)	2026	0.871 0	0.798 9	**0.873 7**	**0.937 5**	**0.983 9**	0.010 4

注:加粗字体表示各列最优结果。“–”表示该项数据未提供、未报告,或无法统计。

下文的联合建模能力。此外,不同数据集下预测结果的一致性进一步验证了ViBound-Net出色的跨域泛化性能。无论是在规则结构场景(如CVC-300)还是复杂异质图像环境(如ETIS-LaribPolypDB)中,



(a) 可见数据集的定性结果



(b) 不可见数据集的定性结果

图 7　不同方法在 5 个数据集上的定性结果

Fig. 7　Qualitative results of different methods on five datasets

((a) the qualitative results for the visible datasets; (b) the qualitative results for the unseen datasets)

ViBound-Net均展现出良好的结构稳定性与边界可控性,为后续临床分析与诊断提供了可靠的分割基础。

为进一步提升模型的可解释性与分析维度,本文补充引入注意力热力图对ViBound-Net的空间响应进行可视化,如图8所示。该热力图由模型关注区域与真实掩膜融合生成,并以渐变色表征不同位置的响应强度。结果显示,ViBound-Net在目标边缘及内部结构区域形成清晰且集中的高激活分布,能够稳定聚焦于关键解剖/病灶区域并抑制背景干扰。这不仅体现了模型对显著区域的有效建模能力,也从侧面验证了多尺度语义融合与边界引导机制在空间结构刻画上的有效性,为后续临床场景中的视觉核验与结果审查提供了直观依据。

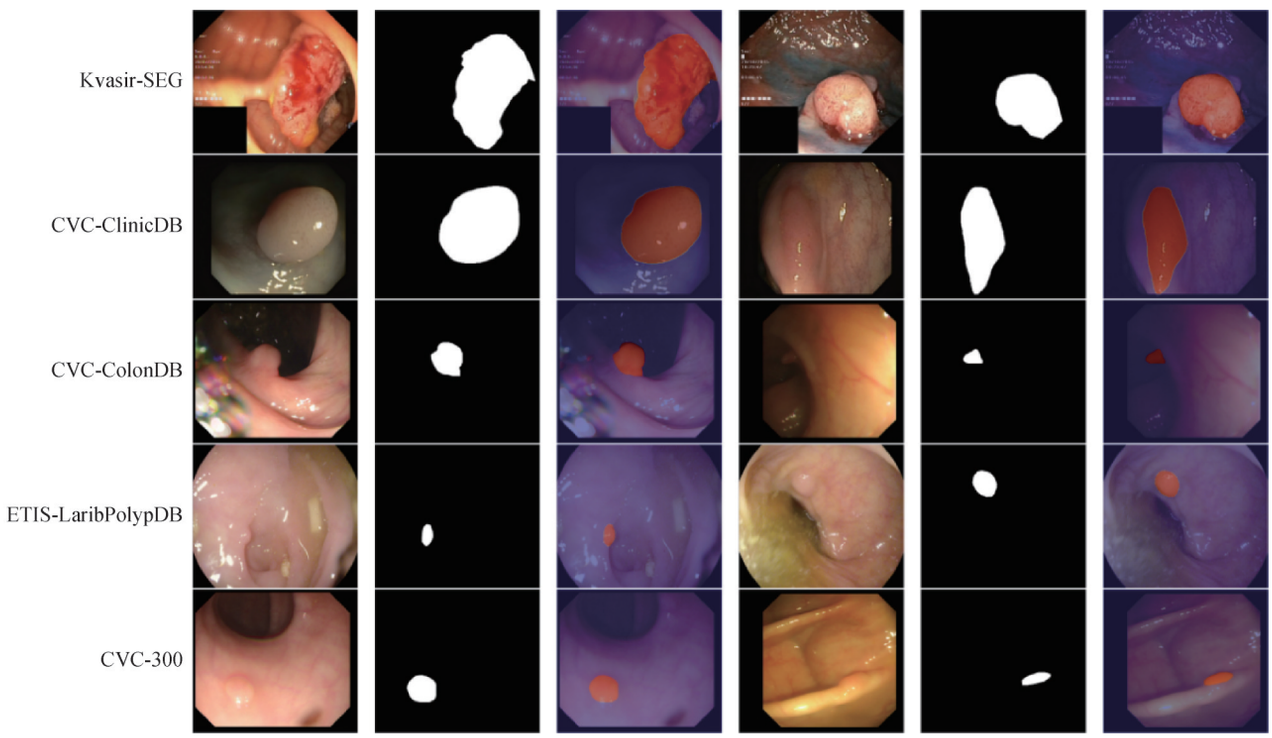


图8　模型在5个数据集上的注意力可视化结果展示

Fig. 8　Visualization of the model's attention responses across five public datasets

2.3　效果与复杂度评估

为系统评估方法的计算代价与实时性,本文在统一评测口径下(见2.1.2节)比较多种代表性方法,指标包括参数量(M)、FLOPs(G)、单帧推理时延(ms)与FPS(按1 000/ms计算),对比结果如表6所示。由表6可见,在相同输入与平台条件下,ViBound-Net以25.11 M参数量/12.60 G FLOPs实现27.3 ms(≈36.6帧/s)的推理速度;相较于FCB-Former与PRCNet,时延与计算负担显著降低;与BFNet、CTNet等轻量化范式相比,则处于同量级实时范围并维持较低计算开销,体现出更优的精度—效率折中与终端侧可部署性。上述结果与引言中所述的临床场景约束一致。

2.4　消融实验

理解各网络结构组件的相对重要性需要进行细致的结构分解和针对性的评估。为此,在5个基准数据集Kvasir-SEG、CVC-ClinicDB、CVC-ColonDB、ETIS-LaribPolypDB和CVC-300上开展了逐模块的消融实验,通过逐步移除关键子模块,并采用mDice、mIoU和MAE对性能进行比较,结果如表7所示。此外,进一步研究了损失函数设计的影响,通过比较单独的Dice损失与复合的Dice + Focal损失来验证其效果,结果如表8所示。

消融实验结果表明,不同模块对分割的稳定性、病灶完整性及边界清晰度具有独特贡献。当移除LGCAF模块时,Kvasir-SEG上的mDice和mIoU分别下降2.4%和2.6%,而MAE上升0.5%。在CVC-

表 6　模型计算复杂度与实时性比较

Table 6　Comparison of model computational complexity and real-time inference performance

模型名称	参数量/M	FLOPs/G	推理时间/ms	FPS/(帧/s)
U-Net(Ronneberger等,2015)	17.27	75.98	85.2	11.7
PraNet(Fan等,2020)	30.50	13.15	29.4	34.0
ACSNet(Zhang等,2020)	29.45	21.75	48.6	20.6
MSNet(Fang等,2022)	27.69	17.00	38.0	26.3
FCBFormer(Sanderson和Matuszewski,2022)	52.95	73.30	164.0	6.1
PolypMixNet(Jia等,2024)	28.70	17.20	38.5	26.0
PRCNet(Li等,2024)	44.98	46.82	104.7	9.6
CTNet(Xiao等,2024)	23.67	10.93	24.4	41.0
BFNet(Yue等,2025)	25.54	11.38	25.4	39.4
ViBound-Net(本文)	25.11	12.60	27.3	36.6

表 7　不同组件有效性的消融实验

Table 7　Ablation study on the effectiveness of different components

方法	Kvasir-SEG			CVC-ClinicDB			CVC-ColonDB		
	mDice	mIoU	MAE	mDice	mIoU	MAE	mDice	mIoU	MAE
基线方法	0.882 5	0.826 1	0.032 7	0.915 4	0.855 8	0.017 1	0.716 1	0.644 7	0.065 5
w/o BSFE	0.883 3	0.826 9	0.035 8	0.914 2	0.859 7	0.012 9	0.722 0	0.645 4	0.048 9
w/o LGCAF	0.884 4	0.827 2	0.032 2	0.923 3	0.867 0	0.013 0	0.732 1	0.656 2	0.037 8
w/o MRAFB	0.896 4	0.836 0	0.030 0	0.925 4	0.876 1	0.012 9	0.736 1	0.659 7	0.037 3
w/o CCSAU	0.892 0	0.834 5	0.030 8	0.930 7	0.879 5	0.012 9	0.739 8	0.657 6	0.051 4
ViBound-Net(本文)	0.908 1	0.852 7	0.027 2	0.936 8	0.887 8	0.012 6	0.741 1	0.662 1	0.037 1

方法	ETIS-LaribPolypDB			CVC-300		
	mDice	mIoU	MAE	mDice	mIoU	MAE
基线方法	0.660 4	0.598 3	0.085 1	0.859 5	0.795 6	0.019 0
w/o BSFE	0.636 4	0.559 4	0.048 9	0.846 7	0.785 4	0.017 4
w/o LGCAF	0.627 7	0.553 8	0.029 1	0.842 8	0.766 6	0.014 0
w/o MRAFB	0.658 7	0.596 3	0.020 1	0.853 2	0.790 5	0.018 7
w/o CCSAU	0.711 5	0.645 0	0.030 5	0.861 3	0.792 3	0.011 7
ViBound-Net(本文)	0.745 4	0.657 6	0.013 5	0.871 0	0.798 9	0.010 4

表 8　Dice与复合Dice有效性的消融实验

Table 8　Ablation study of the effectiveness of Dice and combined Dice

损失函数	Kvasir-SEG		CVC-ClinicDB		CVC-ColonDB		ETIS-LaribPolypDB		CVC-300	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
Dice	0.872 8	0.809 3	0.905 8	0.834 6	0.742 8	0.661 2	0.658 7	0.596 3	0.853 2	0.790 5
Dice + Focal	0.908 1	0.852 7	0.936 8	0.887 8	0.741 1	0.662 1	0.745 4	0.657 6	0.871 0	0.798 9

ClinicDB和CVC-ColonDB上也观察到类似趋势，进一步验证了LGCAF在增强全局感知能力方面的作用。相比之下，移除BSFE对细节恢复造成了显著影响，尤其是在ETIS-LaribPolypDB数据集上，mDice下降了10.9%。当移除MRAFB和CCSAU时，同样观察到性能下降，说明它们在多尺度特征融合与注意力重校准中的重要作用。

除了网络拓扑外，由损失设计所塑造的优化动态同样是影响模型有效性的关键因素，特别是在低对比度或边界模糊的场景中。与单独使用Dice损失相比，混合的Dice + Focal损失在Kvasir-SEG上的mDice和mIoU分别提升了3.5%和4.3%；在CVC-ClinicDB上分别提升了3.1%和5.3%。这些结果表明，复合损失显著增强了模型对类别不平衡及困难样本的响应能力。

3 讨 论

ViBound-Net在结直肠息肉分割任务中的实验结果表明，模型在应对息肉边界模糊与形态多样性方面具备显著优势。得益于局部—全局上下文融合注意力（LGCAF）模块与边界引导尺度感知特征增强（BSFE）模块的引入，ViBound-Net显著增强了对复杂结构与微小边缘区域的建模能力，尤其在边界细节恢复与小目标识别方面优于多数现有方法。

具体而言，ViBound-Net在多个公开数据集上均取得了竞争性结果，特别是在CVC-ClinicDB和Kvasir-SEG数据集上，mDice和mIoU分数显著提升，充分展现了方法在典型结构下的边界保持能力与语义分割一致性。同时，在CVC-300数据集上，模型在不同图像分辨率和复杂背景条件下均表现出较强的跨域泛化能力。

然而，本文亦观察到，在部分具有挑战性的场景中（如ETIS-LaribPolypDB数据集），模型在处理低对比度、小尺寸息肉或高噪声图像时的鲁棒性略有下降，具体表现为局部区域边界预测的不稳定性或细节信息的部分缺失。本文认为这一现象主要源于跨数据集成像风格的领域差异（domain gap），以及轻量化结构在极端场景中对深层语义信息捕获能力的限制。

未来的工作可从以下几个方向进一步提升模型性能：1）融合多尺度边界引导监督策略，增强对复杂息肉结构的建模能力；2）引入自适应解码路径，提升极小尺寸息肉目标的恢复精度；3）结合风格迁移或领域适应机制，强化模型在跨中心、多设备采集数据场景下的鲁棒性与泛化能力。

4 结 论

本文提出的ViBound-Net模型在结直肠息肉分割任务中展现了优异的性能，特别是在边界保持、细节恢复和小目标检测方面。通过引入局部—全局上下文融合注意力（LGCAF）模块、边界引导尺度感知特征增强（BSFE）模块以及多尺度残差注意力融合块（MRAFB）和紧凑通道—空间注意单元（CCSAU），ViBound-Net有效地克服了息肉图像中的边界模糊和形态复杂问题，显著提升了分割精度。特别地，结合Dice和Focal损失的复合损失函数显著提高了模型对难分区域的分割能力，优化了模型在类别不平衡和小目标检测中的表现。尽管如此，实验也表明该模型在低对比度区域和强噪声条件下的结构细节识别能力仍面临一定挑战。上述现象揭示了当前模型在感知深度控制与上下文一致性建模方面的潜在局限。未来研究可围绕跨模态感知融合、显式语义引导机制及数据分布自适应推理策略等方向，进一步提升其在复杂医学图像环境下的鲁棒性与临床适配性。

参考文献（References）

Bernal J, Sánchez F J, Fernández-Esparrach G, Gil D, Rodríguez C and Vilariño F. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics, 43: 99-111 [DOI: 10.1016/j.compmedimag.2015.02.007]

Bernal J, Sanchez J and Vilarino F. 2012. CVC-ColonDB: a database for assessment of polyp detection [EB/OL]. [2025-09-29]. https://vi.cvc.uab.es/colon-qa/cvccolondb/

Brenner H, Stock C and Hoffmeister M. 2014. Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: systematic review and meta-analysis of randomised controlled trials and observational studies. BMJ, 348: #g2467 [DOI: 10.1136/bmj.g2467]

Cai L H, Chen L J, Huang J H, Wang Y F and Zhang Y B. 2024. Know your orientation: a viewpoint-aware framework for polyp segmentation. Medical Image Analysis, 97: #103288 [DOI: 10.1016/j.

media.2024.103288]

Dong B, Wang W H, Fan D P, Li J P, Fu H Z and Shao L. 2023. Polyp-PVT: polyp segmentation with pyramid vision transformers. CAAI Artificial Intelligence Research, 2: #9150015 [DOI: 10.26599/AIR.2023.9150015]

Dumitru R G, Peteleaza D and Craciun C. 2023. Using DUCK-Net for polyp image segmentation. Scientific Reports, 13(1): #9803 [DOI: 10.1038/s41598-023-36940-5]

Fan D P, Cheng M M, Liu Y, Li T and Borji A. 2017. Structure-measure: a new way to evaluate foreground maps//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 4548-4557 [DOI: 10.1109/ICCV.2017.487]

Fan D P, Gong C, Cao Y, Ren B, Cheng M M and Borji A. 2018. Enhanced-alignment measure for binary foreground map evaluation//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: ijcai.org: 698-704 [DOI: 10.24963/ijcai.2018/97]

Fan D P, Ji G P, Zhou T, Chen G, Fu H Z, Shen J B, et al. 2020. PraNet: parallel reverse attention network for polyp segmentation//Proceedings of the 23rd International Conference on Medical Image Computing and Computer-Assisted Intervention. Lima, Peru: Springer: 263-273 [DOI: 10.1007/978-3-030-59725-2_26]

Jha D, Smedsrud P H, Riegler M A, Halvorsen P, De Lange T, Johansen D, et al. 2020. Kvasir-SEG: a segmented polyp dataset//Proceedings of the 26th International Conference on MultiMedia Modeling (MMM). Daejeon, Korea (South): Springer: 451-462 [DOI: 10.1007/978-3-030-37734-2_37]

Jha D, Tomar N K, Sharma V and Bagci U. 2023. TransNetR: transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing//Medical Imaging with Deep Learning (MIDL). Nashville, USA: PMLR: 1372-1384

Jia X, Shen Y T, Yang J H, Song R, Zhang W, Meng M Q H, et al. 2024. PolypMixNet: enhancing semi-supervised polyp segmentation with polyp-aware augmentation. Computers in Biology and Medicine, 170: #108006 [DOI: 10.1016/j.compbiomed.2024.108006]

Li J, Wang J W, Lin F W, Heidari A A, Chen Y, Chen H L, et al. 2024. PRCNet: a parallel reverse convolutional attention network for colorectal polyp segmentation. Biomedical Signal Processing and Control, 95: #106336 [DOI: 10.1016/j.bspc.2024.106336]

Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017. Focal loss for dense object detection//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 2999-3007 [DOI: 10.1109/ICCV.2017.324]

Mao X, Li H Y, Li X X, Bai C B and Ming W J. 2025. C^2E-Net: cascade attention and context-aware cross-level fusion network via edge learning guidance for polyp segmentation. Computers in Biology and Medicine, 185: #108770 [DOI: 10.1016/j.compbiomed.2024.108770]

Margolin R, Zelnik-Manor L and Tal A. 2014. How to evaluate foreground maps?//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, USA: IEEE: 248-255 [DOI: 10.1109/CVPR.2014.39]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Sanderson E and Matuszewski B J. 2022. FCN-transformer feature fusion for polyp segmentation//Proceedings of the 26th Annual Conference on Medical Image Understanding and Analysis. Cambridge, UK: Springer: 892-907 [DOI: 10.1007/978-3-031-12053-4_65]

Siegel R L, Giaquinto A N and Jemal A. 2024. Cancer statistics, 2024. CA: A Cancer Journal for Clinicians, 74(1): 12-49 [DOI: 10.3322/caac.21820]

Silva J, Histace A, Romain O, Dray X and Granado B. 2014. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. International Journal of Computer Assisted Radiology and Surgery, 9(2): 283-293 [DOI: 10.1007/s11548-013-0926-3]

Tan M X and Le Q V. 2021. EfficientNetV2: smaller models and faster training//Proceedings of the 38th International Conference on Machine Learning (ICML). Vienna, Austria: PMLR: 10096-10106

Tang L F, Zhang H, Xu H and Ma J Y. 2023. Deep learning-based image fusion: a survey. Journal of Image and Graphics, 28(1): 3-36 (唐霖峰, 张浩, 徐涵, 马佳义. 2023. 基于深度学习的图像融合方法综述. 中国图象图形学报, 28(1): 3-36) [DOI: 10.11834/jig.220422]

Targ S, Almeida D and Lyman K. 2016. ResNet in ResNet: generalizing residual architectures [EB/OL]. [2025-09-29]. https://arxiv.org/pdf/1603.08029.pdf

Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: convolutional block attention module//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]

Wu Z Y, Lyu F M, Chen C L Z, Hao A M and Li S. 2024. Colorectal polyp segmentation in the deep learning era: a comprehensive survey [EB/OL]. [2025-09-29]. https://arxiv.org/pdf/2401.11734.pdf

Xiao B, Hu J W, Li W S, Pun C M and Bi X L. 2024. CTNet: contrastive transformer network for polyp segmentation. IEEE Transactions on Cybernetics, 54(9): 5040-5053 [DOI: 10.1109/TCYB.2024.3368154]

Yue G H, Wu S J, Li G, Zhao C, Hao Y, Zhou T W, et al. 2025. Boundary-guided feature-aligned network for colorectal polyp segmentation. IEEE Transactions on Circuits and Systems for Video Technology, 35(7): 6993-7004 [DOI: 10.1109/TCSVT.2025.3542969]

Zhang R F, Li G B, Li Z, Cui S G, Qian D H and Yu Y Z. 2020. Adaptive context selection for polyp segmentation//Proceedings of the 23rd International Conference on Medical Image Computing and Computer-Assisted Intervention. Lima, Peru: Springer: 253-262 [DOI: 10.1007/978-3-030-59725-2_25]

Zhao X, Zhang L and Lu H. 2021. Automatic polyp segementation via multi-scale substration network//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer: 120-130

Zhao X Q, Jia H P, Pang Y W, Lyu L, Tian F, Zhang L H, et al. 2023. M^2SNet: multi-scale in multi-scale subtraction network for medical image segmentation [EB/OL]. [2025-09-29]. https://arxiv.org/pdf/2303.10894.pdf

Zhou T, Dong Y L, Huo B Q, Liu S and Ma Z J. 2021. U-Net and its applications in medical image segmentation: a review. Journal of Image and Graphics, 26(9): 2058-2077 (周涛, 董雅丽, 霍兵强, 刘珊, 马宗军. 2021. U-Net网络医学图像分割应用综述. 中国图象图形学报, 26(9): 2058-2077) [DOI: 10.11834/jig.200704]

Zhou T, Zhang Y Z, Zhou Y, Wu Y and Gong C. 2023. Can SAM segment polyps? [EB/OL]. [2025-09-29]. https://arxiv.org/pdf/2304.07583.pdf

作者简介

吴琪琪，女，硕士研究生，主要研究方向为计算机视觉。E-mail：231210702102@stu.just.edu.cn

邓星，通信作者，女，副教授，主要研究方向为大数据与机器学习、计算机视觉和图计算及应用。E-mail：xdeng@just.edu.cn

邵海见，男，副教授，主要研究方向为计算机视觉、自然语言处理、虚拟现实、增强现实、量子神经网络计算、智能计算、边缘计算、云计算与图计算。E-mail：jsj_shj@just.edu.cn

王飞，男，博士，讲师，主要研究方向为计算机视觉和自然语言处理。E-mail：feiw@just.edu.cn