基于级联注意力的结肠息肉图像分割算法研究

周孟然[a], 陆 鹏[b]
安徽理工大学 a. 电气与信息工程学院; b. 计算机科学与工程学院, 安徽 淮南 232000

摘 要:目的 针对现有 Transformer 模型在息肉图像分割中存在注意力分散以及作为编码器提取的多级特征在融合时易产生信息丢失导致的分割精度不高的问题,提出一种新的分割模型 PVT-CAMNet。**方法** 在该模型中,使用金字塔式 Transformer(Pyramid Vision Transformer, PVT)作为编码器,接着设计了多尺度特征注意力提取模块(Multi-scale Feature Attention Extraction,MFAE)和层间注意力聚合模块(Inter-layer Attention Aggregation, IA)。其中,PVT 通过其自注意力机制保证了模型的泛化能力,MFAE 使用不同大小的滤波器多尺度提取特征,旨在缓解注意力分散问题;IA 交互融合不同层级特征,有效解决多级特征融合产生的信息丢失问题;最后引入全局上下文模块(Global Context, GC)使模型更好地理解特征图之间的像素依赖关系。**结果** 在 Kvasir、CVC-ClinicDB、CVC-ColonDB 和 ETIS 数据集上进行了评估,相较于最优基线模型,mDice、mIoU 分别提高了 1.76%、0.81%、1.51%、1.74%、3.15%、2.65% 和 1.73%、3.84%。**结论** PVT-CAMNet 的学习性能和泛化性能均优于其他先进方法,在息肉图像分割上具有一定的应用价值。

关键词:息肉图像分割;多尺度注意力提取;层间注意力聚合;全局上下文
中图分类号:TP391.4 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2026.0001.001

Research on Colon Polyp Image Segmentation Algorithm Based on Cascaded Attention

ZHOU Mengran[a], LU Peng[b]
a. School of Electrical and Information Engineering; b. School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232000, Anhui, China

Abstract: Objective Aiming at the problems of scattered attention in existing Transformer models for polyp image segmentation and the low segmentation accuracy caused by information loss during the fusion of multi-level features extracted by the encoder, a new segmentation model named PVT-CAMNet is proposed. **Methods** In this model, the Pyramid Vision Transformer (PVT) was used as the encoder. Then, a Multi-scale Feature Attention Extraction (MFAE) module and an Inter-layer Attention Aggregation (IA) module were designed. Among them, the PVT ensured the generalization ability of the model through its self-attention mechanism. The MFAE used filters of different sizes to extract features at multiple scales to alleviate the problem of scattered attention. The IA interactively fused features at different levels to effectively solve the problem of information loss caused by the fusion of multi-level features. Finally, a Global Context (GC) module was introduced to enable the model to better understand the pixel dependency relationship between feature maps. **Results** Evaluations were carried out on the Kvasir, CVC-ClinicDB, CVC-ColonDB, and ETIS datasets.

收稿日期:2024-01-26 **修回日期:**2024-05-13 **文章编号:**1672-058X(2026)01-0001-10
基金项目:安徽省科技重大专项(201903A07020013)资助.
作者简介:周孟然(1965—),男,安徽淮南人,博士,教授,从事矿山机电系统监测、光电信息处理与煤矿安全检测研究.
通信作者:陆鹏(1999—),男,安徽亳州人,硕士,从事图像处理研究. Email:1577172319@qq.com.
引用格式:周孟然, 陆鹏. 基于级联注意力的结肠息肉图像分割算法研究[J]. 重庆工商大学学报(自然科学版),2026,43(1):1-10.
ZHOU Mengran, LU Peng. Research on colon polyp image segmentation algorithm based on cascaded attention[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2026, 43(1): 1-10.

Comparing the performance of the proposed PVT-CAMNet model with that of the optimal baseline model, the mDice values of PVT-CAMNet increased by 1.76%, 1.51%, 3.15%, and 1.73% respectively, and the mIoU values of PVT-CAMNet increased by 0.81%, 1.74%, 2.65%, and 3.84% respectively, on these four datasets. **Conclusion** PVT-CAMNet is superior to other advanced methods in both learning performance and generalization capability, demonstrating significant application value in polyp image segmentation.

Keywords: polyp image segmentation; multi-scale attention extraction; inter-layer attention aggregation; global context

结直肠癌(Colorectal Cancer, CRC)是世界三大高发癌症之一。2020 年我国 CRC 总体发病率已跃升至恶性肿瘤的第 2 位,CRC 的死亡率也位居第 5 位[1]。迄今为止,CRC 的具体病因尚未查明,并且由于息肉形状各异,扁平且与周围组织具有高度相似性,人工进行结肠息肉图像的分割将花费大量的人力、物力,并且难以在短时间内完成分割工作,给医护人员带来了极大的负担[2]。

为应对上述挑战,在早期的研究工作中,结肠息肉图像的分割方法通常依赖于边缘检测、阈值分割等传统机器学习技术[3]。近年来,随着深度学习的兴起,基于深度学习的研究方法在医学图像分割中表现出了巨大的潜力[4]。具有对称结构的 U-Net[5]取得了代表性的成功。受 U-Net 的启发,U-Net++[6]和 PraNet[7]网络相继提出。随着不断探索和创新,Transformer[8]进入研究者的视野,其最初是为了解决自然语言处理任务而提出的一种新型网络架构。最近,Dosovitski[9]及其团队提出了视觉变换器(Vision Transformer),将 Transformer 引入到图像领域并取得了不错的效果。受此启发,Chen 等[10]提出 TransUNet 网络,在主干网络中引入 Transformer 结构,提取全局上下文输入序列,但是其网络模型参数量太大;Wang 等[11]提出 SSFormer,使用卷积算子等方法强调关键局部特征,但是对小型息肉数据集的分割效果不是很理想;Dong 等[12]提出 Polyp-PVT 使用金字塔式 Transformer (Pyramid Vision Transformer, PVT) 作为主干网络,提取不同级别的息肉特征,提高了分割精度。

以上方法虽然对息肉图像的分割结果起到了一定的改善,但是仍然存在一些不足:传统方法只能粗略地对息肉图像进行分割,而且分割效果不是很好;基于卷积神经网络的方法虽然在一些数据集上取得了不错的效果,但是由于卷积神经网络的局限性,加之息肉形态的多样性和数据集的稀缺性,导致模型的泛化能力较低;由文献[11]可知,随着 Transformer 结构的深化,整体特征不断混合和收敛,这往往会导致注意力分散。此外,PVT 输出的各级特征有着很大的鸿沟,不正确的融合方式会产生信息丢失问题,并且由于低级信息的冗余性和杂乱性,导致直接使用低级语义信息会使目标预测过于平滑而导致边界模糊。

针对上述问题,本文提出了一个结合 3 种注意力的结肠息肉图像分割模型 PVT-CAMNet。在模型中主要设计了多尺度特征注意力提取模块(Multi-scale Feature Attention Extraction,MFAE)、层间注意力聚合模块(Inter-layer Attention Aggregation, IA)、并引入全局上下文模块(Global Context,GC)。MFAE 通过使用多尺度卷积以及空间通道混合注意力来限制 PVT 中的注意力分散,IA 模块通过自底向上的注意力交互使相邻级别的特征充分融合,达到减少各级特征融合时产生信息丢失问题的目的,GC 模块实现了特征图的全局交互,提高模型对特征图中像素依赖关系的理解。最终,通过逐层上采样获得最后的分割结果。在 4 个数据集上进行实验,实验结果表明:PVT-CAMNet 具有较强的学习和泛化能力。

1 系统结构设计

1.1 PVT-CAMNet 模型结构

PVT-CAMNet 模型结构如图 1 所示,由 PVT 编码器 Encode,级联注意力模块 (Cascade Attention Module, CAM) 和逐层上采样的解码器模块 Decode 构成。其中级联注意力模块包括多尺度特征注意力提取模块 MFAE、层间注意力聚合模块 IA 和全局上下文模块 GC。原始图片由 PVT 编码器分级处理后输出多尺度特征信息$\{E_1、E_2、E_3、E_4\}$,随后输入到级联注意力模块得到局部和全局信息,最后将聚合的特征送入到逐层上采样的解码器中。具体过程:将编码器提取的多尺度特征信息$\{E_1、E_2、E_3、E_4\}$先进行 3×3 卷积,将通道数统一调整为 64,随后将处理好的特征输入到级联注意力模块。MFAE 分别对 4 个不同尺寸的特征进行多尺度的注意力聚合,全方位突出相关信息,抑制不相关特征,IA 进行层间交互,使网络更好地关注局部上下文信息,然后将多尺度信息聚合输入到 GC 完成局部到全局的特征学习,最后经过逐层上采样模块获得最终分割结果。

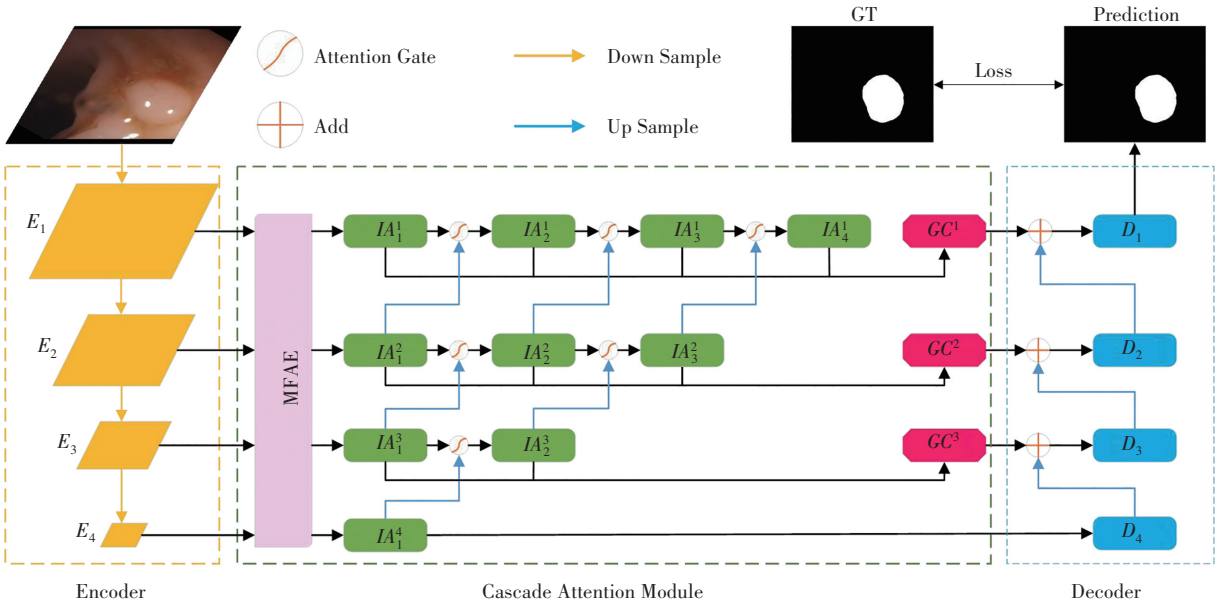

图 1　PVT-CAMNet 模型结构图

Fig. 1　Structure diagram of PVT-CAMNet model

1.2　PVT 编码器

Transformer 是近几年新提出的一种新型的神经网络架构。其不同于卷积神经网络在内核中提取信息进行权重参数训练,而是使用注意力机制获得类似的特征,通过点积操作自适应提取特征来训练其中的权重参数,这可以使 Transformer 具有有效的全局接收场,减少学习偏差。受 Polyp-PVT 的启发,由此引入 PVT 作为息肉图像分割网络的编码器多尺度提取特征,使网络能够综合不同尺度下的信息,更全面地理解息肉图像中的复杂特征,增加网络的泛化能力。

PVT 每一个阶段结构如图 2 所示,在第一阶段,$H\times W\times 3$ 的图像将被分为 4×4 的图像块送入模型以获得高分辨率的特征图,随后将展平的图像块送入线性投射层并获取 C_1 通道的嵌入特征块,然后嵌入片元和位置嵌入一同送入 L_1 层的 Transformer 编码器中,经过改进的多头自注意力机制(Spatial Reduction Attention, SRA)和一个前向传播,输出特征图 E_1,其尺寸为原来的四分之一。在同样的操作下,将前一阶段得到的特征图送入后续阶段得到 E_2、E_3、E_4 特征图,其相对于输入图像的步幅分别为 8、16、32,最终得到 4 个层级的特征金字塔{E_1、E_2、E_3、E_4}。在这些不同层级的特征中,E_1 提供了分割目标的低级语义信息,包含大量的细节信息,E_2、E_3、E_4 则提供了高级语义信息,包含了分割目标的全局信息。

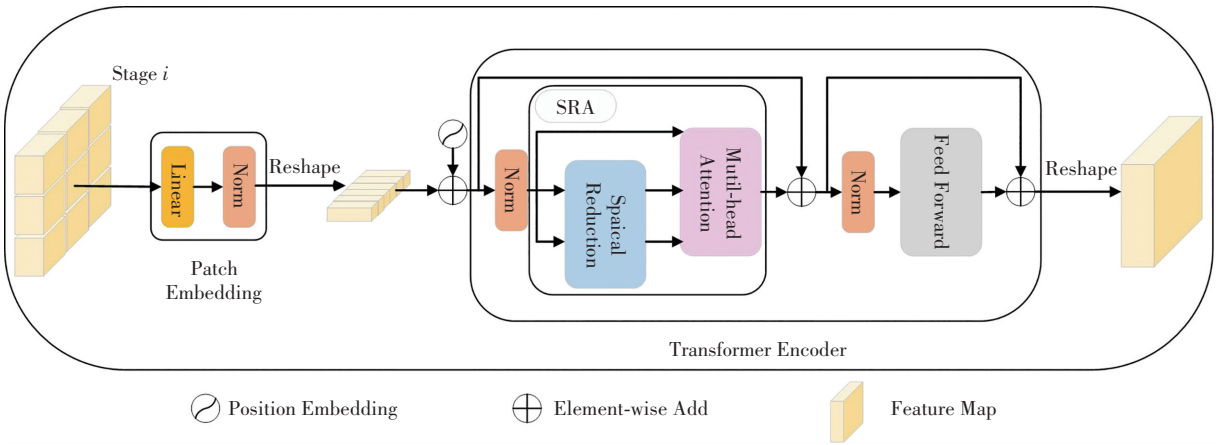

图 2　Transformer stage 模块结构图

Fig. 2　Structure diagram of Transformer stage module

1.3　CAM 级联注意力模块

1.3.1　MFAE 模块

在 PVT 处理图像的过程中,将会频繁进行自我注意力操作。然而,图像经过多次自我注意力操作后,可能导致特征的注意力分散,使编码器提取的特征不够突出,从而影响整体性能。因此,设计了多尺度特征注

意力提取模块 MFAE,通过使用不同大小的卷积核和通道空间交叉注意力[13],扩大局部感受野,突出相关特征,使注意力重新集中在相邻的有限元分析上,缓解注意力分散问题。

MFAE 模型结构如图 3 所示,在经过编码器输出的 4 个特征的通道数统一为 64 后,进入 MFAE 模块。该模块主要包含 4 个不同尺度的卷积操作和空间通道混合注意力。具体过程如下:首先,将统一后的特征通过多尺度卷积操作,使用尺寸为 1、3、5、7 的卷积核提取特征,接着采用空洞卷积,其空洞比为 1、3、5、7,以进一步扩大感受野,然后应用空间通道混合注意力,以增强特征的表征能力,最后将经过注意力机制加强的特征进行拼接,并与经过 1×1 卷积的原始特征相加构成残差链接,以防止梯度消失,得到最终的输出特征。公式表达如式(1)—式(5)所示。

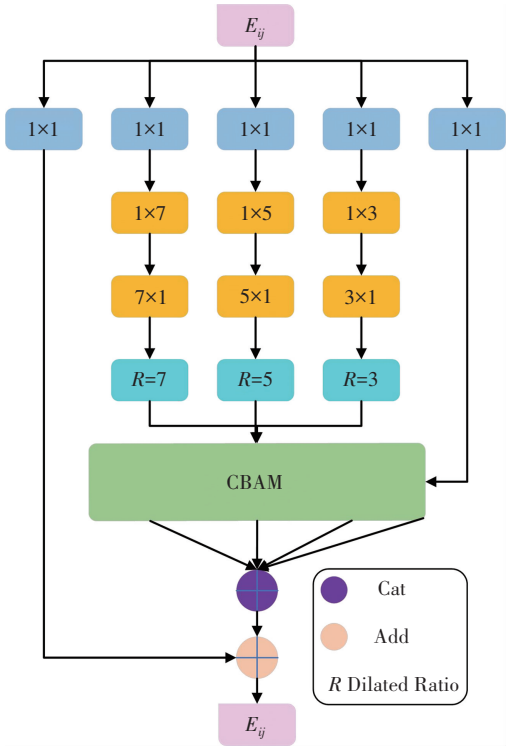

图 3 MFAE 模块结构图

Fig. 3 Structure diagram of MFAE module

$$E_{i1}=\mathrm{SA}(\mathrm{CA}(C_{1\times1}(E_{i1}))) \tag{1}$$

$$E_{i2}=\mathrm{SA}(\mathrm{CA}(C_{1\times1}(C_{1\times3}(C_{3\times1}(E_{i2}))))) \tag{2}$$

$$E_{i3}=\mathrm{SA}(\mathrm{CA}(C_{1\times1}(C_{1\times5}(C_{5\times1}(E_{i3}))))) \tag{3}$$

$$E_{i4}=\mathrm{SA}(\mathrm{CA}(C_{1\times1}(C_{1\times7}(C_{7\times1}(E_{i4}))))) \tag{4}$$

$$E_{ij}=\mathrm{ReLU}(\mathrm{Cat}(E_{i1},E_{i2},E_{i3},E_{i4})+C_{1\times1}(E_{i1})) \tag{5}$$

其中,E_{i1}、E_{i2}、E_{i3}、E_{i4} 表示编码器输出的 4 个不同尺度的特征图,SA (Spatial Attention)表示空间注意力,CA (Channel Attention)表示通道注意力,$C_{1\times1}$、($C_{1\times3}$、$C_{3\times1}$)、($C_{1\times5}$、$C_{5\times1}$)、($C_{1\times7}$、$C_{7\times1}$)分别表示卷积核为 1、3、5、7 的卷积操作,Cat 表示通道维度上的拼接操作,ReLU 为激活函数。

1.3.2 IA 模块

为了缓解特征融合容易产生信息丢失的问题,设计了层间注意力聚合模块 IA,其主要由多个门控注意力 (Attention Gate, AG)[14] 交互构成。此模块接收 MFAE 输出的 4 个特征,自底向上使相邻级别特征之间进行信息交互,实现信息的跨层传递,结合局部和全局信息,使得模型能够同时关注息肉的局部细节和周围的全局上下文,有助于提高分割结果的一致性和完整性。

定义一个基本层间注意力单元 IA(I_A),假定使用 F_a 和 F_b 来表示相邻级别的特征图,其计算过程如式(6)和式(7)所示:对 F_a 和 F_b 进行 1×1 的卷积统一通道数,再做逐元素的相加,然后经过 ReLU 激活函数,再进行 1×1 的卷积和 Sigmoid 得到注意力权重,最后使用注意力权重对特征图进行加权。

$$I_A=\mathrm{Conv}(F_a\odot F_b) \tag{6}$$

$$\odot=F_b\times(\mathrm{Sigmoid}(\mathrm{Conv}(\mathrm{ReLU}(\mathrm{Conv}(F_a)+\mathrm{Conv}(F_b))))) \tag{7}$$

其中,⊙是门控注意力,F_a 和 F_b 为相邻级别特征图,Sigmoid 和 ReLU 为激活函数,Conv 表示卷积核为 1 的卷积运算。IA 结构中的 AG 单元如图 4 所示,其可以捕获 F_a 和 F_b 的互补信息并且具有出色的上下文能力,从而为解码器提供更丰富的信息。

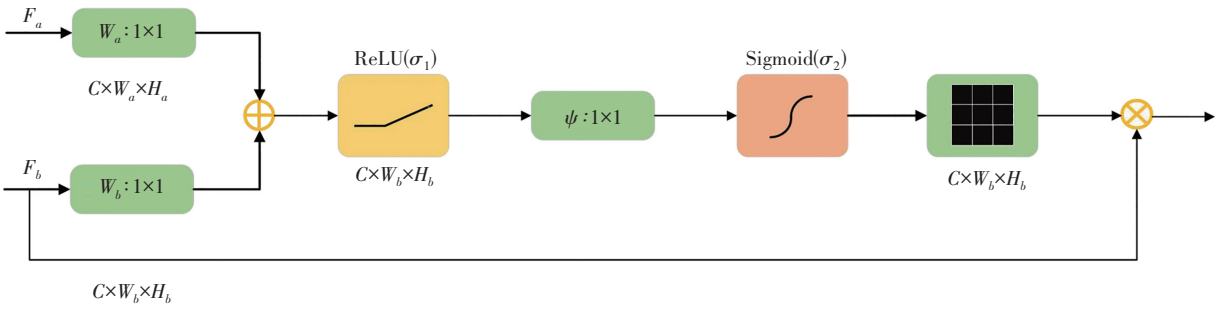

图 4 Attention Gate 结构图

Fig. 4 Structure diagram of Attention Gate

1.3.3　GC 模块与 Decoder

为了获得跨多个特征级别的高阶互补信息,全局上下文模块 GC[15] 水平和垂直接收多个 IA 模块的输出特征来计算一系列具有不同阶数和感受野的特征。GC 模块结构如图 5 所示。

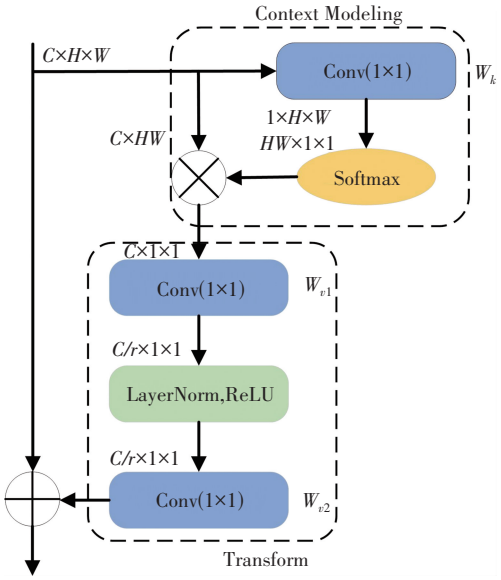

图 5　GC 模块结构图

Fig. 5　Structure diagram of GC module

GC 模块聚合了相应级别和任意级别之间的特定尺度特征和层间交互特征生成的互补增强特征(GC^i)。在此模块中,每个像素都可以与整个图像的其他像素进行交互,从而允许网络利用全局上下文信息来提取更丰富的特征表示。具体过程如下:

(1) 采用 1×1 卷积和 Softmax 进行全局池化获得注意力权重从而进行注意力池化;

(2) 使用 1×1 卷积 Wv 进行特征变换将通道数降低为原来的 $1/r$ (r 取 32,48,64),此操作可以实现模型轻量化;

(3) 利用加法将全局上下文特征聚合到每个位置的特征上。计算过程如式(8)—式(9)所示:

$$X_i = \mathrm{GC}\left(\sum_{n=1}^{5-i} IA_n^i \mid n = 1,2,3,4\right) \tag{8}$$

$$Z_i = X_i + \boldsymbol{W}_{v2}\mathrm{ReLU}\left(\mathrm{LN}\left(\boldsymbol{W}_{v1}\sum_{j=1}^{N_p}\frac{e^{W(k^{x^j})}}{\sum_{j=1}^{N_p} e^{W(k^{x_m})}}x_j\right)\right) \tag{9}$$

其中,$\partial j = \dfrac{e^{W(k^{x^j})}}{\sum_{j=1}^{N_P} e^{W(k^{x_m})}}$ 代表注意力池化的权重,$(\cdot) = \boldsymbol{W}_{v2}\mathrm{ReLU}(\mathrm{LN}(\boldsymbol{W}_{v1}(\cdot)))$ 表示一个 Transformer 块,LN (Layer Normalization) 代表层归一化操作,$\boldsymbol{W}_{v1}$ 和 $\boldsymbol{W}_{v2}$ 代表特征变换矩阵。

解码器如图 1 右侧所示,由 4 个阶段构成,具体过程为 D_4 以双线性插值的方式尺寸扩大为原来的 2 倍,然后与 GC 模块输出的特征 GC^3 相加得到 D_3,以此类推得到 D_2,D_1。计算过程如式(10)所示:

$$D_i = \mathrm{Conv}(\mathrm{Bilinear}(D_i - 1) + GC^i \tag{10}$$

其中,Conv 代表 3×3 卷积,Bilinear 代表双线性插值操作。

1.4　损失函数

本文所使用的损失函数如式(11)—式(13)所示:

$$f_{\mathrm{Loss}} = L_{\mathrm{IoU}}^w + L_{\mathrm{BCE}}^w + L_f \tag{11}$$

$$L_f = l_f^1 + l_f^2 + l_f^3 + l_f^4 \tag{12}$$

$$l_f^i = \| F_P^i - F_G^i \|_2, i = 1,2,3,4 \tag{13}$$

其中,L_{IoU}^w 与 L_{BCE}^w 分别表示加权 IoU 损失和二元交叉熵损失,这两种损失函数在分类任务中被广泛采用。此外 L_f 则是受到了 MSNet[16] 的启发,使用在 ImageNet 预训练的分类网络 VGG-16 分别提取预测值 F_P 和真实标签 F_G 的多尺度特征,并计算他们之间的欧几里得距离 l_f^i,从而实现像素级别的监督。

2　仿真实验与结果分析

2.1　实验设置

本次实验在 4 个公开的息肉数据集上对 PVT-CAMNet 模型和当下 6 个最具代表性模型进行了学习能力和泛化能力的评估。在模型学习能力的测试中,将来自 CVC-ClinicDB[17] 和 Kvasir[18] 中 80%的图片用于训练,10%用于验证,10%用于测试。在模型泛化能力的测试中,使用 CVC-ClinicDB 和 Kvasir 中 80%的图片用于训练,在 CVC-ColonDB[19] 和 ETIS[20] 上进行测试。四个数据集详细信息如表 1 所示。

表 1　各个数据集详细信息

Table 1　Detailed information of each dataset

数据集	图片数量	图片尺寸
CVC-ClinicDB	612	384×288
Kvasir	1 000	不固定
CVC-ColonDB	380	574×966
ETIS	196	1 225×966

2.1.1　评估指标

在此实验中,使用 3 个评估指标来度量分割性能,分别为平均 Dice 系数(mean Dice coefficient, mDice)、平均交并比(mean Intersection over Union, mIoU)和平均绝对误差(Mean Absolute Error, MAE)。mDice 和

mIoU 的得分越高,表明分割结果的精度越高,而 MAE 的得分越低则意味着更高的分割准确度。这些指标的计算公式分别为式(14)、式(15)和式(16)。

$$V_{\text{mDice}} = \frac{1}{n}\sum_{i=1}^{n}\frac{2 \times N_{\text{TP}}}{2 \times N_{\text{TP}} + N_{\text{FP}} + N_{\text{TN}}} \tag{14}$$

$$V_{\text{mIoU}} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}} + N_{\text{TN}}} \tag{15}$$

$$V_{\text{MAE}} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - Y_i'| \tag{16}$$

其中,N_{TP} 表示正确分类为前景像素的样本数,N_{TN} 表示正确分类为背景像素的样本数,N_{FP} 表示被错误分类为前景像素的样本数,N_{FN} 表示被错误分类为背景像素的样本数。Y_i 表示原始样本的标签值,而 Y_i' 表示模型的输出结果。

2.1.2 实验配置与模型超参数

实验基于 Pytorch 框架,使用 NVIDIA-3080 12G 显卡训练模型。使用随机梯度下降(Stochastic Gradient Descent,SGD)作为模型优化器,批次大小(batch_size)为 16,初始学习率(lr)为 0.02,衰减率(decay)为 0.000 5,训练周期(epoch)为 100。在训练期间,将图像随机裁剪为 224、256、288、320、352 大小,并且使用随机翻转,旋转和侵蚀等数据增强的方法,增加网络的鲁棒性。

2.2 结果评估与分析

为了验证所提出的 PVT-CAMNet 模型在结肠息肉图像分割任务中的性能,本文对此模型与当前领域内的 6 种最具代表性的分割模型进行了详细的学习能力评估和泛化能力评估。6 种模型即 U-Net、U-Net++、PraNet、TransUNet、SSFormer 和 Polyp-PVT。通过在多个评价指标上进行全面分析,深入了解每个模型的优劣势。这两种评估方法能够更全面地了解所提出的PVT-CAMNet 模型在结肠息肉图像分割任务中的性能表现。

2.2.1 学习能力评估

为了验证本文提出的 PVT-CAMNet 模型的学习能力,在 CVC-ClinicDB 和 Kvasir 两个息肉图像数据集上对 PVT-CAMNet 与其他 6 种模型进行了学习能力的实验对比,各网络模型分割的性能指标如表 2 所示,加粗数据代表最优值。

表 2 学习能力实验结果

Table 2 Experiment results of learning ability

模 型	Kvasir 数据集			CVC-ClinicDB 数据集		
	mDice ↑	mIoU ↑	MAE ↓	mDice ↑	mIoU ↑	MAE ↓
U-Net	0.821	0.756	0.055	0.824	0.767	0.019
U-Net++	0.824	0.753	0.048	0.797	0.741	0.022
PraNet	0.901	0.862	0.030	0.902	0.858	0.009
TransUNet	0.869	0.816	0.040	0.847	0.798	0.017
SSFormer	0.907	0.858	0.028	0.921	0.873	0.009
Polyp-PVT	0.911	0.860	0.028	0.927	0.878	0.009
PVT-CAMNet	**0.927**	**0.867**	**0.027**	**0.941**	**0.888**	**0.008**

从表 2 中数据可以看出,本文提出的模型在 Kvasir 数据集中相较于最优基线模型 Polyp-PVT 的 mDice,mIoU 分别提高了 1.76%、0.81%,MAE 则降低了 3.57%。在 CVC-ClinicDB 数据集中相较于最优基线模型 Polyp-PVT 的 mDice,mIoU 分别提高了 1.51%、1.14%,MAE 则降低了 11%。本文所提出模型与其他 6 个模型在 CVC-ClinicDB 和 Kvasir 数据集上的实际分割效果图如图 6 所示,可以看出本文所提出模型在这两个数据集上的分割效果相较于其他 6 个模型都更为接近标签图,结合表中数据和实际分割效果来看,PVT-CAMNet 模型相较于其他 6 个模型具有更优异的学习能力。

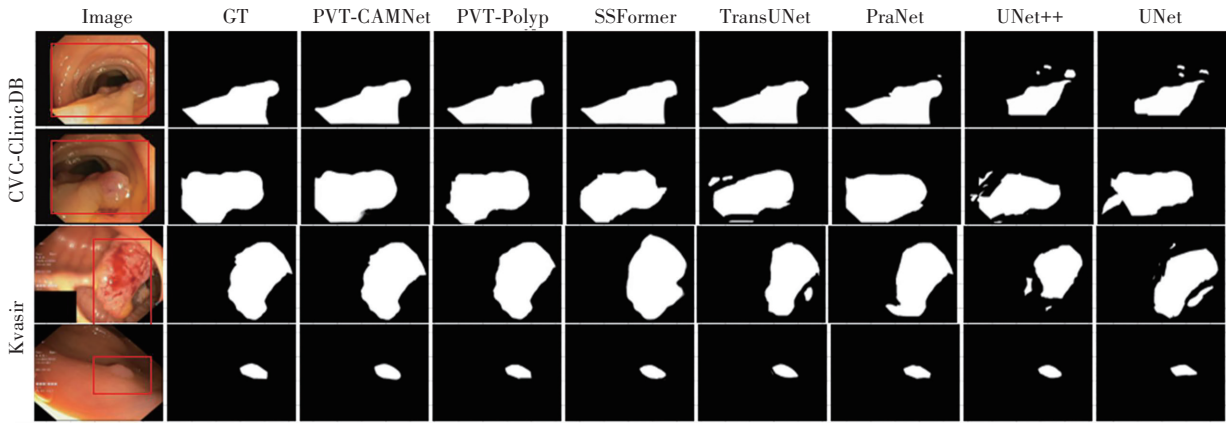

图 6 PVT-CAMNet 与对照模型在学习能力实验中效果对比图

Fig. 6 Comparison of PVT-CAMNet and control model in learning ability experiments

2.2.2　泛化能力评估

为了评估 PVT-CAMNet 模型的泛化能力,在 Kvasir 和 CVC-ClinicDB 数据集中随机抽取 1 450 张图片进行训练,然后在 CVC-ColonDB 和 ETIS 数据集上进行测试。此项工作可以评估模型在未知数据集上的泛化能力,评估结果如表 3 所示,加粗数据代表最优值。

表 3　泛化能力实验结果

Table 3　Experimental results of generalization ability

模　型	CVC-ColonDB 数据集			ETIS 数据集		
	mDice ↑	mIoU ↑	MAE ↓	mDice ↑	mIoU ↑	MAE ↓
U-Net	0.519	0.449	0.061	0.406	0.343	0.036
U-Net++	0.490	0.413	0.064	0.413	0.342	0.035
PraNet	0.716	0.645	0.043	0.630	0.576	0.031
TransUNet	0.717	0.645	0.044	0.573	0.512	0.033
SSFormer	0.784	0.702	0.038	0.742	0.667	0.021
Polyp-PVT	0.794	0.716	0.031	0.752	0.677	0.020
PVT-CAMNet	**0.819**	**0.735**	**0.027**	**0.765**	**0.703**	**0.020**

从表 3 中数据可以看出,在 CVC-ColonDB 数据集中,PVT-CAMNet 相较于最优基线模型 Polyp-PVT 的 mDice、mIoU 分别提高了 3.15%和 2.65%,MAE 降低了 12.9%。在 ETIS 数据集中,PVT-CAMNet 相较于最优基线模型 Polyp-PVT 的 mDice、mIoU 分别提高了 1.73%和 3.84%,MAE 与 Polyp-PVT 持平。两种模型的 mDice、mIoU 不同而 MAE 相同,这可能是因为:MAE 主要用来评估像素级精度,更关注整体误差,而 mDice 和 mIoU 主要评估两个集合之间的相似度,它们更关注分割结果中区域级别的准确性。由表 2 和表 3 中数据可计算出,Polyp-PVT 的 mDice、mIoU、MAE 在 4 个数据集上方差分别为 5.6×10^{-3}、7.7×10^{-3} 和 7.3×10^{-5};而 PVT-CAMNet 的 mDice、mIoU、MAE 在 4 个数据集上方差分别为 5.4×10^{-3}、6.4×10^{-3} 和 6.0×10^{-5}。由此可得出 PVT-CAMNet 相较于 Polyp-PVT 模型,3 个指标方差更小,更稳定。

本文所提出模型与其他 6 个模型在 CVC-ColonDB 和 ETIS 数据集上的分割实际效果如图 7 所示,可以看出本文提出的模型分割效果相较于其他 6 个模型都更接近真实标签图。进一步综合模型学习能力和泛化能力的全面测试结果来看,PVT-CAMNet 相对于其他 6 种模型在结肠息肉图像分割上具有明显优势。

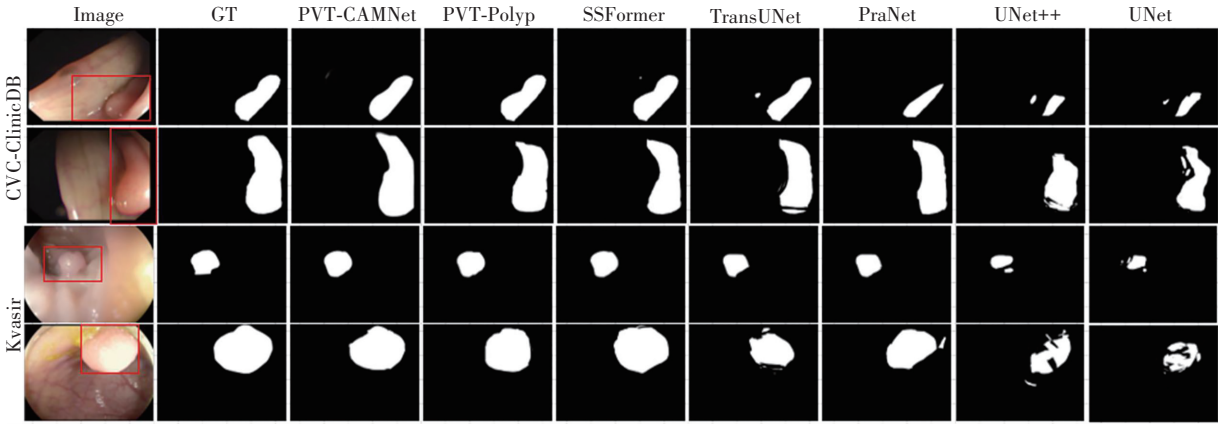

图 7　PVT-CAMNet 与对照模型在泛化能力实验中效果对比图

Fig. 7　Comparison of PVT-CAMNet and control model in generalization ability experiment

2.2.3　模型性能对比

为评估本文方法 PVT-CAMNet 的参数量和计算量,在 CVC-ClinicDB 数据集上进行了测试,测试结果如表 4 所示,其中浮点运算量(Giga Floating-point Operations Per second, GFLOPs)通常用来表示模型计算量。由表 4 数据可知,本文方法的参数量 Params 和浮点运算量 GFLOPs 都相对较少,说明本文方法能快速地完成训练和测试。相比于基于卷积神经网络结构的 U-Net++,虽然参数量增加较多,但是浮点运算量 GFLOPs 明显降低。相较于基于 Transformer 结构的 Polyp-PVT 网络,本文方法的参数量有所增加,但浮点运算量 GFLOPs 降低 24.95%。综合对比分析,本文方法 PVT-CAMNet 在兼顾模型参数量的同时,实现了最低计算量。

表 4　各模型训练时性能对比

Table 4　Performance comparison of each model during training

模　型	Params/10^7 ↓	GFLOPs/10^{10} ↓
U-Net	7.85	10.79
U-Net++	9.16	26.72
PraNet	30.50	13.11
TransUNet	105.5	60.75
SSFormer	29.31	20.18
Polyp-PVT	25.08	10.58
PVT-CAMNet	28.21	7.94

2.2.4 不同参数对实验结果影响

为了验证在实验中对 PVT-CAMNet 模型所设定的一些实验参数有效性,在 CVC-ClinicDB 和 Kvasir 数据集上进行了验证,结果如图 8、表 5 和表 6 所示,其中 l_r 为初始学习率,d 为衰减率 decay。

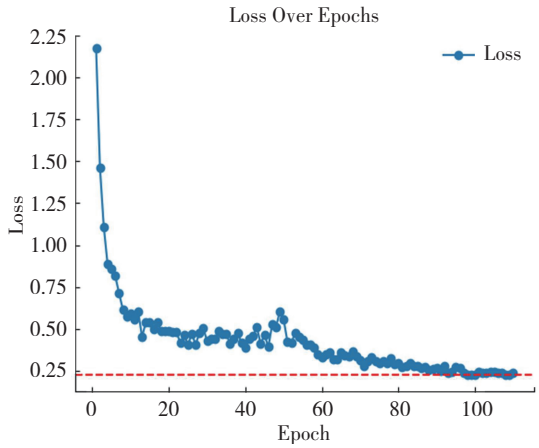

图 8 PVT-CAMNet 训练损失曲线

Fig. 8 PVT-CAMNet training loss curve

表 5 在 CVC-ClinicDB 数据集上验证不同参数对分割精度结果的影响

Table 5 Verification of the impact of different parameters on segmentation accuracy results on CVC-ClinicDB dataset

参 数	CVC-ClinicDB 数据集		
	mDice ↑	mIoU ↑	MAE ↓
$l_r=0.01$	0.935	0.872	0.009
$l_r=0.03$	0.933	0.870	0.010
$d=0.000\ 4$	0.930	0.878	0.009
$d=0.000\ 6$	0.920	0.871	0.010
PVT-CAMNet	**0.941**	**0.888**	**0.008**

表 6 在 Kvasir 数据集上验证不同参数对分割精度结果的影响

Table 6 Verification of the impact of different parameters on segmentation accuracy results on Kvasir dataset

参 数	Kvasir 数据集		
	mDice ↑	mIoU ↑	MAE ↓
$l_r=0.01$	0.916	0.863	0.028
$l_r=0.03$	0.906	0.854	0.031
$d=0.000\ 4$	0.912	0.859	0.028
$d=0.000\ 6$	0.924	0.866	0.027
PVT-CAMNet	**0.927**	**0.867**	**0.027**

图 8 展示了在训练时 PVT-CAMNet 模型损失下降趋势,由图 8 可知,当训练周期 epoch 为 100 时,损失已趋于收敛,故将 epoch 设置为 100。表 5 和表 6 则展示了在训练时,初始学习率 l_r 和衰减率 d 不同取值对模型分割精度的影响。由两表数据可知,对于初始学习率 l_r 无论取 0.01 或 0.03,相较于 PVT-CAMNet 中设定的初始学习率 l_r 为 0.02,分割精度都有所下降,衰减率 decay 同样如此。综合图 8、表 5 和表 6 中数据来看,模型训练周期 epoch、初始学习率 l_r 和衰减率 d 分别设置为 100,0.02,0.000 5,模型分割精度最优。

2.2.5 消融实验

为了验证在 PVT-CAMNet 模型所设计模块的有效性,做了两个关于单独模块的消融实验。实验一在 CVC-ClinicDB 数据集上进行学习能力测试的消融实验,结果如表 7 所示;实验二在 CVC-ClinicDB 和 Kasir-SEG 数据集上训练,在 ETIS 数据集上进行泛化能力测试的消融实验,结果如表 8 所示。表中“W/”代表缺失对应模块。当缺失 PVT 编码器时,以 ResNet50 代替。

表 7 在 CVC-ClinicDB 数据集上的消融实验结果

Table 7 Ablation experiment results on CVC-ClinicDB dataset

模 型	mdice ↑	mIoU ↑	MAE ↓
W/PVT	0.915	0.869	0.011
W/MFAE	0.933	0.880	0.009
W/IA	0.921	0.873	0.010
W/GC	0.937	0.882	0.009
PVT-CAMNet	**0.941**	**0.888**	**0.008**

表 8 在 ETIS 数据集上的消融实验结果

Table 8 Ablation experiment results on ETIS dataset

模 型	mdice ↑	mIoU ↑	MAE ↓
W/PVT	0.630	0.565	0.042
W/MFAE	0.750	0.699	0.029
W/IA	0.744	0.685	0.031
W/GC	0.758	0.692	0.028
PVT-CAMNet	**0.765**	**0.703**	**0.020**

通过分析表 7 和表 8 的消融实验结果可以得出:

(1) PVT 编码器使模型在 CVC-ClinicDB 和 ETIS 数据集上 mDice 和 mIoU 分别提升了 3.50%、2.19% 和

21.4%、24.4%,而 MAE 则降低了 27.3%和 52.4%。

(2) MFAE 模块使模型在 CVC-ClinicDB 和 ETIS 数据集上 mDice 和 mIoU 分别提升了 0.86%、0.91%和 2.00%、0.57%,而 MAE 则降低了 11.1%和 31.03%。

(3) IA 模块使模型在 CVC-ClinicDB 和 ETIS 数据集上 mDice 和 mIoU 分别提升了 2.17%、1.72%和 2.82%、2.63%,而 MAE 则分别降低了 20.0%和 28.57%。

(4) GC 模块使模型在 CVC-ClinicDB 和 ETIS 数据集上 mDice 和 mIoU 分别提升了 0.43%、0.68%和 0.92%、1.59%,而 MAE 则分别降低了 11.1%和 28.57%。

图 9 展示了消融实验的实际效果图,可以看出当缺失任何一个模块后,模型的实际分割效果都明显下降。通过消融实验结果来看,4 个模块各自有效,进一步提高了模型的学习和泛化能力。

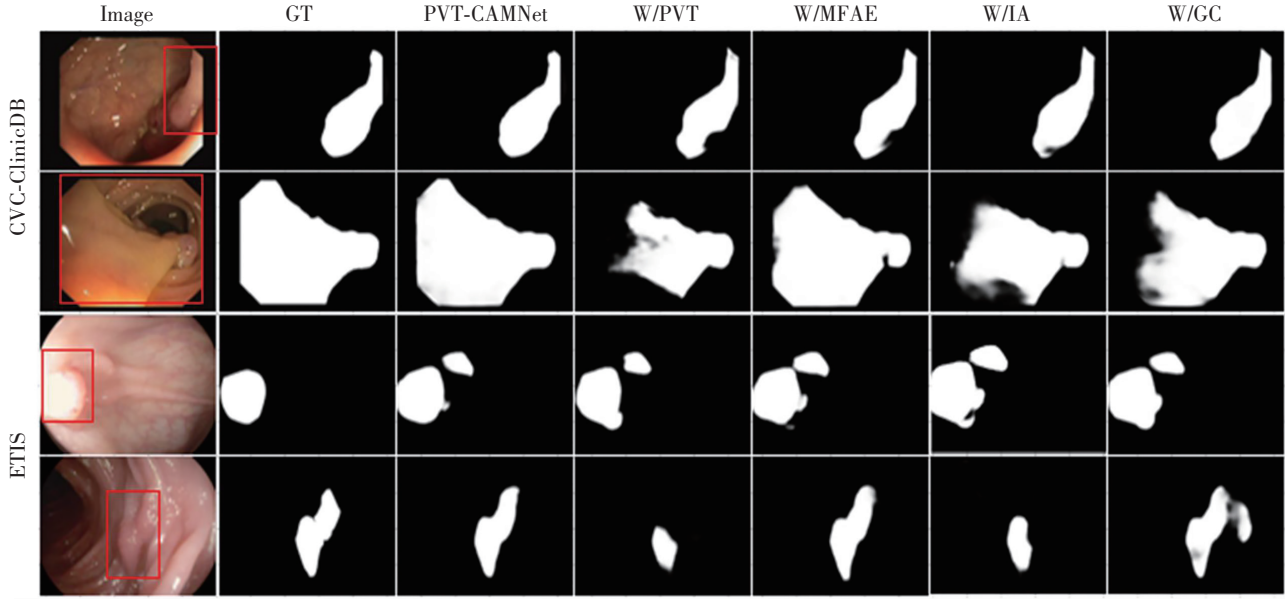

图 9　消融实验实际分割效果图

Fig. 9　Actual segmentation results of ablation experiment

各个模块的优势如下:PVT 通过其自注意力捕获息肉图像远距离依赖,能够提供更多的全局上下文信息,相较于卷积神经网络,极大地增加了网络的泛化能力;MFAE 通过组合 4 种不同大小的滤波器和空间通道混合注意力对编码器提取的特征进一步整合,缓解了模型的注意力分散问题;IA 模块使相邻级别特征进行跨层交互,不仅有利于捕捉息肉的局部细节,而且极大地缓解了特征融合时产生的信息丢失问题;GC 模块通过计算输入特征图中任意两个位置之间的关系来直接捕获远程依赖,从而使模型更好地理解全局上下文。

3　结　论

本文提出以 PVT 作为编码器的级联注意力模型 PVT-CAMNet 用于结肠息肉图像分割。模型首先以 PVT 编码器对结肠息肉图像进行粗细粒度特征提取,接着利用 MFAE 模块通过组合多尺度卷积与空间通道混合注意力,使模型更加关注相关语义信息,缓解注意力分散问题;其次设计了 IA 模块,自底向上,交互融合不同层级特征,在缓解多级特征融合时产生信息丢失问题的同时,也有助于捕捉息肉的形态和纹理信息;然后经过 GC 模块获得跨多个特征级别的高阶互补信息生成增强互补特征;最后经过逐层上采样输出分割图,在公开的 4 个息肉数据集上取得了相较于其他 6 个最具代表性的基线模型更优异的结果。综合来看,通过级联多种注意力机制,PVT-CAMNet 有效地缓解了因 Transformer 在提取特征时存在注意力分散以及在特征融合时信息丢失的问题,并且具有强大的学习能力与泛化能力,该模型结构层次分明。

在今后的研究工作中可以尝试引入更多有效的模块进一步提升分割精度;其次可以尝试将 PVT-CAMNet 应用到其他医学图像分割领域,如肝脏肿瘤分割,皮肤病变分割等;最后,由于使用了多个功能模块,PVT-CAMNet 的参数量也随之增大,因此针对模型轻量化进行改进也是未来的工作方向之一。

参考文献(References):

[1] 王凯悦, 蔡善荣, 张苏展. 不同筛查手段对结直肠腺瘤检出的研究进展[J]. 实用肿瘤杂志, 2022, 37(3): 283-291.
WANG Kai-yue, CAI Shan-rong, ZHANG Su-zhan. Research progress on detection of colorectal adenoma by different screening methods[J]. Journal of Practical Oncology, 2022, 37(3): 283-291.

[2] PUYAL J G B, BHATIA K K, BRANDAO P, et al. Medical Image Computing and Computer Assisted Intervention-MICCAI[M]. Cham: Springer International Publishing, 2020: 295-305.

[3] ALJABRI M, ALGHAMDI M. A review on the use of deep learning for medical images segmentation[J]. Neurocomputing, 2022, 506(1): 311-335.

[4] 李季, 胡锦萍, 乔敏, 等. 一种针对脑部图像分割强度不均匀性的改进方法[J]. 重庆工商大学学报(自然科学版), 2023, 40(1): 34-39.
LI Ji, HUJin-ping, QIAO Min, et al. An improved method for intensity inhomogeneity of brain image segmentation[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(1): 34-39.

[5] RONNEBERGER O, FISCHER P, BROX T. Lecture Notes in Computer Science[M]. Cham: Springer International Publishing, 2015: 234-241.

[6] ZHOU Z, RAHMAN SIDDIQUEE M M, TAJBAKHSH N, et al. UNet++: A nested U-net architecture for medical image segmentation[C]//International Workshop on Deep Learning in Medical Image Analysis, International Workshop on Multimodal Learning for Clinical Decision Support. Cham: Springer, 2018: 3-11.

[7] FAN D P, JI G P, ZHOU T, et al. PraNet: Parallel reverse attention network for polyp segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2020: 263-273.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. 2017: arXiv: 1706.03762. http://arxiv.org/abs/1706.03762.

[9] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale[EB/OL]. 2020: arXiv: 2010.11929. http://arxiv.org/abs/2010.11929.

[10] CHEN J, LU Y, YU Q, et al. TransUNet: Transformers make strong encoders for medical image segmentation[EB/OL]. 2021: arXiv: 2102.04306. http://arxiv.org/abs/2102.04306.

[11] WANG J, HUANG Q, TANG F, et al. Stepwise feature fusion: Local guides global[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2022: 110-120.

[12] DONG B, WANG W, FAN D P, et al. Polyp-PVT: Polyp segmentation with pyramid vision transformers[J]. CAAI Artificial Intelligence Research, 2023: 9150015.

[13] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the Computer Vision-ECCV 2018: 15th European Conference. Munich, Germany, 2018: 3-19.

[14] DAS N, DAS S. Attention-UNet architectures with pretrained backbones for multi-class cardiac MR image segmentation[J]. Current Problems in Cardiology, 2024, 49(1 Pt C): 102129.

[15] CAO Y, XU J, LIN S, et al. GCNet: Non-local networks meet squeeze-excitation networks and beyond[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop. 2019: 1971-1980.

[16] ZHAO X, ZHANG L, LU H. Medical Image Computing and Computer Assisted Intervention-MICCAI[M]. Cham: Springer International Publishing, 2021: 120-130.

[17] SANDERSON E, MATUSZEWSKI B J. FCN-transformer feature fusion for polyp segmentation[C]//Annual Conference on Medical Image Understanding and Analysis. Cham: Springer, 2022: 892-907.

[18] ZHANG Y, LIU H, HU Q. TransFuse: Fusing transformers and CNNs for medical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2021: 14-24.

[19] LOU A, GUAN S, KO H, et al. CaraNet: Context axial reverse attention network for segmentation of small medical objects[C]//Medical Imaging 2022: Image Processing. San Diego, California, 2022, 12032: 81-92.

[20] SILVA J, HISTACE A, ROMAIN O, et al. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer[J]. International Journal of Computer Assisted Radiology and Surgery, 2014, 9(2): 283-293.

责任编辑:陈 芳