基于 R-DCAformer 的结直肠息肉分割模型

高艾国,郑晓亮
安徽理工大学 电气与信息工程学院, 安徽 淮南 232001

摘 要:目的 现有 Transformer 模型虽然在形态复杂的结直肠息肉分割中拥有较高准确率,但是其注意力分散,编码器输出多级语义信息在融合中会产生信息丢失,限制了模型准确率进一步提高,针对此问题,提出一种新的肠道息肉图像分割模型:双通管道聚合网络(Dual-Channel Aggregation Transformer, R-DCAformer)。**方法** R-DCAformer 模型使用金字塔混合的 Transformer(Mix Transformer, MIT)和 Resnet18 充当编码器,设计了双通道聚合(Dual-Channel Aggregation, DCA)模块充当解码器。DCA 解码器由注意力聚合模块(Attention Aggregation, AA)和双通道特征聚合模块(Dual-Channel Feature Fusion,DFF)组成,其中,金字塔 MIT 编码器可以为模型提供充足泛化能力,AA 模块可以通过融合 Resnet18 的额外特征限制模型 MIT 中的注意力分散,DFF 模块则可以缓解多级语义信息融合中的信息丢失问题。**结果** 泛化能力实验中,R-DCAformer 在 CVC-ColonDB 中相比于基线模型中最优的 mDice、mIoU 和 MAE 分别提高了 2.10%、1.65%和 22.5%,在 ETIS 中,相比于基线模型中最优的 mDice、mIoU 和 MAE 分别提高了 2.56%、2.12%和 15%;模型在 CVC-ClinicDB 数据集上,相比于基线模型中的最优 mDice、mIoU 提高了约 0.85% 、1.35%;在 Kvasir-SEG 数据集上,相比于基线模型中的最优 mDice、mIoU 和 MAE 提高了约 1.19% 、1.97%和 17.39%。此外还通过消融实验和注意力图论证了本文所提出模块的有效性。**结论** R-DCAformer 在学习和泛化实验中效果都较为优异,总体上优于对比的基线模型,为结直肠息肉分割提供了新的高性能模型。

关键词:息肉图像分割;深度学习;双通道聚合;注意力聚合;泛化能力
中图分类号:TP393A **文献标识码:**A doi:10.16055/j.issn.1672-058X.2024.0005.006

Colorectal Polyp Segmentation Model Based on R-DCAfomer

GAO Aiguo, ZHENG Xiaoliang
School of Electrical and Information Engineering, Anhui University of Science & Technology, Anhui Huainan 232001, China

Abstract: Objective Although the existing Transformer model has high accuracy in segmenting colorectal polyps with complex morphology, the distraction of the Transformer model and the loss of information in the fusion of its encoder outputting multilevel semantic information limit the further improvement of the model' s accuracy. Based on this, a novel image segmentation model (the Dual-Channel Aggregation Transformer, R-DCAformer) for intestinal polyps was proposed. **Methods** The R-DCAformer model used a pyramid mix Transformer (MIT) and Resnet18 to act as an encoder and a dual-channel aggregation (DCA) module was designed to act as a decoder. The DCA decoder consisted of an attention aggregation (AA) module and a dual-channel feature fusion (DFF) module. In this model, the pyramid MIT encoder provided sufficient generalization ability for the model, the AA module limited the distraction in the model MIT by fusing the additional features of Resnet18, and the DFF module alleviated the problem of information loss in the fusion of multi-

收稿日期:2023-03-05 **修回日期:**2023-05-18 **文章编号:**1672-058X(2024)05-0049-09
基金项目:煤炭安全精准开采国家地方联合工程研究中心开放基金资助(EC2021006);安徽理工大学高层次引进人才科研启动基金资助(2021YJRC02).
作者简介:高艾国(1997—),男,安徽合肥人,硕士研究生,从事模式识别与信息处理自动化研究.
通讯作者:郑晓亮(1979—),教授,博士研究生导师,从事安全监测与监控技术方向的研究. Email:zhengxl@aust.edu.cn.
引用格式:高艾国,郑晓亮. 基于 R-DCAformer 的结直肠息肉分割模型[J]. 重庆工商大学学报(自然科学版),2024,41(5):49—57.
GAO Aiguo, ZHENG Xiaoliang. Colorectal polyp segmentation model based on R-DCAfomer[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2024, 41(5): 49—57.

level semantic information. **Results** In the generalization ability experiment, R-DCAformer improved the optimal mDice, mIoU, and MAE by 2.10%, 1.65%, and 22.5%, respectively, in CVC-ColonDB compared with the optimal ones in the baseline model. The optimal mDice, mIoU, and MAE in ETIS were improved by 2.56%, 2.12%, and 15%, respectively, compared with the optimal ones in the baseline model. The model improved the optimal mDice and mIoU by about 0.85% and 1.35% in the CVC-ClinicDB dataset compared with the optimal ones in the baseline model, and the optimal mDice, mIoU, and MAE on the Kvasir-SEG dataset were improved by about 1.19%, 1.97%, and 17.39%, respectively, compared with those in the baseline model. The effectiveness of the module proposed in this paper was also demonstrated by ablation experiments and attention graphs. **Conclusion** R-DCAformer is more effective in both learning and generalization experiments, and generally outperforms the compared baseline models, providing a new high-performance model for colorectal polyp segmentation.

Keywords: polyp image segmentation; deep learning; dual-channel aggregation; attention aggregation; generalization ability

1 引 言

癌症是人类健康的主要威胁之一,而结直肠癌则是其中一种常发疾病。文献[1]表明定期接受临床医生的结肠镜检查可以帮助患者及时发现息肉,从而进行结肠息肉早期切除,有效降低了罹患大肠癌的风险。然而,文献[2]表明息肉的大小和形状不同,以及息肉和黏膜之间的界限不明确,人工进行准确息肉图像分割将花费大量人力物力,且难以在短时间内完成大量分割,给医护人员增添很多负担。为此,研究者开始使用传统算法解决上述问题,如 GROSS 等[2]将多尺度滤波用于息肉图像分割,虽然可以完成息肉图像的简单分割,但效果很不精确;BERNAL 等[3]使用扇形堆砌能量图算法,可以大致实现对息肉区域分割,但分割失误率较高。

随着深度学习的兴起,研究者使用基于深度学习的分割算法,有效提高了息肉图像分割的精确度。例如,研究者使用 U-Net[4-5] 网络进行息肉分割取得较好的分割效果,但是其网络的特征融合方式过于直接,导致融合过程中信息损失很多。后续,又有研究者提出了 U-Net++[6] 网络,增加了模型中的跳跃连接数量,使得模型能提取多种层次的特征,减小低级和高级特征的信息鸿沟。后续的研究者又进一步提出了 Pranet[7] 网络,通过结构复杂的卷积模块来增强其模型提取特征效果。

随着深度学习的不断发展,Transformer 结构逐渐进入研究者的视野。Transformer[8] 最初是为自然语言处理任务提出来的一个自底向上的模型架构。最近,Dosovitskiy 等[9]提出了视觉变换器(Vision Transformer, VIT),在图像分类任务中取得了很好的效果。文献[10]表明 Transformer 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,使模型具有有效的全局接收场,并减少了模型的学习偏差。因此,Transformer 比 CNN 和多层感知机结构拥有更强大的泛化能力。后期为了进一步提高深度学习模型的泛化能力,TransUNet[11]、TransFuse[12] 和 Polyp-PVT[13] 使用金字塔型的 Transformer 作为模型编码器,使模型可以获得更多维度的特征信息,提高了 Transformer 类模型的分割精度。

然而,传统方法虽然可以实现息肉图像的粗略分割,但分割结果却不尽如人意。基于卷积神经网络(Convolutional Neural Network, CNN)的图像分割模型[5-7]虽在几个息肉图像分割测试中取得了较好的效果,但由于 CNN 模型是自顶向下的建模方法,并且息肉形态的多样性和测试数据集提供的息肉图像数量相对较少,导致这类模型缺乏泛化能力,难以精确地分割由不同结肠镜设备得到的图片。基于 Transformer 的模型[11-13]随着 Transformer 结构的深化,整体特征不断混合和收敛,往往会导致注意分散。此外,金字塔 Transformer 输出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,导致边界模糊。

前人的启发:通过卷积操作可以缓解注意力分散问题[14];全局空间特征有助于大型目标的定位,而局部空间特征对于识别小型目标至关重要[9]。为解决基于 Transformer 模型的注意力分散问题和多级语义信息在融合过程中产生的信息丢失问题,本文提出了一种新的医学图像分割框架:R-DCAformer。其中,设计了一种有效的解码器(Dual-Channel Aggregation, DCA),DCA 解码器包括了注意力聚合模块(Attention Aggregation, AA)和双通道特征聚合模块(Dual-channel feature fusion, DFF)。AA 模块可以通过融合 Resnet18 的额外特征以限制模型编码器(Mix Transformer, MIT[16])中的注意力分

散,得到注意力集中后的各级特征;DFF 模块可以更好地融合各级特征以减少融合过程中的信息损失;最后,将两种通道获取的全局信息和局部信息通过预测层得到最终的结果。

本文的主要贡献:

(1) 提出一种新的医学图像分割框架:R-DCAformer,引入 MIT 作为编码器,为模型提供泛化能力;使用 Resnet18 作为副编码器,提供额外的注意力聚合信息。

(2) 设计一种适配金字塔 MIT 的 DCA 解码器。它是由注意力聚合 AA 模块和双通道特征聚合模块 DFF 模块组成,以缓解常见 Transformer 模型的注意力分散问题和多级信息融合导致的信息丢失问题。

(3) 在 4 个息肉数据集 ETIS[11]、CVC-ClinicDB[12]、CVC-ColonDB[13] 和 Kvasir-SEG[14] 上的大量实验表明:所提出的模型拥有很好的学习能力和泛化能力。

2　方　法

2.1　DCAformer 模型结构

如图 1 所示,本文提出了一种基于金字塔 MIT 的肠道息肉图像分割模型:R-DCAformer。其由 MIT 编码器和 DCA 解码器组成,其中 DCA 解码器包含 AA 模块和 DFF 模块。原始图片由编码器 MIT 分级处理后输出多级特征信息 $\{F_{ti} \mid i \in (1,2,3,4)\}$,这些特征信息先进入 AA 模块与 Resnet18 产生的补充特征信息 $\{F_{ri} \mid i \in (1,2,3,4)\}$ 进行注意力聚合,得到聚合后的多级特征信息 $\{F_{le,i} \mid i \in (1,2,3,4)\}$;然后将多级特征信息输入到 DFF 模块,得到全局和局部信息。具体过程:将所聚合完成的高级语义信息 $\{F_{le,i} \mid i \in (1,2,3,4)\}$ 输入两个特征聚合模块(Stepwise Feature Aggregation,SFA)进行逐级变换、聚合得到全局信息 $F_{2,3}$,即($F_{le,4}$,$F_{le,3}$)先经过第 4 级 SFA 融合得到 $F_{3,4}$,再将($F_{3,4}$,$F_{le,2}$)经过第 3 级 SFA 融合得到全局信息 $F_{2,3}$,此过程中,高级语义特征逐步聚合得到全局特征,缓解了多级信息差异,使模型将各级特征直接变换到最大分辨率并且直接聚合过程中产生的信息丢失问题;通过特征选择(Feature Selection,FS)模块选择性地将高级语义指导信息 F_s 与低级语义特征 $F_{le,1}$ 融合得到局部信息 Z,使局部特征简明有序,可以更好地完成局部特征的学习,增强图像的边界特征;最后将两种通道获取的全局信息 $F_{2,3}$ 和局部信息 Z 通过预测层最终得到结果。

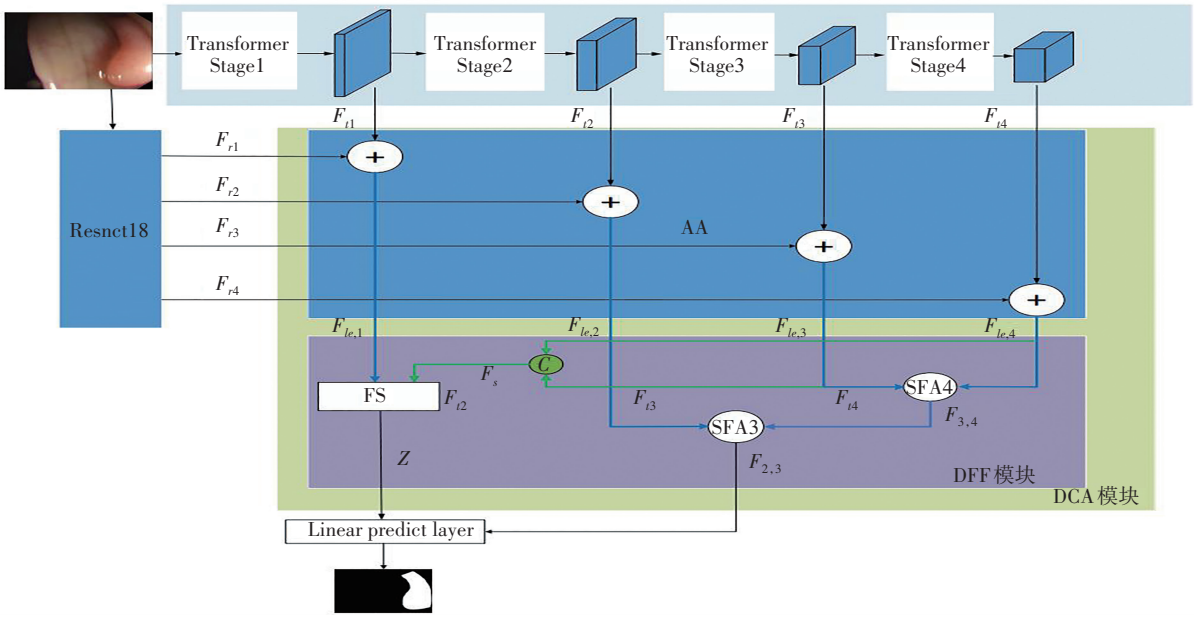

图 1　R-DCAfomer 模型结构图

Fig. 1　Structure of R-DCAfomer model

2.2　MIT 编码器

Transformer 不同于卷积神经网络是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,并减少学习偏差。因此,为了使模型具有足够的泛化能力来进行息肉分割,引入了基于金字塔的 MIT 代替 CNN 作为编码器。金字塔 MIT 相较于常规 VIT[9],有以下 4 点改进:

(1) 金字塔结构。与只能生成单分辨率特征的 VIT 不同,本模块在给定输入图像的情况下,可以生成类似 CNN 的多级特征 $\{F_i \mid i \in (1,2,3,4)\}$。这些特征

包括了目标的低级语义信息 F_1 和高级语义信息 F_2、F_3 和 F_4,从而提供给解码器图片的局部细节信息和完整的全局信息。

(2) 重叠特征提取(Overlapped Patch Merging)。输入图像后,在 VIT 提取特征的过程中,不会重叠提取图像中的像素块。因此,它无法保持这些提取后的特征与其周围特征的连续性,为此,MIT 的提取过程将重叠地提取图像像素块。

(3) 高效的自我注意力机制(Efficient Self-Attention)。VIT 作为编码器的主要计算难点在于使用了大量的自注意层。为此,MIT 使用序列缩减过程来使自注意力机制变得高效。使用序列缩减后,MIT 自我注意力机制的复杂度可以从 $O(N^2)$ 降低到 $O\left(\frac{N^2}{R}\right)$。

(4) 混合的前馈网络(Mix-FFN)。在 VIT 中,当测试分辨率与训练分辨率不同时,需要对位置信息进行插值,这往往会导致精度下降。由实验知道引入 CNN 可以缓解这个问题[14],因此,在 MIT 的前馈网络引入了 CNN,即在前馈网络中直接添加 3×3 的卷积,并引入多层感知机结构来抵消 CNN 中零填充对于位置信息的影响。

MIT 编码器中 Transformer stage 结构如图 2 所示。上级特征先输入重叠特征提取模块,得到具有周围连续特性的信息,再将信息经多头高效的自我注意力机制得到由注意力筛选后的特征,最后将其输入混合的前馈网络得到带有精确位置信息的特征,从而完成一个 Transformer stage。按照图 1 所示,金字塔 MIT 作为编码器经历了 4 次 Transformer stage,对图像进行多级特征 $\{F_i \mid i \in (1,2,3,4)\}$ 提取。在这些特征中,F_1 提供了目标的低级语义信息,其中包含大量细节信息,F_2、F_3 和 F_4 提供了高级语义信息,包含了完整的全局信息。

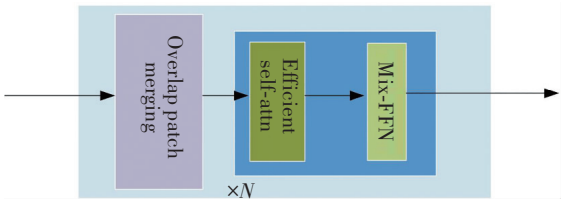

图 2 Transformer stage 模块结构图

Fig. 2 Structure of Transformer stage module

2.3 Resnet18 副编码器

Resnet18 是基于 CNN 的图像任务主干模块,其在内核中提取信息进行权重参数训练,这可以使 Resnet18 具有有效的局部接收场。因此,为了使 MIT 编码器的注意力分散得到缓解,引入 Resnet18 进行特征补充,使得模型更加精准地进行息肉分割。Resnet18 由大量的卷积层和残差层构成,当输入图片尺寸为 3 * 224 * 224 时,具体结构如图 3 所示。

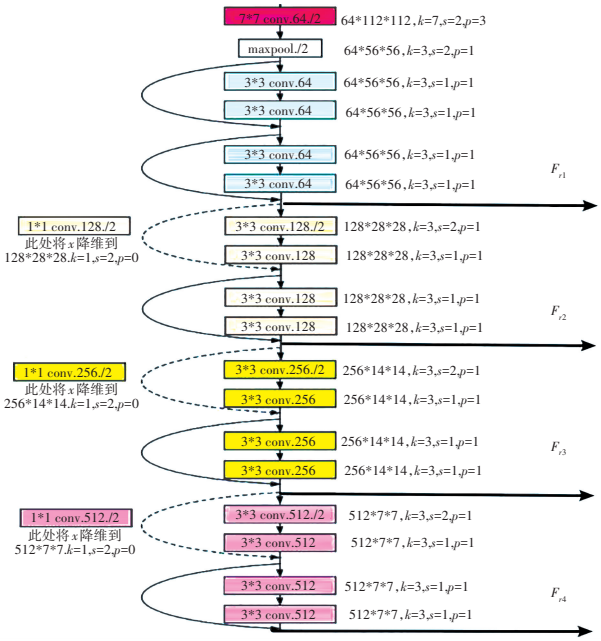

图 3 Resnet18 结构图

Fig. 3 Structure of Resnet18 module

其中,3×3 C 是一个 3×3 卷积,1×1 C 是一个是池化层,3×3 CX 中的 X 代表输出通道数。

2.4 DCA 解码器

2.4.1 AA 模块

使用金字塔 MIT 处理图像的过程中,会进行大量的自我注意力操作。图像在大量的自我注意力操作后,通常会导致特征注意力分散[18]。由于自注意力机制中的注意矩阵可以被视为全局非预设卷积核,因此,在本模型中,使用具有局部感受野卷积操作的 Resnet18 的输出特征补充 MIT 输出特征来设计 AA 模块,从而将注意力重新集中在相邻的有限元分析上,减少注意力分散。具体过程如图 1 所示:模块将 MIT 产生的特征 $\{F_{ti} \mid i \in (1,2,3,4)\}$ 和 Resnet18 的输出特征 $\{F_{ri} \mid i \in (1,2,3,4)\}$ 一一对应聚合,从而将注意力分散问题缓解,其计算推导式为式(1):

$$F_{le,i} = F_{ti} \oplus F_{ri} \tag{1}$$

其中,F_{ri} 是从编码器 MIT 输入 AA 模块的各级特征流;F_{ti} 是从编码器 MIT 输入 AA 模块的各级特征流;$F_{le,i}$ 则是由 AA 模块输出的各级特征流;⊕是聚合操作,具体过程是通过特征通道维度拼接后再通过一个 1×1 卷积恢复特征原来的通道数进行融合,其中 $F_{le,i}$ 对应的 1×1 卷积参数分别为 $(128\times i^i, 64\times i^i, 1, 1)$。

2.4.2 DFF 模块

全局空间特征有助于大型目标的定位,而局部空间特征对于识别小型目标至关重要。为此,设计 DFF

模块,可以更好地融合各级特征,以减少融合过程中的特征损失。在 DFF 模块中,全局信息通过两个 SFA 模块逐级聚合高级语义特征获取,以解决多级语义信息差距在融合过程中产生的信息丢失问题;由于低级语义信息的无规则性和冗余性,因此局部信息通过 FS 模块将低级语义信息和高级语义指导信息选择性地融合来获取,这可以使局部信息更加简练有序。

(1) SFA 模块。对以金字塔 Transformer 为骨干网络的模型进行多层次信息交互融合往往比使用 CNN 进行聚合对模型的效果提升更为显著[22],并且金字塔 MIT 输出的各级特征具有深度差异,这些差异会使模型将各级特征直接变换到最大分辨率,并且在直接聚合的过程中产生信息丢失现象[15]。为此,本模型采用从上到下逐级变换特征尺寸再分次融合的想法来设计全局信息聚合过程,以获取更好的全局特征。如图 1 所示:高级语义信息经历两次 SFA 模块聚合得到全局信息。SFA 模块,如图 4 所示,由上采样操作、特征融合单元和线性融合层组成。其中,第 4 级 SFA 过程:经过 AA 模块输出的特征 $F_{le,4}$ 通过上采样操作将分辨率变换为同 $F_{le,3}$ 一致,通过通道维度上的串联操作将两个特征流聚合,最后再通过 *Linear* 线性映射得到聚合后的特征流 $F_{3,4}$。第 3 级 SFA 过程具体:$F_{3,4}$ 通过上采样操作将分辨率变换为同 $F_{le,2}$ 一致,再通过通道维度上的串联操作将两个特征流聚合,最后再通过 *Linear* 线性映射得到聚合后的全局信息 $F_{2,3}$。其中第 4 级和第 3 级的 SFA 融合的计算推导式为式(2)和式(3):

$$F_{3,4}=Linear(Concat(F_{le,3},(F_{le,4})))\tag{2}$$

$$F_{2,3}=Linear(Concat(F_{le,2},Up(F_{3,4})))\tag{3}$$

其中 *Linear* 线性聚合由核大小为 1 × 1 的卷积运算充当,目的是保持通道数的统一,并使特征融合。SFA4 对应的 *linear* 参数为(768,256,1,1),SFA3 对应的 *linear* 参数为(384,128,1,1);*Concat* 是通道维度上的串联操作;$F_{le,i}$ 是从编码器 MIT 中的输出后再经过 LE 输出的第 i 级特征流;$F_{i-1,i}$ 是由 SFA 聚合 $F_{le,i}$ 和 $F_{le,i-1}$ 后输出的特征流;*Up* 是上采样操作,其中 SFA4 和 SFA3 中 *Up* 的参数 scale_factor 均为 2,其他参数为默认。

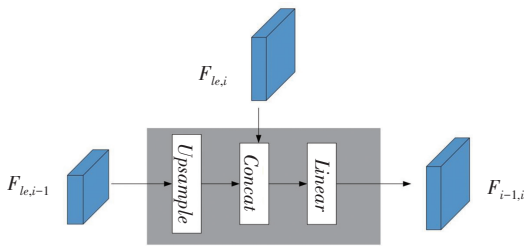

图 4　SFA 模块结构图

Fig. 4　Structure of SFA module

(2) FS 模块。Transformer 模型学习局部特征的能力直接影响其性能[23]。但是,现有的 Transformer 模型往往缺乏充分学习这些局部特征的能力,这使得模型很难更好地去完成对于息肉图片的分割。此外,使用低层特征直接提取局部信息会导致信息混乱和冗余。为了缓解这个问题,设计了特征选择(Feature Selection,FS)模块,选择性地融合低级语义信息与高级语义指导信息,使局部特征简明有序,可以更好地完成局部特征的学习,增强图像的边界特征。设计高级语义指导信息聚合模块来提供高级语义指导信息 F_s,具体过程:AA 模块输出的特征 $F_{le,4}$ 通过上采样操作将分辨率变换为同 $F_{le,3}$ 一致,通过通道维度上的串联操作将两个特征流聚合,然后经过激活函数,再通过 *Linear* 线性映射得到聚合后的高级语义指导信息 F_s,最后通过上采样操作将分辨率变换为与 $F_{le,1}$ 一致,其计算表达式为式(4)。

如图 5 所示,FS 模块在局部特征与全局特征融合前自适应地从两个输入($F_{le,1}$,F_s)中选择语义信息,FS 模块提取局部信息的具体过程:低级语义信息和高级语义指导信息输入到两个自适应校准(Adaptive Calibration,AC)块中,通过使用高级语义指导信息优化低级语义信息,使输出的局部信息更加简洁有序。然后,用两个 AC 模块的输出进行 3×3 卷积后串联,得到局部特征,计算表达式为式(5)。其中,AC 模块聚合信息具体过程:将输入特征 T_1 和 T_2 经过 W_θ 和 W_φ 线性映射后,分别通过点积操作与 T_1 和 T_2 结合,再分别与 T_1 相加和 T'_1 反操作点积,最终相加合并。这种聚合策略实现了不同特征的鲁棒组合,通过使用高级语义指导信息重新优化杂乱的低级语义信息特征,使得所包含的局部信息更加简洁有序,从而更好地利用低级信息特征图进行后续结果的细节优化。其中计算过程如式(7)所示。

$$F_s=Up(Linear(Concat(F_{le,3},Up(F_{le,4}))))\tag{4}$$

$$Z=C_{3\times3}(Concat(AC(F_s,F_{le,1}),AC(F_{le,1},F_s)\tag{5}$$

$$T'_1=W_\theta(T_1),T'_2=W_\varphi(T_2)\tag{6}$$

$$AC(T_1,T_2)=T'_1\odot T_1+T'_2\odot T_2\odot(-(T'_1))+T_1\tag{7}$$

其中,*Linear* 线性映射采用核大小为 1×1 的卷积运算充当,F_s 中 *Linear* 对应参数为(768,64,1,1);*Concat* 是通道维度上的串联操作;*Up* 是上采样操作,求取 F_S 中的 F_l 对应的 scale_factor 参数值为 2,最外层的 *Up* 对应的为 4;$C_{3\times3}$是一个 3×3 卷积,并包含正则化和 ReLU 激活函数层,其中求取 Z 中的 3×3 卷积的参数为(128,64,3,3,1);$F_{le,i}$ 是从编码器 MIT 中的输出后再经过 AA 模块输出后的第 i 级特征流;Z 是 FS 模块输出的局部特

征;T_1、T_2 为输入特征;将两个线性映射 W_θ、W_φ 应用于输入特征;精选特征 T'_1 和 T'_2 由 T_1、T_2 通过 W_θ 和 W_φ 运算获取,其过程可以表述为式(6);⊙是点乘操作;Θ 是通过取反 T'_1 进行的反向操作;⊕是合并操作。

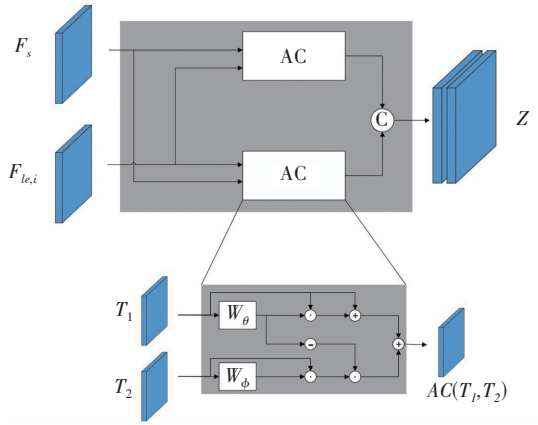

图 5 FS 模块结构图

Fig. 5 Structure of FS moduel

2.4.3 Linear predict layer 模块

通过一个 Linear predict layer 模块来进行最终转化,得到所需结果。其公式如式(8):

$$F_s = Up(C_{3\times3}(Concat(Z, Up(F_{2,3})))) \tag{8}$$

其中,Up 是上采样操作,$F_{2,3}$ 对应的 scale_factor 参数值为 2,最外层 Up 的参数 scale_factor 值为 4;$C_{3\times3}$ 是一个 3×3 卷积,并包含正则化和 ReLU 激活函数层,其中 3×3 卷积的参数为(128,1,3,3,1);$Concat$ 是通道维度上的串联操作。

3 实验与结果分析

3.1 实验设置

3.1.1 评估指标

在实验中使用 mDice、mIoU 和 MAE 作为评估指标。mDice、mIoU 分数值越高,分割精度越高。MAE 的分数越低,分割精度越高。其公式分别为式(9)、式(10)和式(11):

$$V_{\text{mDice}} = \frac{2\times T_P}{2\times T_P + F_P + F_N} \tag{9}$$

$$V_{\text{mIoU}} = \frac{T_P}{T_P + F_P + F_N} \tag{10}$$

$$V_{\text{MAE}} = \frac{1}{m}\sum_{i=1}^{m} |y_i - y'_i| \tag{11}$$

其中,T_P 表示分割结果中正确分类为前景的样本数,T_N 显示分割结果中正确分类为背景像素的样本数,F_P 表示拆分结果中被错误分类为前景的样本数,F_N 表示分割结果中被错误分类为背景像素的样本数,y_i 为原来的样本标签值;y'_i 为模型输出结果,V_{mDice}、V_{mIoU} 和 V_{MAE} 代表 mDice、mIoU 和 MAE 的值。

3.1.2 实验环境和模型超参数

本文在 PyTorch 中实现模型,使用 NVIDIA TESLA V100 GPU 训练模型,使用 AdamW 优化器作为模型优化器,模型的 batch_size 为 32,模型的初始学习率 learning_rate 为 0.000 1,衰减率 learing rate decay 为 0.1,衰减周期为 20 个周期,每次训练 200 个周期,训练模型的损失函数是 Dice 和 BCE 的组合损失函数。在训练期间,将图像大小调整为 3×224×224,并且在训练过程随机使用翻转、缩放、旋转、膨胀和侵蚀对训练数据进行增强,使模型更具有鲁棒性。

3.1.3 实验数据集

本文在 4 个息肉分割数据集上进行实验(ETIS、CVC-ClinicDB、CVC-ColonDB、Kvasir-SEG)。在测试模型学习能力实验中,将来自 Kvasir-SEG 和 CVC-ClinicDB 的图像随机分成 80%用于训练,10%用于验证,10%用于测试。在测试模型泛化能力的实验中,使用 Kvasir-SEG 和 CVC-ClinicDB 数据集的 80%进行训练并在 ETIS 和 CVC-ColonDB 数据集上进行测试,得到的结果体现在未知数据中的效果。各个数据集的详细信息如表 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	图片大小
CVC-ClinicDB	612	384×288
Kasir-SEG	1 000	不固定
ETIS	196	1 225×966
CVC-ColonDB	380	574×966

3.2 结果与分析

为了验证本文模型的分割性能,将本文模型 R-DCAfomer 与目前息肉分割模型中最具代表性的模型进行对比,包括 U-Net 、U-Net++、PranNet 、TransUNet、UCTransNet 和 Polyp-PVT。

3.2.1 学习能力

在测试模型学习能力实验中,将 CVC-ClinicDB 及 Kvasir-SEG 基准数据集分为 80% 训练集,10% 评估集及 10% 测试集。在 Kvasir-SEG 数据集和 CVC-ClinicDB 数据集上的分割性能指标如表 2 所示,表 2 中加粗表示此项指标的最优值。表 2 表明:模型在 CVC-ClinicDB 数据集上,相比于基线模型中的最优 mDice、mIoU 提高了约 0.85% 、1.35%,在 Kvasir-SEG 数据集上,相比于基线模型中的最优 mDice、mIoU 和 MAE 提高了约 1.19% 、1.97%和 17.39%。R-DCAfomer 模型在 Kasir-SEG 数据集上的效果如图 6 所示,可以看出在不同数据集中本模型的分割效果都更为接近标

签图。

由对比可以得出:R-DCAfomer 模型拥有强大的学习能力,相较于其他模型,可以拥有更强的结直肠息肉图片分割能力。

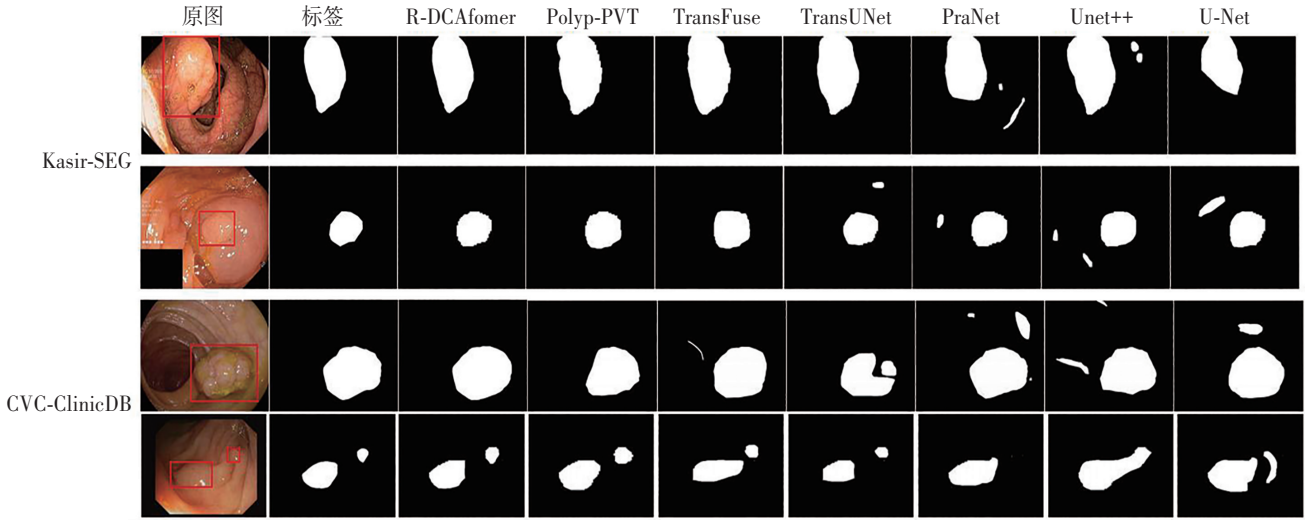

图 6　本文模型与对照模型在学习能力实验中效果对比图

Fig. 6　Comparison of the effects of the model in this article and the control model in the learning ability experiment

表 2　学习能力实验结果

Table 2　Results of learning ability experiment

数据集	CVC-ClinicDB			Kasir-SEG		
模型	mDice	mIoU	MAE	mDice	mIoU	MAE
U-Net	0.823	0.755	0.019	0.818	0.746	0.055
Unet++	0.794	0.729	0.022	0.821	0.743	0.048
PraNet	0.899	0.849	0.009	0.898	0.840	0.030
TransUNet	0.935	0.887	0.008	0.913	0.857	0.028
UCTransNet	0.933	0.860	0.008	0.918	0.960	0.023
Polyp-PVT	0.937	0.889	**0.006**	0.917	0.864	0.023
R-DCAfomer	**0.949**	**0.911**	**0.006**	**0.931**	**0.881**	**0.017**

3.2.2　泛化能力

为了评估 R-DCAfomer 的泛化能力,本文从 Kasir-SEG 和 CVC ClinicDB 基准数据集中随机提取 1 450 幅图像以构建训练集(为了公平性评估,使用与 PraNet 相同的训练集),然后测试模型在 CVC-ColonDB 和 ETIS 数据集上的性能。该测试可以证明模型在未知数据集中的准确预测和泛化能力。表 3 中加粗表示此项指标的最优值。表 3 中的结果表明:模型在 CVC-ColonDB 中,相比于基线模型中最优的 mDice、mIoU 和 MAE 分别提高了 2.10%、1.65%和 22.5%,在 ETIS 中,相比于基线模型中最优的 mDice、mIoU 和 MAE 分别提高了 2.56%、2.12%和 15%,本文模型与对照模型在泛化实验中效果对比如图 7 所示,可以看出在不同数据集中本模型的分割效果都更接近标签图。

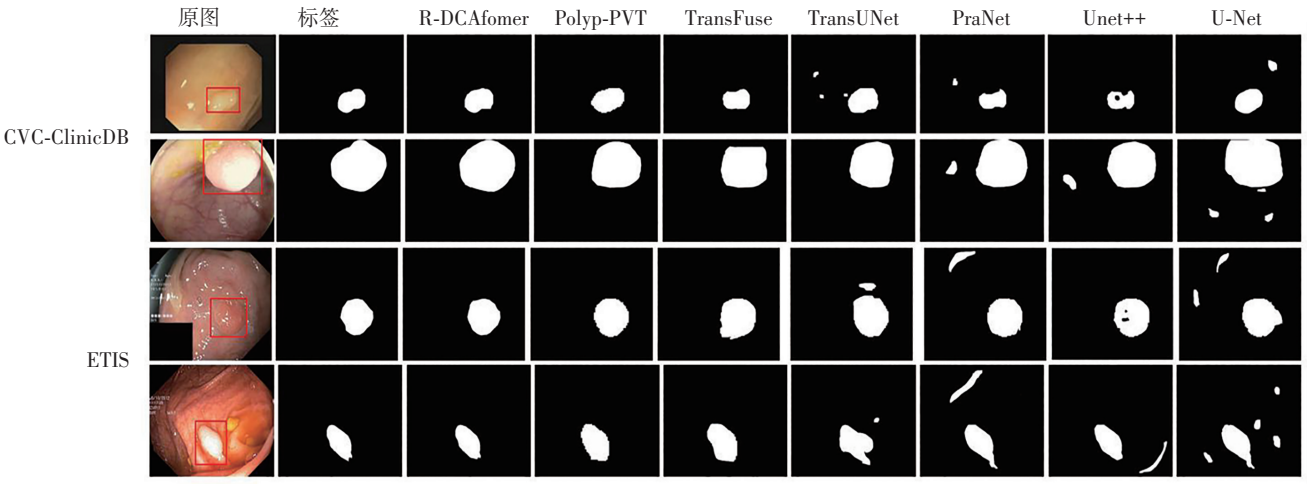

图 7　本文模型与对照模型在泛化实验中效果对比图

Fig. 7　Comparison of the effects of the model in this article and the control model in the generalization experiment

表 3 泛化能力实验结果

Table 3 Results of generalization ability experiment

训练数据集	CVC-ClinicDB& Kasir-SEG					
测试数据集	CVC-ColonDB			ETIS		
模型	mDice	mIoU	MAE	mDice	mIoU	MAE
U-Net	0.512	0.444	0.061	0.398	0.335	0.036
Unet++	0.483	0.410	0.064	0.401	0.344	0.035
PraNet	0.712	0.640	0.043	0.628	0.567	0.031
TransUNet	0.781	0.699	0.036	0.731	0.624	0.021
TransFuse	0.781	0.706	0.035	0.737	0.626	0.020
Polyp-PVT	0.808	0.727	0.031	0.787	0.706	0.013
R-DCAfomer	**0.825**	**0.739**	**0.024**	**0.815**	**0.721**	**0.011**

由不同模型在测试数据集与训练数据集不同的实验结果对比可以得出:R-DCAfomer 模型相较于 CNN 模型,泛化能力有很大进步,对比其他基于 Tranformer 的模型,泛化能力也有一定提升。

3.2.3 消融研究

为验证 R-DCAfomer 模型中各模块对于效果的作用,进行了 2 次实验。

实验 1:在 CVC-ClinicDB 数据集上进行测试学习能力的消融实验。

实验 2:在 CVC-ClinicDB & Kasir-SEG 训练集中,在 CVC-ColonDB 数据集上进行测试泛化能力的消融实验。如表 4 所示,W/AA 表示缺失 AA 模块和 Resnet18 编码器;W/SFA 表示缺失全局信息聚合过程,即使用直接聚合代替两次 SFA 过程;W/FS 表示缺失局部信息聚合过程,即使用直接聚合代替 FS 模块过程,表 4 中加粗为最优指标。

表 4 消融实验实验结果

Table 4 Experimental results of ablation experiment

训练数据集	CVC-ClinicDB			CVC-ClinicDB & Kasir-SEG		
测试数据集	CVC-ClinicDB			CVC-ColonDB		
模型	mDice	mIoU	MAE	mDice	mIoU	MAE
W/AA	0.935	0.887	0.013	0.793	0.698	0.037
W/SFA	0.933	0.887	0.010	0.805	0.719	0.031
W/FS	0.939	0.893	0.009	0.799	0.706	0.035
R-DCAfomer	**0.945**	**0.901**	**0.007**	**0.823**	**0.741**	**0.023**

(1) 对比表 4 中第一行和第四行的实验结果表明:AA 模块在实验 1 中能够提升 1.07% 的 mDice,1.58% 的 mIoU 和 46.13% 的 MAE;实验 2 中能够提升 2.77% 的 mDice,4.87% 的 mIoU 和 21.61% 的 MAE,验证了 AA 模块可以提高模型的学习能力和泛化能力。

(2) 对比表 4 中第二行和第四行的实验结果表明:FS 模块在实验 1 中能够提升 1.50% 的 mDice,1.69% 的 mIoU 和 36.32% 的 MAE;实验 2 中,能够提升 1.75% 的 mDice,2.37% 的 mIoU 和 6.45% 的 MAE,表明 FS 模块能有效地学习局部信息,提高模型的学习能力和泛化能力。

(3) 对比第三行和第四行的实验结果表明:FS 模块在实验 1 中能够提升 0.85% 的 mDice,1.29% 的 mIoU 和 22.13% 的 MAE;实验 2 中能够提升 2.52% 的 mDice,4.42% 的 mIoU 和 17.14% 的 MAE,表明 FS 模块能提高模型的学习能力和泛化能力。

如图 8 所示,模型的注意力热力图可以展示出本模型中的学习权重在图片上的分布情况,由蓝色到红色注意力值逐渐变大。图 8 中可以发现:

(1) MIT 输出特征 $\{F_{ti} \mid i \in (1,2,3,4)\}$ 通过 AA 模块后得到特征 $\{F_{le,i} \mid i \in (1,2,3,4)\}$,其注意力从发散变得集中起来,验证了 AA 模块可以通过融合 Resnet18 提供的额外特征缓解注意力分散问题。

(2) $\{F_{le,i} \mid i \in (1,2,3,4)\}$ 通过 DFF 模块后,模型的特征进一步聚合得到全局信息 F_{23} 和局部信息 Z,观察 F_{23} 和 Z 可以看出,模型的注意力进一步朝关键区域聚合,证明了本模型的 DFF 有效聚合了各个尺度的特征,解决了多级语义信息差距在融合过程中产生的信息丢失问题和低级语义信息的无规则性和冗余性问题。

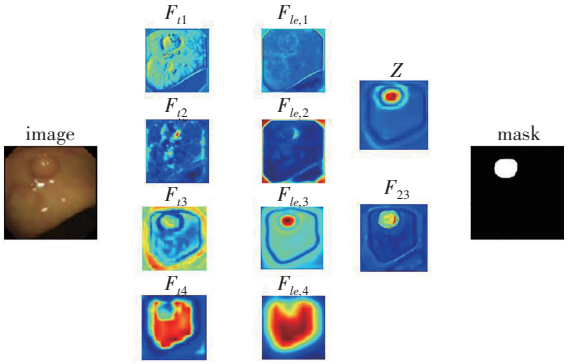

图 8 本模型各个层注意力热力图

Fig. 8 Thermodynamic diagram of attention for each layer of this model

4 结论与展望

4.1 结 论

针对现有模型泛化能力不足、注意力分散和细节模糊问题,提出了一种新的肠道息肉图像分割模型:R-DCAfomer,融合了金字塔 MIT 编码器、Resnet18 编码器和双通道聚合 DCA 解码器,其中 DCA 包括注意力聚合模块 AA 和双通道特征聚合模块 DFF。MIT 编码器提升模型泛化性;Resnet18 编码器和 AA 模块可以限制金字塔 Transformer 骨干注意力分散;DFF 模块可以更好融合图像的语义信息,减少多级语义信息差距在融合过程中产生的问题,缓解细节模糊,增强图像分割能力。结果表明:R-DCAfomer 具有较强的学习能力和泛化能力,定性和定量的结果都表明了 R-DCAfomer 总体

优于其他竞争方法,为结直肠息肉分割任务提供了新的高效模型。

4.2 展　望

对于 R-DCAformer 模型,希望这项研究能为解决医疗图像分割任务提供更多的思路,并在未来的工作中可以使用此模型进行其他医疗图像的分割和分类任务。

参考文献(References):

[1] TAEHUN K, HYEMIN L, DAIJIN K. Uacanet: Uncertainty augmented context attention for polyp segmentation[C]// In Proceedings of the 29th ACM International Conference on Multimedia. NewYork: ACM, 2021: 2167—2175.

[2] GROSS S, KENNEL M T, et al. Polyp segmentation in NBI colonoscopy[C]// Bildverarbeitung Furdie Medizin Heidelberg Germany. Heidelberg: Springer, 2009: 252—256.

[3] BERNALA J, SANCHEZ J, VILARINOA F, et al. Towards automatic polyp detection with a polyp appearance model[J]. Pattern Recognition, 2012, 45(9): 3166—3182.

[4] SRIVASTAVA A, JHAD A, CHANDA S, et al. MSRF-Net: A multi-scale residual fusion network for biomedical image segmentation [J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(5): 2252—2263.

[5] RONN JERGER O, FISCHER P, BROX T. Unet: Convolutional networks for biomedical image segmentation[C]//In International Conference on Medical Image Computing and Computer-assisted Intervention. Cham: Springer International Publishing. 2015: 234—241.

[6] JHA D, SMEDSRUD P A, RIEGLER M A, et al. Resunet++: An advanced architecture for medical image segmentation[C]// Proceedings of the IEEE International Symposium on Multimedia (ISM). Piscataway: IEEE Press, 2019.

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

[8] VASWANI A, SHAZEER N M, PARMAR N. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30(5): 6000—6010.

[9] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]//International Conference on Learning Representations. IEEE, 2021: 663—673.

[10] NASEER M, RANASINGHE K, KHAN S H, et al. Intriguing properties of vision transformers [J]. Advances in Neural Information Processing Systems, 2021, 34(2341): 23296—23308.

[11] CHEN J E, LU Y Y, YU Q H, et al. Transunet: Transformers make strong encoders for medical image segmentation[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 2125—2134.

[12] ZHANG Y D, LIU H Y, 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.

[13] DONG B, WNAG W H, FAN D P, et al. Polyp-PVT: Polyp segmentation with pyramid vision transformers [J]. CAAI Artificial Intelligence Research. 2022, 26(5). 2252—2263.

[14] WANG J, HUANG Q, TANG F, et al. Stepwise feature fusion: Local guides global[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference. NewYork: ACM, 2022: 110—120.

[15] ZHANG Z, ZHANG X, PENG C, et al. Exfuse: Enhancing feature fusion for semantic segmentation[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2018: 269—284.

[16] XIE E Z, WNAG W H, YU Z D, et al. Segformer: Simple and efficient design for semantic segmentation with transformers[J]. Advances in Neural Information Processing Systems, 2021, 34(12): 99—111.

[17] DAVID VAZQUEZ, JORGE BERNAL, F JAVIER SANCHEZ, et al. A benchmark for endoluminal scene segmentation of colonoscopy images[J]. Journal of Healthcare Engineering, 2017, 45(9): 107—129.

[18] 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.

[19] BERNAL J, ANCHEZ F L, ESPARRACH G, et al. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians[J]. Computerized Medical Imaging and Graphics, 2015, 43(12): 99—111.

[20] JHA D, SMEDSRUD P H, RIEGLER M A, et al. Kvasir-seg: A segmented polyp dataset [C]// In International Conference on Multimedia Modeling. Cham: Springer, 2020: 451—462.

[21] ZHOU D Q, KANG B Y, JIN X J, et al. Deepvit: Towards deeper vision transformer [C]//2021 IEEE Globecom Workshops. IEEE, 2021: 1—6.

[22] RAGHU M, SUNTERTHINER T, KORNBLITH S, et al. Do vision transformers see like convolutional neural networks? [J]. Advances in Neural Information Processing Systems, 2021, 34 (301): 3026—3035.

[23] TANG F L, HUANG Q M, WNAG J F, et al. DuAT: Dual-aggregation transformer network for medical image segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 6881—6890.

责任编辑:李翠薇