文章编号　1004-924X(2023)18-2700-13

跨尺度跨维度的自适应 Transformer 网络应用于结直肠息肉分割

梁礼明，何安军，李仁杰，吴　健*

(江西理工大学 电气工程及其自动化学院，江西 赣州 341000)

摘要：针对结直肠息肉图像病灶区域尺度变化大、边界模糊、形状不规则且与正常组织对比度低等问题，导致边缘细节信息丢失和病灶区域误分割，提出一种跨尺度跨维度的自适应 Transformer 分割网络。该网络一是利用 Transformer 编码器建模输入图像的全局上下文信息，多尺度分析结直肠息肉病灶区域。二是通过通道注意力桥和空间注意力桥减少通道维度冗余和增强模型空间感知能力，抑制背景噪声。三是采用多尺度密集并行解码模块来填补各层跨尺度特征信息之间的语义空白，有效聚合多尺度上下文特征。四是设计面向边缘细节的多尺度预测模块，以可学习的方式引导网络去纠正边界错误预测分类。在 CVC-ClinicDB、Kvasir-SEG、CVC-ColonDB 和 ETIS 数据集上进行实验，其 Dice 相似性系数分别为 0.942，0.932，0.811 和 0.805，平均交并比分别为 0.896，0.883，0.731 和 0.729，其分割性能优于现有方法。仿真实验表明，本文方法能有效改善结直肠息肉病灶区域误分割，具有较高的分割精度，为结直肠息肉诊断提供新窗口。

关　键　词：结直肠息肉；Transformer；多尺度密集并行解码模块；多尺度预测模块

中图分类号：TP391.4　　**文献标识码**：A　　**doi**：10.37188/OPE.20233118.2700

Cross-scale and cross-dimensional adaptive transformer network for colorectal polyp segmentation

LIANG Liming，HE Anjun，LI Renjie，WU Jian*

(*School of Electrical Engineering and Automation，Jiangxi University of Science and Technology，Ganzhou* 341000，*China*)

* *Corresponding author，E-mail：wujian@jxust. edu. cn*

Abstract：To address the problem of large-scale variation，blurred boundaries，irregular shapes，and low contrast with normal tissues in colon polyp images，which leads to the loss of edge detail information and mis-segmentation of lesion areas，we propose a cross-dimensional and cross-scale adaptive transformer segmentation network. First，the network uses transformer encoders to model the global contextual information of the input image and analyze the colon polyp lesion areas at multiple scales. Second，the channel attention and spatial attention bridges are used to reduce channel dimension redundancy and enhance the model′s spatial perception ability while suppressing background noise. Third，the multi-scale dense parallel decoding module is used to bridge the semantic gaps between cross-scale feature information at different layers，effectively aggregating multi-scale contextual features. Fourth，a multi-scale prediction module is de-

收稿日期：2023-03-15；**修订日期**：2023-04-15.

基金项目：国家自然科学基金资助项目(No. 51365017，No. 61463018)；江西省自然科学基金面上项目资助(No. 20192BAB205084)；江西省教育厅科学技术研究重点项目资助(No. GJJ170491，No. GJJ2200848)

signed for edge details, guiding the network to correct boundary errors in a learnable manner. The experimental results conducted on the CVC-ClinicDB, Kvasir-SEG, CVC-ColonDB, and ETIS datasets showed that the Dice similarity coefficients are 0.942, 0.932, 0.811, and 0.805, and the average intersection-over-union ratios are 0.896, 0.883, 0.731, and 0.729, respectively. The segmentation performance of our proposed method was better than that of existing methods. The simulation experiment showed that our method can effectively improve the mis-segmentation of colon polyp lesion areas and achieve high segmentation accuracy, providing a new approach for colon polyp diagnosis.

Key words: colcorectal polyps; transformer; multi-scale dense parallel decoding module; multi-scale prediction module

1　引　言

结直肠癌是世界上最常见的高发癌症之一，具有极高的死亡率[1]。研究报告数据显示，早期结直肠腺瘤息肉可以通过简单手术予以切除，生存率高达95%。也就是说，早发现和准确诊断结直肠腺瘤息肉是有效降低死亡率的关键。但由于不同时期的结直肠息肉大小不一，形状尺度变化大、边界模糊且存在正常组织与病变区域相似度高等复杂特性[2]，使得结直肠癌息肉分割面临众多挑战。

为了解决结直肠息肉难以准确分割的挑战，许多学者提出传统分割方法和基于深度学习的方法[3]，传统方法主要依赖于人为选取特征，以区域生长，阈值图像和统计形状为主[4-6]。受到结直肠息肉图像病变区域复杂特性的影响，传统方法存在较大的局限性。随着深度学习的普及，卷积神经网络（Convolutional Neural Network，CNN）以强大的特征提取能力在图像处理领域开辟了新范式[7]。基于CNN结构的U型网络在生物医学语义分割任务中被广泛运用。U-Net[8]采用对称的编码-解码器结构，并在编码-解码器之间引入跳跃连接，使网络可以很好地将深层语义信息和浅层细粒度信息进行融合。由于卷积操作仅进行局部运算，难以建立远距离特征依赖关系，所以U-Net网络结构在一定程度上仍有较大的改进空间。Jha等[9]采用空洞空间金字塔池化（Atrous Spatial Pyramid Pooling, ASPP）来扩大模型感受野，嵌入SE(Squeeze and Excitation)[10]注意力机制促进通道特征信息之间的依赖关系，有效改善了息肉的分割精度，但在边缘细节上处理并不是很好。Fan等[11]使用并行部分解码器进行特征聚合，采用反向注意力模块构建区域和边界线索之间的显性关系，优化了分割结果边缘，但对小目标分割还存在漏检现象。Lou等[12]提出CaraNet，设计上下文轴向注意力模块和通道级特征金字塔模块来提高模型对小目标的分割性能，但整体流程较为复杂。

Transformer[13]结构通过捕获长距离依赖显性关系，建模全局上下文信息，在医学图像分析领域获得了出色的表现。Dai等[14]结合CNN和Transformer的优点，首次将Transformer结构应用于多模块医学图像分类任务，获得了较好的分类效果。Chen等[15]提出TransUNet，设计双边融合模块来处理不同分支的语义信息，改善了多器官和心脏图像分割任务的精度，但该结构带来了大量的浮点运算和参数，影响模型的实际应用。Wang等[16]采用金字塔视觉Transformer提取输入图像的语义信息，提出渐进式解码器来强调局部特征和强化目标信息，提升了息肉分割精度，但在未知数据集上的泛化性能较差。Wu等[17]结合Swin Transformer结构分成多阶段产生多个不同尺度特征，构建多尺度通道注意力机制和空间反向注意力机制，以提升网络学习和提取结直肠息肉各种形态特征的能力，进一步优化结直肠息肉分割结果，但该结构需要预训练权重才能发挥其效果，导致网络结构不能灵活调节。

针对结直肠息肉分割当前面临的技术挑战，本文提出一种跨尺度跨维度的自适应Transformer分割模型。该模型首先在金字塔视觉Transformer（Pyramid Vison transformer v2, PVTv2）基础上融合CNN结构，不仅能有效提取全局上下文信息，还增强了网络对局部特征的解析能力。然后设计多尺度密集并行解码模块来补充浅层网络细粒度信息和深层网络语义信息

之间的语义空白。最后引入多尺度预测模块以调整不同阶段预测结果，逐步细化息肉区域的边缘结构信息。

2　方　法

2.1　研究基础

卷积神经网络(CNN)通过与图像周围像素进行点运算提高模型的局部感知能力，具有平移不变性和归纳偏差，但受到卷积核大小、网络层数以及计算资源的限制，导致捕捉全局上下文特征信息的能力不足。而Transformer结构中的多头注意力机制能有效建立短距离和远距离的特征间距，使网络能够从全局角度解析语义信息。结直肠息肉病灶区域与正常肠黏膜对比度低、边界模糊，有效处理全局语义信息和局部细节信息能对分割精度提升带来很大的帮助。最近，基于Transformer结构方法在多种视觉任务上取得了与基于CNN结构相当的性能，PVTv2[18]表现得相当出色，其核心Transformer编码模块如图1所示。图1中z_{i-1}表示第i个Transformer编码模块经过卷积前馈层和残差连接后的输出特征图；z_i'表示经线性空间自注意力层和残差连接后的输出特征图。

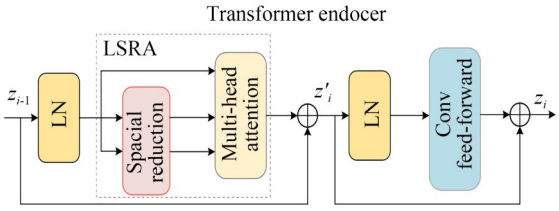

图1　核心Transformer编码模块

Fig. 1　Core Transformer encoding block

每个Transformer编码模块均由层归一化(Layer Normalization，LN)、线性空间自注意力层(Linear Spatial-reduction Attention Layer，LSRA)、卷积前馈层(Conv Feed-Forward，CFF)和残差连接组成，卷积前馈层由全连接层、高斯误差线性单元(Gaussion Error Linear Unit，GELU)和3×3的深度可分离卷积组成。其中，LSRA通过重塑图像结构，缩短远距离特征间距，从而增强网络捕获全局语义信息的能力。LSRA接受三个相同维度的特征向量，分别是查询矩阵$\boldsymbol{Q}$，键矩阵$\boldsymbol{V}$和值矩阵$\boldsymbol{K}$。相比于传统的多头注意力层，LSRA采用平均池化操作降低矩阵$\boldsymbol{K}$和$\boldsymbol{V}$的空间尺寸，在很大程度减少了计算开销。具体的计算过程如下：

$$\mathrm{LSRA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}(\mathrm{head}_0,\cdots,\mathrm{head}_{N_i})\boldsymbol{W}^o, \tag{1}$$

$$\mathrm{head}_j=\mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}_j^Q,LSR(\boldsymbol{K})\boldsymbol{W}_j^K,LSR(\boldsymbol{V})\boldsymbol{W}_j^V), \tag{2}$$

其中：Concat(•)表示连接操作；$\boldsymbol{W}_j^Q\in\boldsymbol{R}^{C_i\times d_{\mathrm{head}}}$，$\boldsymbol{W}_j^K\in\boldsymbol{R}^{C_i\times d_{\mathrm{head}}}$，$\boldsymbol{W}_j^V\in\boldsymbol{R}^{C_i\times d_{\mathrm{head}}}$，$\boldsymbol{W}^o\in\boldsymbol{R}^{C_i\times C_i}$为线性投影参数；$N_i$表示第$i$阶段注意力层的头数，即每个头部的尺度($d_{\mathrm{head}}$)为$\frac{C_i}{N_i}$；$C_i$表示第$i$阶段输出通道数；$LSR(\cdot)$是降低输入序列($\boldsymbol{K}$或$\boldsymbol{V}$)空间维度操作，其计算式为$LSR(x)=\mathrm{Norm}(\mathrm{Reshape}(x,P)\boldsymbol{W}^S)$，其中，$x\in\boldsymbol{R}^{(H_iW_i)\times C_i}$表示输入序列；$P$为线性池化大小，设置为7；$\mathrm{Reshape}(x,P)$是将输入序列$x$重塑大小为$P^2\times C_i$，$\boldsymbol{W}^S\in\boldsymbol{R}^{C_i\times C_i}$是一种线性投影；Norm(•)值归一化层。

自注意力层使用位置编码方式来确定图像上下文信息，其输出结果是输出图像尺寸保持不变。当训练图像尺寸与测试图像尺寸不一致时，需要对图像进行插值操作来保持统一尺寸，而插值操作容易造成局部细节信息的丢失，导致分割结果精度下降。受到文献[19]的启发，为了改善零填充对位置编码泄露的影响，在LSRA后引入卷积前馈层。卷积前馈层计算式为：

$$x_{\mathrm{out}}=\mathrm{DWConv}_{3\times3}(\mathrm{MLP}(\mathrm{Norm}(x_{\mathrm{in}}))), \tag{3}$$

$$x'_{\mathrm{out}}=\mathrm{MLP}(\mathrm{GELU}(x_{\mathrm{out}}))+x_{\mathrm{in}}, \tag{4}$$

其中：MLP(•)表示多层感知器；GELU(•)表示激活函数；$\mathrm{DWConv}_{3\times3}(\cdot)$表示3×3的深度可分离卷积；$x_{\mathrm{in}}$表示自注意力层的输出。

2.2　总体结构

为了减少空间归纳偏差和增强网络对上下文特征信息的有效表示，针对结直肠息肉的特点，本文提出一种跨尺度跨维度的自适应Transformer网络(Cross-Scale and Cross-Dimensional Adaptive Transformer Network，SDAFormer)应用于结直肠息肉分割，如图2所示。其网络结构主要包括编码器、混合注意力机制、多尺度密集

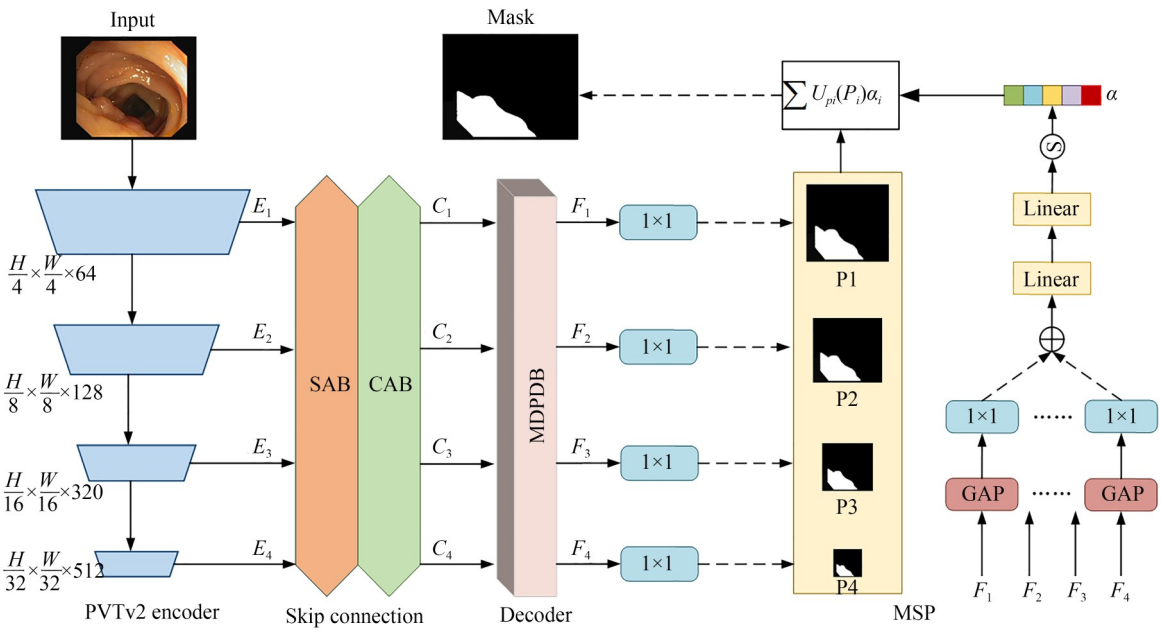

图 2　跨尺度跨维度的自适应 Transformer 网络

Fig. 2　Cross-scale and cross-dimensional adaptive transformer network

并行解码（Multi-scale Dense Parallel Decoding，MDPD）模块和多尺度预测（Multi-Scale Prediction，MSP）模块。其中，编码器采用在 ImageNet 数据集上预训练的 PVTv2 网络模型，逐层提取结直肠息肉图像的空间信息和语义信息，输出多尺度特征图；混合注意力机制通过强调特征图的病灶部分，有效建立不同阶段，不同特征图之间的通道信息联系，以抑制背景噪声影响并为目标区域分配合适的学习权重；多尺度密集并行解码模块用于深层网络信息和浅层网络信息的互补，融合空间信息和语义信息；多尺度预测模块以可学习的方式来获取一组权重系数，对权重系数进行自适应加权加法来纠正预测错误分类结果。

2.3　混合注意力机制

由于结直肠息肉图像中病灶区域的局部特征具有相关性，简单地融合浅层细粒度信息和深层语义信息，容易引入背景噪声等其他无关信息。为了自动调整不同特征之间的依赖关系，对重要特征施加合适的学习权重，在网络跳跃连接处引入空间注意力桥（Spatial Attention Bridge，SAB）模块和通道注意力桥（Channel Attention Bridge，CAB）模块[20]。相比于 CBAM[21]，SAB 更加轻量级，能够在计算资源有限的情况下提供较好的注意力增益。CAB 相对于 SAB 更加重量级，但在特征图的通道数较多时能够提供更好的注意力增益。

2.3.1　空间注意力桥模块

使用空间注意力桥模块来充分利用编码器不同阶段、不同尺度的特征信息，聚焦特征图中病灶部分，抑制背景噪声。算法伪代码（算法 1）表示为：

首先将来自编码器四个阶段的输出特征图分别进行全局平均池化和全局最大池化，平均池化对结直肠息肉病灶区域进行去噪，最大池化用

Algorithm 1：Spatial attention bridge block

Inputs：The input maps of the four channel attention bridge block C_i，i=1，2，3，4

Outputs：S_i，i=1，2，3，4

1：χ^i_{mean}=AvgPool(C_i) /*avg-pooling*/

2：χ^i_{max}=MaxPool(C_i)/*max-pooling*/

3：χ^i_s=Concat(h^i_{mean}，h^i_{max})/*Concatenate the feature map odd*/

4：α=Conv$_{7\times 7}$(h_c)/*7×7 convolution operation*/

5：$\varepsilon=\sigma(\beta)$/*After sigmoid，the feature map become$C\times H\times 1$*/

6：$S_i=\varepsilon * C_i+C_i$/*The feature map of sigmoid with the original feature and then add */

End

于凸显图像待分割的目标区域;然后拼接同一阶段池化后的特征图,将拼接后的特征图输送到权值共享的扩展卷积运算中(扩展卷积运算由一个卷积核大小为 7,扩张率为 3,填充量为 9 的卷积组成),使用 Sigmoid 函数生成空间注意力图,最后将生成的空间注意图以逐元素的方式与原始输入特征图相乘,并引入残差结构。图 3 为空间注意力桥模块示意图。

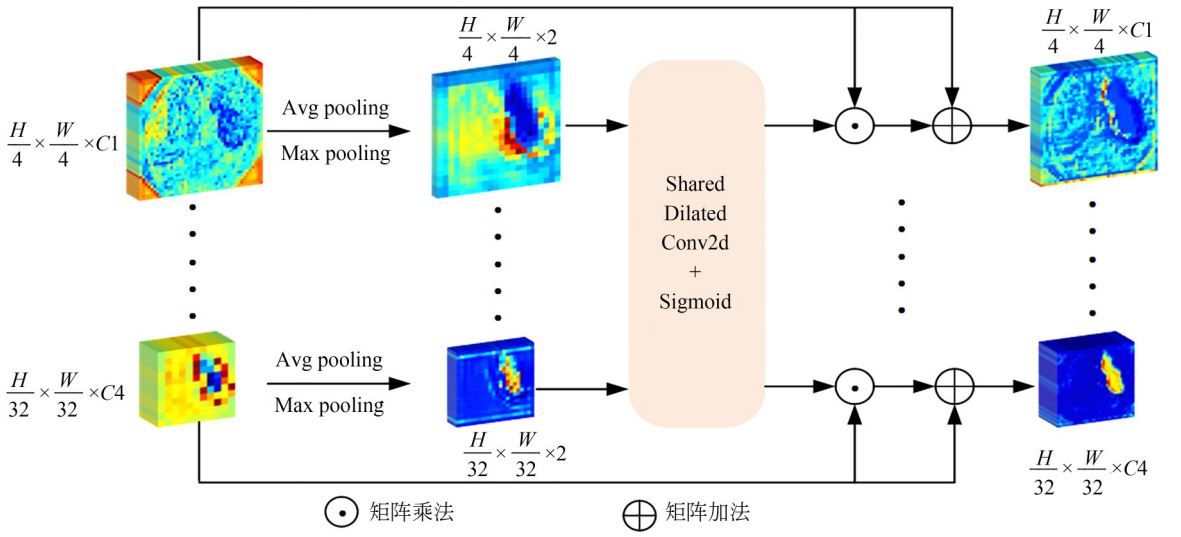

图 3　空间注意力桥模块

Fig. 3　Spatial attention bridge block

空间注意力桥模块的具体表示为:

$$\chi_i = \mathrm{Concate}(\mathrm{Avg}(E_i), \mathrm{Max}(E_i)), \quad (5)$$

$$S_i = E_i + E_i \times \sigma(\mathrm{Conv}_{7d}(\chi_i)), \quad (6)$$

其中:Avg 表示全局平均池化;Max 表示全局最大池化;Conv_{7d} 表示 7×7 卷积;E_i 表示第 i 阶段输出特征图;Concate 表示拼接操作;

2.3.2　通道注意力桥模块

多阶段,多尺度信息的获取对于不同大小目标的分割起着至关重要的作用。在空间注意力桥模块后面引入通道注意力桥模块有利于建立不同阶段特征图之间的长期依赖关系,从而增强重要信息的微观表达能力。算法伪代码(算法 2)表示为:

通道注意力桥模块将多阶段,多尺度信息的融合细分为局部信息融合(一维卷积操作)和全局信息融合(每个阶段都有不同的全连接层),来提供更丰富信息的通道注意力图。首先将来自 SAB 四个阶段特征图分别进行全局平均池化,然后拼接四个阶段池化结果得到 1×1×C 权重值。接着使用由一个 3×3 卷积核和全连接层组成多层感知器网络来促进局部信息和全局信息的交互,使用 Sigmoid 函数生成通道注意力图,最后将生成的通道注意图以逐元素的方式与原始输入特征图相乘,并引入残差结构。图 4 为通道注意力桥模块示意图。

Algorithm 2: Channel attention bridge block

Inputs: The input maps of the four stages $E_i, i=1,2,3,4$

Outputs: $C_i, i=1,2,3,4$

1: $h^i_{\mathrm{mean}} = \mathrm{AvgPool}(C_i)$ /*avg-pooling*/

2: $h_c = \mathrm{Concat}(h^1_{\mathrm{mean}}, h^2_{\mathrm{mean}}, h^3_{\mathrm{mean}}, h^4_{\mathrm{mean}})$/*Concatenate the feature map of avg-pooling*/

3: $\beta = \mathrm{Conv}_{3\times3}(h_c)$/*3×3 convolution operation*/

4: $\gamma = \sigma(\beta)$/*After sigmoid, the feature map become $C \times H \times 1$*/

5: $C_i = \gamma * E_i + E_i$/*The feature map of sigmoid with the original feature and then add */

End

通道注意力桥模块可表示为:

$$T = \mathrm{Concate}(\mathrm{GAP}(S_1), \cdots, \mathrm{GAP}(S_4)), \quad (7)$$

$$C_i = S_i + S_i \times \sigma(\mathrm{FC}_i(\mathrm{Conv}_{1d}(T))), \quad (8)$$

其中:GAP 表示全局平均池化;Concate 表示拼接操作;FC_i 表示第 i 阶段全连接层;Conv_{1d} 表示 3×3 标准卷积。

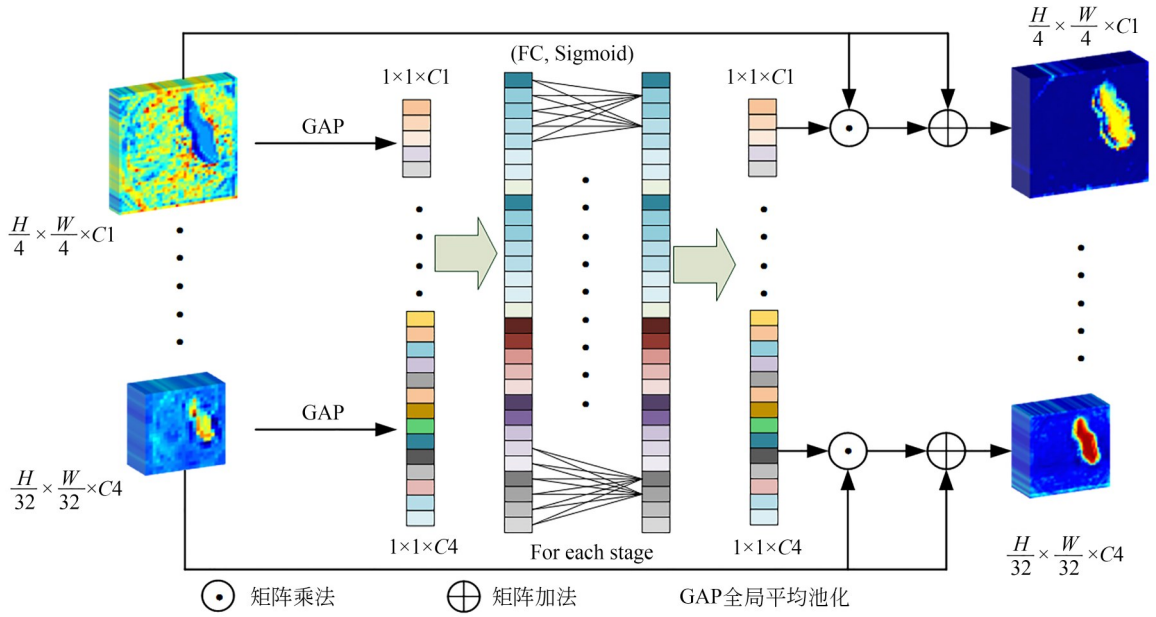

图 4　通道注意力桥模块

Fig. 4　Channel attention bridge block

2.4　多尺度密集并行解码模块

网络解码阶段得到不同尺度特征图，这些特征图蕴含的语义信息和空间细节信息像素相关性是不同的，通过直接上采样操作来恢复图像特征的空间细节，容易造成局部细节信息丢失[22]。因此，本文设计一种新的多尺度密集并行解码模块来对各阶段输出图像进行解码重建，其结构如图 5 所示。

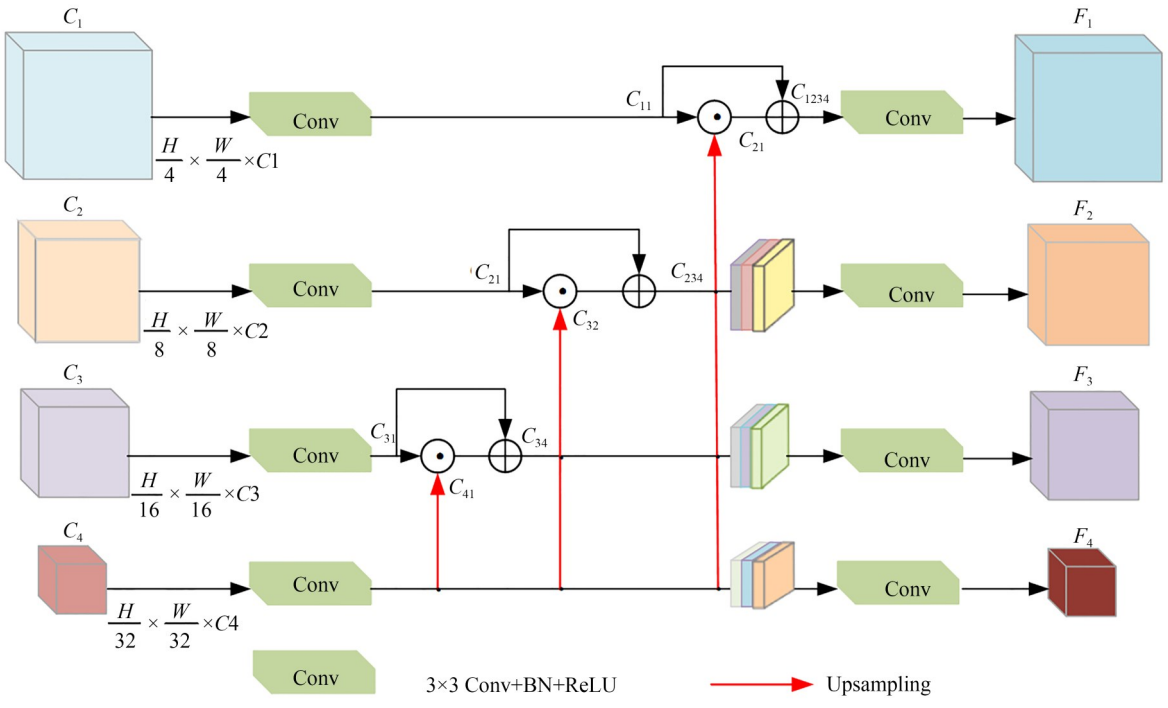

图 5　多尺度密集并行解码模块

Fig. 5　Multi-scale dense parallel decoding block

多尺度密集并行解码模块由标准3×3卷积，批量归一化(BN)层，ReLU激活函数和上采样层组成。模块首先将高级特征图C_4进行双线性插值上采样操作，使其分辨率与特征图C_3相匹配，然后通过两个标准3×3卷积、批量归一化和ReLU激活函数来传递语义结果，得到传递结果C_{41}和C_{31}，再将得到的结果C_{41}与原特征图C_{31}进行矩阵乘法并引入残差结构，最后利用卷积单元来平滑连接特征，得到第三阶段融合特征图。重复上述过程，直至融合所有阶段的输出，最终得到四个阶段的预测输出结果。

2.5 多尺度预测模块

在结直肠息肉病理图像中，病变区域通常具有不同的形状和大小，且与正常组织高度一致。由于肠黏膜特征的相似性，精细的息肉外观特征及容易被忽视。为了促进网络对边缘细节的识别能力，Fan等[11]和Lou等[12]提出反向注意力模块和轴向注意力模块减少目标边缘像素点的误分类。文献[23]提出一种多尺度预测模块，以可学习的方式获取一组权重系数整合不同阶段的预测结果，其结构如图6所示。

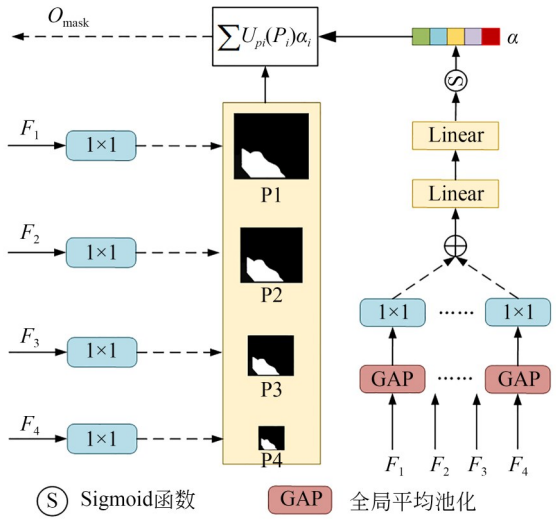

图6 多尺度预测模块

Fig. 6 Multi-scale prediction block

具体来说，在多尺度密集并行解码模块获得输出特征图F_i后，采用四个并行的1×1卷积和上采样(Upsampling)操作获取不同阶段对应的输出二进制掩码P_i。同时，将输出特征图F_i全局平均池化进行空间信息压缩，然后使用4个并行的1×1的卷积来匹配池化后特征的维度，将匹配后的特征图先进行信道加法融合，再通过两个连续的全连接层进一步编码，利用Sigmoid函数得到权重系数，最后采用自适应加权加法得到预测输出掩膜。

3 实验与分析

3.1 数据集和实验设置

为了验证本文模型的有效性，实验采用4个公开的结直肠息肉数据集，即CVC-ClinicDB[24]、Kvasir-SEG[25]，CVC-ColonDB[26]和ETIS[27]。其中CVC-ClinicDB数据库包含612张分辨率大小为384×288的结肠镜图像；Kvasir-SEG数据集包含了1 000张结肠镜图像和分割掩膜；CVC-ColonDB数据集包含了380张分辨率为574×500的结肠镜图像；ETIS数据集包含了196张分辨率为1 226×996的结肠镜图像。实验训练由未经过任何数据增强随机抽取550张CVC-ClinicDB数据图像和900张Kvasir-SEG数据图像组成，测试集由剩下的62张CVC-ClinicDB数据图像、100张Kvasir-SEG数据图像、196张ETIS数据图像和380张CVC-ColonDB数据图像组成。为了方便模型训练和测试，将训练集图像统一分辨率352×352。

本实验在操作系统Windows11进行；建模基于深度学习架构Pytorch 3.9(Facebook Inc.，美国)和计算统一设备架构CUDA 12.1(Nvidia Inc.，美国)。计算机具体配置：显卡(Nvidia GeFore GTX 4070 Ti GPU，Nvidia Inc.，美国)、中央处理器(Intel Core TM i5-13600KF CPU，Inter Inc.，美国)。模型使用以加权二进制交叉熵损失函数和加权交并比损失函数为基础的联合损失，采用自适应矩估计优化器(Adam)，实验迭代次数50，批量大小设置为6，初始学习率5×10^{-5}，动量设置为0.9，并使用多尺度训练策略{0.75，1，1.25}。

3.2 评估指标

本文采用Dice相似性系数、平均交并比(Mean Intersection Over Union，MIoU)、召回率(Sensitivity，SE)、精确率(Precision，PC)、F2得分和平均绝对误差(Mean Absolute Error，MAE)

来对结肠息肉的分割结果进行评估。其具体计算式分别为:

$$Dice=\frac{2|X\cap Y|}{|X|+|Y|},\tag{9}$$

$$MIoU=\frac{|X\cap Y|}{|X|+|Y|-|X\cap Y|},\tag{10}$$

$$SE=\frac{TP}{TP+FN},\tag{11}$$

$$PC=\frac{TP}{TP+FP},\tag{12}$$

$$F2=\frac{5\times SE\times PC}{4\times PC+SE},\tag{13}$$

$$MAE=\frac{1}{N}\sum|Y-X|,\tag{14}$$

其中:X为预测输出图像,Y专家标注的金标签图像,TP为预测结果中正确分类的前景像素数目,FN为预测结果中被错误分类为前景像素数目,FP为预测结果中被错误分类为背景像素数目,N为图像中的像素点数。

3.3　对比试验

为了验证本文方法在结直肠息肉数据集分割方面的优势,分别与经典医学图像分割网络U-Net,EU-Net[28],PraNet和CaraNet以及最近提出基于Transformer结构的医学图像分割网络PolypPVT[21],SSFormer和MSRAFormer进行比较,实验结果如表1~表2所示。表中最优指标加粗表示。

表1　Kvasir和CVC-ClinicDB数据集上不同网络分割结果

Tab. 1　Segmentation results of different networks on Kvasir and CVC-ClinicDB datasets

Dataset	Method	Dice	MIoU	SE	PC	F2	MAE
Kvasir	U-Net	0.818	0.746	0.856	0.857	0.827	0.055
	EUNet	0.908	0.854	0.934	0.911	0.919	0.028
	PraNet	0.898	0.840	0.911	0.916	0.901	0.032
	CaraNet	0.918	0.867	0.912	0.938	0.914	0.023
	PolypPVT	0.917	0.864	0.913	0.947	0.914	0.023
	SSFormer-L	0.918	0.865	0.897	**0.957**	0.904	0.022
	MSRAFormer	0.923	0.873	0.915	0.952	0.917	0.024
	Ours	**0.932**	**0.883**	**0.933**	0.944	**0.931**	**0.021**
CVC-ClinicDB	U-Net	0.823	0.755	0.834	0.839	0.827	0.019
	EUNet	0.902	0.846	0.959	0.880	0.926	0.011
	PraNet	0.899	0.849	0.910	0.907	0.905	0.009
	CaraNet	0.936	0.887	0.955	0.928	0.948	0.007
	PolypPVT	0.937	0.889	0.949	**0.936**	0.945	0.006
	SSFormer-L	0.906	0.855	0.897	0.931	0.898	0.008
	MSRAFormer	0.924	0.874	0.945	0.920	0.932	0.008
	Ours	**0.942**	**0.896**	**0.964**	0.927	**0.954**	**0.006**

表1给出本文模型和其他7种模型在Kvasir和CVC-ClinicDB数据集上的评估结果,表2给出本文模型和其他7种模型在CVC-ColonDB和ETIS上的评估结果,在Dice指标和MIoU指标均为最高。Dice指标是用于衡量分割结果中预测像素数量与总数量的比例,值越高表示分割结果与真实标签越接近,分割质量越高。F2得分是综合考虑召回率和精确度,相比于F1分数,F2分数更加重视召回率,值越高表示能更准确地将目标对象从背景中分割出来。表2显示,在Kvasir和CVC-ClinicDB数据集中,本文方法的Dice和F2得分分别为0.932,0.942和0.931,0.954,相比基于CNN的基础网络CaraNet,分割结果分别提高了1.4%,1.7%和0.6%,0.6%。与基于Transformer结构的PolypPVT相比,Dice指标分别提高了1.5%和0.5%。表2显示,本文方法在CVC-ColonDB和ETIS数据集上的Dice和MIoU指标分别为0.811,0.805和0.731,0.729。实验结果表明,本文方法分割效果更好,泛化性能更加,病变区域误分类更少。

表2　CVC-ColonDB和ETIS数据集上不同网络分割结果

Tab. 2　Segmentation results of different networks on CVC-ColonDB and ETIS datasets

Dataset	Method	Dice	MIoU	SE	PC	F2	MAE
CVC-ColonDB	U-Net	0.512	0.444	0.523	0.621	0.510	0.061
	EUNet	0.756	0.681	0.849	0.758	0.788	0.044
	PraNet	0.712	0.640	0.739	0.755	0.717	0.043
	CaraNet	0.773	0.689	**0.857**	0.753	0.796	0.042
	PolypPVT	0.808	0.727	0.821	0.849	0.809	0.031
	SSFormer-L	0.802	0.721	0.791	0.864	0.787	0.031
	MSRAFormer	0.782	0.707	0.803	**0.874**	0.787	0.028
	Ours	**0.811**	**0.731**	0.823	0.844	**0.813**	**0.027**
ETIS	U-Net	0.398	0.335	0.482	0.439	0.429	0.036
	EUNet	0.687	0.609	0.871	0.635	0.749	0.066
	PraNet	0.628	0.567	0.686	0.628	0.649	0.031
	CaraNet	0.747	0.672	0.811	0.731	0.777	0.017
	PolypPVT	0.787	0.706	0.867	0.774	0.820	0.013
	SSFormer-L	0.796	0.720	0.830	**0.794**	0.807	0.014
	MSRAFormer	0.750	0.679	0.811	0.745	0.777	0.013
	Ours	**0.805**	**0.729**	**0.887**	0.770	**0.842**	**0.012**

表3给出了本文网络和其他7种网络的模型参数性能，以Transformer结构为基础的Polyp-PVT，SSFormer-L，MSRAformer和本文方法在分割性能指标上均高于基于CNN结构的U-Net，EUNet和PraNet。在一定程度上Transformer结构会提升网络的参数量和计算复杂度，相比于其他7种网络本文方法在参数量上获得最优，计算复杂度和单轮训练时长均获得了次优结果，分别为24.99 M，10.01 G和127 s。相比于基于Transformer结构MSRAformer，本文方法参数量和计算复杂度分别降低了63.26%和52.98%。相比于基于CNN结构的EUNet，本文方法参数量和计算复杂度分别降低了20.31%和18.31%。综合分析，本文方法在兼顾参数量、计算量和推理时间的情况下，获得较好的分割结果，实现了网络分割性能的提升。

表3　不同网络性能对比(CVC-ClinicDB)

Tab. 3　Performance comparison of different networks (CVC-ClinicDB)

Method	Parameters/M	GFLOPs	Train/(round·s^{-1})
U-Net	34.53	65.52	309
EU-Net	31.36	12.31	284
PraNet	30.50	6.96	**90**
CaraNet	44.54	11.45	256
Polyp-PVT	25.12	**5.30**	233
SSFormer-L	65.96	17.29	220
MSRAformer	68.03	21.29	199
Ours	**24.99**	10.01	127

图7和图8分别显示了本文方法和上述7种方法在CVC-ClinicDB，Kvasir，ETIS和CVC-ColonDB数据集上的分割结果，从上往下依次为原始输入图像、金标签、U-Net，EU-Net，PraNet，CaraNet，Polyp-PVT，SSFormer，MSRAFormer和本文方法。从图7～图8可以看出，在病灶区域占比较小的病理图像中，EU-Net不能很好地区分背景和前景，出现大量的误检现象。Polpy-PVT使用相似性聚合模块有效分割病灶区域占比小的病理图像，但分割大面积病灶区域时分割结果出现内部不连贯现象。SSFormer和CaraNet减少了分割结果内部不连贯情况，但分割结果边缘不平滑、出现伪影。U-Net和PraNet建模全局信息能力不足，存在大量漏检错检的现象，易将背景错分成前景。由于结直肠息肉病灶区域与正常组织对比度低，如图7第4列和图8第9列所示，U-Net，PraNet和MSRAFormer把病灶

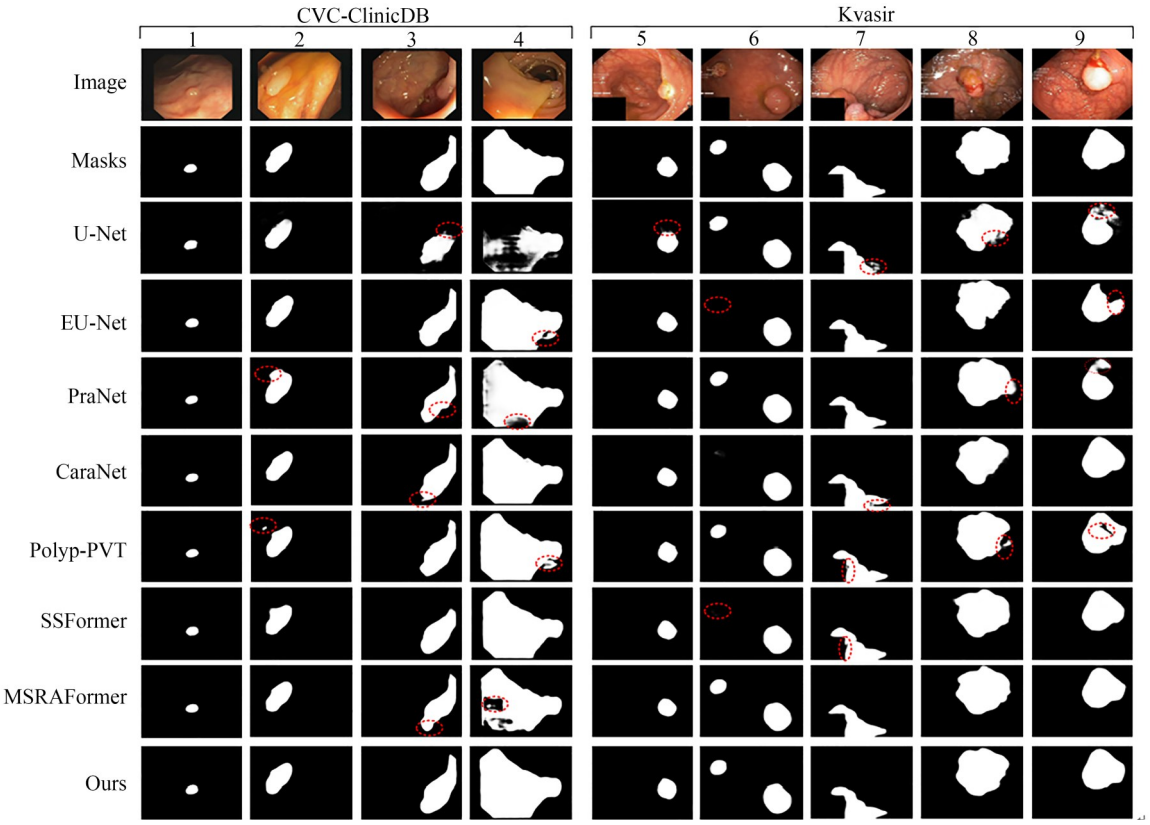

图 7　Kvasir 和 CVC-ClinicDB 数据集上不同网络分割结果

Fig. 7　Segmentation results of different networks on Kvasir and CVC-ClinicDB datasets

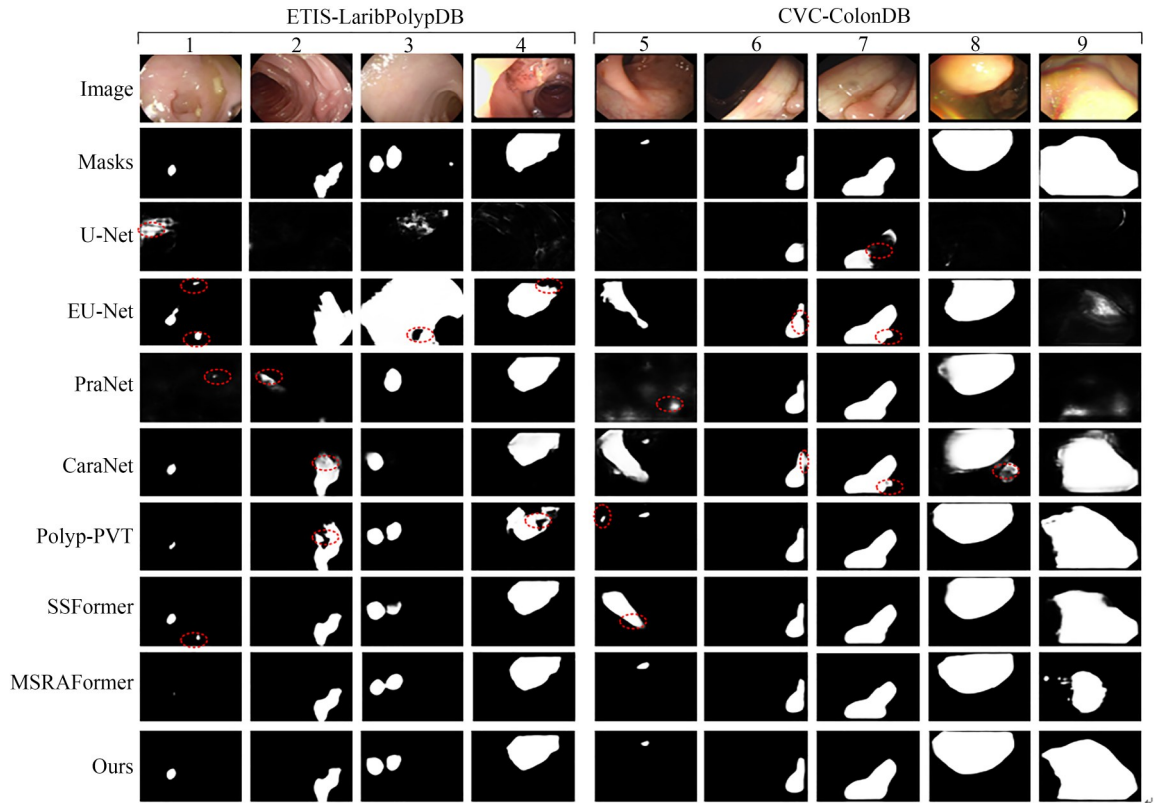

图 8　CVC-ColonDB 和 ETIS 数据集上不同网络分割结果

Fig. 8　Segmentation results of different networks on CVC-ColonDB and ETIS datasets

区域较深的区域当作正常组织，把正常组织较浅的区域错认为病灶部分，误检现象十分严重。同时，由于结直肠息肉的复杂特性，病灶区域常常被淹没在肠黏膜中，病灶区域变化大常常成为分割的难点，如图7第7列和图8第2列所示，EU-Net，Polyp-PVT和CaraNet不能高效定位病灶区域，导致正常组织与病灶边缘分割粗糙，出现伪影。相比之下，本文方法通过Transformer编码器对全局上下文信息进行高效建模，以有效应对结直肠息肉病灶区域尺度变化大的特点，利用混合注意力机制高显病灶区域，减少背景干扰因素的影响，采用多尺度预测模块细化边缘信息，减少分割结果边缘不光滑、伪影现象。综合对比分割结果，本文方法不管是在视觉效果上还是在分割精度上均更胜一筹。

3.4 消融实验

为了探究本文方法各模块对整体模型分割性能的影响，在Kvasir和CVC-ColonDB数据集上进行消融研究。M1在分层Transformer编码器的基础上加入多尺度密集并行解码模块，M2是在M1的基础上添加空间注意力桥模块，M3是在M2的基础上添加通道注意力桥模块，M4是在M3的基础上添加多尺度预测模块，即本文所提方法。消融结果如表4所示，最优指标加粗表示。混合通道注意力机制能有效建立通道信息之间的联系，抑制背景噪声的影响，提升网络的相似性系数和平均交并比。多尺度预测模块对多尺度特征进一步挖掘边缘细节信息，纠正预测分类结果，提升网络像素精度和F2得分，消融实验进一步验证了各模块对整体模型的贡献。

表4 各模块在Kvasir和CVC-ColonDB数据集上消融结果

Tab. 4 Ablation results of each module on the Kvasir and CVC-ColonDB datasets

Dataset	Method	Dice	MIoU	SE	PC	F2
Kvasir	M1	0.906	0.851	0.900	0.931	0.901
	M2	0.921	0.871	0.930	0.931	0.926
	M3	0.928	0.877	**0.934**	0.936	0.928
	M4	**0.932**	**0.883**	0.933	**0.944**	**0.931**
CVC-ColonDB	M1	0.786	0.705	0.7918	0.835	0.785
	M2	0.789	0.706	0.8337	0.803	0.802
	M3	0.810	0.730	**0.841**	0.797	0.806
	M4	**0.811**	**0.731**	0.823	**0.844**	**0.813**

4 结　论

本文提出跨尺度跨维度的自适应Transformer网络应用于结直肠息肉分割，有效地改善了结直肠息肉分割时边缘细节信息丢失和病灶区域误分割等问题。使用Transformer编码器提取结直肠息肉图像病灶特征，通过在网络跳跃连接处引入混合注意力机制，以减少通道维度冗余和增强模型的空间感知能力。同时设计一种新的多尺度密集并行解码模块，充分融合不同尺度的特征信息。最后利用多尺度预测模块激活病灶区域边界的特征响应，进一步优化分割结果。在CVC-ClinicDB和Kvasir-SEG数据集的Dice相似性系数分别为0.942和0.932，相比于基于Transformer结构的SSFormer分别提高了3.6%和1.4%，分割性能优于现有算法，对结直肠息肉的诊断具有一定的临床应用价值。为验证该方法的泛化性和普适性，在CVC-ColonDB和ETIS数据集上进行了验证，实验结果表明本方法在未知数据集上适应能力较强。

参考文献：

[1] LI W, ZHAO Y, LI F, *et al*. MIA-Net: multi-information aggregation network combining transformers and convolutional feature learning for polyp segmentation [J]. *Knowledge-Based Systems*, 2022, 247: 108824.

[2] 徐昌佳, 易见兵, 曹锋, 等. 采用DoubleUNet网络的结直肠息肉分割算法[J]. 光学 精密工程, 2022, 30(8): 970-983.
XU C J, YI J B, CAO F, *et al*. Colorectal polyp segmentation algorithm using DoubleUNet network [J]. *Opt. Precision Eng.*, 2022, 30(8): 970-983. (in Chinese)

[3] 梁礼明, 周珑颂, 冯骏, 等. 基于高分辨率复合网络的皮肤病变分割[J]. 光学 精密工程, 2022, 30(16)2021-2038
LIANG L M, ZHOU L S, FENG J, *et al*. Skin lesion segmentation based on high-resolution composite network [J]. *Opt. Precision Eng.*, 2022, 30(16)2021-2038(in Chinese)

[4] LIANG H, CHENG Z, ZHONG H, *et al*. A region-based convolutional network for nuclei detection and segmentation in microscopy images [J]. *Biomedical Signal Processing and Control*, 2022, 71: 103276.

[5] SHAO D G, XU C R, XIANG Y, *et al*. Ultrasound image segmentation with multilevel threshold based on differential search algorithm [J]. *IET Image Processing*, 2019, 13(6): 998-1005.

[6] 周明全, 杨稳, 林芃樾, 等. 基于最小二乘正则相关性分析的颅骨身份识别[J]. 光学 精密工程, 2021, 29(01):201-210.
ZHOU M Q, YANG W, LIN P Y, *et al*. Skull identification based on least square canonical correlation analysis [J]. *Opt. Precision Eng.*, 2021, 29(1): 201-210. (in Chinese)

[7] 梁礼明, 刘博文, 杨海龙, 等. 基于多特征融合的有监督视网膜血管提取[J]. 计算机学报, 2018, 41(11):2566-2580.
LIANG L M, LIU B W, YANG H L, *et al*. Supervised blood vessel extraction in retinal images based on multiple feature fusion[J]. *Chinese Journal of Computers*, 2018, 41(11): 2566-2580. (in Chinese)

[8] RONNEBERGER O, FISCHER P, BROX T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015: 234-241.

[9] JHA D, SMEDSRUD P H, RIEGLER M A, *et al*. ResUNet: an advanced architecture for medical image segmentation [C]. 2019 *IEEE International Symposium on Multimedia (ISM)*. 9-11, 2019, *San Diego, CA, USA*. IEEE, 2020: 225-2255.

[10] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18-23, 2018, *Salt Lake City, UT, USA*. IEEE, 2018: 7132-7141.

[11] FAN D, JI G, ZHOU T, *et al*. PraNet: Parallel Reverse Attention Network for Polyp Segmentation [EB/OL]. 2020: *arXiv*: 2006.11392. https://arxiv.org/abs/2006.11392.pdf

[12] LOU A G, GUAN S Y, KO H, *et al*. CaraNet: context axial reverse attention network for segmentation of small medical objects[C]. *SPIE Medical Imaging. Proc SPIE* 12032, *Medical Imaging* 2022: *Image Processing, San Diego, California, USA*. 2022, 12032: 81-92.

[13] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is All You Need [EB/OL]. 2017: *arXiv*: 1706.03762. https://arxiv.org/abs/1706.03762.pdf

[14] DAI Y, GAO Y F, LIU F Y. TransMed: transformers advance multi-modal medical image classification [J]. *Diagnostics (Basel, Switzerland)*, 2021, 11(8): 1384.

[15] CHEN J, LU Y, YU Q, *et al*. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation [EB/OL]. 2021: *arXiv*: 2102.04306. https://arxiv.org/abs/2102.04306.pdf

[16] WANG J F, HUANG Q M, TANG F L, *et al*. *Stepwise Feature Fusion: Local Guides Global* [M]. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 110-120.

[17] WU C, LONG C, LI S, *et al*. MSRAformer: Multiscale spatial reverse attention network for polyp segmentation [J]. *Computers in Biology and Medicine*, 2022, 151: 106274.

[18] WANG W H, XIE E Z, LI X, *et al*. PVT v2: improved baselines with pyramid vision transformer [J]. *Computational Visual Media*, 2022, 8(3): 415-424.

[19] ISLAM M A, JIA S, BRUCE N D B. How Much Position Information Do Convolutional Neural Networks Encode? [EB/OL]. 2020: *arXiv*: 2001.08248. https://arxiv.org/abs/2001.08248.pdf

[20] RUAN J C, XIANG S C, XIE M Y, *et al*.

MALUNet: a multi-attention and light-weight UNet for skin lesion segmentation[C]. 2022 *IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*). 6-8, 2022, *Las Vegas*, *NV*, *USA*. IEEE, 2023: 1150-1156.

[21] DONG B, WANG W, FAN D, *et al*. Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers [EB/OL]. 2021: *arXiv*: 2108.06932. https://arxiv.org/abs/2108.06932

[22] 刘媛媛, 周小康, 王跃勇, 等. 改进U-Net模型的保护性耕作田间秸秆覆盖检测[J]. 光学 精密工程, 2022, 30(9): 1101-1112.
LIU Y Y, ZHOU X K, WANG Y Y, *et al*. Straw coverage detection of conservation tillage farmland based on improved U-Net model [J]. *Opt. Precision Eng.*, 2022, 30(9): 1101-1112. (in Chinese)

[23] ZHANG W, FU C, ZHENG Y, *et al*. HSNet: a hybrid semantic network for polyp segmentation [J]. *Computers in Biology and Medicine*, 2022, 150: 106173.

[24] BERNAL J, SÁNCHEZ FJ, FERNÁNDEZ-ESPARRACH G, *et al*. WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation *vs*. saliency maps from physicians[J]. *Computerized Medical Imaging and Graphics*, 2015, 43: 99-111.

[25] JHA D, SMEDSRUD P H, RIEGLER M A, *et al*. *Kvasir-SEG: a Segmented Polyp Dataset*[M]. MultiMedia Modeling. Cham: Springer International Publishing, 2019: 451-462.

[26] TAJBAKHSH N, GURUDU S R, LIANG J M. Automated polyp detection in colonoscopy videos using shape and context information [J]. *IEEE Transactions on Medical Imaging*, 2016, 35(2): 630-644.

[27] 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.

[28] PATEL K, BUR A M, WANG G H. Enhanced U-Net: a feature enhancement network for polyp segmentation[C]. 2021 18*th Conference on Robots and Vision* (*CRV*). 26-28, 2021, *Burnaby*, *BC*, *Canada*. IEEE, 2021: 181-188.

作者简介：

梁礼明(1967—),男,江西吉安人,教授,硕士,江西省高等学校中青年骨干教师,主要从事机器学习和医学影像方面的研究。E-mail: 9119890012@jxust.edu.cn

通讯作者：

吴　健(1991—),男,江西赣州人,硕士,讲师。主要从事医学影像和机器学习等方面研究。E-mail: wujian@jxust.edu.cn