激光与光电子学进展
先进成像

结合自注意力与卷积神经网络的腺体及息肉分割方法

张家宝，肖志勇*
江南大学人工智能与计算机学院，江苏 无锡 214122

摘要 腺体和息肉的自动分割是人工智能辅助结直肠腺癌诊断的基础，但医学图像中的分割目标大小、形状多变，基于单一的卷积神经网络的自动分割方法已陷入瓶颈。基于此，提出了一种卷积神经网络和自注意力相结合的双分支网络（LG UNet），用以提升分割的精度。首先，基于U-Net设计了Local UNet分支，利用卷积神经网络的优势，学习分割目标的局部信息。然后在Global Transformer分支中，利用Transformer全局依赖关系的学习能力来优化分割细节。最后在编码过程中通过交叉融合模块将Local分支和Global分支的特征图进行融合，将两者优势互补。在腺体分割挑战数据集Glas的两个测试子集Test A和Test B上，以Dice系数和交并比（IOU）系数为主要评价指标，LG UNet的测试结果分别为93.62%、88.44%和88.17%、80.49%。在息肉分割数据集Kvasir-SEG上，LG UNet的Dice系数和IOU系数分别为85.63%和77.82%。实验结果表明，结合Transformer和卷积神经网络优势的LG UNet在腺体和息肉分割上取得了更好的性能。

关键词 医用光学；自注意力机制；卷积神经网络；双分支网络；结直肠腺癌；腺体分割；息肉分割
中图分类号 TP391 **文献标志码** A **DOI**：10.3788/LOP212696

Gland and Colonoscopy Segmentation Method Combining Self-Attention and Convolutional Neural Network

Zhang Jiabao, Xiao Zhiyong*
School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, *Jiangsu, China*

Abstract The automatic segmentation of glands and polyps is the foundation for the diagnosis of artificial intelligence-assisted colorectal adenocarcinoma. However, the size and shape of segmentation targets in medical images vary considerably, and the automatic segmentation approach based on a convolutional neural network has thus run into a hindrance. Therefore, a dual branch network (LG UNet) combining convolutional neural network and self attention is proposed to improve the accuracy of segmentation. First, the Local UNet branch was developed based on U-Net, and the convolutional neural network's benefits were employed to elucidate the segmentation target's local information. Subsequently, the segmentation details were optimized using the Transformer's learning ability of global dependencies in the Global Transformer branch. Finally, during the encoding process, feature maps of the Local and Global branches were merged by a cross-fusion module to complement their benefits. The two test subsets of Glas and findings of LG UNet were 93.62% and 88.44% for Test A and 88.17% and 80.49% for Test B, employing the Dice coefficient and intersection and union (IOU) coefficient as the primary examination indexes. Furthermore, the Dice and IOU coefficients in the polyp segmentation dataset Kvasir-SEG were 85.63% and 77.82%, respectively. The experimental findings demonstrate that LG UNet exhibits better performance efficiency in gland and polyp segmentation by combining the benefits of the Transformer and convolutional neural network.

Key words medical optics; self-attention mechanism; convolutional neural network; multi-branch network; adenocarcinoma of the colon; segmentation of glands; segmentation of polyps

1 引 言

结直肠腺癌（CRC）是最常见的一种癌症。自动量化腺体和息肉形态，精确提取定量化的形态学特征，然后用于计算机辅助癌症分级，使分级过程比目前更加客观和可复现，是人工智能在癌症诊断领域中一个

收稿日期：2021-10-09；**修回日期**：2021-10-27；**录用日期**：2021-11-16；**网络首发日期**：2021-11-26
通信作者：*zhiyong.xiao@jiangnan.edu.cn

非常重要的应用。

而近些年深度神经网络在计算机视觉领域的飞速发展，为人工智能在癌症辅助诊断领域中带来极大的进展。目前医学图像分割中使用最多的是基于卷积神经网络(CNN)的U-Net[1]，该网络由对称的编码器-解码器网络组成，并通过跳跃连接保留高分辨率细节。U-Net在医学图像分析中取得了巨大的成功，遵循这一思想的各种变体相继被提出并应用于医学图像分割，如3D U-Net[2]、VNet[3]、nnUNet[4]、Res-UNet[5]、SR-Net[6]、文献[7]的网络、文献[8]的网络。但是，卷积运算由于其固有的局部感知性，很难学习全局信息，这使得目前基于U-Net的医学图像分割方法仍不能完全满足医学应用对分割精度的严格要求。全局能力在医学图像分析中也尤为重要：由于医学图像来源复杂，形状各异，并且噪声多、干扰大，如果缺乏对全局信息的分析，将会导致分割精度的缺失。比如在腺体和息肉数据集中，因为成像方法的原因，腺体与息肉的边界与背景颜色非常接近。一些小腺体、小息肉游离于大目标旁边，精确地分割这些目标边界，需要网络引入更多的全局信息。

因此，研究者们尝试将注意力机制加入U-Net中，以增强网络的全局信息提取能力，如Attention-UNet[9]、Acu-UNet[10]、文献[11]中的网络。而最近随着自注意力在自然语言处理领域的成功，研究者们开始尝试将更好的Transformer[12]带入U-Net中，比如TransUNet[13]、Swin-UNet[14]、UTNet[15]。TransUNet将Transformer加入U-Net中，利用来自CNN的局部高分辨率空间信息和Transformer编码的全局上下文信息实现精确定位。

但引入Transformer的医学分割网络存在一个问题，即Transformer不具备CNN的归纳偏置能力，因此需要在大型数据集上进行预训练，才能在下游任务中取得最佳的性能。而在医学图像分割领域，由于人工标注需要相关领域专家的先验知识，成本高昂，无法提供足量的数据进行预训练。一些研究人员针对该问题提出了改进方案：DeiT[16]提出了几种与数据蒸馏方法相结合的训练策略，使得Transformer只需在中型数据集上预训练，就能得到了一个较好的效果；MedT[17]参考文献[18]中提出的轴注意力模型，在自注意力中加入了4个可学习的门控参数，自动控制位置嵌入量的学习，以适应不同大小的医学数据集，以此提出了一种无须在其他数据集上进行预训练的医学分割网络。

而本文提出了一个双分支的混合网络模型(LG UNet)，将U-Net局部信息提取能力与Transformer对长期依赖关系的学习能力以双分支的形式相结合，去除Transformer对预训练的依赖，使其能在不预训练的情况下进行小数据集的医学图像分析。将LG UNet直接用于腺体和息肉分割，实验结果表明，结合了Transformer的LG UNet在小数据集的医学图像中依然能取得很好的效果。

2 所提方法

2.1 Local UNet分支

U-Net由一个对称的编码器-解码器及跳跃连接组成。在编码器中，每阶段采用两个连续的卷积层来编码目标特征，每个卷积层后面紧跟批次归一化层(BN)和激活函数ReLU，如图1(a)所示。在每阶段最后，通过一个最大池化层来进行降采样，以增强网络中卷积的感受野。解码器中每阶段由同样的两个连续卷积层组成，并且在解码器中将编码器提取的高维特征向上采样到输入分辨率。而且在解码过程中，通过跳跃连接融合来自编码器的不同尺度的低级语义特征，以减轻编码过程中降采样导致的空间信息丢失。

另外，在分析结直肠腺癌自动诊断中的腺体和息肉分割数据集后发现，腺体会占图像中的绝大部分，而且分布密集、边界变化曲折。同时，息肉分割中息肉的尺度大小，形状变化差异性大，占比也普遍偏大，但相对于腺体来说形状要规则很多。两者都是大目标分割，因此为了增强Local UNet分支提取特征的能力，

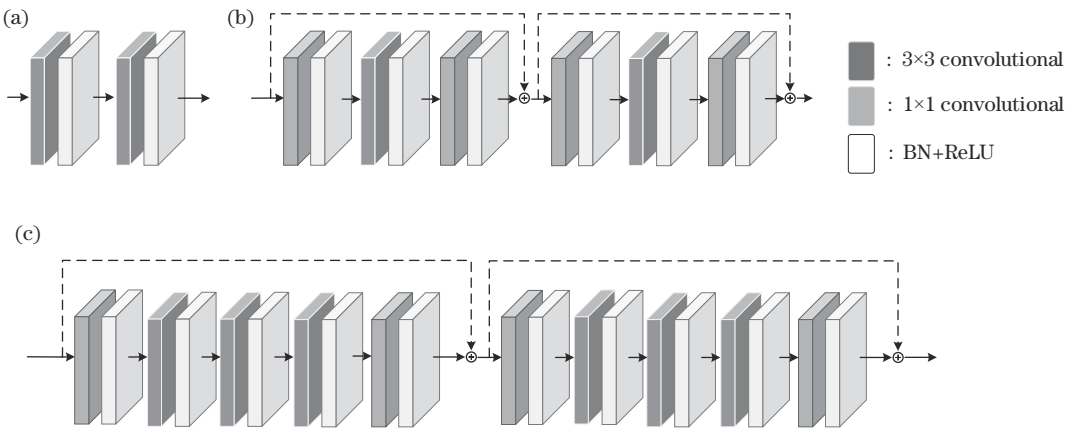

图1 Local UNet编码结构

Fig. 1 Local UNet encoding structure

本研究对U-Net进行了如下改进。首先，为了得到更优的编码性能和效率，引入了ResNet[19]中Bottleneck模块，如图1(b)所示。ResNet中的残差连接可以缓解层数加深后的网络退化问题，使编码器更好地学习深层网络中的信息，并且Bottleneck模块能通过两个1×1的卷积来减少计算量。其次，为了增加编码时的感受野及增强对多尺寸目标的编码能力，受GoogleNet[20]和DC-UNet[21]的启发，将Bottleneck模块中的卷积增加为3个连续的3×3卷积。为了减少参数量和计算代价，Bottleneck模块中的通道数设为输入通道数的1/4，再通过1×1卷积，将通道数调整回预定的输出通道数。由此得到了Local UNet中的编码模块，如图1(c)所示。

解码过程中，采用U-Net中的标准模块。首先将上一层输出的特征图经过一个上采样，然后通过跳跃连接与编码层中相应的特征图拼接。跳跃连接中的低层次特征，能够让网络更好地还原分割目标的边缘信息。最后得到的特征图再经过两次卷积，输入下一层。

在Local UNet分支中，通过叠加卷积层，增大了卷积的感受野。感受野越大，卷积操作能提取的信息范围就越大，对大目标分割就越有优势。而且通过叠加3个3×3的卷积，还能有效增强多尺度特征的提取能力，多尺度特征的学习能力是小目标分割的关键，在腺体和息肉分割中会出现游离于大目标的小型分割目标，这也是影响分割精度的因素之一。最后再减少中间过程的通道数及网络层数，在不增加大量参数和过多计算量的情况下，有效提升了Local分支的特征提取能力。

2.2 Global Transformer分支

在将Transformer应用于医学分割时，除了医学数据集普遍偏小，无法为网络提供足量的预训练外，另一个不可忽视的问题就是计算量。因为Transformer是由多个自注意力模块组成的，而自注意力模块的核心计算公式为

$$\mathrm{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\cdot\boldsymbol{V},\quad(1)$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别是queries、keys、values矩阵，其中$\boldsymbol{Q},\boldsymbol{K}\in\mathbf{R}^{N\times d_{\mathrm{k}}}$、$\boldsymbol{V}\in\mathrm{R}^{N\times d_{\mathrm{v}}}$，$N$表示序列长度，$d_{\mathrm{k}}$、$d_{\mathrm{v}}$表示词嵌入维度大小。因此自注意力的计算量与输入序列的长度成二次方。但在医学分割领域，如果将一个像素点视为一个词，采用最常见的医学分割图像尺寸，即256×256×3的分辨率，其输入规模将近20万，计算代价较高。

因此为了减少计算量，目前的自注意力网络只会在局部计算注意力，或者沿单个轴计算注意力。比如在ViT[22]中，会将输入的2D图像切分成若干个大小固定的小图块，然后将每个小图块展开到一维，并映射到统一的长度D，将其视为一个词嵌入向量输入Transformer中。大致过程可描述为

$$\boldsymbol{z}_0=[\boldsymbol{x}_{\mathrm{class}};\boldsymbol{x}_p^1\boldsymbol{E};\boldsymbol{x}_p^2\boldsymbol{E};\cdots;\boldsymbol{x}_p^N\boldsymbol{E}]+\boldsymbol{E}_{\mathrm{pos}},\quad(2)$$

式中：$\boldsymbol{x}_{\mathrm{class}}$表示用于分类的嵌入向量，$\boldsymbol{x}_p^i\in\mathbf{R}^{1\times(P^2\cdot C)}$表示的是第$i$小图块，$P$表示小图块的大小，$C$表示通道数；$\boldsymbol{E}\in\mathbf{R}^{(P^2\cdot C)\times D}$表示的是映射矩阵，用于将$\boldsymbol{x}_p^i$映射到固定的$D$维，而$D$是每层输入向量设定的恒定维度；$\boldsymbol{E}_{\mathrm{pos}}\in\mathbf{R}^{(N+1)\times D}$是相对位置嵌入量，用于保存每个小图块的位置信息；$N=HW/P^2$，表示小图块的数量。

而在MedT中为了减少计算量，采用轴向自注意力机制，沿图像水平和垂直方向分成两个自注意力机制计算。以水平方向为例，MedT中的自注意力计算模块可描述为

$$\boldsymbol{y}_{ij}=\sum_{w=1}^{W}\mathrm{Softmax}\left(\boldsymbol{q}_{ij}^{\mathrm{T}}\boldsymbol{k}_{iw}+\boldsymbol{G}_{\mathrm{Q}}\boldsymbol{q}_{ij}^{\mathrm{T}}\boldsymbol{r}_{iw}^{q}+\boldsymbol{G}_{\mathrm{K}}\boldsymbol{k}_{iw}^{\mathrm{T}}\boldsymbol{r}_{iw}^{k}\right)\times\left(\boldsymbol{G}_{\mathrm{V1}}\boldsymbol{v}_{iw}+\boldsymbol{G}_{\mathrm{V2}}\boldsymbol{r}_{iw}^{v}\right),\quad(3)$$

式中：$i\in\{1,\cdots,H\}$，$j\in\{1,\cdots,W\}$，H、W分别代表输入特征图的高和宽；$\boldsymbol{q}_{ij}$、$\boldsymbol{k}_{iw}$、$\boldsymbol{v}_{iw}$分别表示queries、keys、values矩阵中相应位置的元素；相对位置编码矩阵$\boldsymbol{r}^q$，$\boldsymbol{r}^k,\boldsymbol{r}^v\in\mathbf{R}^{W\times W}$；可学习的门控参数$\boldsymbol{G}_{\mathrm{K}},\boldsymbol{G}_{\mathrm{Q}},\boldsymbol{G}_{\mathrm{V1}},\boldsymbol{G}_{\mathrm{V2}}\in\mathrm{R}$。以学习得到的参数$G$作为一个权重，对所学得的位置嵌入量进行调控，从而获得更优秀的结果。

此外，为了取得更好的效果，研究者们也在Transformer的基础上进行了一些改动，如Swin Transformer[23]、TNT[24]。Swin Transforme首先将输入图片分成4×4大小的小图块，然后以小图块为单位，构造7×7大小的窗口，这样较以往分成16×16大小的小图块，在局部范围内，Swin Transformer就具有了更高的分辨率。在窗口内使用多头自注意力机制(W-MSA)计算局部注意力，计算量与图像大小呈线性关系。其次，在连续的自注意层之间对窗口进行有规则的移位，构造出了基于移动窗口的多头自注意力计算机制(SW-MSA)。移动的窗口桥接了前一层的窗口，提供了层与层之间的连接，显著提高了编码能力。最后，通过一个Patch Merging层逐步合并相邻的小图块，构造一个分层特征映射，达到一个类似降采样的效果，增加窗口中自注意力计算的感受野。改进之后的Swin Transformer拥有更强的编码能力、更低的计算代价、更大的注意力计算范围。而更大的注意力计算范围，在处理腺体和息肉这种大目标数据集时更有优势，因此本研究选用Swin Transformer代替了普通的Transformer作为Global分支的骨架。

Global Transformer分支的设计参考了文献[7]提出的TransUNet。在编码时由若干个Swin Transformer模块组成一个编码阶段，每个阶段最后通过一个Patch Merging层进行降采样，整个编码网络由3个阶段构成。解码时采用与Local分支一样的设计。但TransUNet中并没有纯粹使用Transformer作为编码器，而是先用CNN编码输入图像，然后再在最后一个编码阶段用Transformer计算全局依赖。这主要有

两个原因：1）直接通过 Transformer 解码后的特征图大小为输入图像的 1/P，如果直接输入网络中进行解码，会损失很多细节；2）Transformer 注重全局信息，如果直接使用 Transformer 解码就会损失一些局部信息，影响分割的细节。而本研究在 Global 分支里使用纯粹的 Swin Transformer 作为编码器，因此为了解决上述问题，在编码中设计了一个交叉融合模块，如图 2 所示。

2.3　交叉融合模块

在交叉融合模块中，上面的 $C\times W\times H$ 表示的是 Global Transformer 分支中的特征维度，下面的 $C'\times 2W\times 2H$ 表示的是 Local UNet 分支中的特征维度。该模块中，每个分支的输出会经过一个 Bottleneck 模块，得到的结果再与另一个分支的输入直接相加，最后得到的结果作为该分支的输出，输入下一层的编码器中。

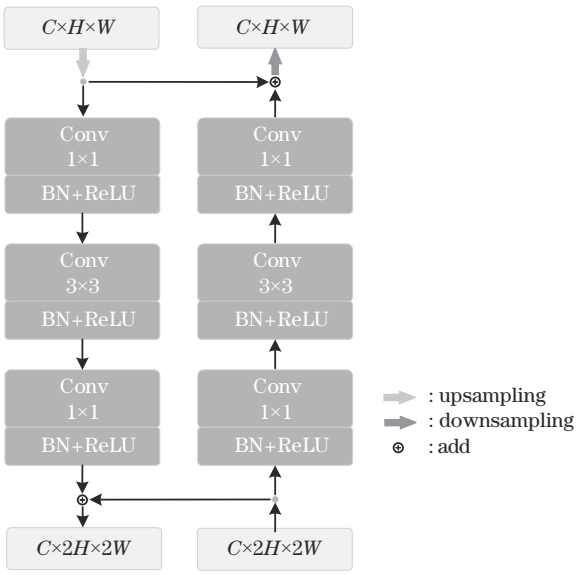

图 2　交叉融合模块结构

Fig. 2　Cross fusion block structure

在交叉融合模块中，将 Local 分支中的局部信息融入 Global 分支中，补充 Transformer 单独编码而丢失的局部信息。同时也利用 CNN 的归纳偏置能力补充 Transformer 分支的编码特征，省去预训练的过程，使 Transformer 在小数据集的医学分割中依旧取得较好的性能。另外又能将 Transformer 解码得到的全局信息补充到 Local UNet 分支中，使 UNet 在解码时能获得更好的性能。

2.4　平均相关性模块

在 Swin Transformer 中为了减少计算量，在划分的窗口中计算局部注意力，然后逐层合并窗口，扩大感受视野。然而，分割操作会影响同一通道的不同分区之间的位置关系，网络会单独查看每个分区，没有考虑通道之间的相关性。因此为了减少分割的影响，本研究在每个 Transformer 模块中加入了平均相关性模块。即将输入通过一个卷积核大小与窗口大小一致的平均池化层，再通过两个连续的卷积层，最后通过自动广播的方式，添加到网络的主体输入中。

2.5　LG UNet 架构简介

所提 LG UNet 的架构如图 3 所示。LG UNet 由 Local UNet 分支和 Global Transformer 分支组成。

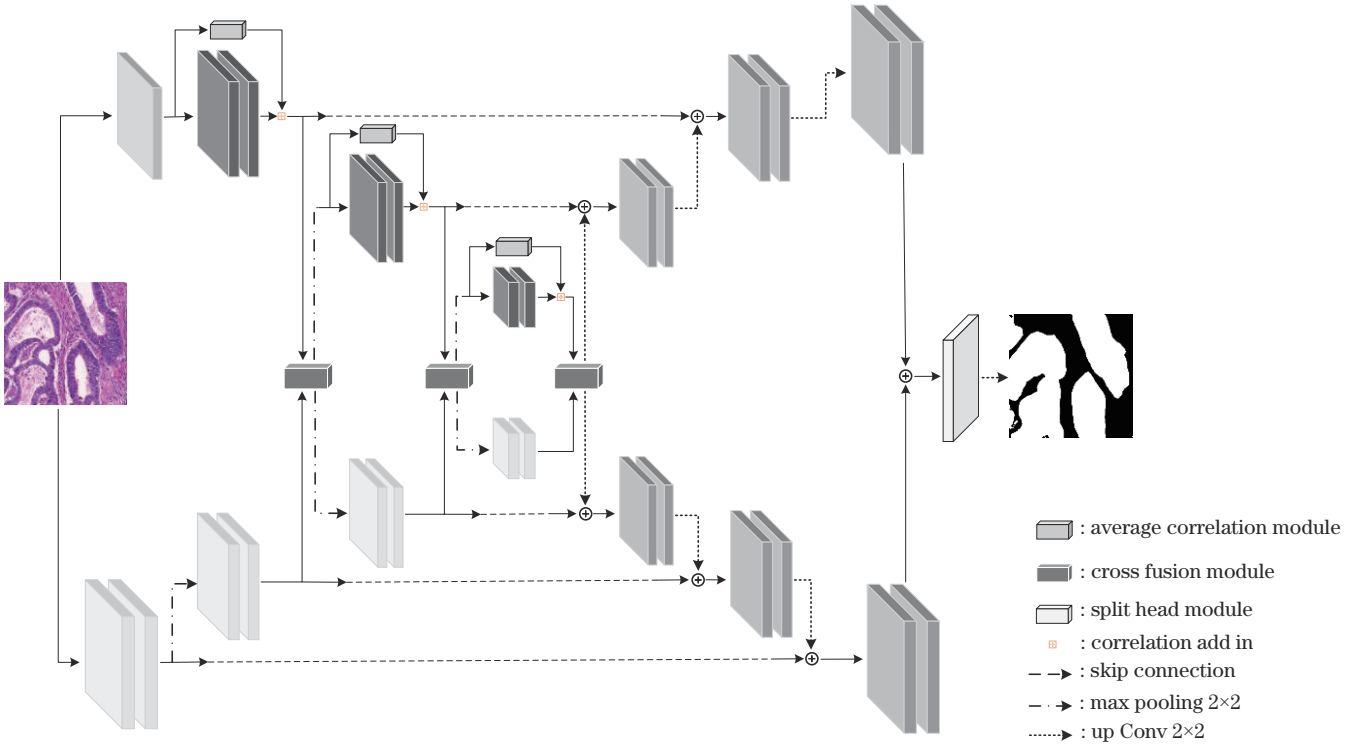

图 3　LG UNet 结构图

Fig. 3　LG UNet structure diagram

Local UNet分支位于网络下方，是在UNet结构的基础上改进而来的，主要作用在于增大卷积的感受野，增强对大目标的检测能力及该结构的特征提取能力。Local UNet分支的编码网络由最初的5层减少为4层，每层的通道数分别为64、256、512、1024。编码时会将上一阶段的输出特征图输入分支融合模块中，与Global分支中提取的特征进行融合，并将其输出特征图传入下一阶段的编码器中继续编码。通过交叉融合，将Global分支得到的全局信息补充到Local分支中，而这种全局信息是精确分割边界的关键。然后在解码层中将编码得到的特征图逐步上采样回最初的分辨率，用于像素级的分割预测。并且在解码过程中通过跳跃连接，将编码最初得到的低维特征图融入解码过程中，进一步增强Local分支中的局部特征及补充编码过程中因降采样而损失的细节。

在Global Transformer中，依照Swin Transformer的组织方式，两个连续的W-MSA和SW-MSA模块组成一个基本的编码模块，整个Global Transformer分支由3个阶段构成，每个阶段包含若干个基本编码模块。每个阶段的输出会输入分支融合模块中，与Local UNet得到的特征图进行融合，再经过一个Patch Merging层下采样后输入下一阶段的编码中。同样通过分支间的交叉融合，Global Transformer分支获得了Local UNet中的局部信息和卷积的归纳偏置能力。解码层中依照TransUNet的设计思路，采用卷积解码，同样具有跳跃连接。

在最后，将两个分支的最终输出特征图相加，输入分割头模块中得到最终的分割结果。分割头模块由一个简单的Bottleneck模块组成。

2.6 损失函数

使用的损失函数为Dice损失函数和交叉熵损失函数的混合函数，其表达式为

$$L = L_{\text{dice}} + 0.5 \times L_{\text{ce}}, \tag{4}$$

式中：L_{dice}表示Dice损失函数；L_{ce}表示交叉熵损失函数；L表示所提网络使用的损失函数。

Dice损失是一种区域相关的计算方式，会考虑点与点之间的联系。而且Dice损失更加注重前景得分，能够让网络更加关注正样本的学习。L_{dice}的表达式为

$$L_{\text{dice}} = 1 - \frac{2|A \cap B|}{|A| + |B|}, \tag{5}$$

式中：A、B分别表示网络输出和真实标签；$|A \cap B|$表示A和B之间的交集；$|A|$和$|B|$分别表示A和B中的元素个数。

但Dice损失的计算方式也导致其不稳定性，在学习过程中容易受到干扰。因此在Dice损失的基础上加入交叉熵损失函数，目的是让网络在训练过程中更加稳定。交叉熵损失函数计算中会单独考虑每个像素，然后求平均。交叉熵损失函数的表达式为

$$L_{\text{ce}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right], \tag{6}$$

式中：N为像素点个数；y_i表示该像素点的真实值；$\hat{y}_i$表示该像素点的预测值。

3 实验结果与讨论

3.1 实验数据与预处理

主要使用结肠组织学图像中的腺体分割挑战数据集Glas[25]和胃肠道息肉分割数据集Kvasir-SEG[26]。Glas数据集包含165张结肠癌诊断中的组织学切片图像，大多数是775 pixel×522 pixel的图像，其中85张为训练数据，80张为测试数据。测试数据集官方又分为子集A（60张图像）和子集B（20张图像）两个不同的子集。本研究结果展示的是在两个测试集上的单独结果。Kvasir-SEG数据集包含1000张胃肠道息肉图像，并给出了对应的息肉分割标签，数据集按照8∶1∶1随机分成训练集、验证集、测试集。

在预处理中，使用常规的水平翻转和垂直翻转来增强训练过程中的样本多样性，另外还加入了色彩调节。对于真实标签图，将其中的像素进行0、1化，0表示背景，1表示前景，就是分割目标。最后将训练图像和对应的真实标签图像重新裁剪为256×256，以便统一网络中的参数。

3.2 实验环境与参数设置

实验中的硬件环境为Intel Core i7处理器、NVIDIA GTX1080ti显卡。选用的编码语言为Python，深度学习框架为PyTorch 1.8.1版本。本实验中采用AdamW[27]优化器，批次大小设为8，训练1000个epoch，初始学习率α为0.001，权重衰减为1×10^{-5}。

在Local UNet分支中，网络设为4层，每层通道数分别设置为64、256、512、1024。在Global Transformer分支中经过3次编码，通道分别为96、192、384。Swin Transformer参数设置中，图块大小patch_size设为4×4，窗口大小window_size设为8×8。网络输出为两个通道，0号通道表示背景，1号通道表示前景，统计时查看前景指标。

3.3 评价指标

选用的评价指标为Dice系数和交并比（IOU）系数，其表达式分别为

$$\text{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \tag{7}$$

$$\text{IOU}(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{8}$$

式中：$|A \cap B|$和$|A \cup B|$分别表示A和B之间的交集和并集。Dice指标和IOU指标都能用来衡量标签和输出结果之间的重叠度，两者重合度越高，分割效果越好。但IOU指标对重叠部分的考虑太过粗糙，因此引入Dice指标作为主要评价指标。

3.4 消融实验

为了验证网络中各个组件的有效性，设计以下几组对比实验，结果如表1所示。

表1 消融实验结果

Table 1 Results of ablation experiment

Model	Dice		IOU	
	Test A	Test B	Test A	Test B
U-Net	0.8826	0.8269	0.7934	0.7216
Local UNet	0.9133	0.8746	0.8447	0.7868
LG-v1	0.9267	0.8822	0.8663	0.8010
LG-v2	0.9302	0.8834	0.8711	0.8028
LG-v3	**0.9362**	**0.8844**	**0.8817**	**0.8049**

表1第2行是基准网络U-Net的结果，分别在Test A、Test B上单独测试Dice和IOU评分。第3行是基于U-Net改进之后的Local分支的结果，在增大编码时的感受野后，网络获得较以往更大的局部信息提取范围，在Glas这样的大目标的数据集中，网络也能取得了非常好的性能。第4行是在Local UNet的基础上加入Global Transformer分支后的结果，LG-v1中没有加入交叉融合模块，而只是在最后通过一个分割头连接。第5行是在LG-v1的基础上，在编码过程中加入交叉融合模块的结果，在编码过程中将Transformer和U-Net的编码信息融合在一起，更好地利用两个分支之间的优势，提取更加完善的分割信息。从表1可以看出，在U-Net的基础上加入Transformer来补充卷积神经网络所缺少的全局信息后，网络的性能提升明显。解码过程中为了避免两个分支编码信息之间的相互影响而取消交叉融合模块，因为多分支网络在训练时，不同路径的组合可以丰富特征空间，提高性能。但是在推理时，复杂的结构会降低推理速度。

第6行是在LG-v2的基础上，加入平均相关性模块的结果。与LG-v2相比，平均相关性模块的加入减少了Swin Transformer中分割操作带来的影响，网络在性能上也获得了0.6个百分点的增幅。

网络输出时选择双通道输出，而不是二分类中常见的单通道输出，是因为在Glas数据集上的对比实验结果表明，双通道结果要优于单通道结果，如表2所示。实验数据使用Glas数据集，实验参数设置采样设定的统一参数。表2中，-2表示双通道输出，-1表示单通道输出。从表2可以看出，在双分支的LG UNet中，双通道输出的效果要远远优于单通道输出时的效果。

表2 双通道与单通道输出结果对比

Table 2 Double channel and single channel output results are compared

Model	Dice		IOU	
	Test A	Test B	Test A	Test B
U-Net -2	0.8826	0.8269	0.7934	0.7216
U-Net -1	0.8819	0.8539	0.7956	0.7606
LG UNet -2	**0.9362**	**0.8844**	**0.8817**	**0.8049**
LG UNet -1	0.9100	0.8663	0.8380	0.7858

3.5 实验结果

为了验证网络的效果，在Glas数据集和Kvasir-SEG数据集上分别与其他算法进行对比实验，结果如表3和表4所示。ResUNet将ResNet与U-Net结合，在U-Net编码过程中加入残差连接和Bottleneck模块。MedT是文献[17]中提出的用于医学分割的纯自注意力网络模型，MedT通过在轴注意力模块中加入门控机制，构成了纯Transformer模块组成编码网络，而且还提出了Local-Global（LoGo）训练方式，类似于双分支的组织模式，将需要大规模预训练的自注意力机制应用于数据量较少的医学图像分割任务中。

表3 Glas数据集上与其他算法的对比结果

Table 3 Comparison results with other algorithms on Glas dataset

Model	Dice			IOU		
	Test A	Test B	Mean	Test A	Test B	Mean
U-Net	0.8826	0.8269	0.8547	0.7934	0.7216	0.7575
ResUNet	0.8940	0.8459	0.8699	0.8146	0.7461	0.7803
KiU-Net[28]	0.8898	0.8527	0.8712	0.8065	0.7565	0.7815
MedT	0.8674	0.8051	0.8362	0.7711	0.6889	0.7300
Rota-Net[29]	0.919	0.849	0.884			
U-Node[30]	0.8930	0.8420	0.8675			
LG UNet	**0.9362**	**0.8844**	**0.9103**	**0.8817**	**0.8049**	**0.8433**

表4 Kvasir-SEG数据集上与其他算法的对比结果

Table 4 Comparison results with other algorithms on Kvasir-SEG dataset

Model	Dice	IOU
	Test A	Test A
U-Net	0.8110	0.7250
KiU-Net[28]	0.7724	0.6689
ResUNet	0.8306	0.7477
ResUNet++	0.8508	0.7699
ColonSegNet[31]	0.8206	
MedT	0.8039	0.7116
LG UNet	**0.8563**	**0.7782**

KiU-Net[28]也是一个双分支网络，由一个KiU-Net分支和一个U-Net分支组成，其中KiU-Net分支在编码过程中会将特征图上采样，然后提取特征，解码时再通过下采样调整回输入分辨率，与U-Net的过程正好相反。Rota-Net[29]是Glas数据集上目前能取得的最好结果。将不同网络在相同的训练参数下进行训练，得到的对比结果如表3所示。另外又对各个网络在Glas数据集、Kvasir-SEG数据集和NoMuSeg数据集的测试集上的实际输出结果进行了对比，结果如图4所示，图中最左边的一列是输入的图像，最右边的一列是对应的标签。

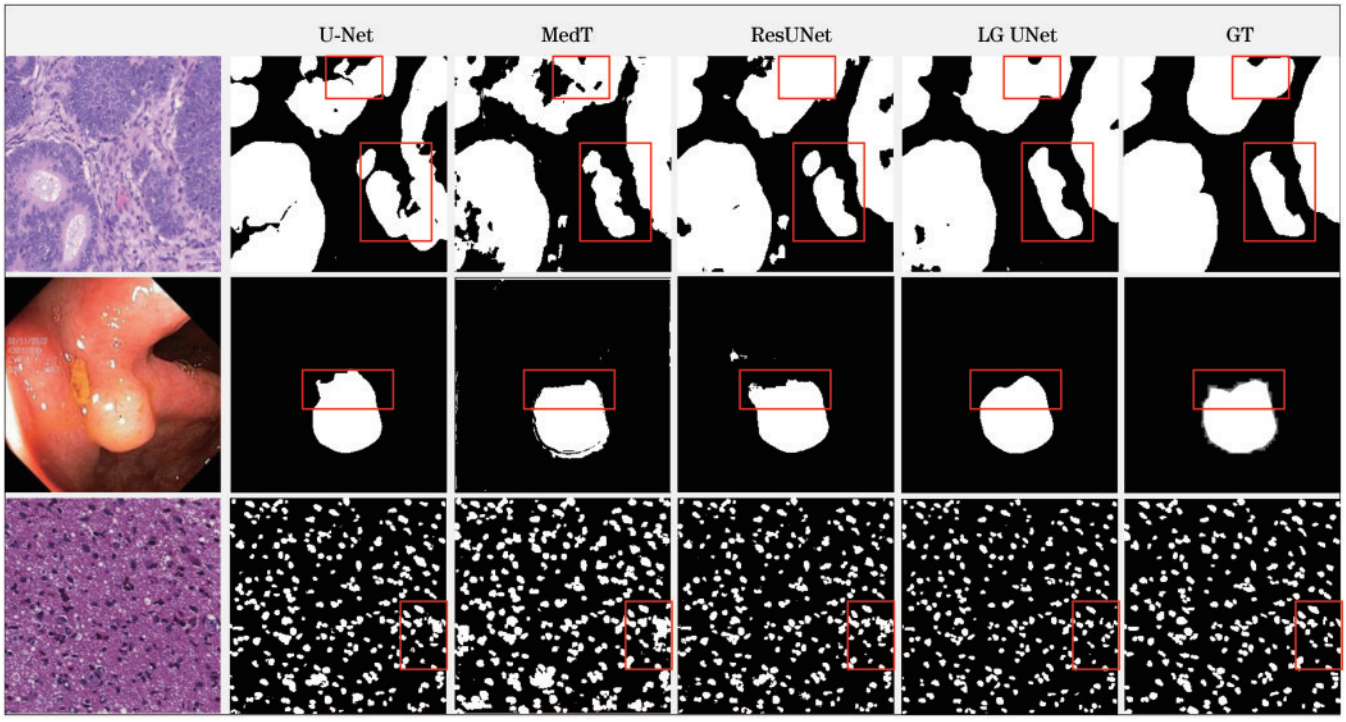

图 4　各网络输出结果比对

Fig. 4　Comparison of output results of each network

由表 3 的结果对比可知，Local 分支在增大了感受野后，效果要比同样加入残差连接的 ResUNet 更佳。而 KiU-Net 这种先上采样的训练方式在中间过程中产生的特征图较大，容易导致计算量过大的问题。而该网络为了减少计算量，将输入图像的分辨率下调，这样反而容易丢失原始图像的一些细节。从图 4 可以看出，在通过自注意力模块导入全局信息后，分割的结果在目标的边界和形状上相比卷积神经网络要更加精细。这是因为图像的背景是分散的，学习与背景相对应的像素之间的长期依赖关系可以帮助网络降低误分率。同样，当分割掩码较大时，学习对应掩码像素之间的长期依赖关系也有助于进行有效预测。而且从 MedT 的结果可知，由纯粹的自注意力机制组成的网络，在缺少预训练时，不能很好地发挥其优势。

表 4 是在 Kvasir-SEG 数据集下进行的对比实验。Kvasir-SEG 数据集中分割难度较大，因为肠道中息肉与肠道并无太过明显的区别，而且息肉的形状变化较大、小目标多，这些都增加了分割的难度。但从表 4 可以看出，LG UNet 依旧取得了不错的效果。

在其他腺体和息肉分割数据集上也同样进行了简单的比较，结果如表 5 所示。ClinicDB[32] 数据集是从结肠镜检查视频中提取的息肉分割数据集，该数据集由 612 张图像和对应的标签组成。NoMuSeg 数据集是从多家医院诊断的不同器官肿瘤患者的组织图像中提取的，由 30 张训练图像和 14 张测试图像组成。在这个两个数据集上，在给定的相同训练参数下训练。与基准网络 U-Net 相比，LG UNet 同样取得了不错的效果。

表 5　其他数据集的结果比较

Table 5　Comparisons on other data sets

Model	ClinicDB		NoMuSeg	
	Dice	IOU	Dice	IOU
U-Net	0.7926	0.7115	0.7930	0.7188
LG UNet	**0.8607**	**0.7946**	**0.8317**	**0.7126**

4　结　　论

针对腺体和息肉分割数据集中分割目标较大的问题，提出了基于 U-Net 的几种改进方法，增大了网络编码时的感受野，让网络能更好地处理大目标分割任务，并由此组建了 Local UNet 分支。又因为腺体和息肉分割中，目标边界与背景颜色太过相近，目标边界曲折，小目标游离在大目标旁边容易误分等原因，引入 Swin Transformer 中的全局能力来精确分割边界。但不同于将自注意力机制作为组件加入 U-Net，或者纯粹由自注意力机制组成分割网络，本研究采用双分支的方式将卷积运算和自注意力的优势通过交叉融合结合在一起。从输出结果可看出，加入全局信息后，分割目标的边缘、形状及位置之类的细节要更优。另外也发现用卷积操作的优势去补充自注意力模块预训练需求的方法是可行的，相比纯自注意力的网络，混合网络效果上要更好。

但目前的工作，并不能在大多数、不同类型的医学图像分割数据集上都取得优异的效果，仅对于大目标的分割任务有明显的提升。比如在分割 NoMuSeg 这

样小而多的目标时，LG UNet的提升并不明显。因此后续仍需继续修改网络的结构，使LG UNet能适应更多类型的医学分割数据集，提升其泛化能力。

参 考 文 献

[1] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[M]//Navab N, Hornegger J, Wells W M, et al. Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science. Cham: Springer, 2015, 9351: 234-241.

[2] Çiçek Ö, Abdulkadir A, Lienkamp S S, et al. 3D U-net: learning dense volumetric segmentation from sparse annotation[M]//Ourselin S, Joskowicz L, Sabuncu M R, et al. Medical image computing and computer-assisted intervention-MICCAI 2016. Cham: Springer, 2016, 9901: 424-432.

[3] Milletari F, Navab N, Ahmadi S A. V-net: fully convolutional neural networks for volumetric medical image segmentation[C]//2016 Fourth International Conference on 3D Vision (3DV), October 25-28, 2016, Stanford, CA, USA. New York: IEEE Press, 2016: 565-571.

[4] Isensee F, Jäger P F, Kohl S A A, et al. Automated Design of Deep Learning Methods for Biomedical Image Segmentation[EB/OL]. (2019-04-17)[2021-09-30]. https://arxiv.org/abs/1904.08128.

[5] Xiao X, Lian S, Luo Z M, et al. Weighted res-UNet for high-quality retina vessel segmentation[C]//2018 9th International Conference on Information Technology in Medicine and Education (ITME), October 19-21, 2018, Hangzhou, China. New York: IEEE Press, 2018: 327-331.

[6] Xiao Z Y, Du N M, Liu J J, et al. SR-Net: a sequence offset fusion net and refine net for undersampled multislice MR image reconstruction[J]. Computer Methods and Programs in Biomedicine, 2021, 202: 105997.

[7] Xiao Z Y, He K H, Liu J J, et al. Multi-view hierarchical split network for brain tumor segmentation [J]. Biomedical Signal Processing and Control, 2021, 69: 102897.

[8] 刘一鸣, 肖志勇. 基于特征融合的肝脏肿瘤自动分割方法[J]. 激光与光电子学进展, 2021, 58(14): 1417001.
Liu Y M, Xiao Z Y. Automatic segmentation algorithm of liver tumor based on feature fusion[J]. Laser & Optoelectronics Progress, 2021, 58(14): 1417001.

[9] Oktay O, Schlemper J, Folgoc L L, et al. Attention U-net: learning where to look for the pancreas[EB/OL]. (2018-04-11)[2021-09-15]. https://arxiv.org/abs/1804.03999.

[10] Hu C, Kang G X, Hou B B, et al. Acu-net: a 3D attention context U-Net for multiple sclerosis lesion segmentation[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, May 4-8, 2020, Barcelona, Spain. New York: IEEE Press, 2020: 1384-1388.

[11] 张文秀, 朱振才, 张永合, 等. 基于残差块和注意力机制的细胞图像分割方法[J]. 光学学报, 2020, 40(17): 1710001.
Zhang W X, Zhu Z C, Zhang Y H, et al. Cell image segmentation method based on residual block and attention mechanism[J]. Acta Optica Sinica, 2020, 40 (17): 1710001.

[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[EB/OL]. (2017-06-17) [2021-09-30]. https://arxiv.org/abs/1706.03762v1.

[13] Chen J N, Lu Y Y, Yu Q H, et al. TransUNet: transformers make strong encoders for medical image segmentation[EB/OL]. (2021-02-01)[20 21-09-30]. https://arxiv.org/abs/2102.04306[LinkOut]

[14] Cao H, Wang Y Y, Chen J, et al. Swin-UNet: UNet-like pure transformer for medical image segmentation [EB/OL]. (2021-05-12)[2021-09-15]. https://arxiv.org/abs/2105.05537.

[15] Gao Y H, Zhou M, Metaxas D N. UTNet: a hybrid transformer architecture for medical image segmentation [M]//de Bruijne M, Cattin P C, Cotin S, et al. Medical image computing and computer assisted intervention-MICCAI 2021. Lecture notes in computer science. Cham: Springer, 2021, 12903: 61-71.

[16] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention [EB/OL]. (2021-01-15)[2021-09-15]. https://arxiv.org/abs/2012.12877v2.

[17] Valanarasu J M J, Oza P, Hacihaliloglu I, et al. Medical transformer: gated axial-attention for medical image segmentation[EB/OL]. (2021-02- 21)[2021-09-30]. https://arxiv.org/abs/2102.1066 2.

[18] Ho J, Kalchbrenner N, Weissenborn D, et al. Axial attention in multidimensional transformers [EB/OL]. (2019-12-20)[2021-09-15]. https://arxiv.org/abs/1912.12180.

[19] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[20] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA. New York: IEEE Press, 2015: 15523970.

[21] Lou A G, Guan S Y, Loew M. DC-UNet: rethinking the U-Net architecture with dual channel efficient CNN for medical image segmentation[J]. Proceedings of SPIE, 2021, 11596: 758-768.

[22] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2020-10-2 2) [2021-09-30]. https://arxiv.org/abs/2010.11929.

[23] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2021: 9992-10002.

[24] Han K, Xiao A, Wu E H, et al. Transformer in transformer[EB/OL]. (2021-02-23)[2021-09-15]. https://

arxiv.org/abs/2103.00112.

[25] Rivière B, Hönig W, Yue Y S, et al. GLAS: global-to-local safe autonomy synthesis for multi-robot motion planning with end-to-end learning[J]. IEEE Robotics and Automation Letters, 2020, 5(3): 4249-4256.

[26] Jha D, Smedsrud P H, Riegler M A, et al. Kvasir-SEG: A segmented polyp dataset[M]//Ro Y M, Cheng W H, Kim J, et al. MultiMedia modeling. Lecture notes in computer science. Cham: Springer, 2019, 11962: 451-462.

[27] Loshchilov I, Hutter F. Fixing weight decay regularization in Adam[EB/OL]. (2017-11-14) [2021-09-15]. https://arxiv.org/abs/1711.05101v1.

[28] Valanarasu J M J, Sindagi V A., et al. KiU-Net: towards accurate segmentation of biomedical images using over-complete representations [EB/OL]. (2020-06-08)[2021-09-15] https://arxiv.org/abs/2006.04878v2.

[29] Graham S, Epstein D, Rajpoot N. Rota-net: rotation equivariant network for simultaneous gland and lumen segmentation in colon histology images[M]//Reyes-Aldasoro C C, Janowczyk A, Veta M, et al. Digital pathology. Cham: Springer, 2019, 11435: 109-116.

[30] Pinckaers H, Litjens G. Neural ordinary differential equations for semantic segmentation of individual colon glands[EB/OL]. (2019-10-23) [2021-09-30]. https://arxiv.org/abs/1910.10470.

[31] Jha D, Ali S, Tomar N K, et al. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning[J]. IEEE Access, 2021, 9: 40496-40510.

[32] Bernal J, Sánchez F J, 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.