doi: 10.13413/j.cnki.jdxblxb.2023402

基于 CNN 和 Transformer 并行编码的腹部多器官图像分割

赵 欣[1], 李 森[1,2], 李智生[2]
(1. 大连大学 信息工程学院, 辽宁 大连 116622;
2. 中国人民解放军 91550 部队, 辽宁 大连 116023)

摘要: 针对现有方法在腹部中小器官图像分割性能方面存在的不足, 提出一种基于局部和全局并行编码的网络模型用于腹部多器官图像分割. 首先, 设计一种提取多尺度特征信息的局部编码分支; 其次, 全局特征编码分支采用分块 Transformer, 通过块内 Transformer 和块间 Transformer 的组合, 既捕获了全局的长距离依赖信息又降低了计算量; 再次, 设计特征融合模块, 以融合来自两条编码分支的上下文信息; 最后, 设计解码模块, 实现全局信息与局部上下文信息的交互, 更好地补偿解码阶段的信息损失. 在 Synapse 多器官 CT 数据集上进行实验, 与目前 9 种先进方法相比, 在平均 Dice 相似系数(DSC)和 Hausdorff 距离(HD)指标上都达到了最佳性能, 分别为 83.10%和 17.80 mm.

关键词: 多器官图像分割; 分块 Transformer; 特征融合
中图分类号: TP391.41 **文献标志码**: A **文章编号**: 1671-5489(2024)05-1145-10

Abdominal Multi-organ Image Segmentation Based on Parallel Coding of CNN and Transformer

ZHAO Xin[1], LI Sen[1,2], LI Zhisheng[2]
(1. *School of Information Engineering, Dalian University, Dalian* 116622, *Liaoning Province, China*;
2. *Chinese People's Liberation Army* 91550, *Dalian* 116023, *Liaoning Province, China*)

Abstract: Aiming at the shortcomings of existing methods in the image segmentation performance of small and medium-sized organs in the abdomen, we proposed a network model based on local and global parallel coding for multi-organ image segmentation in the abdomen. Firstly, a local coding branch was designed to extract multi-scale feature information. Secondly, the global feature coding branch adopted the block Transformer, which not only captured the global long distance dependency information but also reduced the computation amount through the combination of intra-block Transformer and inter-block Transformer. Thirdly, a feature fusion module was designed to fuse the context information from two coding branches. Finally, the decoding module was designed to realize the interaction between global information and local context information, so as to better compensate for the information loss in the decoding stage. Experiments were conducted on the Synapse multi-organ CT dataset, compared with the current nine advanced methods, the average Dice similarity

收稿日期: 2023-09-25.
第一作者简介: 赵 欣(1974—), 女, 汉族, 博士, 副教授, 从事人工智能和数字医学图像处理的研究, E-mail: zhaoxin@dlu.edu.cn.
基金项目: 国家自然科学基金(批准号: 61971424).

coefficient (DSC) and Hausdorff distance (HD) indicators achieve the best performance, with 83.10% and 17.80 mm, respectively.

Keywords: multi-organ image segmentation; block Transformer; feature fusion

腹部器官的CT(computed tomography)影像分割是实现腹部疾病智能诊断的前提和基础. 通过分割出感兴趣的腹部器官区域, 可以使智能诊断系统的注意力集中在特定的器官上, 使系统更易检测到器官内潜在的异常病变, 为后续的病灶自动分析与评估提供依据. 腹部多器官分割对放化疗计划的制定也非常重要, 分割出感兴趣的腹部器官区域可以帮助放疗医生评估病变与周围器官的关系, 以确定适当的放射治疗剂量和照射范围. 腹部多器官分割结果还可与实时手术导航相结合, 为外科医生提供精确的导航和定位信息, 减少手术风险和损伤周围健康组织的可能性. 因此, 设计一种高效、准确的腹部多器官图像分割模型对辅助临床治疗有重要意义.

目前, 腹部多器官图像分割方法可分为传统机器学习方法和深度学习方法两大类. 早期的研究主要集中在单个大器官图像的分割上, 如肝脏或肾脏. 此后, 多器官分割研究引起广泛关注. 传统方法多采用基于器官结构先验知识或多图谱匹配的方法分割腹部器官图像. 但这些方法需要人工参与, 导致图像分割结果依赖于人为操作, 准确性较低. 近年来, 深度学习在医学图像分割领域取得极大进展, 目前已有许多基于深度学习的腹部多器官图像分割方法[1-4], 如文献[5]在Unet基础上通过使用可分离卷积代替普通卷积和一种自定义加权训练损失, 在腹部图像中完成了肝脏、肾脏、肺和膀胱的分割. 由于卷积操作的局部感受野限制, 因此单独使用卷积神经网络(CNN)进行腹部多器官分割可能无法充分捕捉到全局上下文信息. 而腹部器官之间的相互关系和空间布局对准确分割至关重要. 仅使用局部感受野的特征提取无法捕获到器官的全局结构和彼此之间的相互关系. 为克服该问题, 研究人员提出了一些解决方案, 其中常用的方法是引入跳跃连接或多尺度特征融合机制. 跳跃连接可以在不同层级上连接局部和全局特征, 使网络能同时获得局部和全局上下文信息. 多尺度特征融合则通过在不同尺度上进行特征提取, 并将这些特征进行融合, 以获得更全面的信息. 如Sinha等[1]提出了一种多尺度引导注意力机制算法(MS-Dual), 使用多个ResNet提取不同尺度的特征, 并在跳跃连接处用注意力机制挖掘这些特征完成了肝脏和肾脏的分割. 文献[4]提出了UCTransNet网络模型, 该模型以U-Net为基础, 在跳跃连接处先用通道式交叉融合转换器(CCT)整合多尺度上下文信息, 再利用通道交叉注意力(CRES)解决语义层次的不一致问题, 以提升分割性能. 自然语言处理中的Transformer已被证明在学习计算机视觉的全局上下文特征方面有效[5], 于是有研究人员开始关注Transformer在腹部多器官图像分割中的研究[6-10], 在肝脏、脾脏等大型器官的图像分割上取得了进一步提升. 如Chen等[6]提出了TransUNet网络结构, 使用Transformer将CNN特征图编码为输入序列, 利用解码器对编码特征进行上采样, 并与高分辨率CNN特征图组合, 实现精准定位. Cao等[8]提出了Swin-UNet网络模型, 该网络主要是基于Swin Transformer block构建了编码器-解码器体系结构, 在编码器中实现了从局部到全局的自注意力机制; 在解码器中, 将全局特征上采样到输入分辨率, 并进行分割预测. Huang等[10]提出了MISSFormer网络结构, MISSFormer将输入图像划分为重叠块, 通过多层次编码器生成多尺度特征, 并在跳跃连接处通过Transformer获取局部和全局信息, 最后通过解码器进行上采样, 得到判别性分层多尺度特征. Transformer模型更注重全局信息的建模, 能更好地捕捉长距离的依赖关系, 在处理腹部多器官图像的全局特征方面更有效. 但其在处理图像局部方面能力相对较弱, 虽然通过适当的位置编码能提供一定程度的局部信息, 但相对于传统的CNN模型, Transformer在处理局部信息方面仍显不足[11].

CNN具有局部感知性, 可更好地捕获局部模式和顺序信息. 通过将CNN和Transformer相结合, 可以更充分地引入局部信息. 因此, 本文提出一种基于CNN和Transformer相结合的腹部多器官图像分割模型BCNet(block former and CNN parallel encoding of segmentation networks), 通过设计兼顾全局与局部特征提取的双分支并行编码结构融合局部和全局信息. 本文工作的创新主要体现在提出一种基于Transformer编码和带有注意力引导与残差连接的多尺度卷积编码的双线并行编码框架,

以有效融合全局和局部信息，从而更好地实现腹部多器官图像分割.

1　本文方法

如图 1 所示，BCNet 包括两个编码分支：局部特征编码分支和全局特征编码分支. 局部特征编码分支采用局部特征提取模块进行特征编码，专注局部特征提取. 全局特征编码分支侧重全局特征提取，采用分块 Transformer，通过块内 Transformer 和块间 Transformer 的组合，既捕获了全局的长距离依赖信息又降低了计算量. 模型首先基于全局与局部上下文交互信息，特征融合模块从空间维度增强 CNN 编码分支的局部细节，并抑制无关区域，从通道维度提升了 Transformer 分支的全局信息；然后将其与交互信息进行残差连接，从而实现了特征信息的有效融合；最后设计解码模块作为解码结构，解决了普通卷积对上下文信息利用不足的问题，从而提高了图像分割性能.

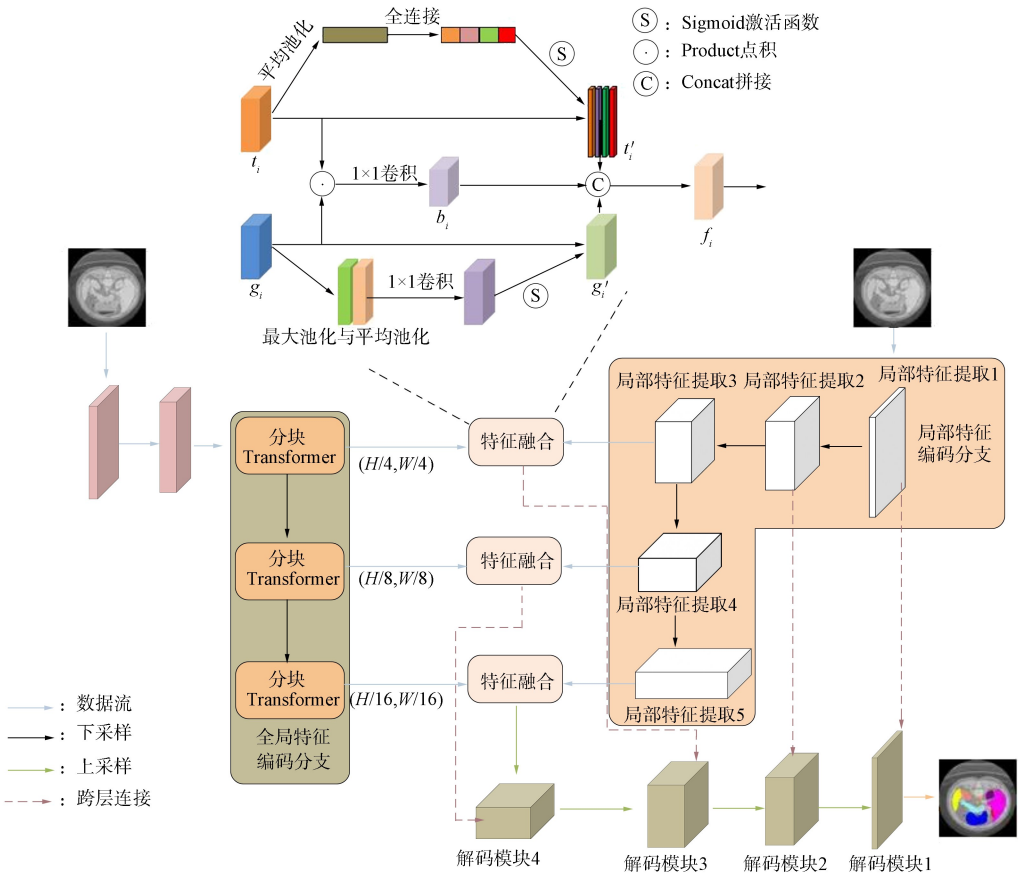


图 1　本文网络模型

Fig. 1　Proposed network model

1.1　局部特征编码模块

为提取腹部器官图像的多尺度特征，设计局部特征编码模块，该模块利用分组卷积与残差连接相结合的方式，在更细粒度级别表示多尺度特征，并通过坐标注意力[12]，使其既可以捕获多尺度特征，又可以捕获位置敏感信息，如图 2 所示. 首先，将特征图 $\boldsymbol{X}$ 分成 4 组，除第一组外其他组都有相应的 3×3 卷积核，每个 3×3 卷积操作都可以通过残差连接融合上一组的特征信息；其次，将所有特征图进行拼接，实现在更细粒度级别表示多尺度特征 $\boldsymbol{y}$，为在多尺度特征图基础上捕获位置敏感信息，两个一维全局池化操作分别将垂直和水平方向的输入特征聚合成两个独立的方向感知映射，随后将垂直和水平两个方向的注意力进行矩阵相乘，获取特征图 $\boldsymbol{y}$ 坐标注意力 p，并将坐标信息 p 嵌入到特征图 $\boldsymbol{y}$ 中，从而生成权重特征图 $\boldsymbol{X}'$；最后，将特征图 $\boldsymbol{X}'$和原图 $\boldsymbol{X}$ 进行线性融合，从而生成多尺度残差坐标注意力特征图.

1.2 分块 Transformer

传统 CNN 在处理每个像素时只考虑局部的感受野，无法建立图像中像素之间的长距离依赖关系，这种处理模式不能很好地解读腹部各器官在位置上的相对依赖关系. 但 Transformer 技术可以通过自注意力机制在全局范围内对图像进行建模，从而更好地理解像素之间的语义关系，提高图像分割的准确性. 因此，本文在模型中引入 Transformer 技术进行腹部多器官图像的全局依赖特征提取，设计一种基于分块的 Transformer 模块(B-Transformer)，通过将输入数据分成若干块，每块内进行块内 Transformer 操作，并在块间进行块间 Transformer 操作，以实现更高效的计算和更好的并行性，如图 3 所示. 全局编码分支的前两层使用步长为 2、大小为 3×3 的卷积对输入图像进行处理，以获取初步的浅层特征图，然后传输给分块 Transformer 模块，如图 1 所示. 在图 3 中，经过浅层处理后的特征图首先被输入到块内 Transformer 的特征提取流程，计算每个分块内像素间的长距离依赖关系，处理结果与浅层输入拼接后经过 3×3 的卷积进行特征融合，然后再传输给块间 Transformer 的特征提取流程，进而生成全局上下文信息.

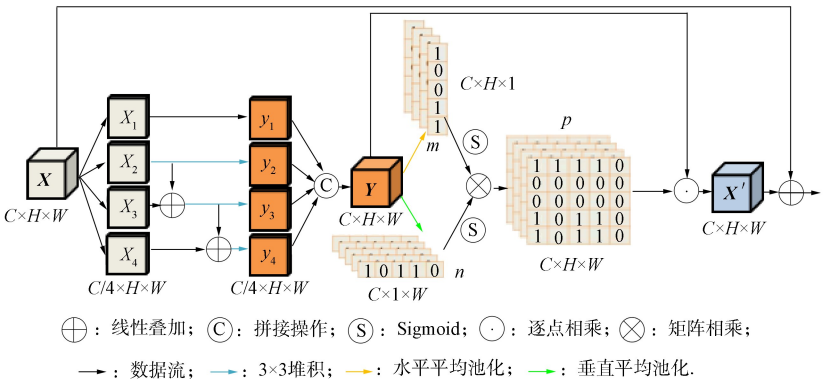


图 2 局部特征编码模块

Fig. 2 Local feature coding module

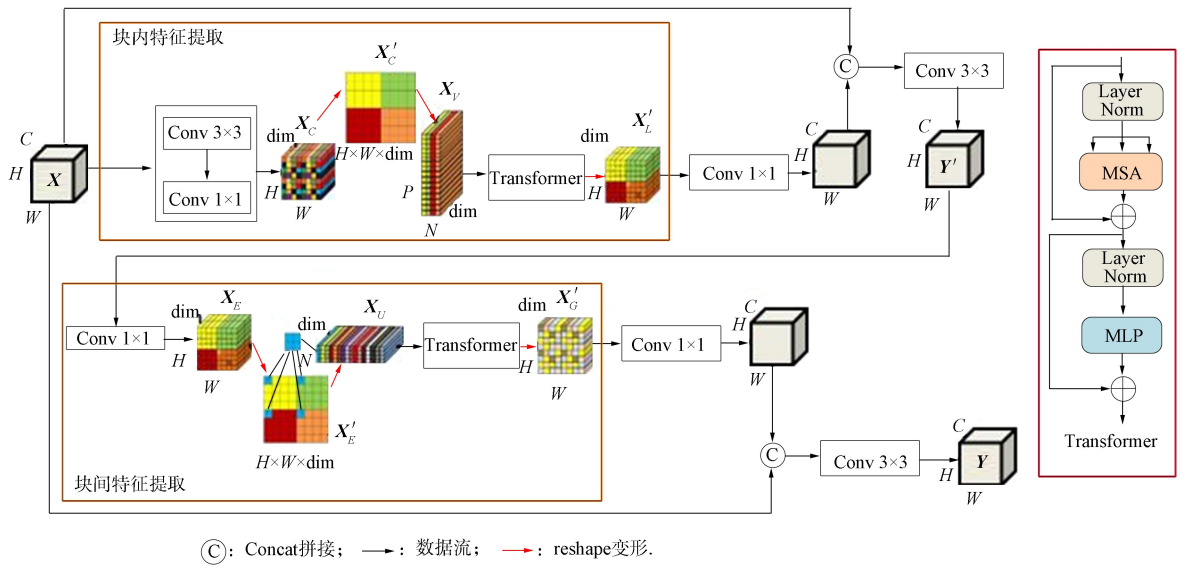


图 3 分块 Transformer 模型

Fig. 3 Block Transformer model

1.2.1 块内 Transformer

块内 Transformer 是在每个块内部进行的操作，其结构与传统的 Transformer 相似，包含自注意力机制和前馈神经网络. 块内 Transformer 将注意力放在当前块内的元素之间，使其可以在每个块内部并行计算，提高计算效率. 块内 Transformer 首先对输入特征图 $\boldsymbol{X}\in\mathbb{R}^{C\times H\times W}$ 进行 3×3 卷积，以获

取特征图中的局部空间信息和顺序信息，并利用 1×1 卷积扩大特征图的通道生成 $\boldsymbol{X}_C \in \mathbb{R}^{\dim \times H \times W}$，以增强表达能力. 对 $\boldsymbol{X}_C$ 进行分块(Patch)生成 $\boldsymbol{X}'_C \in \mathbb{R}^{H \times W \times \dim}$ 后，再将 $\boldsymbol{X}'_C$ 展开为 N 个不重叠的序列化集合 $\boldsymbol{X}_V \in \mathbb{R}^{N \times P \times \dim}$. 若 Patch 的高和宽分别为 h 和 w，则 $P=h \times w$，$N=\dfrac{H \times W}{P}$，其中 N 表示 Patch 的数量. 每个 Patch 用 Transformer[5] 计算获得块内像素间的长距离依赖信息 $\boldsymbol{X}_L(n) \in \mathbb{R}^{N \times P \times \dim}$，用公式表示为

$$\boldsymbol{X}_L(n)=\text{Transformer}(\boldsymbol{X}_V(n)), \quad 1 \leqslant n \leqslant N, \tag{1}$$

其中 $\boldsymbol{X}_V(n)$表示某个指定 Patch 的空间信息，$\boldsymbol{X}_L(n)$表示对应 Patch 的块内像素间的长距离依赖信息.

1.2.2　块间 Transformer

块间 Transformer 是在块之间进行的操作，其目的是在不同的块之间建立联系，使块之间的信息可以传递和交流，从而使不同块之间的信息得到有效整合并生成全局依赖表达. 块内 Transformer 的输出 $\boldsymbol{X}_L$ 经过与输入特征图 $\boldsymbol{X}$ 拼接，生成融合信息 $\boldsymbol{Y}' \in \mathbb{R}^{C \times H \times \dim}$. 为增强表达能力，首先对 $\boldsymbol{Y}'$进行 1×1 卷积将其映射到高维空间，生成 $\boldsymbol{X}_E \in \mathbb{R}^{\dim \times H \times W}$，为方便操作，通过对 $\boldsymbol{X}_E$ 进行 reshape 生成 $\boldsymbol{X}'_E \in \mathbb{R}^{H \times W \times \dim}$ 后，再将 $\boldsymbol{X}'_E$ 展开为 P 个不重叠的相对位置序列化集合 $\boldsymbol{X}'_U \in \mathbb{R}^{P \times N \times \dim}$，而每个相对位置序列化集合表示所有 Patch 某个相对位置对应像素的空间信息集合. 若 Patch 的高和宽分别为 h 和 w，则 $N=\dfrac{H}{h} \times \dfrac{W}{w}$，表示相对位置特征块的大小，$P=hw$，表示相对位置特征块的数量. 每个相对位置特征块用 Transformer[5] 计算获得所有 Patch 相对位置对应像素之间的长距离依赖信息 $\boldsymbol{X}_G(p) \in \mathbb{R}^{P \times N \times \dim}$，用公式表示为

$$\boldsymbol{X}_G(p)=\text{Transformer}(\boldsymbol{X}_U(p)), \quad 1 \leqslant p \leqslant P, \tag{2}$$

其中 $\boldsymbol{X}_U(p)$表示所有 Patch 某个相对位置对应像素的空间信息，$\boldsymbol{X}_G(p)$表示所有 Patch 相对位置对应像素之间的全局上下文信息. 由于在块内 Transformer 每个像素已经与块内的所有像素建立了相关性，所以 $\boldsymbol{X}_G$ 中的每个像素都可以通过 Patch 相对位置对应像素间接与 $\boldsymbol{X}_G$ 中的其他像素建立相关性，从而实现建模全局上下文. 最后与输入特征图 $\boldsymbol{X}$ 拼接，生成融合信息 $\boldsymbol{Y} \in \mathbb{R}^{C \times H \times W}$.

1.3　特征融合模块

设计特征融合模块的目的是为有效结合来自 CNN 编码的局部上下文信息及 Transformer 编码的全局上下文信息，以提高图像分割性能. 如图 4 所示，t_i 表示 Transformer 输入的特征图，g_i 表示 CNN 输入的特征图. t_i 有两条分支，一条是经过 SE 通道注意力[13] 与 g_i 经空间注意力[14] 做拼接处理；另一条是与 g_i 做点积运算，以获取来自两条编码分支的交互信息. 最后，使用 1×1 卷积对融合信息进行升通道，生成 b_i，并与拼接部分做残差处理，从而生成的融合信息 f_i 可有效地捕获当前空间分辨率的全局和局部上下文信息. 在此过程中，t_i 使用 SE 通道注意力主要是为从全局角度对 Transformer 输入的特征图进行把握，得到 t'_i；而 g_i 使用空间注意力主要是为从细节角度对 CNN 编码的特征图进行把握，得到 g'_i.

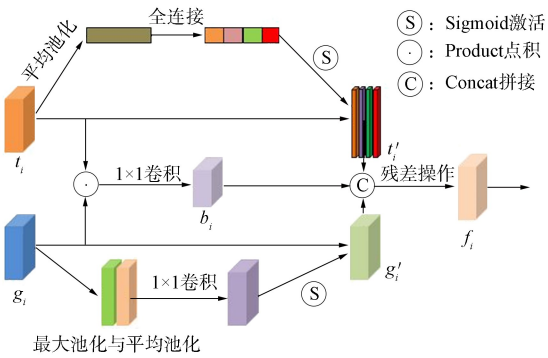


图 4　特征融合模块

Fig. 4　Feature fusion module

1.4　解码模块

由于腹部各器官图像形态差异较大，所以在解码过程中，需要充分利用上下文信息，因此本文设计了基于上下文 Transformer[13] 的解码模块，以实现特征图的局部上下文与全局上下文的有效融合. 模块的输入来自双编码分支对应层的特征融合与上采样图经过跨层连接生成的特征图 $\boldsymbol{X}$. 输入的特征图 $\boldsymbol{X}$ 分别经过 3 条不同的路径处理，如图 5 所示. 由图 5 可见：第一条路径首先对输入进行 3×3 卷积，以实现输入的静态上下文表示 L；与第二条路径的原始输入拼接后，通过两个连续的 1×1 卷积和

reshape生成动态多头注意力矩阵，该矩阵由原始输入 $\boldsymbol{X}$ 与局部上下文信息 L 交互得到，从而在局部上下文信息的引导下，各像素之间的关联性均体现在每个3×3的张量上，再对每个3×3张量进行平均池化生成自注意力图；第三条路径对输入首先进行3×3卷积，然后经过与自注意力图点乘，以实现输入的全局上下文表示 M. 最后，将静态上下文表示 L 与全局上下文 M 通过叠加进行特征信息融合，以增强在解码过程中的特征表达能力.

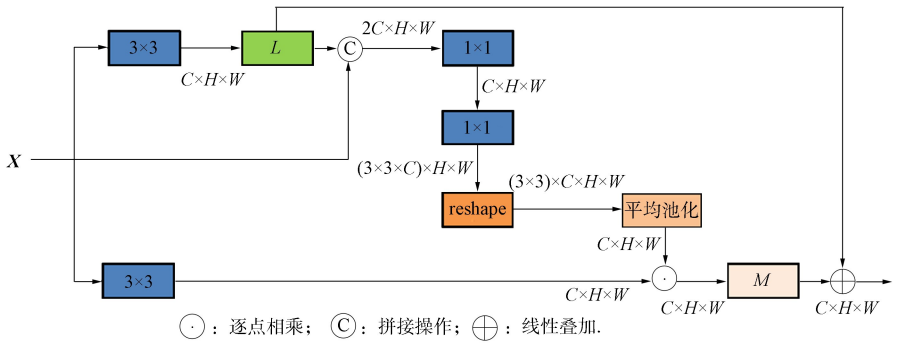


图5 解码模块

Fig. 5 Decoding module

2 实验与结果讨论

2.1 数据集

实验采用的数据集是MICCAI 2015多图谱腹部标记挑战赛的Synapse多器官CT数据集，其中共有30名患者提供3 779张轴向腹部临床CT切片. 本文随机选取18个样本(共2 211张轴向切片)作为训练集，12个样本作为测试集，在训练过程中随机选择6个样本作为验证集. 为防止模型因数据样本过少而出现过拟合，对训练集中的图像分别进行水平翻转和90°,180°,270°旋转，将训练集中样本扩大为原来的5倍.

2.2 实验设置

本文实验使用的设备是NVIDIA GeForce RTX 3090 GPU显卡，内存为22 GB，深度学习框架为Pytorch1.9.1. 输入图像大小设为224×224，在Transformer分支编码器上patch size分别设为8,4,2，batch size设为4. 为更好地解决训练样本不平衡问题，网络模型在训练过程中使用 FL_{Softmax} 训练BCNet：

$$FL_{\text{Softmax}}=\sum_{c=1}^{n}(-\alpha_c(1-p_c)^{\gamma}\log(p_c)), \tag{3}$$

其中 c 表示当前样本的类别，α_c 表示类别 c 对应的权重，p_c 表示输出概率分布对类别 c 的概率值，超参数 γ 设为2. 采用学习率为0.01，权重衰减为0.000 1的SGD优化器进行训练，批处理量(batch size)设为4，迭代次数设为300.

2.3 评价指标

为验证本文网络模型的有效性，采用DSC(平均Dice相似系数)和HD(平均Hausdorff距离)两个性能指标对模型图像分割结果的准确性进行评估，其中：Dice是常用于评价医学图像分割的标准指标，标签与分割结果的重合面积越大，数值越大；Hausdorff距离是在度量空间中任意两个集合之间定义的一种距离，表示分割结果与标签两个点集之间最短距离的最大值. 这两个评价指标分别表示为

$$\text{Dice}=\frac{2|A\cap B|}{|A|+|B|}, \tag{4}$$

$$\text{HD}(A,B)=\max\{\max_{a\in A}\min_{b\in B}d(a,b),\max_{b\in B}\min_{a\in A}d(a,b)\}, \tag{5}$$

其中 A 和 B 分别表示标签和预测的结果，$d(a,b)$ 表示 a,b 两点之间的欧氏距离. 本文采用DSC作为评价指标，对8个腹部器官(主动脉、胆囊、左肾、右肾、肝脏、胰腺、脾脏、胃)图像进行测试.

2.4　结果与分析

为验证本文网络模型的性能和分割结果，将 BCNet 与 9 种先进的医学图像分割方法进行比较. 表 1 列出了各方法的性能指标，所有方法的结果均取自原文献. 由表 1 可见，BCNet 在评估指标平均 DSC 和 HD 上取得了最佳性能，分别为 83.10%和 17.80 mm. 与传统的 UNet 算法相比，BCNet 在平均 DSC 上提高了 6.25 个百分点. 此外，传统的 U 型编码器-解码器结构(如 U-Net 和 Attention UNet)由于卷积操作的制约，特征提取感受野大小受卷积核大小限制，难以捕获全局特征信息，从而导致肝脏、肾脏、脾脏、胃等较大器官的图像分割效果不如 UNet 与 Transformer 相结合的网络模型. 将 UNet 与 Transformer 相结合的网络模型在处理局部信息方面仍显不足，导致在小型器官(如胆囊、主动脉)的图像分割方面性能并不理想. 而 BCNet 成功地解决了 Transformer 忽略局部特征的缺点，并结合了传统卷积编码的优势，在腹部多器官图像分割方面性能优异. 特别是在中小型器官(如胆囊、左肾、右肾)的分割方面，与 MISSFormer 相比，BCNet 的分割精度分别提高了 4.10，3.92，6.78 个百分点.

表 1　不同方法在 Synapse 多器官 CT 数据集上的分割效果对比

Table 1　Comparison of segmentation effects of different methods on Synapse multi-organ CT dataset

模型	平均 DSC/%	HD/mm	DSC/%							
			主动脉	胆囊	左肾	右肾	肝脏	胰腺	脾脏	胃
UNet	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.48	86.67	75.58
AttentionUNet[15]	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
TransUNet[6]	77.48	31.69	87.32	63.16	81.87	77.02	94.08	55.86	85.08	75.62
UCTransNet[4]	78.23	26.74	88.86	66.97	80.18	73.17	93.16	56.22	87.84	79.43
MTUnet[16]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
Swin-UNet[8]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
TransClaw UNet[9]	78.09	26.38	85.87	61.38	84.83	79.36	94.28	57.65	87.74	73.55
MISSFormer[10]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
CoT-TransUnet-50[17]	78.24	23.75	88.69	62.56	83.33	76.91	94.57	55.23	86.35	78.28
本文	83.10	17.80	88.01	72.75	89.13	88.78	95.50	64.34	86.64	79.68

由于部分文献采用的数据集划分是 21 个样本作为训练集，9 个样本作为测试集，训练过程中选择 6 个样本作为验证集，并用 DSC 作为评价指标. 现将 BCNet 做相同的数据集划分进行对比实验，并对训练集进行数据增强. 表 2 列出了重新划分数据集后不同方法在 Synapse 多器官 CT 数据集上的分割效果对比. 由表 2 可见，BCNet 在保证肝脏、肾脏、脾脏精度的同时，在主动脉、胆囊、胰腺以及胃部的器官图像分割上有进一步提升，并且平均 DSC 相比 UNETR 提升了 1.55 个百分点.

表 2　重新划分数据集后不同方法在 Synapse 多器官 CT 数据集上的分割效果对比

Table 2　Comparison of segmentation effects of different methods on Synapse multi-organ CT dataset after repartitioning dateset

%

模型	平均 DSC	DSC							
		主动脉	胆囊	左肾	右肾	肝脏	胰腺	脾脏	胃
nnUNet[18]	84.73	87.7	70.4	91.0	89.4	94.8	68.0	94.2	82.4
CoTr w/o CNN encoder[19]	84.48	87.6	70.5	90.9	89.4	94.8	67.1	94.1	81.5
CoTr*[19]	87.95	91.4	68.7	92.9	92.4	96.2	78.3	94.3	89.4
CoTr[19]	87.83	92.0	70.0	93.6	92.1	96.3	77.5	95.8	85.4
UNETR[20]	89.05	89.0	75.0	94.1	92.4	97.1	76.7	96.8	91.3
本文	90.60	92.8	86.4	92.4	90.6	96.8	82.4	92.1	91.3

2.5　视觉效果

为清楚地显示本文方法(BCNet)的效果，可视化了本文方法与其他方法的分割效果，如图 6 所示. 由图 6 可见，传统的卷积编码器-解码器结构对边界信息不够敏感，易出现过度分割的问题. 例如，在第一行中，胃被 UNet，Attention UNet 过度分割；在第五行中，肝脏被其他模型过度分割. UNet 与

Transformer 相结合的网络模型虽然对边界信息较敏感，但由于一些特征的丢失，易出现分割不足的问题. 例如，在第四行中，Swin-Unet 未能将胆囊预测出来，而 BCNet 正确预测了胆囊；在第二行中，CoT-TransUNet,Swin-UNet 对右肾的分割出现误分，而 BCNet 正确预测了右肾；在第一行和第三行中，CoT-TransUNet,Swin-UNet,TransUNet 对胰腺的预测出现缺失，而 BCNet 正确预测了胰腺，并保留了很好的边界信息. 实验结果表明，本文方法对中小型腹部器官图像的分割效果较好，将 CNN 局部上下文信息与 Transformer 全局上下文信息进行融合，使 BCNet 更注重边界信息，实现了更好的边缘预测.

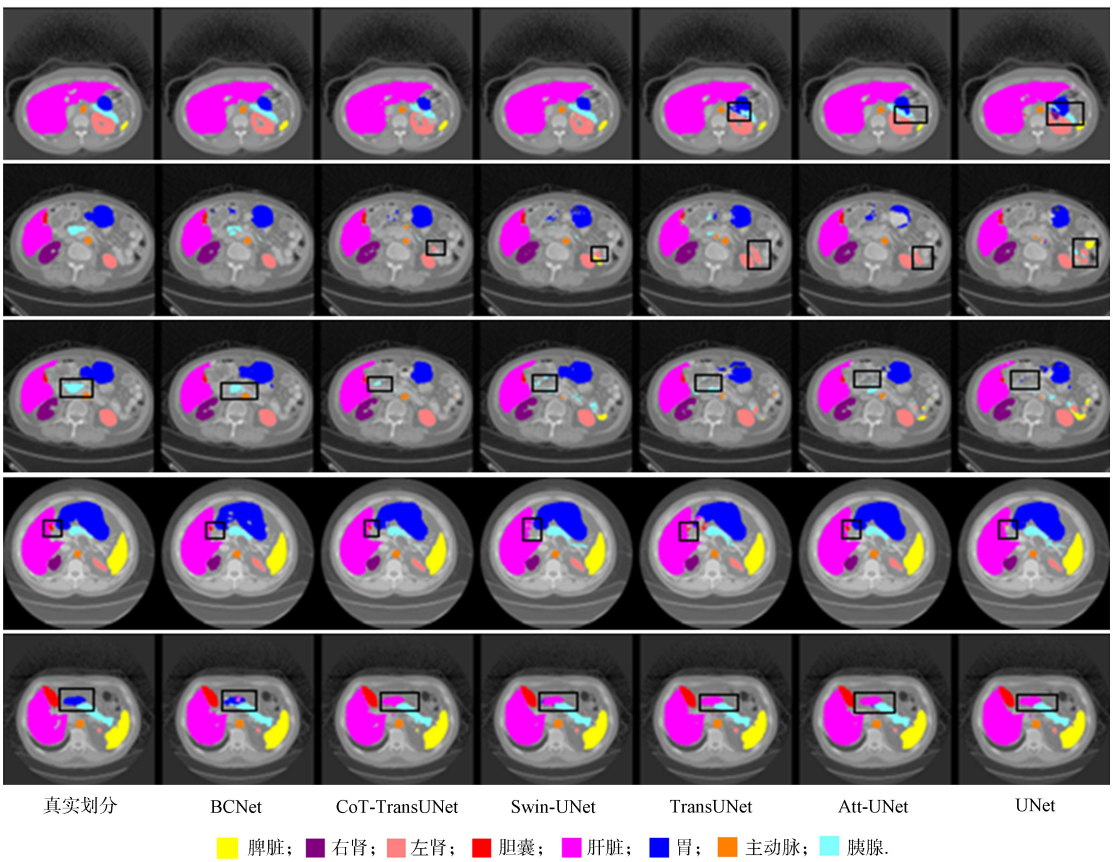


图 6 不同方法对腹部器官图像的分割对比结果

Fig. 6 Comparison of segmentation results of abdominal organ images of different methods

2.6 消融实验

为进一步研究本文算法各模块的性能，设计消融实验验证局部特征提取模块(Res)、分块 Transformer(Bvit)、解码模块(CoT)3 个关键组件的有效性. 表 3 列出了本文算法各组件消融实验结果. 为更好地提取腹部各器官图像的多尺度特征，在实验 1 中，以 UNet 作为基本模型验证其对结果的影响. 实验 2～4 中，在基本模型上添加了 Res,Bvit 和 CoT，以验证该模块的性能. 在实验 5～7 中，进行了 Res,Bvit 和 CoT 的成对组合，以验证模块之间相互作用对基本模型的影响. 在实验 8 中，将 3 个模块组合在一起，以验证 Res,Bvit 和 CoT 之间的相互作用可提高基本模型的预测精度.

由表 3 可见，实验 2 在 Synapse 多器官 CT 数据集上的性能优于实验 1，说明使用多尺度残差坐标注意力模块进行编码，不仅可以进一步提取多尺度特征，而且可以从位置信息角度提升图像分割效果. 通过比较实验 1 和实验 4 以及实验 5 和实验 8 可以发现，平均 DSC 显著提高，平均 Hausdorff 距离显著降低. 这主要是因为 CoTNet 相较于普通卷积可以充分利用局部上下文信息，从而更好地完成解码过程. 通过比较实验 1 和实验 3 表明，以并行方式结合 Transformer 和 CNN 对图像进行编码比单一使用 CNN 对图像进行编码效果更好. 由实验 6 和实验 8 可见，通过引入 Bvit 可以在不损失大型器官图像分割准确性的情况下，极大提高中小型器官图像的分割性能，尤其像胰腺这种边界模糊的

器官.

表 3　本文算法各组件消融实验结果

Table 3　Experimental results of ablation of each component of proposed algorithm

实验编号	Res	Bvit	Cot	平均 DSC/%	HD/mm	DSC/%						
						主动脉	胆囊	左肾	右肾	肝脏	胰腺	脾脏	胃
1	—	—	—	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.48	86.67	75.58
2	√	—	—	79.47	24.47	86.06	66.80	85.82	88.96	95.20	55.19	84.33	73.45
3	—	√	—	81.57	21.76	88.46	71.29	87.30	88.81	94.93	62.94	83.81	75.09
4	—	—	√	81.37	21.29	87.68	72.14	86.15	87.94	95.54	62.70	83.65	75.16
5	√	√	—	82.12	19.78	88.03	72.58	87.15	88.29	95.51	62.35	85.35	77.70
6	√	—	√	81.67	20.79	87.47	73.21	88.40	89.69	95.76	60.26	84.22	74.36
7	—	√	√	82.68	19.62	89.30	73.99	86.67	87.64	95.27	62.78	83.97	79.52
8	√	√	√	83.10	17.80	88.01	72.75	89.13	88.78	95.50	64.34	86.64	79.68

表 4 列出了不同模型的性能及计算复杂度. 由表 4 可见，在与目前几种主流的网络进行比较后，相比于其他引入注意力机制的网络，例如 TransUNet 和 Swin-UNet，BCNet 的参数数量更少，推理时间更快，计算复杂度(GFLOPs)更低，更轻量化. 说明分块 Transformer 不仅可以保证模型图像分割的精确度，而且可以减少计算量.

表 4　不同模型的性能及计算复杂度

Table 4　Performance and computational complexity of different models

模型	参数数量/M	推理速度/ms	计算复杂度
UNet	31.13	223	55.84
TransUnet[6]	105.32	246	38.52
Swin-Unet[8]	96.34	238	42.68
BCNet	60.53	213	32.46

综上所述，针对现有方法在腹部中小器官图像分割性能方面存在的不足，本文提出了一种新的融合 CNN 与 Transformer 双重编码结构的网络模型(BCNet). BCNet 由局部特征编码分支和全局特征编码分支组成，通过特征互补模块实现融合. 使用解码模块进行解码，克服了普通卷积对上下文信息利用不足的问题，提高了分割性能，可为辅助医生诊断提供有效的帮助.

参　考　文　献

[1] SINHA A, DOLZ J. Multi-scale Self-guided Attention for Medical Image Segmentation [J]. IEEE Journal of Biomedical and Health Informatics, 2020, 25(1): 121-130.

[2] BERZOINI R, COLOMBO A A, BARDINI S, et al. An Optimized U-Net for Unbalanced Multi-organ Segmentation [C]//2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). Piscataway, NJ: IEEE, 2022: 3764-3767.

[3] LU H C, TIAN S W, YU L, et al. Medical Image Segmentation Using Boundary-Enhanced Guided Packet Rotation Dual Attention Decoder Network [J]. Technology and Health Care, 2022, 30(1): 129-143.

[4] WANG H N, CAO P, WANG J Q, et al. Uctransnet: Rethinking the Skip Connections in U-Net from a Channel-Wise Perspective with Transformer [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022: 2441-2449.

[5] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale [EB/OL]. (2020-10-22)[2023-02-01]. https://arxiv.org/abs/2010.11929.

[6] CHEN J N, LU Y Y, YU Q H, et al. Transunet: Transformers Make Strong Encoders for Medical Image Segmentation [EB/OL]. (2021-02-08)[2023-03-10]. https://arxiv.org/abs/2012.04306.

[7] PETIT O, THOME N, RAMBOUR C, et al. U-Net Transformer: Self and Cross Attention for Medical Image Segmentation [C]//Machine Learning in Medical Imaging: 12th International Workshop. Berlin: Springer, 2021:

267-276.

[8] CAO H, WANG Y Y, CHEN J, et al. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation [C]//European Conference on Computer Vision. Berlin: Springer, 2022: 205-218.

[9] YAO C, HU M H, ZHAI G T, et al. Transclaw U-Net: Claw U-Net with Transformers for Medical Image Segmentation [EB/OL]. (2021-07-12)[2023-03-12]. https://arxiv.org/abs/2107.05188.

[10] HUANG X H, DENG Z F, LI D D, et al. Missformer: An Effective Medical Image Segmentation Transformer [EB/OL]. (2021-09-15)[2023-01-15]. https://arxiv.org/abs/2109.07162.

[11] LIU Y, ZHANG Y, WANG Y X, et al. A Survey of Visual Transformers [J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(6): 7478-7498.

[12] HOU Q B, ZHOU D Q, FENG J S. Coordinate Attention for Efficient Mobile Network Design [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 13713-13722.

[13] LI Y H, YAO T, PAN Y W, et al. Contextual Transformer Networks for Visual Recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 1489-1500.

[14] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional Block Attention Module [C]//Proceedings of the European Conference on Computer Vision (ECCV). New York: ACM, 2018: 3-19.

[15] OKTAY O, SCHLEMPER J, FOLGOC L L, et al. Attention U-Net: Learning Where to Look for the Pancreas [EB/OL]. (2018-04-11)[2022-12-15]. https://arxiv.org/abs/1804.03999.

[16] ZHENG P F. MT-ONet: Mixed Transformer O-Net for Medical Image Segmentation [C]//2022 International Conference on Sensing, Measurement & Data Analytics in the Era of Artificial Intelligence (ICSMD). Piscataway, NJ: IEEE, 2022: 1-4.

[17] 杨鹤, 柏正尧. CoT-TransUNet: 轻量化的上下文 Transformer 医学图像分割网络 [J]. 计算机工程与应用, 2023, 59(3): 218-225. (YANG H, BAI Z Y. CoT-TransUNet: Lightweight Contextual Transformer Medical Image Segmentation Network [J]. Journal of Computer Engineering & Applications, 2023, 59(3): 218-225.)

[18] 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)[2023-01-15]. https://arxiv.org/abs/1904.18128.

[19] XIE Y T, ZHANG J P, SHEN C H, et al. Cotr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation [C]//Medical Image Computing and Computer Assisted Intervention. Berlin: Springer, 2021: 171-180.

[20] HATAMIZADEH A, TANG Y, NATH V, et al. Unetr: Transformers for 3D Medical Image Segmentation [C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway, NJ: IEEE, 2022: 574-584.

(责任编辑: 韩 啸)