计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(6):21−32 [doi: 10.15888/j.cnki.csa.009872] [CSTR: 32024.14.csa.009872]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

MCCNET: 特征增强的双分支多器官图像分割模型①

郭俊林[1], 陈平华[1], 陈一嘉[1], 詹晗晖[2]

[1](广东工业大学 计算机学院, 广州 510006)
[2](北京师范大学-香港浸会大学联合国际学院, 珠海 519087)
通信作者: 陈平华, E-mail: phchen@gdut.edu.cn

摘　要: 针对腹部 CT 图像多器官分割面临的不同器官大小形态不一、相邻器官边界难以确认以及低对比度等挑战问题, 提出一种特征增强的双分支多器官分割模型. 模型总体采取编码器-解码器结构: 编码器采取主/从双分支结构, 主分支使用 Mamba 捕捉多器官全局依赖信息, 从分支使用 CNN 逐层提取多器官局部信息, 同时设计级联上下文模块将从分支局部细节特征补充到主分支中; 解码器设计多尺度特征融合模块和深度特征增强模块, 多尺度特征融合模块对跨层级特征信息进行融合, 增强多器官边界分割锐度, 深度特征增强模块应用交叉注意机制提高器官前景与背景的对比度, 减少背景信息对分割的干扰. 在 Synapse 和 ACDC 两组公开数据集上的实验结果表明, 与近几年主要基线模型相比, 所提模型的 Dice 相似系数 (*DSC*)、HD95 指标均具有一定的提升.

关键词: 多器官图像分割; Mamba; 卷积神经网络; 交叉注意力机制

引用格式: 郭俊林,陈平华,陈一嘉,詹晗晖.MCCNET: 特征增强的双分支多器官图像分割模型.计算机系统应用,2025,34(6):21–32. http://www.c-s-a.org.cn/1003-3254/9872.html

MCCNET: Feature-enhanced Dual-branch Multi-organ Image Segmentation Model

GUO Jun-Lin[1], CHEN Ping-Hua[1], CHEN Yi-Jia[1], ZHAN Han-Hui[2]

[1](School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China)
[2](Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519087, China)

Abstract: To address the challenges in multi-organ segmentation of abdominal CT images, such as varying organ sizes and shapes, difficulties in distinguishing boundaries between adjacent organs, and low contrast, this study proposes a feature-enhanced dual-branch multi-organ image segmentation model. The model adopts an encoder-decoder architecture, with a master-slave dual-branch structure in the encoder. The master branch leverages Mamba to capture global dependencies among organs, while the slave branch employs CNN to hierarchically extract local features of multiple organs. A cascade context module is introduced to transfer detailed local features from the slave branch to the master branch. In the decoder, a multi-scale feature fusion module integrates cross-level feature information to enhance boundary sharpness in multi-organ segmentation, and a deep feature enhancement module applies a cross-attention mechanism to improve the contrast between organ foregrounds and backgrounds, mitigating the interference of background noise. Experimental results on two public datasets, Synapse and ACDC, demonstrate that the proposed model achieves notable improvements in Dice similarity coefficient (*DSC*) and HD95 indexes compared to recent baseline models.

Key words: multi-organ image segmentation; Mamba; convolutional neural network (CNN); cross-attention mechanism

① 基金项目: 广东省重点领域研发计划 (2023B1111050010)
收稿时间: 2024-11-27; 修改时间: 2024-12-17; 采用时间: 2025-01-07; csa 在线出版时间: 2025-03-24
CNKI 网络首发时间: 2025-03-25

医学图像分割旨在从医学影像中自动提取目标区域, 为临床诊断和治疗提供精确的结构信息. 作为医学图像分割的关键分支, 多器官分割旨在同时分割多个不同器官, 其广泛应用于复杂疾病诊断、放射治疗计划和外科路径模拟等场合[1]. 相较于单器官分割, 多器官分割面临更多挑战, 包括各器官形态的不确定性、器官之间边界模糊以及低对比度区域的背景干扰等问题, 这对模型的鲁棒性和精度提出了更高的要求.

传统的单器官分割方法主要是机器学习方法, 依赖于手工提取特征, 如阈值分割、区域生长和支持向量机 (support vector machine, SVM) 等. 手工提取特征方法难以适应复杂的形态变化和背景干扰. 随着深度学习的发展, 基于深度学习的多器官分割方法逐渐成为研究的热点[2]. 目前, 基于深度学习的多器官分割方法大致可分为基于卷积神经网络 (convolutional neural network, CNN) 的方法、基于 Transformer 的方法、基于 CNN 与 Transformer 混合方法、基于 Mamba 的方法等 4 种类型. 基于 CNN 的方法通过卷积层自动提取图像特征, CNN 通过层级结构捕捉局部特征并逐层构建更复杂和抽象的特征表示, 然而, 这种方法在处理大尺寸器官或需要捕捉长距离依赖关系的场景时表现出不足. 基于 Transformer 的方法利用自注意力机制捕捉图像中的长距离依赖和全局特征, 这使得它在构建更复杂和抽象的特征表示方面表现出色, 然而却在提取小器官的细节特征时表现不足, 同时对计算资源和数据量要求较高[3]. 基于 CNN 与 Transformer 的混合方法结合了 CNN 的局部特征提取能力和 Transformer 的全局特征捕捉优势, 能够同时捕捉图像中的局部细节和长距离依赖关系, 从而构建更加丰富且多层次的特征表示, 然而, 简单的特征融合机制难以协调不同器官的特征差异, 尤其在边界模糊和低对比度区域时, 小器官的细微特征容易被大器官的全局信息所掩盖, 从而影响分割精度. 最近, 基于 Mamba 的多器官分割方法引起了研究人员的广泛关注, Mamba[4]作为一种现代状态空间模型 SSM (state space model) 引起研究人员的广泛研究. 它不仅在经典 SSM 研究的基础上建立了长距离依赖关系, 还表现出与输入规模线性相关的计算复杂度, 因此 Mamba 在语言理解、一般视觉任务等多个领域都得到大量应用研究. 这些研究表明, Mamba 在捕获长程依赖和动态权重调整方面具有显著优势, 尤其在密集预测任务中. 然而, 与基于 Transformer 的方法相比, Mamba 的选择机制在一定程度上限制其性能[5].

考虑到已有研究无法同时解决多器官分割中存在的器官间形态差异、分割边界模糊与低对比度等问题, 本文提出了一种特征增强的双分支多器官分割模型 MCCNET (Mamba and CNN collaborative network). 模型采取编码器-解码器架构, 编码器采取主/从双分支结构, 主分支使用 Mamba 提取全局信息和高层语义特征, 从分支使用 CNN 提取细节特征和低层次空间信息. 设计级联上下文模块 (cascading context module, CCM) 将从分支局部细节特征补充到主分支中, 实现主/从编码器特征的整合, 并增强模型对不同尺度器官的感知能力. 在解码器中, 引入多尺度特征融合模块 (multi-scale feature aggregation, MCA) 和深度特征增强模块 (deep feature augmentation, DFA). MCA 模块通过融合多尺度特征, 使模型充分利用不同层级特征的边界信息, 提升对器官边界的感知能力; DFA 模块进一步提取和增强特征图中的重要信息, 并抑制无关的背景噪声. 通过上述方法, 模型可有效缓解多器官分割问题, 实现多器官影像的精确分割.

综上, 本文贡献有以下几方面.

(1) 针对不同器官相对大小存在较大差异问题, 提出主/从双分支编码器结合级联上下文模块, 提升网络对不同大小、形态器官的感知能力.

(2) 针对相邻器官的空间界限难以确认以及低对比度问题, 提出多尺度特征融合模块 MCA 和深度特征增强模块 DFA, 提升网络的分割精度和鲁棒性.

(3) 在多器官分割公开数据集上进行对比实验, 验证了模型的有效性和竞争性.

1 相关工作

近年来, 医学图像分割领域涌现了许多先进的深度学习方法. 本文重点介绍 4 类典型的多器官分割技术, 包括基于 CNN 的分割方法、基于 Transformer 的分割方法、基于 CNN 与 Transformer 混合的分割方法, 以及 Vision Mamba 及其相关技术. 这些方法在处理复杂的多器官分割任务时表现出了卓越的性能, 为相关研究提供了重要支持.

1.1 基于 CNN 的方法

医学图像分割在计算机视觉和医疗诊断中起着重要作用, 目前基于 CNN 的方法在多器官分割领域已经

表现出了很好的性能[6]. 全卷积网络 (fully convolution network, FCN) 和 U-Net 是其中两个具有代表性的模型. FCN[7]通过将全连接层替换为卷积层, 使得网络可以处理任意大小的输入图像, 并进行像素级预测. U-Net[8]通过编码器-解码器结构, 结合跳跃连接在不同层次的特征之间传递信息, 从而有效捕捉器官的边界信息, 显著提高多器官分割的准确性. U-Net 的系列变体, 例如 U-Net++[9]、Attention U-Net[10]和 EGE-UNet[11] , 通过不同的优化策略进一步提高了网络性能, 使其在处理多器官分割任务时表现优异. 然而, CNN 由于其固有的感受野限制, 难以全面捕捉大尺寸器官之间的全局上下文信息, 进而影响多器官分割任务精度.

1.2 基于 Transformer 的方法

Transformer 最初用于自然语言处理 (natural language processing, NLP) 任务中的序列建模, 其基于自注意力机制的全局建模能力可以有效捕捉长距离依赖关系. 在视觉任务中, 卷积神经网络 (CNN) 擅长捕捉局部特征, 但在处理全局信息时存在局限. 研究者们意识到 Transformer 的强大潜力, 并将其应用到视觉任务中, 开启了视觉 Transformer (vision Transformer, ViT)[12]的研究潮流.

在医学图像分割领域 Swin Transformer[13]通过引入层次化的分层结构, 提升了多尺度特征建模的能力, 使其在处理不同大小和形状的器官时具有良好的表现. Swin-Unet[14]是将 Swin Transformer 与经典 U-Net 结构结合的模型. 在编码器部分, Swin Transformer 被用来提取全局和局部特征; 而在解码器部分, U-Net 的跳跃连接被保留, 确保在分割过程中能够充分利用不同层次的特征信息. 这种结合方式使得 Swin-Unet 在处理器官边界模糊或低对比度区域时具有良好的表现, 因为其能够在保持全局信息的同时, 也注重局部细节的精确性. Swin-Unet 的多层级特征建模能力对于同时处理多个形态差异较大的器官尤为有效. 然而, 基于 Transformer 的方法虽然在捕捉全局信息和长距离依赖关系方面表现出色, 但由于计算资源需求较高, 且在局部细节上表征不足, 给多器官分割任务带来了挑战.

1.3 基于 CNN 与 Transformer 混合的方法

相较于基于 CNN 方法和基于 Transformer 的方法, 基于 CNN 与 Transformer 混合的方法结合了 CNN 对局部特征的提取能力与 Transformer 在全局信息捕捉上的优势, 从而在处理多尺度、多形态的器官时展现出更为优异的表现.

TransUNet[15]将 Transformer 与 U-Net 架构相结合, 利用 Transformer 的全局自注意力机制弥补 CNN 在捕捉长距离依赖上的不足, 通过 Transformer 的全局特征建模能力增强对复杂边界和低对比度器官的分割效果. 这种方法尤其适合器官边界模糊的情况, 能够显著提高分割精度. 然而, 由于不同器官间的层次关系和特征表达缺乏一致性, 尤其是在处理边界模糊或低对比度区域时, 模型在特征融合过程中难以有效捕捉这些器官的细微差异, 导致分割效果不理想. 此外, 在小器官特征处理中, 其局部细节容易被大器官的全局特征所掩盖, 导致细节保留不足, 进而影响对小器官的精确分割.

1.4 Vision Mamba

近年来, 深度学习技术在医学图像分割领域取得了巨大的进步[16]. 然而精确的多器官分割需要将局部特征和全局特征相结合. CNN 和 ViT 这两种主流的做法在长相关性建模方面都有着自身的局限性. CNN 因其固有的局部感受野, 这限制了其捕捉长距离依赖关系. ViT 展现了其在处理全局上下文和长距离依赖性的能力, 然而 ViT 受到其注意力机制的限制, 在长序列建模中存在二次时间复杂度.

最近状态空间模型 (SSM) 证明了其在长序列建模中的有效性和效率. 与 Transformer 相比, SSM 可随序列长度线性或接近线性地扩展, 同时保持对长距离依赖建模的能力, 这使得其在处理连续长序列数据分析如基因组分析中获得了显著效果. 此外视觉 Mamba (vision Mamba, ViM)[17]和视觉状态空间模型 (visual state space model, VMamba)[18]探索了 Mamba 在视觉领域的有效性. ViM 提出了一种新的通用视觉骨干网络, 网络采用双向 Mamba 块, 通过位置嵌入标记图像序列, 并利用双向状态空间模型压缩视觉表示. VMamba 建立了一个基于 Mamba 的视觉骨干, 在没有牺牲全局感受野的前提下, 实现了与图像分辨率线性相关的复杂度. 此外 VMamba 通过引入交叉扫描模块 (cross-scan module, CSM) 解决因序列与图像之间的差异而导致的方向敏感性问题.

1.5 基于 Mamba 的方法

对于医学图像分割, U-Mamba[19]提出了一种新型的通用网络架构, 它结合了 CNN 的局部特征提取能力和 SSM 处理长序列能力, 设计了一种混合 CNN-SSM

块. U-Mamba 为医学图像分割有效建模长距离依赖关系提供了新途径. SegMamba[20]提出了一种基于 SSM 的 Mamba 架构, 有效捕捉了 3D 影像全体积特征的长距离依赖性. VM-UNet[21]引入了视觉状态空间块 (visual state space, VSS) 作为基础模块, 以捕获广泛的上下文信息, 并构建了不对称的编码器-解码器结构. 这些模型取得的效果, 展现了 Mamba 在医学图像分割中的潜力.

2 MCCNET 模型结构

MCCNET 模型结构如图 1 所示. 与传统的仅使用单一分支的编码器相比, 提出了一种主/从双分支编码器结构. 在主分支编码器中, Mamba 能够有效地捕捉全局上下文和长距离依赖并保持线性复杂度, 特别是在处理肝脏和肾脏等大器官时展现出优势, CNN 构成的从分支编码器专注于提取局部特征, 尤其在处理胰腺和胆囊等小器官时, CNN 能够更准确捕捉细节信息. 为加强主/从分支联系, 从分支编码器提取的特征通过级联上下文模块 CCM 传递到主分支编码器中, 以进一步增强模型的整体特征表达能力. 主/从双分支编码器结构能够有效平衡大小器官之间的形态差异, 提升模型在多器官分割任务中的整体性能. 经主/从双分支编码器后, 输出的特征图会传递到解码器部分. 在解码器部分, 引入了多尺度特征融合模块 MCA 和深度特征增强模块 DFA. MCA 模块融合多尺度特征, 使模型在保留细节信息的同时, 提升对整体结构的感知能力. DFA 模块进一步提取和增强特征图中的重要信息, 同时抑制无关的背景噪声. 下面将就上述各个模块进行详细介绍.

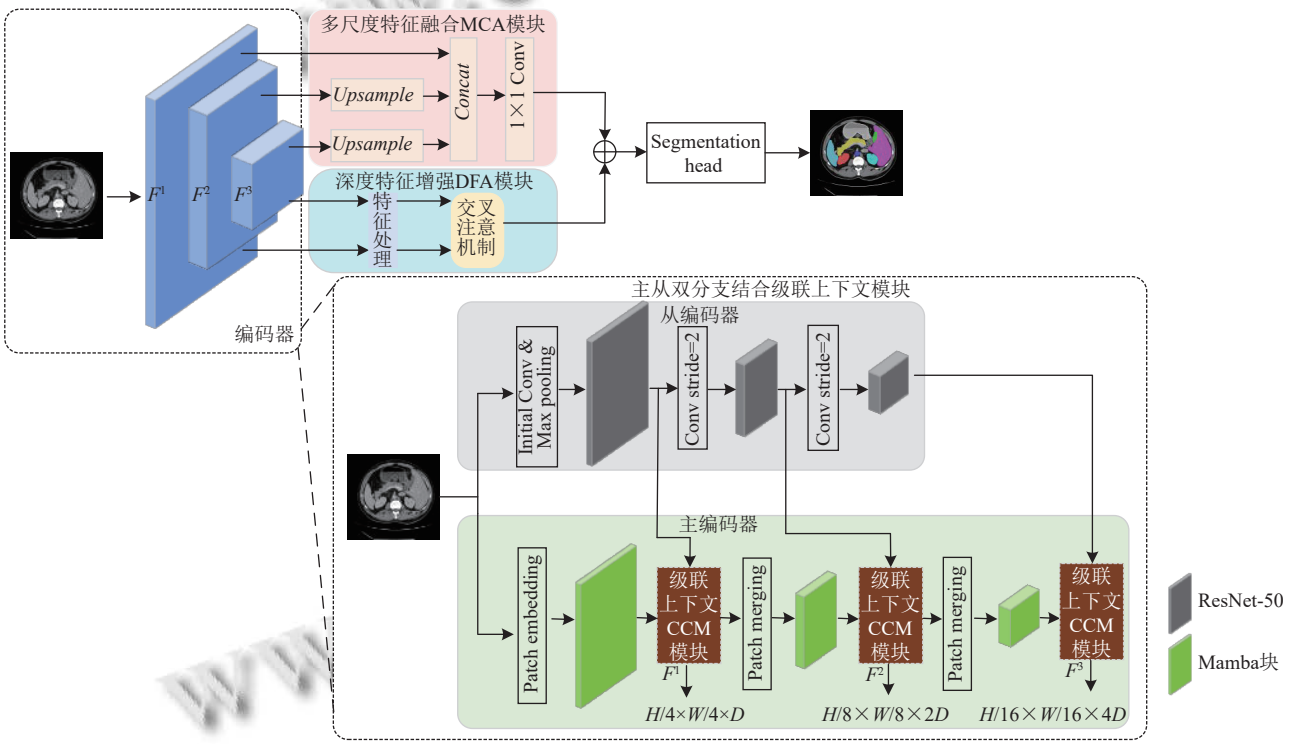


图 1 MCCNET 网络总体结构

2.1 主/从双分支结合级联上下文模块

如图 1 所示, 编码器由并行的主/从双分支结构构成. 主编码器用于提取多器官图像下不同分辨率的主特征图. 首先, 输入图像经过 Patch embedding 操作, 将其划分为多个不重叠的小图像块, 并映射到指定的维度; 随后, 特征图依次通过 Mamba 块、CCM 模块和 Patch merging 操作, 生成多分辨率特征图, 其中, Mamba 块为主编码器特征提取的核心组件, 负责全局关系建模, 由 VMamba 的 VSS Block 组成; CCM 模块用于将相应分辨率下从特征图中的细节信息补充到主编码器的特征图中, 以弥补主编码器特征图中缺失的细节信息; Patch merging 则负责缩小特征图维度并调整通道数, 每次将特征图维度缩小为原来的一半同时将通道数扩大至原来的两倍.

Mamba 块的处理流程图如图 2 所示. 特征图经过层归一化后被送入两个分支, 在第 1 个分支中输入经线性层、深度可分离卷积和 SiLU 激活函数后馈送到核心模块 SS2D (2D-selective-scan), 随后经过 Layer norm 对特征图进行归一化操作; 在第 2 个分支中输入依次经线性层和 SiLU 激活函数处理. 随后将两个分支的输出使用逐元素相乘的操作进行合并. 合并后的特征经过线性层对特征进行混合, 并通过残差连接直接馈送到输出, 形成最终的 Mamba 块输出.

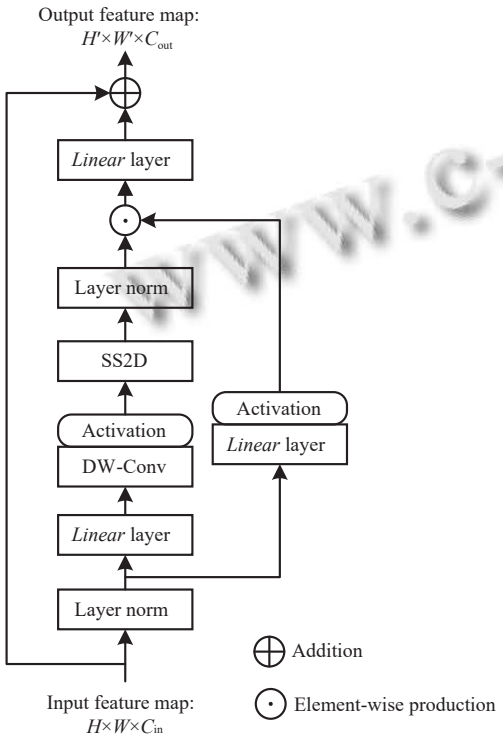


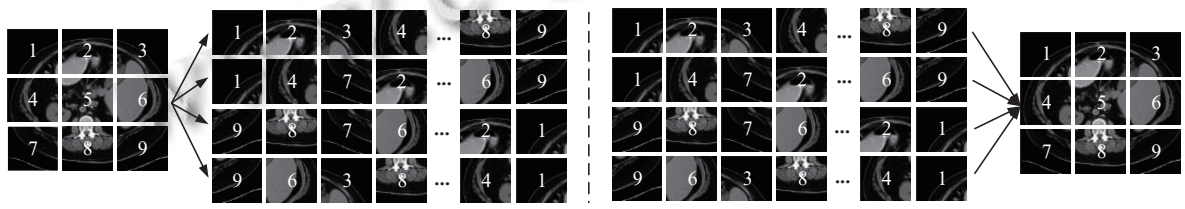
图 2 Mamba 块处理流程

SS2D 由 3 个部分组成: 扫描扩展操作、S6 块和扫描合并操作. 如图 3(a) 所示, 扫描扩展操作先将特征图沿 4 个不同的方向展开为序列; 随后, 这些序列由 S6 模块进行特征提取; 最后, 如图 3(b) 所示, 扫描合并操作对来自 4 个方向的序列求和并合并, 最终将输出特征图恢复至与输入相同大小. S6 模块在 S4[22]的基础上根据输入调整 SSM 参数, 引入选择机制, 这使得模型在区分和保留关键信息的同时过滤掉不相关的信息. S6 模块的伪代码如算法 1 所示. 输入序列 $x\in R^{B\times L\times D}$ 通过线性投影操作动态计算 SSM 参数 Δ,B,C, 这使得 S6 模块能够感知输入数据中所嵌入的上下文信息, 从而确保权重具有动态适应性. 随后, 通过指数映射 $\exp(\Delta A)$ 和离散化公式 $(\Delta A)^{-1}(\exp(\Delta A)-I)\cdot\Delta B$ 对连续状态演化进行离散化处理, 得到状态矩阵 $\overline{A}$ 和控制矩阵 $\overline{B}$. 在时间步 t, 状态向量 h_t 通过递推关系 $h_t=\overline{A}h_{t-1}+\overline{B}x_t$ 进行更新, 而输出 y_t 由状态向量 h_t 和线性映射项 Dx_t 共同生成. 最终, 所有时间步的输出序列 y 被拼接得到 $y=[y_1,y_2,\cdots,y_t,\cdots,y_L]$.

算法 1. S6 模块伪代码

输入: $x\in R^{B\times L\times D}$ (批量大小, 序列长度, 维度); 参数: A、D; 操作: $Linear(\cdot)$, 线性投影层.
输出: $y\in R^{B\times L\times D}$.

1. $\Delta,B,C=Linear(x),Linear(x),Linear(x)$
2. $\overline{A}=\exp(\Delta A)$
3. $\overline{B}=(\Delta A)^{-1}(\exp(\Delta A)-I)\cdot\Delta B$
4. $h_t=\overline{A}h_{t-1}+\overline{B}x_t$
5. $y_t=Ch_t+Dx_t$
6. $y=[y_1,y_2,\cdots,y_t,\cdots,y_L]$
7. return y



(a) Scan expanding (b) Scan merging

图 3 SS2D 中的扫描扩展操作和扫描合并操作

从编码器由 ResNet-50[23]组成, 用于提取多器官图像下不同分辨率的从特征图. 对于输入图像 $X\in R^{H\times W\times C}$, 其中空间维度为 H 和 W, 通道为 C. 从编码器包含 3 层, 每层依次将特征图分辨率缩小至 $H/4\times W/4$、$H/8\times W/8$ 和 $H/16\times W/16$, 同时通道数依次增加至 D、$2D$ 和 $4D$. 此外 CNN 提取的 3 个层次特征图依次通过 1×1 的卷积操作连接至相应分辨率的主编码器中, 从而补充主编码器中丢失的信息并恢复局部空间信息.

CCM 模块的处理流程如图 4 所示. 由于主编码器擅长提取全局上下文和长程依赖、从编码器则关注局

部细节特征, 这二者特征具有互补性. 同时由于主/从编码器分别处理不同层次的特征, 直接将两种特征简单相加容易导致特征不一致, 从而影响网络性能, 故通过 CCM 模块的主/从注意力机制分别学习深层特征, 使主/从分支特征在级联前分别适应各自特性, 以确保更高的表达一致性. CCM 模块的输入分别为主特征图和从特征图, 从特征图用于增强主特征图. 主/从编码器输入的特征图经过 Layer norm 后输入到主/从注意力机制中, 在主/从注意力机制中分别使用不同的方式学习深层特征, 考虑到主分支特征是由具有长程特性的 VSS 获得的, 故使用卷积学习局部细节; 从分支特征是通过具有局部性质的卷积运算获得的, 故使用基于 window 的多头注意建模长程依赖. CCM 模块中使用多个跳跃连接让模块更多的注意主分支特征.

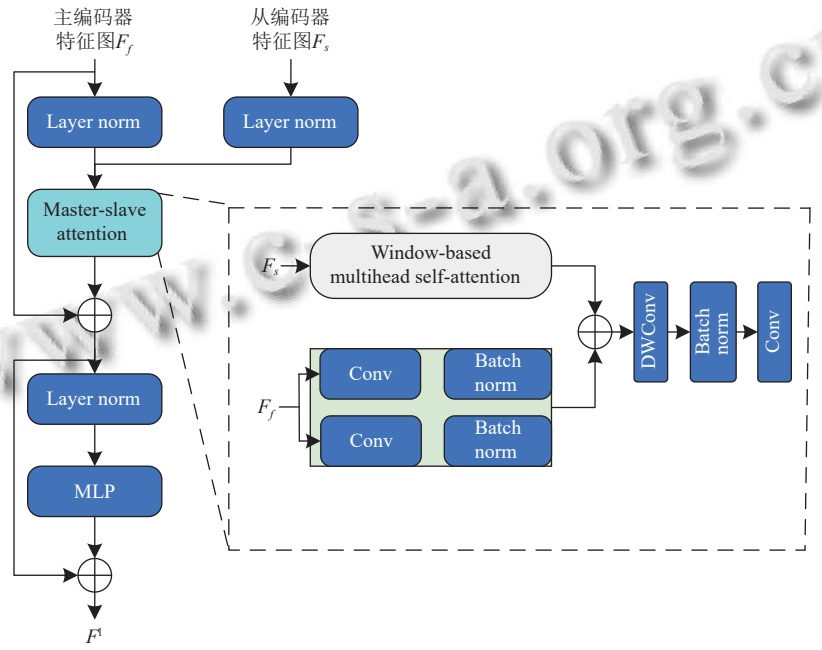


图 4 CCM 模块处理流程

2.2 多尺度特征融合模块

如图 1 所示, 针对相邻器官空间界限难以确认问题, 提出了多尺度特征融合模块 MCA. 具体来说, 主/从双分支结合级联上下文模块提取了 3 个不同尺度的特征图F^1、F^2和F^3, 分别代表浅层、中层和深层特征. 为了在保留低层特征细节信息的同时增强高层特征的语义表达, MCA 模块对深层次特征图F^2和F^3采用双线性插值进行上采样, 使其空间维度与F^1对齐. 上采样后的特征图分别记为$F^{2'} = Upsample(F^2)$, $F^{3'} = Upsample(F^3)$. 接着, 将上采样后的特征图$F^{2'}$和$F^{3'}$与浅层特征图F^1进行通道维度上的拼接, 形成融合特征图F^{cat}, 表示为:

$$F^{\text{cat}} = Concat([F^1, F^{2'}, F^{3'}]; axis = channel) \quad (1)$$

其中, $Concat$表示拼接操作. 最后为了压缩通道数并减轻计算负担, 对拼接后的特征图F^{cat}进行1×1的卷积操作得到最终输出F^{conv}.

2.3 深度特征增强模块

深度特征增强模块 DFA 旨在进一步提取和增强特征图中的重要信息, 同时抑制不相关的背景噪声. 具体来说, 将主/从双分支结合级联上下文模块提取的最浅层特征 (F^1) 有效集成至最深层特征 (F^3) 中, 最终得到富含语义和细节信息的深层特征.

DFA 模块的处理流程如图 5 所示. 浅层特征F^1和深层特征F^3分别经过全局平均池化 (global average pooling, GAP) 后, 得到两个类令牌信息CLS^s和CLS^l. 类令牌信息起着重要作用, 它包含了输入特征的核心信息. 得到两个类令牌后, 分别经过 Transformer Encoder 捕捉特征图中更广泛的空间依赖关系, 并细化特征表示. 最后使用交叉注意机制将浅层特征对齐至深层特征, 得到富含语义和细节信息的深层特征, 从而突出关键器官区域, 同时抑制背景中不相关的信息. 在交叉注意机制中, 浅层特征的类令牌信息CLS^s首先被投影到

F^3维度上, 投影是为了更好地共享跨级别令牌信息. 投影后的类令牌表示为$CLS^{s'}$, $CLS^{s'}$与F^3进行通道上的拼接, 作为交叉注意机制的 Key 和 Value、$CLS^{s'}$作为 Query, 计算注意力分数. 最终的输出Z^s可以用式 (2) 表示为:

$$y^s = f^s(CLS^s) + MCA(LN([f^s(CLS^s)\|F^1])) \quad (2)$$

$$Z^s = [F^3 \| g^s(y^s)] \quad (3)$$

其中, f^s表示将CLS^s投影到F^1维度; MCA表示交叉注意机制; LN表示层归一化; $\|$表示拼接操作; g^s表示将特征维度从F^1维度投影到F^3维度; Z^s表示富含语义和细节信息的深层特征.

最后, DFA 模块生成的特征图采用双线性差值方式恢复特征图维度; MCA 模块生成的特征图则使用 1×1 卷积调整通道数. 处理后的 DFA 特征图和 MCA 特征图进行逐元素相加, 得到联合掩码特征, 该联合掩码特征随后通过上采样恢复至原图像大小, 并最终通过分割头生成精确的分割输出.

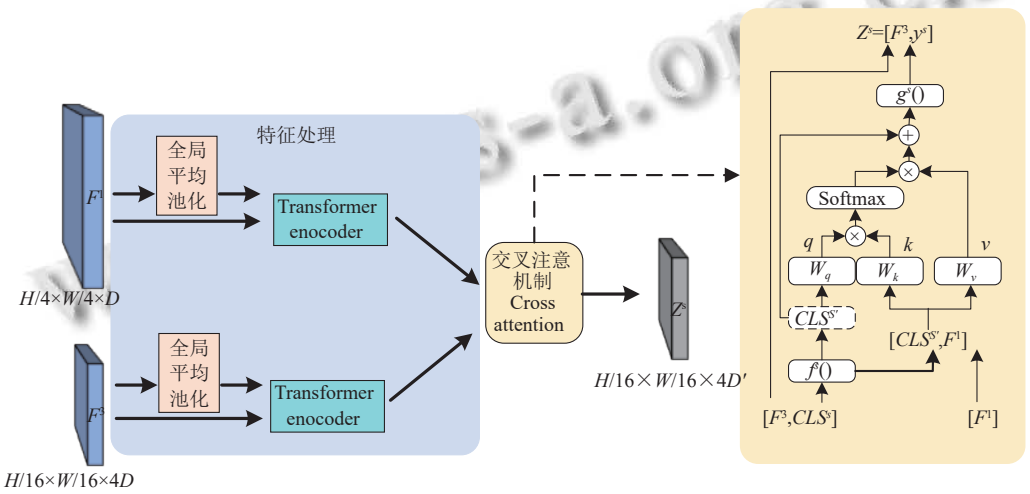


图 5 DFA 模块处理流程

2.4 损失函数

本文采用了交叉熵损失函数 (cross entropy loss, CE) 和 Dice 损失函数 (Dice loss) 构成的联合损失来优化模型. 具体而言, 联合损失函数$\mathcal{L}$由两部分组成:交叉熵损失函数$\mathcal{L}_{CE}$和 Dice 损失函数$\mathcal{L}_{Dice}$. 交叉熵损失函数的表达式为:

$$\mathcal{L}_{CE} = -\sum_i [y_i \log(p_i) + (1-y_i)\log(1-p_i)] \quad (4)$$

其中, y_i表示真实标签, p_i表示预测概率. Dice 损失函数的表达式为:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_i y_i p_i + \varepsilon}{\sum_i y_i + \sum_i p_i + \varepsilon} \quad (5)$$

其中, ε是一个小的常数, 用于避免分母为 0 的情况. 最终的联合损失函数被定义为:

$$\mathcal{L} = 0.4 \times \mathcal{L}_{CE} + 0.6 \times \mathcal{L}_{Dice} \quad (6)$$

这一联合损失函数通过平衡像素级的分类误差和目标区域的形状相似性, 来提升模型在多器官分割任务中的性能.

3 实验结果与分析

3.1 数据集与评估指标

本文采用 Synapse 和 ACDC 两个公开的多器官分割数据集进行实验来评估所提模型的性能效果.

Synapse 数据集收集了 30 例患者的 3 779 张腹部轴向临床 CT 图像. 为确保实验一致性, 与 TransUNet 的数据集划分保持一致, 18 例患者数据用于模型训练, 12 例患者的数据用于模型测试.

ACDC 数据集由 100 个心脏磁共振成像序列组成, 每个序列包含收缩末期和舒张末期的图像, 覆盖左心室、右心室和心肌等心脏结构. 为确保实验一致性, 同样参考 TransUNet 的划分策略, 将 70 个样本用于训练, 10 个样本用于验证, 20 个样本用于测试.

为了分析模型在多器官分割数据集上的实验表现, 本文使用平均 Dice 相似系数 (Dice similarity coe-

fficient, *DSC*) 和平均 95% Hausdorff 距离 (Hausdorff distance, HD95) 作为评估指标来衡量分割结果的精度和边界质量.

DSC 是一种常见的评价分割效果的指标, 其定义如式 (7):

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{7}$$

其中, X和Y分别表示预测的分割结果和真实的分割区域, $|X \cap Y|$表示他们的交集区域的大小. *DSC* 的值介于 0–1 之间, 值越大表示分割结果与真实标签越接近.

Hausdorff 距离衡量的是两个边界之间的最大最小距离, 其定义为:

$$H(X,Y) = \max\left\{\max_{x\in X}\min_{y\in Y} d(x,y), \max_{y\in Y}\min_{x\in X} d(y,x)\right\} \tag{8}$$

其中, $d(x,y)$表示点x和点y之间的欧几里得距离. Hausdorff 距离越小, 说明分割结果的边界与真实边界越接近. HD95 是通过 Hausdorff 结果值乘以 95%, 目的是消除离群值的一个非常小的子集的影响.

3.2 实施细节

本实验基于 PyTorch 1.13 深度学习框架, 使用 Python 3.8 版本, CPU Intel i911900K/F, 内存 64 GB, 显存为 24 GB 的 NVIDIA GeForce RTX 3090. 实验开始阶段对数据集进行随机旋转和翻转预处理, 以增强数据的多样性和鲁棒性.

模型的输入图像大小为 224×224, 初始学习率为 0.01, 批量大小设置为 10, 使用 SGD 优化器进行优化, 动量设置为 0.9, 权重衰减为 0.000 1. 此外, 主/从编码器都在 ImageNet 上进行了预训练.

3.3 基线模型介绍

为验证 MCCNET 模型的有效性, 本文选择了如下几种基线模型进行对比, 这些基线模型包括基于 CNN 的方法、基于 Transformer 的方法、基于 CNN 与 Transformer 混合的方法和基于 Mamba 的方法.

DARR[24]: 一种无监督域适配方法, 用于多器官分割. 通过学习器官相对位置的空间关系, 结合超分辨率网络标准化不同域分辨率, 并在测试时自适应优化.

U-Net: 采用编码器-解码器结构, 通过跳跃连接将不同层次的特征结合.

Att-UNet[25]: 提出一种基于 CNN 的注意力门模块, 用于自动聚焦多器官图像中的关键区域.

Swin-Unet: 基于 Swin Transformer 的医学图像分割模型, 利用分层的窗口自注意力机制来捕捉多尺度特征, 结合 U-Net 的编码器-解码器结构.

TransUNet: 结合 Transformer 和 U-Net 的混合架构, 通过 Transformer 捕捉长程依赖关系, 同时利用 U-Net 的跳跃连接和卷积操作增强局部特征.

MT-UNet[26]: 一种混合 Transformer 和 U-Net 的分割模型, 提出混合 Transformer (MTM) 模块, 通过同时捕捉局部和全局依赖关系以及一个可学习的高斯矩阵, 提升复杂器官分割性能.

PVT-CASCADE[27]: 设计了 CASCADE 解码器, 结合了多级特征表示, 利用注意力机制和卷积模块进行特征提取和融合.

VM-UNet: 一种基于纯状态空间模型的分割模型, 通过视觉状态空间块 (visual state space block, VSS) 捕获长距离上下文信息, 实现了线性计算复杂度.

3.4 实验结果

在展示所提模型与基线模型在公开数据集上的对比结果之前, 先对所提模型的训练过程进行分析. 图 6 展示了模型在 Synapse 训练过程的损失收敛曲线, 包括总损失 (total loss)、交叉熵损失 (cross entropy loss) 和 Dice 损失 (Dice loss).

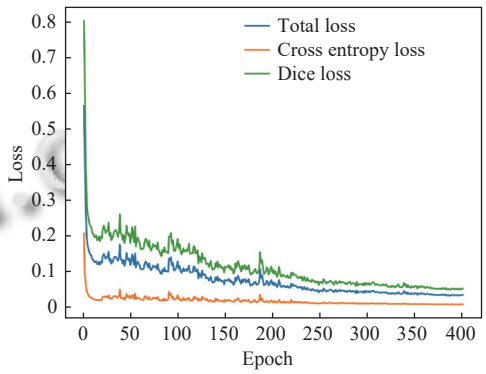


图 6 所提模型在 Synapse 数据集上的损失收敛曲线

可以看到, 训练初期所有损失值迅速下降, 表明模型在早期阶段有效地学习到了 CT 图像的基本特征. 随着训练的进行, 损失值逐渐趋于平稳, 尤其在第 250 个 epoch 后, 总损失及各项损失不再显著下降, 显示出模型训练过程的稳定性和良好的收敛效果. 此外, Dice 损失在训练初期数值较高且波动较大, 反映出模型在优化目标区域边界的复杂性, 而后期随着训练深入, 该损失逐渐平稳, 说明模型对分割性能的进一步优化趋于

收敛. 整体而言, 图 6 表明模型的训练过程稳定且收敛良好, 总损失及其组成部分均呈现出合理的下降趋势, 训练后期的稳定性验证了模型在多器官分割任务中的有效性.

基于良好的收敛性与稳定性, 本文进一步将所提模型与基线模型在两个公开数据集上的对比实验结果进行量化分析, 结果如表 1 和表 2 所示, 其中加粗数值为最优结果.

表 1 Synapse 数据集上与其他算法性能对比

模型	年份	平均*DSC* (%)	HD95 (mm)	*DSC* (%)							
				主动脉	胆囊	左肾	右肾	肝脏	胰腺	脾脏	胃
DARR	2020	69.77	—	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50 U-Net	2021	74.68	36.87	87.47	66.36	80.60	78.19	93.74	56.90	85.87	74.16
U-Net	2015	76.85	39.70	**89.07**	69.72	77.77	68.60	93.43	53.98	86.67	75.58
R50 Att-UNet	2021	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
Att-UNet	2019	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
TransUNet	2021	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet	2022	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
MT-UNet	2022	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
PVT-CASCADE	2023	81.06	20.23	83.01	70.59	82.23	80.37	94.08	**64.43**	90.10	**83.69**
VM-UNet	2024	81.08	19.21	86.40	69.41	**86.16**	**82.76**	94.17	58.80	89.51	81.40
MCCNET (ours)	2024	**81.17**	**16.49**	86.56	**71.34**	85.73	79.24	**94.62**	58.23	**91.70**	81.98

表 2 ACDC 数据集上与其他算法性能对比

模型	平均*DSC* (%)	*DSC* (%)		
		左心室	心肌	右心室
R50 U-Net	87.55	87.10	80.63	94.92
R50 Att-UNet	86.75	87.58	79.20	93.47
TransUNet	89.71	88.86	84.53	95.73
Swin-Unet	90.00	88.55	85.62	**95.83**
MT-UNet	90.43	86.64	89.04	95.62
PVT-CASCADE	91.46	88.90	89.97	95.50
MCCNET (ours)	**92.60**	**91.72**	**91.08**	95.00

表 1 展示了所提模型与其他基线网络在 8 个腹部器官的平均 *DSC* 和 HD95 方面的比较. 所提模型在平均 *DSC* 系数上达到了 81.17%, 在平均 HD95 上降低至 16.49 mm, 两个评价指标上均展现出最佳性能. 具体而言, 无论是在小型腹部器官 (胆囊、脾脏) 还是较大型器官 (肝脏) 所提模型相较于其他模型均展现出最佳性能. 与 VM-UNet 相比, 所提模型在平均 *DSC* 指标上提升了约 0.1%, 虽提升不大, 但在 HD95 评价指标上却提高了约 2.7 mm, 这展现了主/从双分支结合级联上下文模块相较于纯 Mamba 作为编码器的优势.

为了更加直观地展示分割效果, 图 7 展示了不同网络在腹部多器官任务中的可视化效果. 对比模型包括 Swin-Unet、TransUNet、VM-UNet 和 U-Net. 可以看出, 由于 Transformer 在捕捉细节信息方面的不足, Swin-Unet 在分割胆囊器官时出现了误分割现象. 类似地, 基于纯 Mamba 的 VM-UNet 在分割胆囊器官时也出现了误分割现象, 这是由于 Mamba 更适合长距离关系的建模, 但在细节信息的捕捉上较弱. 而由于卷积本身的局限性, 难以捕捉长距离依赖, U-Net 在分割肝脏和胃等大器官时, 明显存在过分割问题. 本文提出的模型不仅能够有效处理不同大小和形态的器官, 还加强了相邻器官之间的边界识别, 并缓解了低对比度的问题. 从可视化结果来看, 所提模型的分割效果更加准确, 不同器官之间的边界更加清晰.

表 2 展示了所提模型与其他基线模型在 ACDC 数据集上的平均 *DSC* 比较. 所提模型在平均 *DSC* 上达到了 92.60%, 在所有的基线模型中展现出最佳性能. 相较于 PVT-CASCADE, 所提模型在平均 *DSC* 指标上提升约 1.1%. 实验结果表明, 所提模型能够有效地识别和分割心脏结构.

3.5 消融实验

为了验证主/从双分支结合级联上下文模块、多尺度特征融合模块和深度特征增强模块对于提升多器官分割性能大小, 本文进行了消融实验, 并在 Synapse 数据集上进行实验和评估.

多尺度特征融合与深度特征增强模块的验证: 为了验证 MCA 模块和 DFA 模块在提升分割精度和处理边界模糊问题中的有效性, 本文设计了针对二者的消融实验, 通过逐步移除或添加这些模块来评估它们的实际贡献. 实验结果如表 3 所示.

首先, 在 MCA 和 DFA 模块均启用的情况下, 模型取得了最佳的分割效果, 平均 *DSC* 达到了 81.17%, HD95 仅为 16.49 mm. 这表明, MCA 和 DFA 模块的结合可以有效地解决相邻器官的空间界限模糊和低对比度问

题, 从而提高分割精度. 当仅启用 MCA 模块时, 模型的平均 *DSC* 保持在较高水平, 但 HD95 值显著增加. 这说明 MCA 模块在融合多尺度特征以处理器官边界问题方面起到重要作用, 但缺乏 DFA 模块的深度特征增强, 分割效果仍存在不足, 尤其是在低对比度区域. 相反地, 仅启用 DFA 模块时, 尽管 HD95 值有所改善, 但平均 *DSC* 下降到 76.14%, 表明 DFA 模块对低对比度问题的改善有限, 而缺少 MCA 模块的多尺度特征融合, 使得模型在处理复杂边界时的能力下降. 最后当 MCA 模块和 DFA 模块均未启用时, 而是简单采用将特征图通过上采样方式还原至原 CT 图像分辨率, 模型表现最差, 平均 *DSC* 下降至 74.75%, HD95 值增加至 19.64 mm, 验证了这两个模块在提升分割精度和鲁棒性的重要性.

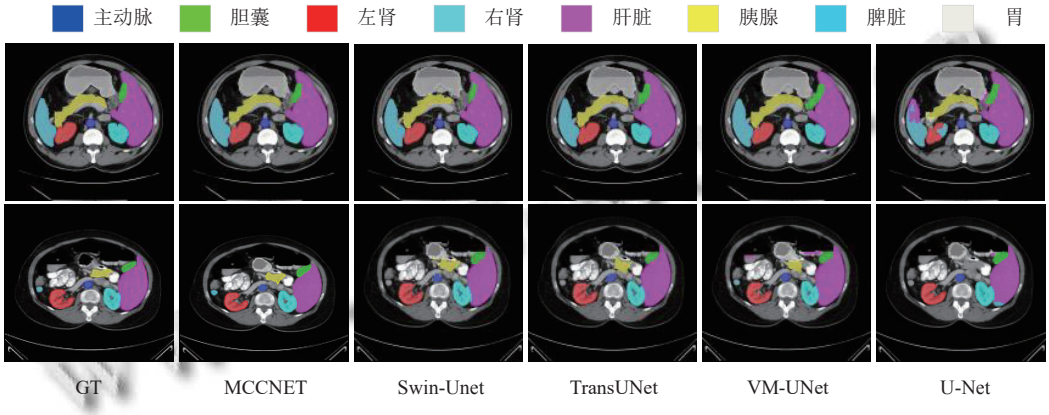


图 7 不同网络在 Synapse 数据集上的分割可视化效果

表 3 MCA 块与 DFA 块在 Synapse 数据集上的消融研究

模块		平均*DSC* (%)	HD95 (mm)	*DSC* (%)							
MCA	DFA			主动脉	胆囊	左肾	右肾	肝脏	胰腺	脾脏	胃
√	√	81.17	16.49	86.56	71.34	85.73	79.24	94.62	58.23	91.70	81.98
√	×	79.25	20.99	85.86	66.95	83.74	77.62	94.12	58.16	89.50	78.06
×	√	76.14	17.68	80.36	62.13	80.58	76.39	92.67	50.98	87.04	79.01
×	×	74.75	19.64	79.92	61.52	77.27	73.29	92.64	49.15	87.57	76.66

主/从双分支结合级联上下文模块的验证: 为了验证主/从双分支结合级联上下文模块相较于单一编码器架构在对不同大小、形态器官的感知能力上的优势. 本文设计了针对主/从双分支结合级联上下文模块的消融实验. 通过逐步移除 CCM 模块、从编码器和主编码器, 以评估这些组件在分割不同大小器官中的作用. 实验结果如表 4 所示.

首先, 完整模型在所有组件都启用的前提下, 达到了最优的分割效果. 这表明主/从双分支结合级联上下文模块的设计能够有效应对不同器官大小的挑战, 尤其是在肾脏、胃等大器官和胰腺、胆囊等小器官分割中. 当移除 CCM 模块时, HD95 值显著升高, 表明 CCM 模块在提升全局上下文信息的同时, 显著增强了小器官的分割能力, 尤其是对于胰腺和脾脏等器官. 移除从编码器导致小器官 (如胰腺和胆囊) 的分割性能下降, 这表明从编码器对于小器官的细节捕捉的重要性, 从编码器的存在显著提升了小器官的分割性能. 而移除主编码器导致整体感知能力下降, 尤其是在大器官 (如胃和肾脏) 的分割精度显著降低. 这进一步证明了主/从双分支结合级联上下文模块在多器官分割中的关键作用.

表 4 主/从双分支结合级联上下文模块在 Synapse 数据集上的消融研究

方法	平均*DSC* (%)	HD95 (mm)	*DSC* (%)							
			主动脉	胆囊	左肾	右肾	肝脏	胰腺	脾脏	胃
完整模型	81.17	16.49	86.56	71.34	85.73	79.24	94.62	58.23	91.70	81.98
−CCM	80.49	20.66	86.61	72.48	83.86	79.95	94.43	57.69	88.67	80.23
−从编码器	79.61	17.33	83.64	70.95	86.35	81.97	94.13	56.25	88.17	75.41
−主编码器	78.17	22.15	86.23	63.57	83.08	73.69	94.65	59.86	90.00	74.26

4 结论与展望

本文针对不同器官大小存在较大差异问题, 提出特征增强的双分支多器官分割模型. 主分支利用 Mamba 建模全局依赖信息, 从分支利用卷积神经网络建模局部空间信息, 同时从分支特征通过级联上下文模块 CCM 引入至主分支中, 用于补充主分支中缺少的局部空间信息. 此外针对相邻器官空间界限难以确认以及低对比度问题, 本文引入了多尺度特征融合模块 MCA 和深度特征增强模块 DFA, 从而显著提升了分割的精度. 实验结果表明, 与现有基线方法相比, 本文方法在多器官分割数据集上展现出了更高的分割性能和较强的竞争力. 未来的工作将聚焦于进一步优化模型的泛化能力, 并探索其在更复杂的医学图像分割任务中的应用潜力.

参考文献

1 Fu YB, Lei Y, Wang TH, *et al*. A review of deep learning based methods for medical image multi-organ segmentation. Physica Medica, 2021, 85: 107–122. [doi: 10.1016/j.ejmp.2021.05.003]

2 Zhou HY, Guo J, Zhang Y, *et al*. nnFormer: Volumetric medical image segmentation via a 3D Transformer. IEEE Transactions on Image Processing, 2023, 32: 4036–4045. [doi: 10.1109/TIP.2023.3293771]

3 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

4 Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces. arXiv:2312.00752, 2023.

5 Xu R, Yang S, Wang YH, *et al*. Visual Mamba: A survey and new outlooks. arXiv:2404.18861, 2024.

6 Zhang DW, Huang GH, Zhang Q, *et al*. Cross-modality deep feature learning for brain tumor segmentation. Pattern Recognition, 2021, 110: 107562. [doi: 10.1016/j.patcog.2020.107562]

7 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3431–3440.

8 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference. Munich: Springer, 2015. 234–241.

9 Zhou ZW, Siddiquee MMR, Tajbakhsh N, *et al*. UNet++: A nested U-Net architecture for medical image segmentation. Proceedings of the 4th International Workshop. Granada: Springer, 2018. 3–11.

10 Oktay O, Schlemper J, Le Folgoc L, *et al*. Attention U-Net: Learning where to look for the pancreas. arXiv:1804.03999, 2018.

11 Ruan JC, Xie MY, Gao JS, *et al*. EGE-UNet: An efficient group enhanced UNet for skin lesion segmentation. Proceedings of the 26th International Conference. Vancouver: Springer, 2023. 481–490.

12 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. 2021.

13 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10002.

14 Cao H, Wang YY, Chen J, *et al*. Swin-Unet: Unet-like pure Transformer for medical image segmentation. Proceedings of the 2022 European Conference on Computer Vision. Tel Aviv: Springer, 2022. 205–218.

15 Chen JN, Lu YY, Yu QH, *et al*. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306, 2021.

16 Guo JS, Zhou HY, Wang LS, *et al*. UNet-2022: Exploring dynamics in non-isomorphic architecture. Proceedings of the 2022 International Conference on Medical Imaging and Computer-aided Diagnosis. Leicester: Springer, 2022. 465–476.

17 Zhu LH, Liao BC, Zhang Q, *et al*. Vision Mamba: Efficient visual representation learning with bidirectional state space model. Proceedings of the 41st International Conference on Machine Learning. Vienna, 2024.

18 Liu Y, Tian YJ, Zhao YZ, *et al*. VMamba: Visual state space model. Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, 2024.

19 Ma J, Li FF, Wang B. U-Mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv:2401.04722, 2024.

20 Xing ZH, Ye T, Yang YJ, *et al*. SegMamba: Long-range sequential modeling Mamba for 3D medical image segmentation. Proceedings of the 27th International Conference on Medical Image Computing and Computer Assisted Intervention. Marrakesh: Springer, 2024. 578–588.

21 Ruan JC, Li JC, Xiang SC. VM-UNet: Vision mamba UNet for medical image segmentation. arXiv:2402.02491, 2024.

22 Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.

23 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

24 Fu SH, Lu YY, Wang Y, *et al*. Domain adaptive relational reasoning for 3D multi-organ segmentation. Proceedings of the 23rd International Conference. Lima: Springer, 2020. 656–666.

25 Schlemper J, Oktay O, Schaap M, *et al*. Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis, 2019, 53: 197–207. [doi: 10.1016/j.media.2019.01.012]

26 Wang HY, Xie SA, Lin LF, *et al*. Mixed Transformer U-Net for medical image segmentation. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022. 2390–2394.

27 Rahman MM, Marculescu R. Medical image segmentation via cascaded attention decoding. Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2023. 6211–6220.

(校对责编: 张重毅)