计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(6):41−52 [doi: 10.15888/j.cnki.csa.009876] [CSTR: 32024.14.csa.009876]

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

基于注意力和类相关性的膝关节软骨 MRI 分割①

王　顺, 张俊虎, 李海涛, 李　辉

(青岛科技大学 信息科学技术学院, 青岛 266061)
通信作者: 张俊虎, E-mail: jzhang@qust.edu.cn

摘　要: 本研究针对膝关节软骨 MRI 分割中标注数据稀缺的问题, 提出了一种多层次膝关节软骨图像分割网络. 该网络采用医学注意力机制, 并融合条件随机场, 形成了具有层次化注意力的架构. 通过将数据流分为全局流和局部流, 本网络能够同时捕获图像的全局特征和局部细节, 从而提升分割的准确性. 此外, 为了降低计算负担, 我们引入了轴向注意力机制, 有效地简化了计算过程并减少了模型参数. 通过层次化分割策略和条件随机场的整合, 网络能够更深入地挖掘类别间的相互依赖性, 提高了对关键特征的捕获能力. 在两个公共数据集 K-Space 和 MOST 上的实验验证了所提方法的有效性. 实验结果表明, 即使在数据标注有限的情况下, 本方法也能实现高精度的膝关节软骨图像分割. 与当前先进方法相比, 本研究的方法在 Dice 相似系数 (DSC) 和 95% Hausdorff 距离 (HD95) 等评价指标上均展现出显著优势.

关键词: 类相关性; 多类分割; 膝关节软骨图像分割

引用格式: 王顺,张俊虎,李海涛,李辉.基于注意力和类相关性的膝关节软骨 MRI 分割.计算机系统应用,2025,34(6):41–52. http://www.c-s-a.org.cn/1003-3254/9876.html

Knee Cartilage MRI Segmentation Based on Attention and Class Relevance

WANG Shun, ZHANG Jun-Hu, LI Hai-Tao, LI Hui

(College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: This study proposes a multi-level knee cartilage image segmentation network to address the challenge of limited annotated data in knee cartilage MRI segmentation. This network incorporates a medical attention mechanism and integrates conditional random fields to create a hierarchical attention framework. By separating the data stream into global and local streams, the network captures both global features and local details of the image, thus enhancing segmentation accuracy. In addition, an axial attention mechanism is introduced to reduce computational load, effectively simplifying the calculation process and minimizing model parameters. The integration of hierarchical segmentation strategies and conditional random fields enables a deeper exploration of the interdependence between categories, improving the network's ability to capture key features. The proposed method's effectiveness is validated through experiments on two public datasets, K-Space and MOST. Experimental results demonstrate that even with limited data annotation, the method achieves high-precision knee cartilage image segmentation. Compared to current state-of-the-art methods, the proposed approach shows significant improvements in evaluation metrics such as the Dice similarity coefficient (DSC) and 95% Hausdorff distance (HD95).

Key words: class relevance; multi-class segmentation; image segmentation of knee cartilage

① 基金项目: 国家自然科学基金 (61702295); 山东省重点研发计划 (科技示范工程) (2021SFGC0701); 青岛市海洋科技创新专项 (22-3-3-hygg-3-hy)
收稿时间: 2024-11-02; 修改时间: 2024-11-28, 2024-12-25; 采用时间: 2025-01-10; csa 在线出版时间: 2025-04-30
CNKI 网络首发时间: 2025-05-06

疲劳或经验不足可能会导致影像科医生在诊断中出现失误, 即便是一名经验丰富的医学影像科医生, 出错率也可能达到 5%. 因此, 利用人工智能技术辅助 MRI 的分割与诊断, 对于提高诊断准确性和减轻医生工作负担至关重要, 这也是推动计算机辅助诊断技术发展的一个重要方向.

然而, 训练一个高精度的分割模型需要大量有标注的高质量数据. 目前, 公开的膝关节影像数据库规模都较小, 且样本多集中于欧美人群, 缺乏针对中国或亚洲人群的大型数据库, 这限制了为国内人群定制辅助诊断工具的开发. 膝关节 MRI 的自动分割技术面临挑战, 因为软骨组织细小、形态复杂, 且与周围组织对比度较低. 传统分割技术在精确度和处理速度上难以满足要求, 虽然基于深度学习的分割技术近年来有所突破, 但计算成本高昂.

现有的自回归模型在处理高维数据时要么需要过多计算资源, 要么需要在分布表达或便于实现方面做出妥协, 以减少资源需求[1]. 为了解决这些挑战, 本研究在 Axial-DeepLab[2]网络的基础上, 结合 U-Net[3]和层次分类, 提出了一种引入层次化分割方法的网络架构, 通过全局流和局部流两个分支来处理数据. 本文基于轴向注意力 (axial attention), 它是自注意的一种简单泛化, 自然地与编码器和解码器设置中张量的多个维度对齐, 允许在解码过程中并行计算大部分上下文, 而不引入任何独立性假设. 这种设计不仅提升了学习效率, 简化了计算过程, 还降低了内存需求. 并且通过条件随机场整合更有效地利用了类别间的相互关系, 缓解了类别偏差的问题. 本文架构在保持对联合概率分布完全表达的同时, 利用标准深度学习框架进行实现.

在多标签分类任务中, 本研究提出了一种层次网络训练策略, 将膝关节图像按包含关系划分为 3 个层级, 通过网络的局部流分阶段进行学习, 以精确识别和分类图像中的不同标签. 这种分层处理使得网络能够更加精确地识别和分类图像中的不同标签.

总之, 膝关节软骨的精确图像分割对于早期发现和治疗膝骨关节炎至关重要. 尽管存在标注样本不足和多标签数据处理困难的问题, 但通过挖掘标签间联系和采用创新的分级分割技术, 有望在数据有限的情况下实现更高的分割精度, 为临床诊断提供强有力的辅助.

1 相关工作

1.1 医学图像分割

医学图像分割是计算机辅助诊断领域的一个重要分支, 它涉及将医学图像中的不同组织和结构分离出来, 以便进一步分析和诊断. 近年来, 随着深度学习技术的发展, 医学图像分割取得了显著进展. U-Net[3]及其变体由于灵活性和优化的模块化设计, 成为最广泛应用的图像分割体系结构. Transformer[4]模型也在医学图像分析领域的应用中取得了显著进展, 尤其是在医学图像分割任务中, 例如引入了轴向注意力和位置嵌入的 MedT[5]和结合了 CNN 局部特征提取能力的 Trans-UNet[6]. 基于 YOLOv5 的 Med-YOLO[7]也表现出优异性能, 它在中等和大型结构的检测上表现出色, 尤其是在心脏、肝脏和胰腺等结构的检测上.

1.2 层次多标签分类

机器学习研究的主要关注点是为典型的分类问题引入模型, 在这些问题中, 一个对象与一组不相交类中的一个类相关联. 然而存在一些任务, 其中类别并非不相交的, 而是组织成层次结构, 即层次分类 (hierarchical classification, HC). 在 HC 中, 对象与给定的超类及其相应子类相关联. 根据任务, 对应关系可以是所有子类或其中一部分子类. 形式化类之间关系的层次结构可以采用树的形式, 也可以采用有向无环图 (DAG) 的形式.

在更具有挑战性的场景中, 存在每个对象可以与类层次结构的多个不同路径相关联的 HC 问题, 即分层多标签分类 (hierarchical multi-label classification, HMC). 典型 HMC 问题包括文本分类 (Mayne 等[8])、图像注释 (Dimitrovski 等[9]) 以及生物信息学任务. 近年来, HMC 还被应用于多任务学习 (Yannis 等[10]) 及利用标签相关性 (Xu 等[11]).

然而, 自动膝关节图像分割任务依然面临挑战, 主要原因是软骨组织尺寸细小、形状复杂, 并且与周围组织之间对比度不高. 传统的分割技术在精确度和处理速度上存在局限, 而深度学习技术的应用虽然在某些方面取得了进展, 却存在计算成本较高的问题.

1.3 类相关性

类相关性 (class relevance) 是指在多标签分类问题中, 不同标签之间存在的关联性或依赖性. 在多标签分类任务中, 一个实例 (如一篇文章、一张图片或一个生物样本) 可以同时被分配多个标签. 类相关性体现在这些标签不是相互独立的, 而是存在某种程度的联系.

在层次化多标签分类 (hierarchical multi-label classification, HMLC) 领域中, 类相关性是一个重要概念, 它涉及标签之间的层次结构和相互关系. Xu 等[11]提出一种利用标签相关性进行层次化多标签分类的方法, 通过将标签映射到潜在向量空间中, 使得相关性较强的标签在空间中彼此接近, 从而捕捉标签间的相关性. Xu 等[12]提出了一个新的层次化多标签文本分类框架, 考虑了类别层次结构中的垂直和水平类别相关性.

众所周知, 分割 (segmentation) 是一种特殊的分类 (classification), 分割其实就是对每一个像素进行分类. 本文将分类领域中的类相关性和 Wehrmann 等[13]提出的 HMCN 引入分割领域, 以求在更少数据条件下完成高精度分割.

2 模型架构

在本研究中, 我们提出了一种层次条件随机场轴向注意力 (hierarchical conditional random field axial attention, HCAA) 网络架构, 用于三维 MRI 的分割任务. 该网络设计的核心在于其独特的结构, 能够有效地从三维 MRI 数据中提取全局和局部特征. 具体而言, 网络的输入为三维 MRI 数据, 该数据首先被送入网络的全局特征提取模块, 以捕获图像的整体特征.

在经过初步的全局特征提取之后, 网络将数据流分为两个并行的处理路径. 一方面, 部分数据继续流向更深层次的全局特征提取层, 以进一步深化对全局特征的理解; 另一方面, 另一部分数据则被引导至局部特征提取层, 专注于挖掘图像的局部细节和纹理信息. 这种双路径并行处理策略不仅提高了特征提取的效率, 而且增强了网络对图像不同尺度特征的捕捉能力, 从而为精确的图像分割提供了强有力的支持.

图 1 展示了 HCAA 网络的结构示意图, 其中详细描绘了数据流的路径以及各层之间的连接关系, 进一步阐明了网络如何通过层次化和轴向注意力机制实现对三维 MRI 数据的高效分割.

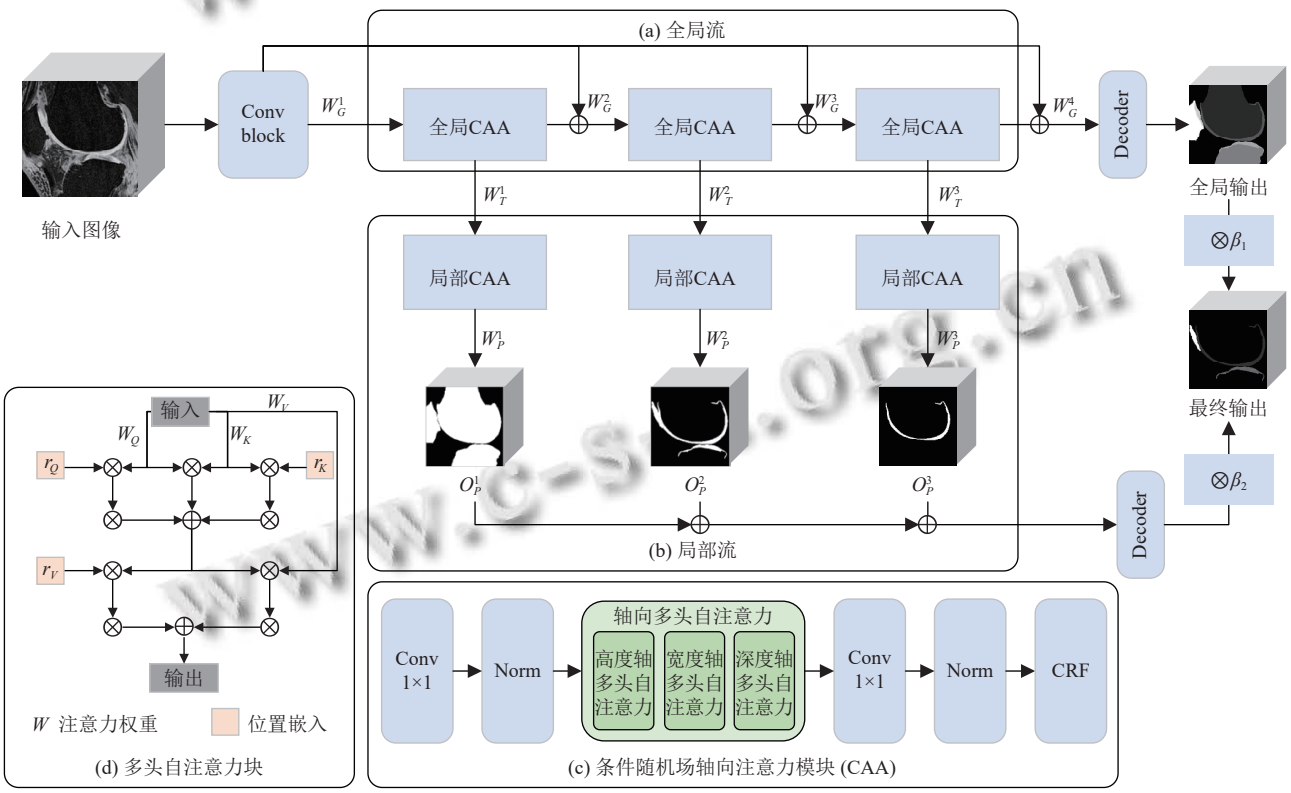

图 1 本文提出的 HCAA 网络

全局流和局部流提取出的特征数据分别相加, 与权重参数相乘后进行叠加, 生成最终的输出图像. 通过这种设计, 网络能够有效整合全局和局部信息, 从而提高分割精度.

在本研究中, 该网络结构旨在优化局部层次结构中的梯度传播效率, 并增强局部损失函数的性能. 具体来说, 我们为局部层次结构中的每个层级引入一个中间输出层. 这一设计的核心目的在于提升局部损失函

数的梯度传播效率, 从而更有效地编码相应层次中类别之间的局部信息.

通过在每个局部层级中设置中间输出层, 我们能够确保在特征提取过程中, 局部信息得到充分保留和传递. 这不仅有助于提高网络对局部特征的敏感度, 而且有助于增强模型对局部结构变化的适应能力.

此外, 本模型中, 全局输出层扮演着至关重要的角色. 我们采用了精细优化的损失函数, 以密切监控整个网络层次结构内类别间的相互依赖性如何变化. 通过这种方式, 全局输出层能够对整个网络的层次结构进行全局优化, 确保类别间的相互依赖关系得到准确捕捉和建模. 这一全局输出层的设计旨在捕捉网络前向传播过程中积累的所有关系, 并通过反向传播过程中各层级的类别梯度来实现.

为了进一步提升预测的准确性, 我们为模型增加了一种违规惩罚机制, 该机制能够有效识别并过滤掉违反常理的预测结果, 同时激励模型进行更为精确的层次化预测. 该机制将在第 2.3 节中详细介绍.

2.1 条件随机场轴向注意力层

条件随机场轴向注意力层如图 1(c) 所示, 根据文献[2], 将每个轴向注意力块分为二维高、宽、深度轴多头自注意力. 在此基础上, 每个二维自注意力又可分为两个一维自注意力, 该分解过程如图 2 所示, 具体步骤在本节详细介绍. 每个一维自注意力的计算如图 1(d).

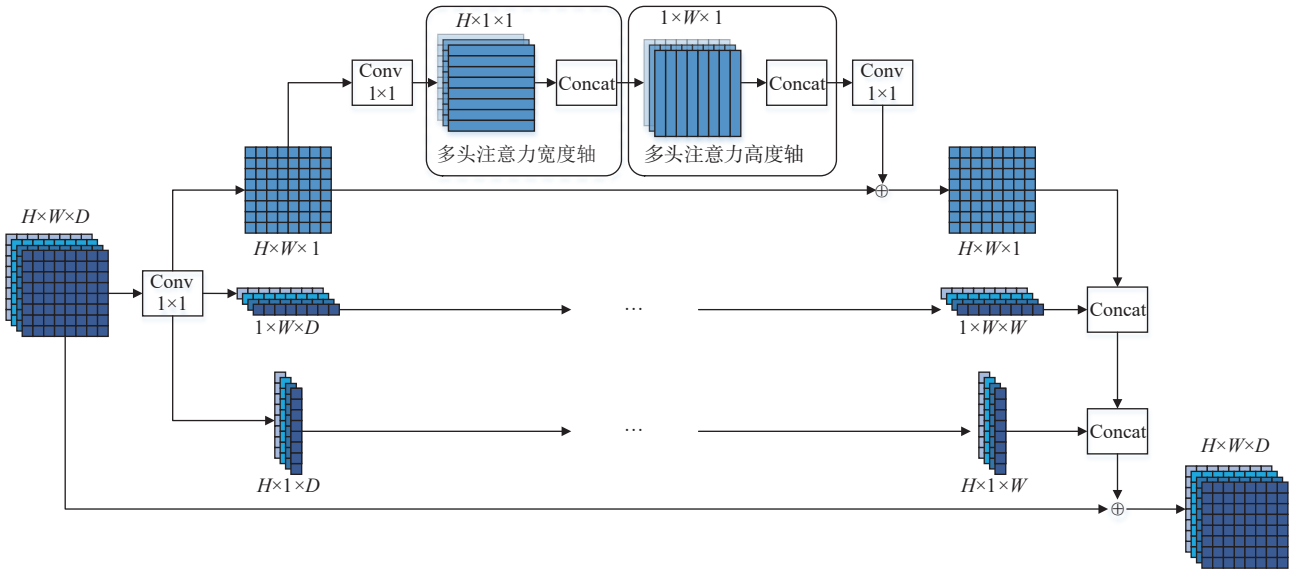

图 2 轴向注意力层分解

对于常规自注意力的计算, 以一张给定的 3D 图像 $x \in \mathbb{R}^{C_{\text{in}} \times H \times W \times D}$ 为例, 其高为 H, 宽为 W, 深度为 D, C_{in} 为输入通道数, 输出特征图 $y \in \mathbb{R}^{C_{\text{out}} \times H \times W \times D}$ 的注意力输出计算如下:

$$y_{hwd} = \sum_{a,b,c \in \mathcal{N}} Softmax_{abc}\left(q_{hwd}^{\mathrm{T}} k_{abc}\right) v_{abc} \tag{1}$$

其中, $\mathcal{N}$ 是位置格点, $q = W_Q x$ 表示特征图 x 的查询, $k = W_K x$ 表示键, $v = W_V x$ 表示值, $h \in \{1, \cdots, H\}$, $w \in \{1, \cdots, W\}$, $d \in \{1, \cdots, D\}$, $W_Q, W_K, W_V \in \mathbb{R}^{C_{\text{in}} \times C_{\text{out}}}$ 为可学习的投影矩阵.

自注意力机制能够捕捉到整个特征图上的相关性, 而不仅是局部区域的信息, 这使得网络能够理解更广泛的上下文. 不过这种机制在处理维度在二维以上的数据时计算成本非常高, 因为它的时间复杂度是与输入数据的立方成正比的 ($O(h^3 w^3 d^3)$). 因此, 它通常只适用于卷积神经网络的高层阶段, 这些阶段的特征图已经过下采样, 或者用于处理小尺寸的图像. 此外, 虽然全局池化操作可以减少计算量, 但它忽略了位置信息, 这对于视觉任务中识别结构和形状特征至关重要.

为了提升自注意力机制在处理空间位置信息方面的效能, 本研究引入相对位置编码[14]. 该方法的核心在于增强模型对元素间空间关系的敏感度. 具体实施策略包括计算元素在高度、宽度和深度 3 个维度上的相对偏移量, 即元素间在这 3 个方向上的距离差 (表示为 a–h、b–w、c–d). 随后, 将这些偏移量与预先定义的嵌入向量相结合. 嵌入向量的维度设定为通道总数的

1/3, 这一设计旨在平衡模型的参数复杂度与性能. 通过将 3 个方向上的嵌入向量进行串联, 我们能够构建出一个综合的相对位置向量$r_{a-h,b-w,c-d}$. 该向量能够为模型提供丰富的空间关系信息, 从而在自注意力机制中实现对位置信息的精确编码, 以此来丰富模型对位置信息的感知能力. 该方法可以帮助模型更好理解序列中各元素间的相对位置关系, 从而提高对结构和形状特征的识别能力, 尤其在视觉任务中. 该过程计算如下:

$$y_{hwd}=\sum_{a,b,c\in\mathcal{N}_i(h,w,d)} Softmax_{abc}\left(q_{hwd}^{\mathrm{T}}k_{abc}+q_{hwd}^{\mathrm{T}}r_{a-h,b-w,c-d}^{q}+k_{abc}r_{a-h,b-w,c-d}^{k}\right)\left(v_{abc}+r_{a-h,b-w,c-d}^{v}\right) \tag{2}$$

其中, $\mathcal{N}_i$是以像素x_{hwd}为中心提取的大小为$i\times i\times i\times C_{\mathrm{in}}$的区域.

为了解决自注意力机制计算复杂的问题, 我们采用轴向注意力机制. 该机制不是对整个特征图的所有元素同时进行注意力计算, 而是沿着特定的轴 (如高度、宽度或深度) 进行计算. 由于单一轴的长度通常远小于整个张量中元素的总数, 因此轴向注意力机制在计算和内存使用上都比传统的自注意力机制更加高效.

具体地, 根据文献[2], 轴向注意力机制通过将图像的轴向注意力层视为一种基于位置的二维自注意力操作, 此操作分别针对图像的高度、宽度和深度 3 个维度进行应用. 以高度轴为例, 注意力层的计算方法为:

$$y_{hwd}=\sum_{a,b,c\in\mathcal{N}_{1\times i\times i}(h,w,d)} Softmax_{abc}\left(q_{hwd}^{\mathrm{T}}k_{abc}+q_{hwd}^{\mathrm{T}}r_{a-h,b-w,c-d}^{q}+k_{abc}r_{a-h,b-w,c-d}^{k}\right)\left(v_{abc}+r_{a-h,b-w,c-d}^{v}\right) \tag{3}$$

为了进一步降低计算成本, 在已有的轴向注意力层 (式 (3)) 的基础上, 我们再次采用分解策略, 将二维的高度轴向注意力层细化为两个一维的注意力层, 分别沿着宽度和深度轴上进行展开计算. 以宽度轴为例, 这个过程可以表示为:

$$y_{hwa}=\sum_{a,b,c\in\mathcal{N}_{1\times 1\times i}(h,w,d)} Softmax_{abc}\left(q_{hwd}^{\mathrm{T}}k_{abc}+q_{hwd}^{\mathrm{T}}r_{a-h,b-w,c-d}^{q}+k_{abc}r_{a-h,b-w,c-d}^{k}\right)\left(v_{abc}+r_{a-h,b-w,c-d}^{v}\right) \tag{4}$$

上述分解过程和一维自注意力计算过程如图 2 和图 1(d) 所示. 这种设计策略将复杂的三维注意力计算分解为一系列简单的一维计算, 从而提高了计算效率且减少了资源消耗.

为了充分挖掘不同类别之间的相关性, 我们在所有编码层中集成了条件随机场 (CRF) 模型. 对于一系列图像$Y=[y_1,y_2,\cdots,y_n]$, 对应的积极正类为$Z=[z_1,z_2,\cdots,z_n]$, $Z(y)$代表所有合法类序列的集合, 而这一过程所对应的概率可以通过式 (5) 进行计算:

$$P(z\mid y)=\frac{\sum_{i=1}^{n}\mathrm{e}^{f(z_{i-1},z_i,y)}}{\sum_{z'}^{Z(y)}\sum_{i=1}^{n}\mathrm{e}^{f\left(z'_{i-1},z'_i,y\right)}} \tag{5}$$

其中, $f(z_{i-1},z_i,y)$计算z_{i-1}到z_i的转移得分以及z_i的得分. 该过程以最大化概率$P(z\mid y)$为优化目标, 在解码过程中, 采用维特比算法来寻找概率最高的路径.

在医学图像分割领域, 实验用的数据集通常较小, 很难学习位置偏差, 因此在编码长距离交互时并不总是准确的. 在学习到的相对位置编码不够准确的情况下, 将它们添加到相应的键、查询和值张量中会导致性能下降. 本文提出的带有条件随机场的轴向注意力层能够控制非局部上下文编码中可能产生的位置偏差, 轴向注意力的位置敏感可以在合理的计算开销内精确捕捉到长距离交互作用, 该设计能够提升网络在样本较少情况下的表现[2]. 条件随机场使用给定观察序列的整个标签序列的联合概率的单个指数模型[15], 加强了模型对标签关系的学习, 弥补了位置偏差的缺失.

2.2 层次多类训练

在医学图像分割任务中, 常因为现有的公开数据集标注样本数量有限和质量不高而受到限制. 在多标签分类领域, 标签之间的相互关联性是一个关键特征, 它对于提高分类任务的准确性至关重要. 为了充分利用这些标签间的层级联系和相似性, 本研究提出了一种创新的分层训练网络架构. 以膝关节图像分割为例, 我们采用了一种分层的标签划分策略, 如图 3 所示, 将膝关节图像中的标签划分为 3 个层级, 以实现更精细化的图像分割.

在实施过程中, 该网络架构将局部流细分为 3 个层级, 每个层级负责处理不同层次的分割任务. 第 1 层中, 网络专注于在背景 (Obj) 中识别并分割出标签 B 和 C, 为后续的层级分割提供基础. 进入第 2 层, 网络在已识别出 B 和 C 的背景下, 进一步细化这两个标签之间的边界. 最后, 在第 3 层中, 网络在已经明确区分 B 和 C 标签的背景下, 进一步细化其他标签. 具体来

说, 在 B 标签背景下, 网络分割出 DF (股骨远端)、TB (胫骨) 和 PB (髌骨); 而在 C 标签背景下, 网络分割出 FC (股骨软骨)、TC (胫骨软骨) 和 PC (髌骨软骨).

通过这种分层训练网络架构, 我们能够有效挖掘并利用多标签分类任务中标签间的层级联系和相似性, 从而提高膝关节图像分割的准确性和效率. 这种方法不仅适用于膝关节图像分割, 还可以推广到其他多标签分类任务中, 为相关领域的研究提供了新的视角和工具. 通过这种分层策略, 网络能够更精确地识别和分割图像中的不同标签, 从而提高分割的准确性和效率.

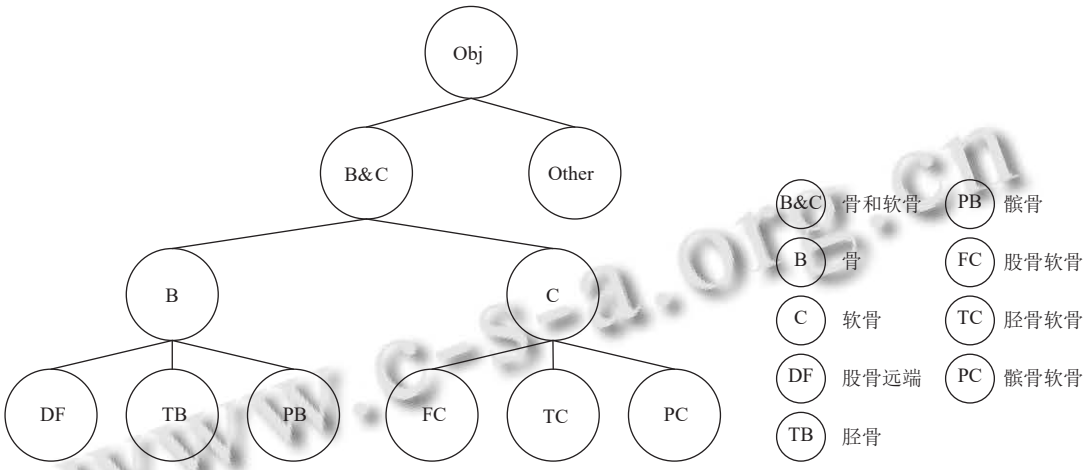

图 3 膝关节图像类关系示例

根据文献[5], 轴向自注意力机制的引入, 使得网络能够更好地利用位置信息而不会增加太多的计算成本. 显然, 在局部区域上使用 Transformer 是更快的, 但对于医学图像分割任务, 尤其是像膝关节软骨这样细小区域的分割是不够的. 逐块训练限制了网络学习区域之间的标签联系和位置关系. 为了提高网络对图像的整体理解, 我们使用了两个分支, 一个全局分支处理图像的原始分辨率, 一个局部分支处理图像块. 该部分从多个网络输出传播梯度, 包括每个分层级别的一个本地输出, 并使用相应级别中类别的本地损失函数来反向传播来自这些类别的梯度. 全局输出捕获了整个网络中向前传播的所有累积关系, 并通过所有层次的类的梯度进行反向传播[13].

在本文的网络架构中, 信息流动遵循两种路径: 全局流和局部流. 全局流从输入层出发, 经过所有全连接层, 最终到达全局输出层. 局部流同样从输入层开始, 但在通过全局全连接层后, 与局部全连接层相结合, 以生成局部输出. 这是为最大限度提高标记数据层次结构的学习能力而专门设计的, 为了生成最终预测, 所有本地输出都会被串联起来, 并与全局输出进行池化以得出一致的预测. 全局流负责携带来自第 i 级的信息到第 i+1 级层次结构. 它受到局部输出的影响, 这些局部输出通过反向传播每个级别特定于类别的梯度来强化全局信息流中的本地级别关系. 这种设计使得网络能够同时考虑全局和局部的特征, 以提高预测的准确性和全面性. 全局流的职能是将当前层的信息传递到网络的更高层次, 同时它也受到局部输出的调节; 这些局部输出通过反向传播机制, 在每个类别层级中增强全局信息流内的局部类别联系.

具体地, 给定一张三维图像 $x \in \mathbb{R}^{|F| \times C_{\text{in}}}$, C_{in} 表示输入维度, $|F|$ 表示特征数, E_G^1 表示网络全局流中的首个激活层 (即图 1 中的第 1 个全局 CAA), 其计算过程如下:

$$E_G^1 = ReLU\left(W_G^1 x + b_G^1\right) \tag{6}$$

其中, $W_G^1 \in \mathbb{R}^{|E_G^1| \times |F|}$ 是权重矩阵, $b_G^1 \in \mathbb{R}^{|E_G^1| \times 1}$ 是偏置向量, 它们直接从输入中学习全局信息, 随后的第 i 个全局层表示为:

$$E_G^i = ReLU\left(W_G^i\left(E_G^{i-1} + x\right) + b_G^i\right) \tag{7}$$

其中, $W_G^i \in \mathbb{R}^{|E_G^1| \times |E_G^{i-1}|}$, $b_G^i \in \mathbb{R}^{|E_G^i| \times 1}$. 令 I 表示网络总层数, 此处借助权重矩阵 $W_G^{|I|+1} \in \mathbb{R}^{|E_G^I| \times |C_I|}$、最后一层全局层输出 E_G^I 和偏置向量 $b_G^{|I|+1} \in \mathbb{R}^{|C_I| \times 1}$, 计算出全局流的分层输出 O_G. 其中 $|C_I|$ 表示最后一层全局输出的维度, O_G 表示为:

$$O_G = Sigmoid\left(W_G^{|I|+1} E_G^I + b_G^{|I|+1}\right) \tag{8}$$

对于局部流的处理, 其计算过程与全局层相似, 利用该层的全局激活函数、变换矩阵W_T^i和变换偏置向量b_T^i, 可以得到第 i 个局部层的激活表达式:

$$E_P^i = ReLU(W_T^i E_G^i + b_T^i) \tag{9}$$

局部层结构比全局层略复杂, 每一个局部层都带有一个单独的局部预测输出:

$$O_P^i = Sigmoid(W_P^i E_P^i + b_L^i) \tag{10}$$

在网络处理完所有局部层并生成预测之后, 这些局部预测结果会被整合, 形成整体的局部预测. 为了全面地融合全局视角与局部细节, 网络进一步将全局预测与整合后的局部预测相结合. 最终, 网络输出的预结果由这两部分信息联合决定, 表达为:

$$O_{\text{final}} = \beta_1 O_G + \beta_2 \sum_{i=1}^{I} O_P^i \tag{11}$$

其中, β_1和β_2是输出权重参数, 用来控制全局层和局部层之间的学习权重, 为了增强对局部特征的捕捉和学习, 其初值分别为 0.4 和 0.6, 以实现对细节信息的深入挖掘和利用.

分层多标签分类机制允许模型以分层的方式提取和处理类别信息, 从而更深入地挖掘类别间的相关性, 进一步提升数据的利用效率.

2.3 损失函数

在网络训练时, 我们分别对全局流和局部流应用了 Focal loss[16]来进行网络的训练. 全局流和局部流的损失分别表示为:

$$\mathcal{L}_G = -\alpha_1 \sum_{i=1}^{N} (1 - p_i)^{\alpha_2} \log(p_i) \tag{12}$$

$$\mathcal{L}_L = -\alpha_1 \sum_{j=1}^{I} \sum_{i=1}^{N} \left(1 - p_j^i\right)^{\alpha_2} \log\left(p_j^i\right) \tag{13}$$

其中, I 表示局部层的最大层数, N 表示总样本数.

在该损失函数中, 引入了两个关键的平衡系数来优化模型的性能, α_1代表正负样本的权重调整因子, 其值应随着正样本的稀缺程度而增加. α_2为用于调节样本难易程度的权重, 当α_2设置为 0 时, 损失函数将不再考虑样本的难易程度, 而是平等地对待所有样本, 此时损失函数将简化为标准的交叉熵损失函数. 这种简化有助于在不需要额外关注难分类样本时, 保持模型的泛化能力. 根据文献[16], 默认$\alpha_1 = 0.25$ ($g = 1$), $\alpha_2 = 2$. p_i和p_j^i是受基准真值指示参数$g \in \{-1, +1\}$控制的概率项, p_i由式 (14) 给出, p_j^i同理:

$$p_i = \begin{cases} O, & \text{if } g = 1 \\ 1 - O, & \text{otherwise} \end{cases} \tag{14}$$

其中, O 表示当前层的概率输出. 如果子节点n_c的预测生成概率超过了其父节点n_p, 这种情况被视为违规行为. 为了对此进行约束, 我们引入了一个额外的损失项, 用以惩罚此类越级行为:

$$\mathcal{L}_{V_i} = \max\left\{0, n_c - n_p\right\}^2 \tag{15}$$

总的损失函数组成如下:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_G + \mathcal{L}_L + \alpha_3 \mathcal{L}_{V_i} \tag{16}$$

其中, α_3为控制越级惩罚的平衡系数, 该参数初值为 0.5, 并在训练过程中根据余弦退火策略[17]进行调整.

3 实验

3.1 实验数据集

我们选取 K-Space[18]和 MOST[19]数据集来进行消融实验和对比实验. K-Space 数据集源自 fastMRI[20]项目, 该项目由 FAIR 和纽约大学朗格尼医学中心联合开展, 旨在通过 AI 技术提升 MRI 扫描的效率. MOST 数据集则是由 NIH 资助的公共数据集, 专注于收集包括膝关节在内的身体部位影像, 以支持相关研究.

这两个数据集中每幅图像均详细标注了 6 个区域: 股骨、股骨软骨、胫骨、胫骨软骨、髌骨和髌骨软骨. 表 1 展示了 K-Space 数据集上一张膝关节 MRI 的标签分布. 这些数据集图像按照 7:2:1 的比例划分为训练集、验证集和测试集. 此外, 我们对每张三通道的图像进行了尺寸调整, 统一裁剪为 256×256×128 像素, 并进行了归一化处理, 以便用于模型训练和评估.

3.2 实验细节及评价指标

训练过程中使用了大小为 128×128×128 的随机采样块. 实验的训练和测试均在 Ubuntu 20.04 操作系统环境下进行, 硬件配置包括一块 NVIDIA GeForce RTX 3090 显卡. 整个模型的训练和测试过程采用 Python 3.7 编程语言, 并利用 PyTorch 深度学习框架实现.

在本研究提出的 HCAA 模型架构中, 我们特别将条件随机场轴向注意力机制集成至全局分支的核心部分, 以强化模型对全局特征的捕捉能力. 与此同时, 为降低模型的参数复杂度, 我们在局部分支中有意省略

了位置编码. 尽管 HCAA 模型采用了多分支结构, 但我们在全局分支的设计中仅包含 3 个编码器和解码器层, 令每个局部分支仅对其相应的局部图像区域进行处理, 从而实现对局部特征的精细化分析. 通过这种设计, HCAA 模型在保持参数数量可控的同时, 能够有效平衡全局和局部特征的提取, 以提高模型的整体性能.

表 1 一张膝关节 MRI 的标签分布, 样本取自 K-Space 集

类别	像素数	体积 (mm^3)	强度平均值±标准差	占比 (%)
背景	9 100 683	1.39E+6	46.1815±33.0406	75.88
股骨	1 517 137	2.315E+5	15.4458±8.3527	12.65
胫骨	1 013 228	1.546E+5	17.3246±7.7035	8.45
髌骨	146 976	2.242E+4	18.3014±11.5719	1.23
股骨软骨	133 182	2.032E+4	105.2334±26.6202	1.11
胫骨软骨	58 479	8 922	97.4307±28.7349	0.49
髌骨软骨	23 835	3 636	115.8951±21.1733	0.2

在训练设置上, 网络通过 SGD 优化器进行了 2 000 次迭代训练, 初始学习率设为 0.01, 每 800 次迭代后学习率衰减 0.1, 批处理大小为 5.

为了确保比较的公正性, 我们没有使用任何后处理或模型集成技术. 在评估分割性能时, 我们关注了 Dice 相似系数 (DSC)、95% Hausdorff 距离 (HD95) 和模型参数量 (Params, 以百万计). 这些指标共同用于衡量模型分割结果的准确性.

3.3 消融实验结果及分析

本节针对 K-Space 数据集进行了广泛的实验分析, 以探究训练图像数量对模型性能的影响. 如图 4 所示, 当训练样本从 500 张减少至 100 张时, 模型的平均 DSC 仅表现出微小下降, 降幅不足 0.3%. 然而, 随着训练样本量的进一步减少, 平均 DSC 的下降趋势变得显著. 基于此发现, 我们决定采用 100 张图像作为网络训练的最优样本数量, 以平衡模型性能与计算资源的效率.

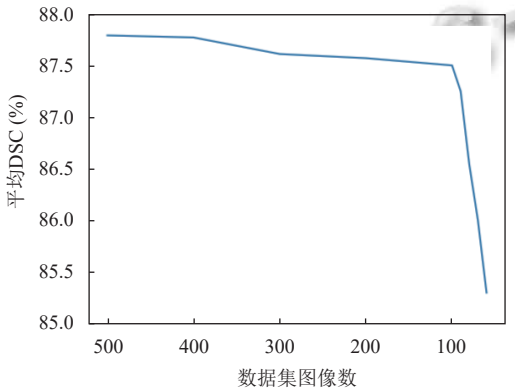

图 4 训练样本数量对模型平均 DSC 的影响

为了全面评估轴向注意力和层次化网络在实际应用中的有效性, 本研究在 K-Space 数据集上选取了 100 幅三维膝关节磁共振成像 (MRI) 图像进行了一系列实验分析. 在实验中, 我们比较了 5 种不同网络架构的性能表现: (1) 传统的三维 U-Net 网络[3]; (2) 整合了轴向注意力机制的 U-Net (以下简称为 Axial-DeepLab)[2]; (3) 采用层次化结构的三维 U-Net 网络 (以下简称为 H-UNet); (4) 层次轴向注意力网络 (以下简称为 HAA); (5) 集成了条件随机场的层次轴向注意力网络 (以下简称为 HCAA).

如表 2 所示, 在 K-Space 数据集上, 我们对比了本节所列 5 种方法的分割性能. 实验结果表明, 相较于 U-Net, H-UNet 在 3 个主要区域的 DSC 上实现了约 2% 的提高. 进一步地, Axial-DeepLab 在 DSC 上的表现提升更为显著, 达到了 2%–3%. 此外, HAA 方案相较于前两种方法, 在 DSC 上约有 1% 的提升, 同时 HD95 指标也有所降低, 表明层次化分割策略和轴向注意力机制的引入对于改进模型分割精度具有显著积极作用. 这是因为, 与 U-Net 相比, 使用基于 Attention 机制的神经网络来提取图像特征, 可以显著提升特征提取效率, 减少数据损失, 能够在数据较少的情况下仍保持强大的提取性能. 而层次多类分割机制的加入, 使得网络能够利用类之间的隐含关系, 提升数据利用效率.

表 2 消融实验结果

Methods	DSC (%)			HD95 (voxel)			Params (M)
	TC	FC	PC	TC	FC	PC	
U-Net	83.13	84.37	83.92	8.82	7.33	7.54	16.21
Axial-DeepLab	85.19	87.55	85.86	7.03	6.22	6.82	**4.45**
H-UNet	85.49	86.71	85.03	5.88	6.37	6.79	17.39
HAA	86.12	88.51	86.09	4.73	5.2	5.49	5.93
HCAA	**86.5**	**89.42**	**86.58**	**4.26**	**5.11**	**5.28**	5.93

在参数量方面, 为弥补注意力机制带来的庞大运算量, 我们引入了经过改进的轴向注意力机制, 显著减少了计算和内存需求, 验证了本方法在资源效率上的

优势. 与 HAA 相比, HCAA 在预测精度上也有小幅提升, 这进一步证实了条件随机场在增强模型学习效果方面的价值. 表 2 中, TC、FC 和 PC 分别指代胫骨软骨、股骨软骨和髌骨软骨.

图 5 呈现了 U-Net、Axial-DeepLab、H-UNet 以及本研究提出的 HAA 和 HCAA 方法的预测结果的直观比较. 通过可视化对比可以明显看到不同方法在图像分割中的性能差异. 图 5 用实线框标注了易出现欠分割的区域, 用虚线框标注了易过分割的区域. 观察结果表明, 轴向注意力机制有效地减少了模型在某些情况下的欠分割问题, 而引入层次分割机制和条件随机场显著增强了网络对类别间依赖关系的学习能力, 从而有效降低了过分割现象的发生. 这一改进不仅提升了模型的分割精度, 还优化了对复杂结构的识别能力.

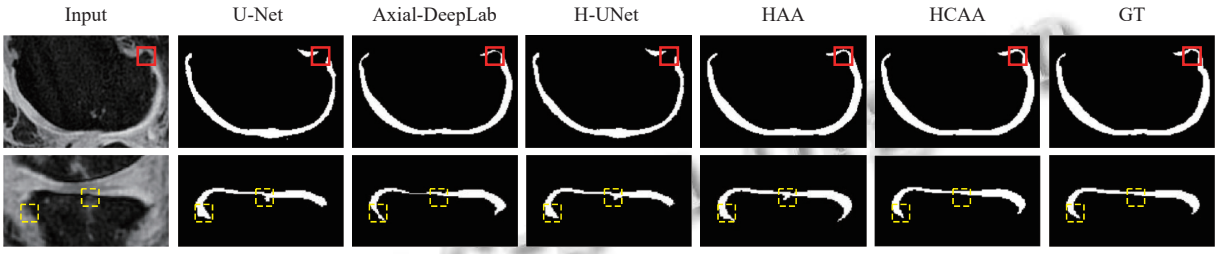

图 5 在 K-Space 测试集上的预测结果示例

这些视觉对比结果清晰地展示了不同方法在处理分割任务时的性能差异, 以及本研究所提出的技术如何通过结合轴向注意力、层次化处理和条件随机场来优化分割结果, 使得模型在保持高准确度的同时, 也能更好地适应不同的分割场景.

如图 6 所示, 我们绘制了网络训练过程中的收敛曲线, 在后处理阶段, 我们适当地对曲线进行平滑处理, 过滤掉了一些异常波动点, 使得曲线能够更好地展示模型的整体训练情况和收敛趋势. 可以看到, 当迭代次数达到 2 000 次时, 模型训练趋于收敛.

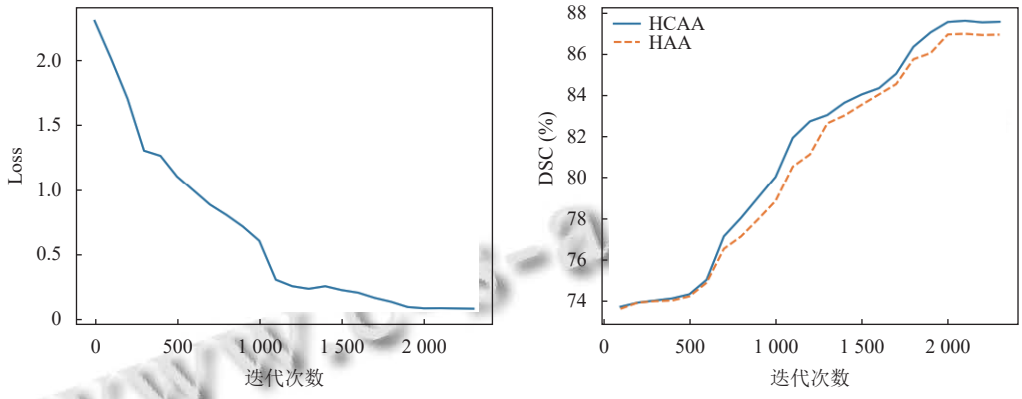

图 6 Loss 及 DCS 收敛曲线

3.4 对比实验结果及分析

本节将所提方法与多种当前流行的技术进行对比: 基于卷积的方法, 包括全卷积网络 (FCN)[21]、U-Net[3] (作为基准模型) 以及 Res-UNet[22]. 此外, 还探讨了几种基于注意力机制的模型, 包括 Axial-DeepLab[2] (作为基准模型)、医学 Transformer (MedT)[5]和 Swin-Unet[23]. 表 3 汇总了这些模型在 DSC、HD95 和 Params 这 3 个指标上的性能表现. 其中, Axial-DeepLab、MedT 和 HCAA 训练使用的数据集包含 100 张图像, 其他方法在训练时使用了 500 张图像, 此外, 我们额外添加了 Axial-DeepLab 在 500 张图像下的实验. 与表 3 总结的其他方法相比, 本文方法在股骨软骨 (FC)、胫骨软骨 (TC) 和髌骨软骨 (PC) 区域的 DSC 上均实现约 1% 的显著提升. 在模型参数方面, 尽管引入的层次分割机制使 HCAA 的参数量 (Params) 略高于轴向注意力 U-Net 和 MedT, 但与其他方法相比, 参数量仍然显著减少. 这一结果表明, 我们的方法在降低计算和内存需求的同时, 成功提升了分割精度.

表 3 本文提到的 7 种方法在 K-Space 数据集上的对比实验结果

Methods	Data quantity	DSC (%)			HD95 (voxel)			Cost
		TC	FC	PC	TC	FC	PC	Params (M)
U-Net[3] (baseline)	500	83.13	84.37	83.92	8.82	7.33	7.54	16.21
Axial-DeepLab[2] (baseline)	500	85.75	88.13	86.21	6.4	5.82	6.66	4.45
	100	85.19	87.55	85.66	7.03	6.22	6.82	4.45
Res-UNet[22]	500	84.78	87.21	85.14	7.64	5.9	7.22	8.27
FCN[21]	500	81.67	83.15	82.25	12.67	8.91	10.59	12.5
Swin-Unet[23]	500	85.37	88.03	85.79	6.33	5.97	5.81	27.15
MedT[5]	100	86.14	88.39	85.92	5.92	5.39	5.55	**4.2**
HCAA	100	**86.5**	**89.42**	**86.58**	**4.26**	**5.11**	**5.28**	5.93

图 7 展示了本文提到的 7 种方法在 K-Space 测试集上的分割结果示例. 可以看出, 尽管不同年龄、种族、健康状况和性别的人群在膝关节的形状、大小和形态上存在细微差异, 本文提出的 HCAA 分割网络预测结果与分割掩码之间的一致性更高. 这进一步证明了我们方法在处理实际临床数据中的有效性和可靠性.

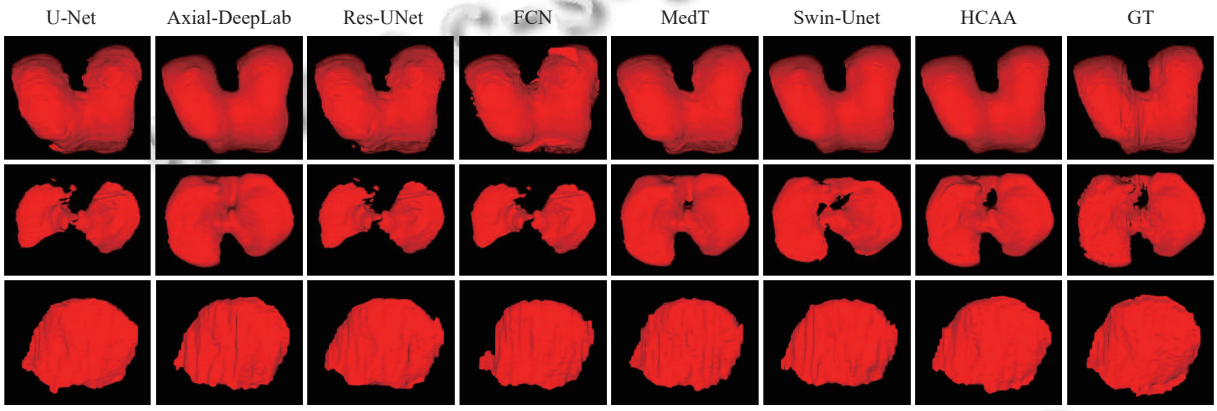

图 7 本文提到的 7 种方法在 K-Space 测试集上的分割结果示例

在 MOST 数据集上, 我们实施了与 K-Space 数据集相似的实验流程, 实验结果虽有波动, 但总体趋势与表 3 所示的实验结果保持一致, 具体数据详见表 4. 实验结果揭示了传统的全卷积网络和 U-Net 在处理软骨分割这类细节丰富且复杂的任务时的局限性, 这主要是因为可用于特征学习的训练样本数量不足, 限制了神经网络的归纳学习能力. 本文方法通过融合自注意力机制与类相关性, 显著增强了网络在面对有限样本数量时的分割能力. 该创新性方法使模型在数据稀缺环境下依然能够实现卓越的分割性能. 此外, 轴向注意力机制的整合使得网络在维持较低参数量的同时, 依然能够保持较高分割精度. 实验证明, 本研究所提出的方法在准确性、网络参数量和计算成本之间实现了最优的平衡.

表 4 本文提到的 7 种方法在 MOST 数据集上的对比实验结果

Methods	Data quantity	DSC (%)			HD95 (voxel)			Cost
		TC	FC	PC	TC	FC	PC	Params (M)
U-Net[3] (baseline)	500	82.06	83.52	82.59	9.9	8.68	9.42	16.21
Axial-DeepLab[2] (baseline)	500	85.97	87.6	86.32	5.23	5.85	5.71	4.45
	100	85.46	87.32	85.71	6.29	6.01	6.34	4.45
Res-UNet[22]	500	85.29	87.43	85.37	6.82	5.55	6.85	8.27
FCN[21]	500	79.76	82.82	80.83	14.31	10.67	12.9	12.5
Swin-Unet[23]	500	85.49	88.14	85.61	5.63	6.02	5.65	27.15
MedT[5]	100	86.1	88.06	86.05	5.17	5.12	5.37	**4.2**
HCAA	100	**86.32**	**89.09**	**86.62**	**4.92**	**5.03**	**5.33**	5.93

这些发现表明, 本研究提出的方法能够有效地应对样本稀缺的挑战, 尤其是在医学图像分割领域, 对于提高模型的实用性和临床应用潜力具有重要意义.

4 结论与展望

在面对数据量有限的膝关节软骨图像分割挑战时, 本研究提出了一种新颖的层次化多类分割和自注意力机制相结合的方法, 有效应对了类别关系利用不足和细小结构分割效果不佳的问题. 自注意力机制的优势在于其能够捕捉输入数据中的长距离依赖性, 这种能力使得它在处理噪声数据或小规模数据集时, 相较于传统的卷积神经网络可能更为有效, 尤其是在特征学习方面. 因此, 本研究中的网络设计能够在数据稀缺的环境中维持高性能.

通过整合层次化分割策略与条件随机场, 本研究提出的网络模型能够更加深入地挖掘类别间的相互依赖性. 这一机制不仅提高了对图像关键信息的学习能力, 还有效排除了异常值的干扰, 从而强化模型对关键特征的捕获能力. 总体而言, 与现有技术相比, 本研究提出的方法在分割精度上取得了显著提升, 在常用的评估指标上也展现了明显的进步, 对临床诊断具有潜在的辅助作用和实际应用价值.

展望未来, 我们期望进一步提升模型的稳定性和泛化能力, 探索如何有效融合多模态数据, 以提供更丰富的图像信息表示, 从而提高分割的准确度. 同时, 也致力于研究和开发更加适合医学领域的网络架构, 以期在未来的医学图像分析任务中取得更好的成果.

参考文献

1 Ho J, Kalchbrenner N, Weissenborn D, *et al*. Axial attention in multidimensional Transformers. arXiv:1912.12180, 2019.

2 Wang HY, Zhu YK, Green B, *et al*. Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 108–126.

3 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241.

4 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.

5 Valanarasu JMJ, Oza P, Hacihaliloglu I, *et al*. Medical Transformer: Gated axial-attention for medical image segmentation. Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention. Strasbourg: Springer, 2021. 36–46.

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

7 Sobek J, Inojosa JRM, Inojosa BJM, *et al*. MedYOLO: A medical image object detection framework. Journal of Imaging Informatics in Medicine, 2024, 37(6): 3208–3216. [doi: 10.1007/s10278-024-01138-2]

8 Mayne A, Perry R. Hierarchically classifying documents with multiple labels. Proceedings of the 2009 IEEE Symposium on Computational Intelligence and Data Mining. Nashville: IEEE, 2009. 133–139.

9 Dimitrovski I, Kocev D, Loskovska S, *et al*. Hierarchical annotation of medical images. Pattern Recognition, 2011, 44(10-11): 2436–2449. [doi: 10.1016/j.patcog.2011.03.026]

10 Yannis M, Heng YS, Axel C, *et al*. Two-stage learning-to-defer for multi-task learning. arXiv:2410.15729, 2024.

11 Xu ZK, Zhang BF, Li DY, *et al*. Hierarchical multilabel classification by exploiting label correlations. International Journal of Machine Learning and Cybernetics, 2022, 13(1): 115–131. [doi: 10.1007/s13042-021-01371-z]

12 Xu LL, Teng SJ, Zhao RY, *et al*. Hierarchical multi-label text classification with horizontal and vertical category correlations. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. ACL, 2021. 2459–2468. [doi: 10.18653/v1/2021.emnlp-main.190]

13 Wehrmann J, Cerri R, Barros R. Hierarchical multi-label classification networks. Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018. 5075–5084.

14 Ramachandran P, Parmar N, Vaswani A, *et al*. Stand-alone self-attention in vision models. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2019. 7.

15 Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., 2001. 282–289.

16 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense

object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318–327. [doi: 10.1109/TPAMI.2018.2858826]

17 Loshchilov I, Hutter F. SGDR: Stochastic gradient descent with warm restarts. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017.

18 Knoll F, Zbontar J, Sriram A, *et al*. fastMRI: A publicly available raw K-Space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning. Radiology: Artificial Intelligence, 2020, 2(1): e190007. [doi: 10.1148/ryai.2020190007]

19 Segal NA, Nevitt MC, Gross KD, *et al*. The multicenter osteoarthritis study: Opportunities for rehabilitation research. American Academy of Physical Medicine and Rehabilitation, 2013, 5(8): 647–654. [doi: 10.1016/j.pmrj.2013.04.014]

20 Zbontar J, Knoll F, Sriram A, *et al*. fastMRI: An open dataset and benchmarks for accelerated MRI. arXiv:1811.08839, 2018.

21 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 3431–3440. [doi: 10.1109/CVPR.2015.7298965]

22 Xiao X, Lian S, Luo ZM, *et al*. Weighted Res-UNet for high-quality retina vessel segmentation. Proceedings of the 9th International Conference on Information Technology in Medicine and Education (ITME). Hangzhou: IEEE, 2018. 327–331.

23 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 (ECCV) Workshops. Tel Aviv: Springer, 2022. 205–218.

(校对责编: 王欣欣)