小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2024-0522
2025 年 12 月 第 12 期
Vol.46 No.12 2025

并行双分支融合网络及其在医学图像分割中的应用

余辰婷[1],王朝立[1],孙占全[1],冯小晨[2],张雅颖[3]

[1](上海理工大学 光电信息与计算机工程学院,上海 200093)
[2](海军军医大学 第一附属医院 人工智能影像,上海 200433)
[3](同济大学 附属上海市第四人民医院 人工智能影像 脑卒中,上海 200081)
E-mail:clwang@usst.edu.cn

摘 要: 医学成像技术能够清晰展示患者的解剖结构,辅助医生非侵入性观察患者体内结构和功能.近年来,基于 CNN 和 Transformer 的图像分割算法在医学图像处理领域得到了广泛应用.但两者的结合方式往往过于简单,不能充分发挥其各自的优势.本文提出了一种新型的双分支融合网络(PDBF),该网络在继承编码器-解码器基本结构的基础上,设计了由深度可分离卷积分支和窗口自注意力分支组成的并行模块.这一双分支结构能够同时提取 Transformer 窗口内和窗口间的特征信息,从而有效扩大感受野.此外,模块中引入了跨分支的双向注意力融合机制,用以弥补因权重共享导致的通道或空间维度上的信息缺失问题.以 DSC 和 HD95 为评价指标,本文在 BCV、ACDC 及私有胰腺肿瘤数据集上的对比实验结果表明,PDBF 与其他医学图像分割网络相比,可以取得更好的分割效果.

关 键 词: 医学图像分割;CNN;Transformer;注意力机制
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2025)12-2967-09

Parallel Dual Branch Fusion Network and Its Application in Medical Image Segmentation

YU Chenting[1],WANG Chaoli[1],SUN Zhanquan[1],FENG Xiaochen[2],ZHANG Yaying[3]

[1](Institute of Optical-Electrical and Computer Engineering,University of Shanghai for Science and Technology,Shanghai 200093,China)
[2](Artificial Intelligence Imaging,The First Affiliated Hospital of Naval Medical University,Shanghai 200433,China)
[3](Cerebral Apoplexy,Artificial Intelligence Imaging,Shanghai Fourth People's Hospital Affiliated to Tongji University,Shanghai 200081,China)

Abstract: Medical imaging techniques are capable of clearly displaying the anatomical structures of patients,assisting doctors in non-intrusive observation of internal structures and functions. In recent years,image segmentation algorithms based on Convolutional Neural Networks (CNN) and Transformer have been widely applied in the field of medical image processing. However,the integration of these two methods is often too simplistic,failing to fully leverage their respective strengths. This paper introduces a novel Parallel Dual Branch Fusion network (PDBF),which based on the fundamental encoder-decoder architecture,designs a parallel module composed of a depth-wise separable convolution branch and a window-based self-attention branch. This dual-branch structure simultaneously extracts feature information within and between Transformer windows,thereby effectively expanding the receptive field. Moreover,the module incorporates a cross-branch bidirectional attention fusion mechanism to compensate for the loss of information in channel or spatial dimensions caused by weight sharing. Using the Dice Similarity Coefficient (DSC) and Hausdorff Distance 95 (HD95) as evaluation metrics,comparative experimental results on the BCV,ACDC,and a private pancreas tumor dataset demonstrate that the PDBF achieves superior segmentation performance compared to other medical image segmentation networks.

Keywords: medical image segmentation;CNN;Transformer;attention mechanism

0 引 言

医学成像技术是一种重要的医疗检查手段,是疾病诊断的重要参考依据.据世界卫生组织统计,医学诊断中 70% ~ 80% 依靠医学成像技术[1].在临床实践中,医疗诊断要求医生具备丰富的经验和精力,但在大量切片中确定器官和病灶位置的需求极大增加了医生手动注释的负担[2].另一方面,不同的医生在专业水平上难免存在差异,对于同样的病情,不同医生的诊断结果可能不同,甚至出现误诊的情况[3].随着计算机算力的高速发展,基于深度学习的图像分割网络作为一种可以快速而精确的从医学图像中自动提取器官或病灶,从而辅助医生进行决策的解决方案被提出.CNN 在图像分割领域已经显示出卓越的性能,其深层次的网络结构能有效学习图像的局部特征,这对医学图像分割至关重要.2015 年由 Ronneberger[4] 所提出的 U-Net 网络作为深度学习在医学图像分割领域的基石,它利用对称的编码器-解码器结构,构造

收稿日期:2024-12-05 收修改稿日期:2025-01-20 基金项目:国家自然科学基金项目(62173232)资助. 作者简介:**余辰婷**,女,2000 年生,硕士研究生,研究方向为深度学习、医学图像分割;**王朝立**(通信作者),男,1965 年生,博士,教授,研究方向为非线性控制、机器人控制、图像分割等;**孙占全**,男,1977 年生,博士,副教授,研究方向为人工智能、医学图像分割;**冯小晨**,男,1995 年生,住院医师,研究方向为腹部 CT/MRI、人工智能;**张雅颖**,女,1991 年生,博士,主治医师,研究方向为中枢神经系统 CT/MRI、人工智能.

了数据像素点从输入域到输出域的映射,并在当年的 ISBI 竞赛中获得多个第 1 名. Han[5] 提出的 ResU-net 使用残差模块代替了 U-Net 模型中的所有卷积模块,在肝脏 CT 图像分割中取得了很好的效果. Oktay 等人[6] 提出的集成注意力门(Attentiongate,AG)的 attentionU-Net,通过在网络中嵌入特殊模块而非改进网络本身结构来提高分割性能. 纯卷积神经网络具有自动提取特征、适应性强、处理速度快等优点,但由于其每个卷积核的感受野有限,使得卷积网络缺乏全局上下文建模的能力.

近年来,Transformer 架构在自然图像处理领域展现出了强大的性能优势,其核心在于通过全局注意力机制建立任意两像素点之间的关联,从而具备了极强的上下文建模能力. 这一特性使 Transformer 在捕捉长距离依赖和复杂图像特征方面优于传统的卷积神经网络. Transformer 结构最早由 Vaswani 等人[7] 在 2017 年提出,用于机器翻译领域. 2020 年,Dosovitskiy 等人[8] 提出了第一个完全基于自注意力机制的图像分类 Transformer 模型 ViT,这也是第一个使用 Transformer 来代替标准卷积的方法. 之后,Liu 等人[9] 基于自注意力机制,以层次化构建方式建立了通用的视觉骨干网络 Swin Transformer,改善了 VIT 模型计算量大,模型本身无法编码位置的问题. Mixformer[10] 针对跨窗口的信息交互方式进行改进,通过卷积提取窗口间信息从而舍弃了移动窗口结构.

然而,与自然图像相比,医学图像数据集通常存在规模较小、样本数量有限、原始尺寸较大的特点,这种差异给 Transformer 的直接应用带来了挑战. 一方面,Transformer 的自注意力机制导致计算复杂度随输入图像大小呈二次增长,使得其在大规模训练时计算成本高昂;另一方面,Transformer 缺乏卷积神经网络所具备的归纳偏置,在小样本条件下的泛化能力较差. 为了克服这些问题,在 CNN 架构的基础上加入 Transformer 的混合结构作为医学图像分割的折衷解决方案被提出,该方案同时平衡了计算效率与模型性能. 2021 年,首个将 Transformer 应用于医学图像分割领域的模型 TransUNet[11] 被提出,该网络模型依赖于经过预训练的 ViT 模型,通过把编码器中的深层卷积层与 Transformer 层进行简单级联从而实现了二维图像分割性能的提升. 在此基础上,TransBTS[12] 直接处理体积三维医学图像而非二维切片,实现跨空间维度的特征提取. Hatamizadeh 等人[13] 则提出了使用多头自注意力替换了编码器中卷积层的 UNETR,它将 3D 医学图像分割任务设计为一维序列到序列的预测问题. CoTr[14] 通过 Transformer 桥接编码器的所有阶段以捕获多尺度上的全局依赖性. 与以上更关注于探索 Transformer 与 CNN 混合应用的可行性,在结构上仅通过简单的替换或串联实现的研究相对应,TransFuse[15] 为了同时保留全局信息与细节信息,在编码器中对卷积与 Transformer 并行排列,提出了并行分支结构. 考虑到信息粒度的一致性,TransFuse 在融合分支间特征时选择了逆序融合. X-Net[16] 在 TransFuse 的基础上改进,使用两条完整且独立的 CNN 与 Transformer 分支同时提取局部和全局特征. PHTrans[17] 保留了并行分支结构,转而构建单独的可堆叠双分支子模块,建立分层的局部-全局表示. MS-Dual[18] 则通过两条额外的平行引导分支对不同尺度下的特征进行细化,强化有利于医学图像分割任务的关联特征,降低无关特征对分割结果的影响.

目前上述文献所讨论的 CNN 与 Transformer 混合的网络模型,在网络结构实现上可以分为 3 类:简单地在 U 型结构的不同层次插入 Transformer 层、在单层中将卷积与 Transformer 串联以及 CNN 与 Transformer 组成独立双分支结构. 这 3 类组合结构都不能完全发挥出 CNN 与 Transformer 结合的优势. 本文在此基础上提出了一个用于医学图像分割的并行 CNN 与 Transformer 双分支融合网络(PDBF). PDBF 遵循了经典的编码器-解码器结构,通过并行应用深度可分离卷积和窗口自注意力,分别处理窗口内和跨窗口的信息,扩大接受域. 其次,为了克服深度可分离卷积和多头自注意力具有的权重共享所带来的限制,PDBF 引入了分支间的双向注意力融合. 它可以补足 CNN 和 Transformer 分支所缺失的信息,增强空间和通道维度的建模能力.

本文工作的主要成果如下:1)针对医学图像分割问题,提出了一种新的分割网络 PDBF,并行耦合了 CNN 与 Transformer 这两种特征的提取重建方法,使网络兼顾局部细节和全局上下文的统一理解,进而提升对整体输入数据的表征能力;2)进一步的,在分支间引入了双向的注意力融合机制,利用分支间不同维度上的富裕信息进行相互补足,从而克服深度可分离卷积和窗口自注意力因权重共享所带来的限制;3)新提出的 PDBF 网络在公共数据集和私有胰腺肿瘤数据集上的实验结果均优于目前先进的图像分割竞争方法.

1 网络结构

1.1 总体框架

PDBF 结构图如图 1 所示,本文参考了 3D U-Net 的主体框架,采用 U 型编码器-解码器结构设计,同级编码器和解码器之间通过跳跃连接进行特征拼接. 相较于经典的卷积 U-Net 网络,本文希望能在同一层中同时构建局部特征和全局特征,因此提出了一个特殊的并行双分支注意力融合块. 该融合块包含深度可分离卷积和窗口自注意力两个分支,分别处理不同尺度的局部特征. Transformer 分支专注于窗口内信息,而卷积分支利用空间连续性提取跨窗口信息. 两者结合实现了不同感受域窗口下的特征整合,增强了对全局特征的表征,以适应复杂分割任务. 同时为了弥补不同分支在空间域或通道域上的缺失,融合块中引入了跨分支的双向注意力融合机制. 不同分支之间通过对应的注意力机制进行增强,完善不同分支的特征信息表达. 考虑到局部窗口自注意力分支中所应用的自注意力机制的计算复杂度与输入图像的像素大小成正比,若直接将原始图像数据的像素值作为 token 序列输入至计算网络模型中,会对图形处理单元(GPU)提出较高的性能要求. 在网络的具体实现上,本文采用了在浅层网络部分引入堆叠的级联卷积和下采样的方案. 通过下采样操作减小特征图的空间尺寸后,堆叠的级联卷积逐层提取高分辨率的底层特征,从而进一步优化特征图所包含的局部信息表征.

PDBF 整体网络结构由堆叠卷积模块(Stacked Conv Module)和双融合模块(BiFusion Module)组成. 堆叠卷积模块内包含了两层级联的纯卷积块(Conv Block)和下采样操作,初步提取图像浅层特征信息并减小或还原特征图大小,便

于后续计算;而双融合模块则由若干层并行双分支注意力融合块(Dual Branch Fusion Block)序列组成,单个序列中包含两个相同的串联双分支注意力融合块,负责进一步提取融合

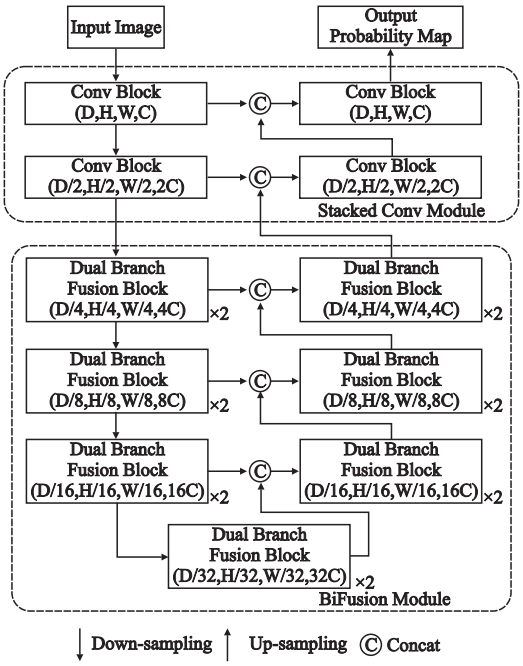

图 1　PDBF 网络架构

Fig. 1　Architecture of PDBF

图像的深层特征信息. 设输入图像块的体积像素尺寸大小为 $x \in \mathbb{R}^{D\times H\times W\times 1}$,其中 D、H、W 分别为输入图像块的深度、高度和宽度. 在编码器中,输入图像块首先通过浅层堆叠卷积模块,使通道数将上升至基础通道数 C,再逐层进行局部特征提取并下采样. 堆叠卷积模块内包含两层级联纯卷积块和下采样,通过堆叠卷积模块得到的特征映射像素尺寸大小将为 $x \in \mathbb{R}^{\frac{D}{4}\times\frac{H}{4}\times\frac{W}{4}\times 4C}$. 之后,纯卷积层初步提取的图像特征经由若干并行双分支注意力融合块组成的双融合模块,充分对局部和全局特征的分层表示进行建模,经过融合块多次下采样后的输出特征尺寸为 $x \in \mathbb{R}^{\frac{D}{32}\times\frac{H}{32}\times\frac{W}{32}\times 32C}$. 解码器部分的网络结构设置与编码器对称,深层网络的输出经上采样还原后与对应层级的编码器输出进行跳跃连接操作. 跳跃连接的输出将作为对应层的并行双分支交互模块或纯卷积块的输入,继续进行后续的图像空间维度恢复和像素级别的分类.

1.2　并行双分支注意力融合块

为了避免计算复杂度对网络造成限制,并行双分支注意力融合块仅部署在网络的深层阶段. 具体结构如图 2 所示.

1.2.1　并行结构

由于网络在特征提取过程中操作的累积效应,深层网络阶段所处理的特征映射已呈现出抽象化形态,这强迫网络更专注于特征的位置、结构等全局信息. 受结构所限,纯卷积结构的感受野偏小,只能捕捉局部特征,如边缘、纹理等;而与 CNN 互补的 Transformer 能够捕获长距离的全局依赖关系,有助于网络建立并理解特征像素点之间的相互作用和整体结构. 但在实际应用中,直接使用 Transformer 计算量过于庞大. 因此,将完整的特征图分为若干非重叠的窗口,并对单个窗口域进行多头自注意力的操作作为提高 Transformer 计算效率的替代方案被提出,这种方法以牺牲接受野为代价大幅减小了计算量. 之后的研究以此为基础,采取多种方法增加跨窗口信息交互,模拟跨窗口的连接,包括移位[9]、卷积[19]或展开[20]等. 这些方法虽然能够分别捕捉窗口内和跨窗口的信息,但它们通常在网络中串行排列,跨窗口的信息交互依赖于窗口内信息的再提取,这导致了局部关系和全局关系交织较少.

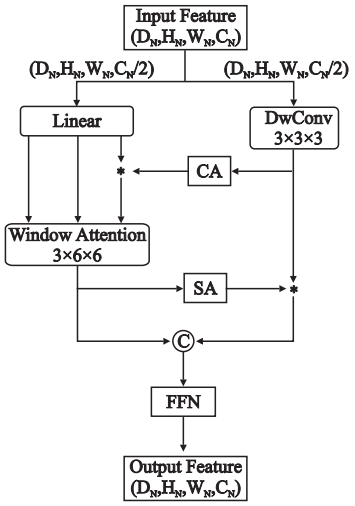

图 2　并行双分支注意力融合块

Fig. 2　Parallel dual branch attention fusion block (dual branch fusion block)

结合卷积及跨窗口自注意力的特点和性质,本文设计了一种并行双分支结构. 该结构包括深度可分离卷积分支和局部窗口 Transformer 分支,二者分别独立提取不同窗口大小下局部特征的分层表示进行建模. 对于 Transformer 分支,该分支仅通过窗口自注意力,进行限定窗口范围内的信息提取. 对于深度可分离卷积分支,卷积操作由于其固有的空间连续性,通过卷积核在特征图上的滑动操作,可以自然地捕捉到相邻区域间的连续性,从而有效地提取窗口间的信息. 因此,平行卷积分支不仅实现了在进行多头自注意力特征提取的同时对窗口间信息的建模,而且确保了模型在处理复杂医学图像分割任务时,能综合理解局部细节与全局上下文信息,从而增强了网络模型对输入数据的整体表征能力. 此外,通过分支间的特征融合,网络能够在不同感受窗口尺寸下整合局部特征,从而获得更为全面的全局特征信息,扩大接受域. 得益于此,模块无需额外的移动窗口结构,减小了计算消耗.

并行双分支注意力融合模块由两条平行路径组成,分别包含基于不同窗口大小的窗口自注意力机制和深度卷积操作. 在深度可分离卷积分支中,采用了 3×3×3 大小的卷积核,以平衡计算效率和精度. 而在局部窗口 Transformer 分支中,窗口大小根据不同数据集的特性进行调整. 图 2 以 BCV 数据集为例,采用 3×6×6 的窗口大小. 设 $x \in \mathbb{R}^{D_N\times H_N\times W_N\times C_N}$ 作为并行双分支注意力融合块的输入,通过线性层对其进行映射,生成两个独立的张量,维度均为 $D_N \times H_N \times W_N \times \frac{C_N}{2}$,分别作为两条分支的输入. 两个分支的输出经过归一化处理后

进行拼接融合,随后传递至前馈神经网络(FFN)升维,在高维空间中探索更丰富的特征表示. 最终,得到包含局部信息的全局特征,该全局特征的维度与输入维度保持一致.

相比传统的卷积网络或 Vision Transformer,该模块中的并行结构通过并行应用小尺寸卷积核和局部窗口 Transformer,PDBF 能够同时建模窗口内和窗口间的关系,建立跨窗口的连接模型,从而以更小的计算代价捕获更全面的特征表达.

1.2.2 双向注意力融合

深度可分离卷积的过滤器独立作用于各自的通道且仅在单通道内滑动并提取特征,该设计减少了模型参数和计算量,同时保证了有效的特征提取能力,但使得每个过滤器在各自通道的空间维度上共享权重. 而基于查询(Query)、键(Key)和值(Value)框架的 Transformer 则采取相反的策略. 它通过在空间维度上动态计算权重,并在通道维度上共享权重来实现特征建模. Transformer 通过 Query 和 Key 矩阵之间计算相似度得到权重矩阵,从而使模型能够在处理任意位置的输入时综合考虑序列中的所有位置. 由于该权重矩阵通过比较序列中不同位置的元素计算获得,主要反映的是序列内元素在空间维度上的相互关系,而非通道维度上的关联.

通常,共享权重的应用会对共享维度的建模能力产生限制. 为了克服这一问题,常见的做法是生成与数据特征相对应的权重. 因此,本文在分支间双向注意力融合机制中,设计结合空间注意力和通道注意力,实现了跨平行分支的信息融合. 在网络架构设计上,针对两个分支各自的特性,采取相匹配的差异化处理策略. 从卷积分支出发,深度可分离卷积处理后的特征经过通道注意力的筛选与提炼,生成与当前特征相匹配的通道权重并作为 Transformer 分支的额外输入,以提升窗口自注意力在通道维度上的建模能力. 得到的 Transformer 分支输出也将同时作为空间注意力的输入,用以增强同一阶段深度卷积分支已获得的提取特征. 值得注意的是,在 Transformer 分支中,由于自注意力在通过 Query 矩阵和 Key 矩阵计算权重矩阵这一过程中仅考虑空间域上像素与像素之间的联系,而不考虑通道间的关联,因此将 Q、K 矩阵与通道注意力权重结合是无效的,本文选择仅在 Value 矩阵上进行注意力融合操作.

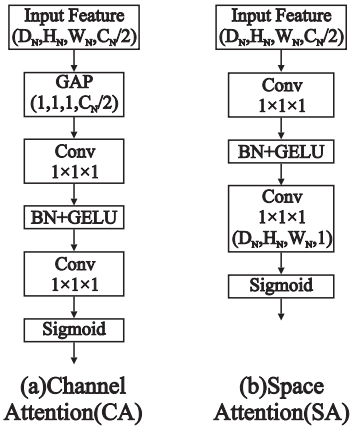

图 3 通道注意力块 CA 和空间注意力块 SA

Fig. 3 Channel attention & space attention

本文借鉴 SE 层[21]和 CBAM[22],设计了通道注意力块 CA 和空间注意力块 SA,如图 3 所示.

对于通道注意力块 CA,通过对输入特征进行全局平均池化,将每个通道的空间信息浓缩为一个标量值. 该标量值作为一个简洁的通道描述符,编码了每个通道的全局概要信息. 随后,利用两个串联的 1 × 1 × 1 卷积层,并插入批归一化(BN)和激活函数(GELU),以轻量级的方式捕捉通道间的相互关系,学习不同通道的相对重要性. 最终通过 sigmoid 激活函数在通道维度上生成注意力权重,为每个通道分配一个 0 ~ 1 之间的权重值,以表征其在当前输入中的重要性. 对于空间注意力块 SA,输入特征首先通过一系列 1 × 1 × 1 卷积层进行融合和降维,生成综合的空间特征图. 在此过程中,特征的通道数被缩减至 1,从而引导模型关注特征图的空间信息而非通道信息. 通过这一操作,每个空间位置均可获得一个综合所有通道信息的单一值,用以评估该位置的空间重要性. 最终,利用与通道注意力类似的 sigmoid 层对空间特征图进行处理,生成空间注意力图.

2 实验验证

2.1 数据集

The Multi-Atlas Labeling Beyond the Cranial Vault (BCV)[23],出自 MICCAI 2015 举办的 Workshop,由范德堡大学医学中心(Vanderbilt University Medical Center)提供. 包含共 30 份已标注腹部 CT 扫描,数据集由 85 ~ 198 张切片组成,切片厚度在 2.5mm ~ 5.0mm 之间,像素大小为 512 × 512. 分为 18 个训练样本和 12 个测试样本. 其完整数据集包含 13 个待分割目标,为了确保所进行的分割性能对比研究的客观性和可比性,且为了与其他独立研究者的数据集进行一致性的评估,本文从完整的 BCV 数据集中选取了 8 个类别(主动脉、胆囊、脾脏、左肾、右肾、肝脏、胰腺、脾脏、胃)进行研究.

The Automated Cardiac Diagnosis Challenge(ACDC)[24],出自 2017 年 MICCAI 心脏自动诊断挑战赛. 数据集包含 100 例试者心脏舒张末期和收缩末期的磁共振影像以及对应的左心室心内膜、心外膜和右心室心外膜的手工标注轮廓,由 6 ~ 21 张切片组成,切片厚度在 5.0mm ~ 10.0mm 之间,像素大小由 154 × 154 ~ 428 × 512. 分为 70 个训练样本、10 个验证样本和 20 个测试样本,分割目标为左心室(LV),右心室(RV)和心肌(MYO).

由长海医院所提供的胰腺肿瘤数据集,包含 117 例患者的 3D CT 成像,由 130 ~ 300 张切片组成,切片厚度在 0.8mm ~ 2.0mm 之间,像素大小为 512 × 512. 分为 94 个训练样本和 23 个测试样本,分割目标为胰腺的肿瘤区域(包括偏良性的 SCN 浆液性囊性肿瘤与偏恶性的 MCN 粘液性囊性肿瘤).

2.2 评价指标

为了有效评估分割算法的性能,通常采用度量网络预测的分割结果与由人类专家标注的真实分割结果之间相似度的方法. 这种评估方式能够客观衡量算法性能的优劣. 本文选择使用 DSC[25](Dice Similarity Coefficient)和 HD95[26](Hausdorff Distance-95%)作为网络分割性能的评价指标. 这两个评价指标从不同的角度反映了分割质量,DSC 关注的是预测结果与真实标签的整体一致性,而 HD95 则专注于衡量分割

轮廓与真实边界之间的最大偏差[27].

2.2.1　DSC

DSC 是衡量两个集合相似度的指标,反映了预测分割结果与真实标签之间的相似程度. DSC 的计算考虑了正确分割的像素数量以及两个集合的总体大小. DSC 值越接近 1,表明分割结果与真实标签的重合度越高,即两个集合的相似度越大.

其定义为:对于两个集合 X 和 Y:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{1}$$

其中,$|X|$ 和 $|Y|$ 表示两个集合中所包含的元素的数量.

对于多分类的医学图像分割任务,针对每个分割类别单独进行 DSC 的计算,并基于类别种类求出平均 DSC 作为当前网络对整个数据集的分割性能表现的评判准则. 对于每个类别 i,DSC 的计算公式为:

$$DSC = \frac{2 \times TP_i}{2 \times TP_i + FP_i + FN_i} \tag{2}$$

其中,TP_i 为正确预测为类别 i 的像素数量(真正例),FP_i 为错误预测为类别 i 的像素数量(假正例),FN_i 为未正确预测为类别 i 的像素数量(假反例).

2.2.2　HD95

HD95 是衡量分割轮廓准确性的指标,反映了预测结果中分割轮廓的整体准确性. HD95 关注于分割结果中的点到真实标签边界最大距离的 95 百分位数,即分割轮廓在整体上与真实边界的接近程度. HD95 值越低,说明分割轮廓与真实边界的差异越小,整体准确性越高.

对于 HD95,定义如下:

$$HD_{95}(P,G) = max\{hd_{95}(P,G), hd_{95}(G,P)\} \tag{3}$$

其中:

$$hd_{95}(P,G) = \max_{p \in P}\{\min_{g \in G} \| p - g \|\} \tag{4}$$

$$hd_{95}(G,P) = \max_{g \in G}\{\min_{p \in P} \| g - p \|\} \tag{5}$$

P 为预测体素的点集合,G 为真实体素的点集合,$\| \cdot \|$ 是点集 P 和点集 G 间的距离范式. $hd_{95}(P,G)$ 和 $hd_{95}(G,P)$ 分别为预测体素到真实体素及真实体素到预测体素之间的最大第 95 百分位距离.

2.3　损失函数

2.3.1　Dice 损失函数

Dice 损失函数[28]依据 Dice 系数构建,它对类别不平衡问题具有一定的鲁棒性,这表明 Dice 损失函数在分割结果中对于小目标的检测具有更高的敏感度.

Dice 损失函数的定义如下:

$$DiceLoss(P,G) = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i} \tag{6}$$

其中,P 为预测的图像,G 为真实的标签,p_i 和 g_i 分别为预测图像 P 和真实标签 G 中第 i 个像素的数值,N 为图像中的总像素数量.

2.3.2　交叉熵损失函数

交叉熵损失函数[29]用于衡量模型预测的像素概率分布与真实标签的像素分布之间的不一致性,它能够帮助模型更为精确地学习并捕捉到各个类别的边界,促进模型识别出图像中的微小变化.

交叉熵损失函数的定义如下:

$$L_{CE}(P,G) = -\frac{1}{M}\sum_{i=1}^{M}\sum_{C=1}^{C}(G_{ic}\log(p_{ic})) \tag{7}$$

其中,P 为预测的图像,G 为真实的标签,M 为训练样本的总数,C 为训练样本中的类别总数,G_{ic} 为第 i 个样本中第 C 个类别的标签所对应的 one-hot 编码,p_{ic} 是模型的输出经由 softmax 函数转换后所得到的概率分布,它代表了第 i 个样本中属于第 C 个类别的概率值.

2.4　实验环境及参数设置

本文实验使用单个 RTX 3090 GPU 进行训练,基于 Python3.8 及 Pytorch 深度学习框架实现 PDBF. 在训练阶段,使用 nnUNet[30]的训练框架,采用 nnU-Net 的跨步策略进行下采样和上采样,网络基础通道数 C 为 24,由浅到深不同阶段的并行双分支注意力融合块所使用的多头自注意头数为[3,6,12,12]. 对于 BCV、ACDC 和私有胰腺肿瘤数据集,局部窗口自注意力的窗口大小分别设置为[3,6,6]、[2,8,7]和[5,6,5]. 在训练阶段,分别从 BCV、ACDC 和私有胰腺肿瘤数据集的原始扫描中随机裁剪大小为、和的子卷作为输入. 损失函数为联合使用的 Dice 损失和交叉熵损失函数.

3　实验结果

3.1　对比实验

本文将 PDBF 与目前比较先进的方法进行了比较,对比目标网络主要以具有混合框架的医学图像分割网络为主.

在 BCV 数据集上的分割结果如表 1 所示. 其中,Swin-Unet、TransUNet、LeViT-Unet、MISSFormer、nnFormer、UNETR 和 PHTrans 的分割结果引用自原论文,见参考文献[11,13,17,31-34],其余网络的分割结果引用自参考文献[17]. 进一步的,Swin-Unet、TransUNet、LeViT-Unet 和 nnFormer 的分割结果使用了在 ImageNet 上预训练的权重初始化网络,而其余网络则在 BCV 数据集上从头开始训练.

本文所提出的网络模型 PDBF 在多项评估指标上表现优异,达到了最佳或接近最佳的水平,充分展现了其在医学图像分割任务中的显著优势. 在总体分割性能方面,PDBF 在多器官分割平均 DSC 和 HD 两个指标上均取得了最优异的效果,分别为 89.21%(DSC ↑)和 8.14(HD ↓). 比之前的最佳模型在平均 DSC 上高出 0.66%,在 HD95 上降低了 0.54,这证明了本文所提出的 PDBF 网络在总体分割精度上具有更优的性能,且网络能够有效减少边界误差,对复杂解剖结构的捕获和分割更为准确. 在器官级别的性能对比中,PDBF 在脾脏(Spl)和胆囊(Gal)等关键器官的分割中均达到了最高的 Dice 系数,证明了网络对器官形状和解剖特征的高效捕捉能力. 此外,在较难分割的小器官胰腺(Pan)上,PDBF 的性能也优于其余网络模型,展现了 PDBF 对低对比度和小目标的良好适应性.

在 ACDC 数据集上的分割结果如表 2 所示. 其中 Swin-Unet、TransUNet、LeViT-Unet、MISSFormer、nnU-Net、PHTrans 和 UNETR 的分割结果引用自原论文,见参考文献[11,13,17,30-32,34],除 MS-Dual 外的其余网络分割结果引用自参

考文献[17]. 进一步的,Swin-Unet、TransUNet、LeViT-Unet 和 nnFormer 使用了在 ImageNet 上预训练的权重初始化网络,而其余网络则在 ACDC 数据集上从头开始训练.

表 1 BCV 数据集分割结果

Table 1 Segmentation results on BCV dataset

Methods	DSC ↑(%)	HD95 ↓(mm)	Aot(%)	Gal(%)	Kid(L)(%)	Kid(R)(%)	Liv(%)	Pan(%)	Spl(%)	Sto(%)
Swin-Unet * [31]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.6
TransUNet * [11]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
LeViT-Unet * [32]	78.53	16.84	87.33	62.23	84.61	80.25	93.11	59.07	88.86	72.76
MISSFormer[34]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
CoTr[14]	86.33	12.63	92.10	81.47	85.33	86.41	96.87	80.20	92.21	76.08
nnFormer * [33]	86.45	14.63	89.06	78.19	87.53	87.09	95.43	81.92	89.84	82.58
nnU-Net[30]	87.75	9.83	**92.83**	80.66	84.86	89.78	**97.17**	82.00	92.39	82.31
UNETR[13]	79.42	29.27	88.92	69.80	81.38	79.71	94.28	58.93	86.14	76.22
Swin-UNETR[35]	85.78	17.75	92.78	76.55	85.25	89.12	96.91	77.22	88.70	79.72
PHTrans[17]	88.55	8.68	92.54	80.89	85.25	**91.30**	97.04	83.42	91.20	**86.75**
PDBF	**89.21**	**8.14**	92.79	**83.32**	**87.72**	88.74	96.83	**83.61**	**95.48**	85.16

注:标 * 处表示网络使用了 ImageNet 进行预训练

在该数据集上,PDFB 在平均 DSC 以及不同解剖结构的分割指标(RV、MYO 和 LV)上均取得了最优的分割结果,其平均 Dice 系数达到了 91.97%,超过了所有对比方法,显示了其卓越的全局分割能力. 对于较复杂结构的分割任务右心室(RV)和心肌(MYO),PDBF 的 DSC 评价指标分别达到了 90.21% 和 89.86%,均为所有对比方法的最高值. 这表明 PDBF 能够有效捕捉和重建复杂形状及边界信息.

表 2 ACDC 数据集分割结果

Table 2 Segmentation results on ACDC dataset

Methods	DSC ↑(%)	RV(%)	MYO(%)	LV(%)
Swin-Unet * [31]	90.00	88.55	85.62	95.83
TransUNet * [11]	89.71	88.86	84.53	95.73
LeViT-Unet * [32]	90.32	89.55	87.64	93.76
MISSFormer[34]	90.86	89.55	88.04	94.99
nnFormer * [33]	91.62	90.27	89.23	95.36
nnU-Net[30]	91.36	90.11	88.75	95.23
UNETR[13]	88.61	85.29	86.52	94.02
MS-Dual[18]	91.14	89.16	89.74	94.51
PHTrans[17]	91.79	90.13	89.48	95.76
PDBF	**91.97**	**90.21**	**89.86**	**95.83**

注:标 * 处表示网络使用了 ImageNet 进行预训练

表 3 胰腺肿瘤数据集分割结果

Table 3 Segmentation results on the pancreatic tumor dataset

Methods	DSC ↑(%)	HD95 ↓(mm)
Swin-Unet	67.03	29.68
TransUNet	63.49	35.97
Swin-UNETR	73.18	25.26
CoTr	74.29	16.42
nnU-Net	72.89	31.27
UNETR	71.18	38.51
nnFormer	74.36	15.99
MS-Dual	71.94	28.5
PHTrans	73.22	29.17
PDBF	**76.76**	**13.39**

在私有胰腺肿瘤数据集上的分割结果如表 3 所示,所有网络都没有使用预训练权重.

在分割精度方面,PDBF的DSC评价指标达到了76.76%,相比于其它网络模型具有明显提升. 在分割边界准确性方面,PDBF 的 HD95 评价指标降低至 13.39,相比其他网络模型展现出更加卓越的边界拟合能力. 这表明,对于大小和形状多变、位置复杂的小型待分割目标,PDBF 在提升全局分割质量的同时,能够更好地捕捉目标边界的精细细节,从而适应医学图像中复杂且多样的形态特征.

针对以上 3 个不同数据集的分割结果对比表明,本文所提出的 PDBF 网络可以与其他已提出的先进方法竞争.

3.2 定性评价

图 4 展示了在 BCV 数据集上,Swin-UNETR、nnFormer、PHTrans 和本文所提出的 PDBF 网络的分割结果与原始标签的定性可视化对比. PDBF 在胰腺和胆囊等小器官上预测的边界更加平滑,漏分割现象更少;在肝脏等大器官上,其预测轮廓也更加完整.

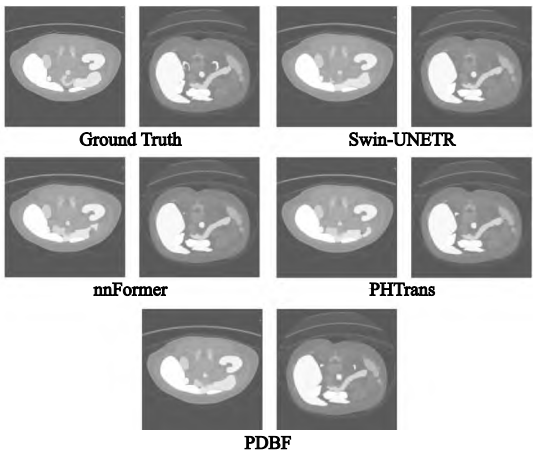

图 4 基于 BCV 数据集的可视化对比

Fig. 4 Visual comparison based on BCV dataset

图 5 展示了在 ACDC 数据集上,nnFormer、PHTrans 和 PDBF 的分割结果与原始标签的对比. 通过可视化对比可以看出,PDBF 在所有分割目标(LV、RV、MYO)上均展现出更好的边界细节保留能力和复杂形状适应性. 尤其在边界模糊或形态复杂的区域,PDBF 相对更少存在过分割或欠分割现象.

图 6 展示了在胰腺肿瘤数据集上,Swin-UNETR、nnU-Net、PHTrans 和 PDBF 的分割结果与原始标签的对比. PDBF 网络的分割结果边界更加贴近真实标签,对肿瘤区域的识别遗漏更少.

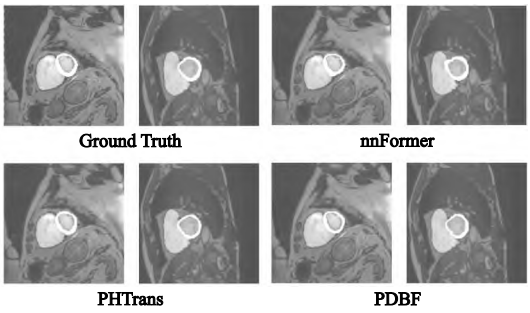

图 5　基于 ACDC 数据集的可视化对比

Fig. 5　Visual comparison based on ACDC dataset

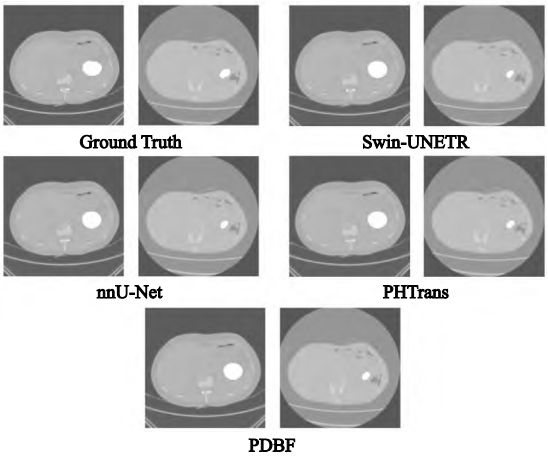

图 6　基于胰腺肿瘤数据集的可视化对比

Fig. 6　Visual comparison based on the pancreatic tumor dataset

3.3　消融实验

基于 BCV 数据集,本文以 3D Swin-Unet 网络作为基线网络. 在此基础上本文逐步整合了 PDBF 的各组成部件进行消融实验,以探索不同的组成部分对网络分割性能的影响.

表 4 给出了结构消融实验的定量结果,后一步的额外添加都将在前一步的操作基础上进行. “ + Stacked Conv Module”代表在网络的浅层使用堆叠卷积模块替代原本的跨行卷积操作. “ + BiFusion Module w/o Bidirectional Attention Fusion”则表示将 3D Swin-Unet 中原本的 Swin Transformer Block 替换为本文所提出的双融合模块,但仅包含深度卷积与局部窗口的并行双分支结构而不包含分支间的双向注意力融合. “ + Bidirectional Attention Fusion”则表示在前一步的基础上,继续添加空间和通道的分支间双向注意力融合,组成完整的双融合模块. 以上的实验结果足以证明,堆叠卷积模块的引入显著增强了网络在特征提取早期对细粒度细节的捕获能力. 此外,使用 CNN 与 Transformer 的并行双分支结构可以取得比串行连接的 Swin Transformer Block 更好的分割效果,分支间的双向注意力融合进一步优化了特征的信息表达,使网络获得更好的性能提升. 与基线网络 3D Swin-Unet 相比,本文所提出的 PDBF 在平均 DSC 上提升了 5.56%,在 HD 指标上降低了 10.5. 结果表明,使用并行结构及分支间的双向注意力融合来进行特征的信息提取融合是有效的.

本文还增加了并行双分支注意力融合块中局部窗口和移动窗口之间的对比研究. “ + shifted windows”表示在原并行双分支注意力融合块的基础上,参考 Swin Transformer Block 结构,将分支中的局部窗口自注意力修改为移动窗口自注意力. 结果显示,增加移动窗口结构无法提供比目前的局部窗口自注意力更优异的改进性能,这说明了本文的并行结构可以提供充足的跨窗口信息聚合. 考虑到增加移动窗口所需要的额外计算量,出于平衡计算消耗和分割性能之间的平衡,本文选择不保留移动窗口结构.

表 4　PDBF 网络架构的消融研究

Table 4　Ablation study on the architecture of PDBF

Methods	DSC ↑ (%)	HD95 ↓ (mm)	Aot(%)	Gal(%)	Kid(L)(%)	Kid(R)(%)	Liv(%)	Pan(%)	Spl(%)	Sto(%)
3D Swin-Unet	83.65	18.64	88.07	73.03	83.26	85.54	94.84	78.18	87.03	79.28
+ Stacked Conv Module	86.14	16.42	91.83	79.90	86.04	86.97	96.32	76.31	90.54	81.19
+ BiFusion Module w/o Bidirectional Attention Fusion	88.08	11.79	93.5	83.16	86.51	87.89	97.7	83.75	89.35	82.74
+ Bidirectional Attention Fusion	**89.21**	**8.14**	92.79	83.32	87.72	88.74	96.83	83.61	95.48	85.16
+ shifted windows	89.23	8.09	93.23	82.99	86.39	88.81	97.58	83.74	95.29	85.84

表 5　双融合模块深度的消融研究

Table 5　Ablation study on the depth of the BiFusion module

depth heads of Multi-headed Self-attention	DSC ↑ (%)	HD95 ↓ (mm)	Aot(%)	Gal(%)	Kid(L)(%)	Kid(R)(%)	Liv(%)	Pan(%)	Spl(%)	Sto(%)
depth = 2[6,12]	88.55	9.59	93.02	80.57	87.79	88.04	97.32	83.71	91.87	86.05
depth = 3[3,6,12]	88.68	8.51	92.52	81.86	87.28	88.51	96.06	84.22	92.37	86.63
depth = 4[3,6,12,12]	**89.21**	**8.14**	92.79	83.32	87.72	88.74	96.83	83.61	95.48	85.16
depth = 5[3,6,6,12,12]	87.62	10.59	92.84	79.25	86.17	89.17	96.79	82.33	93.09	81.3

表 5 验证了不同双融合模块深度对分割性能的影响. “depth”表示双融合模块的深度,即全部并行双分支注意力融合块的层数,“heads of Multi-headed Self-attention”表示对应层中双分支融合块内的多头自注意力的头数. 随着网络深度的增加,数据特征变得更加复杂. 适当增加多头自注意力机制的头数可以更高效地并行处理信息,允许网络在每一层内部

执行多样化的信息处理策略. 每个头关注输入数据的不同子空间,从而更细致地分析数据,增强每一层的特征提取能力,增加模型的灵活性. 因此本文设置多头自注意力机制的头数随网络深度而递增. 结果表明,随着双融合模块深度的适度增加,分割性能随之增强. 当双融合模块的深度为 4、多头自注意头数设置为[3,6,12,12]时分割性能最佳.

表 6 多头自注意力头数的消融研究

Table 6 Ablation study on the number of heads in the Multi-headed Self-attention

heads of Multi-headed Self-attention	DSC ↑ (%)	HD95 ↓ (mm)	Aot(%)	Gal(%)	Kid(L)(%)	Kid(R)(%)	Liv(%)	Pan(%)	Spl(%)	Sto(%)
[3,6,6,12]	88.82	8.31	92.94	81.39	87.73	90.02	97.31	82.76	95.5	82.88
[3,6,12,12]	**89.21**	**8.14**	92.79	83.32	87.72	88.74	96.83	83.61	95.48	85.16
[3,6,12,24]	88.36	10.74	92.82	82.4	85.57	88.24	96.98	83.15	92.49	85.25

本文在双融合模块深度的基础上进一步对比了不同的多头自注意力头数配置方案对分割结果的影响. 表 6 中“heads of Multi-headed Self-attention”表示当双融合模块深度为 4 时,各层并行双分支注意力融合块中的多头自注意力的不同头数对比. 根据表 6 的实验结果,当多头自注意头数设置为[3,6,12,12]时网络取得了最好的性能.

4 结 论

本文提出了一种基于局部窗口自注意力和深度卷积的并行双分支融合架构(PDBF),用于医学图像的分割. 通过耦合局部窗口和深度卷积的并行设计,PDBF 能够有效地扩展接受域,而无需使用移动窗口. 双向的注意力融合则增强了两个分支在通道和空间维度上的建模能力. 在 BCV、ACDC 和私有数据集上进行的大量实验表明,本文提出的方法优于其他先进的方法. 作为一种通用架构,PDBF 具有高度的灵活性. 根据不同的目标任务,两个分支可以灵活替换为更适合的卷积和 Transformer 模块,为下游医学图像任务带来了新的可能性.

References:

[1] Sandström S, Pettersson H, Åkerman K. The WHO manual of diagnostic imaging: radiographic technique and projections[M]. Chicago: The Radiological Society of North America, 2003.

[2] Negi A, Raj A N J, Nersisson R, et al. RDA-UNET-WGAN: an accurate breast ultrasound lesion segmentation using wasserstein generative adversarial networks[J]. Arabian Journal for Science and Engineering, 2020, 45(8): 6399-6410.

[3] Lee L K, Liew S C, Thong W J. A review of image segmentation methodologies in medical image[C]//Advanced Computer and Communication Engineering Technology: Proceedings of the 1st International Conference on Communication and Computer Engineering (ICOCOE), 2015: 1069-1080.

[4] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer AssistedIntervention (MICCAI), 2015: 234-241.

[5] Han X. Automatic liver lesion segmentation using a deep convolutional neural network method[J]. arXiv preprint arXiv: 1704.07239, 2017.

[6] Oktay O, Schlemper J, Folgoc L L, et al. Attention u-net: learning where to look for the pancreas[J]. arXiv preprint arXiv: 1804.03999, 2018.

[7] Vaswani A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017: 5998-6008, doi: 10.48550/arXiv.1706.03762.

[8] Dosovitskiy A. An image is worth 16 × 16 words: transformers for image recognition at scale[J]. arXiv preprint arXiv: 2010.11929, 2020.

[9] Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 10012-10022.

[10] Chen Q, Wu Q, Wang J, et al. Mixformer: mixing features across windows and dimensions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 5249-5259.

[11] Chen J, Lu Y, Yu Q, et al. Transunet: transformers make strong encoders for medical image segmentation[J]. arXiv preprint arXiv: 2102.04306, 2021.

[12] Wenxuan W, Chen C, Meng D, et al. transbts: multimodal brain tumor segmentation using transformer[C]//Medical Image Computing and Computer Assisted Intervention (MICCAI), 2021: 109-119.

[13] 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 (WACV), 2022: 574-584.

[14] Xie Y, Zhang J, Shen C, et al. Cotr: efficiently bridging cnn and transformer for 3d medical image segmentation[C]//Medical Image Computing and Computer Assisted Intervention (MICCAI), 2021: 171-180.

[15] Zhang Y, Liu H, Hu Q. Transfuse: fusing transformers and cnns for medical image segmentation[C]//Medical Image Computing and Computer Assisted Intervention (MICCAI), 2021: 14-24.

[16] Li Y, Wang Z, Yin L, et al. X-net: a dual encoding decoding method in medical image segmentation[J]. The Visual Computer, 2023, 39(11): 1-11.

[17] Liu W, Tian T, Xu W, et al. Phtrans: parallelly aggregating global and local representations for medical image segmentation[C]//Medical Image Computing and Computer Assisted Intervention (MICCAI), 2022: 235-244.

[18] 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.

[19] Yuan Y, Fu R, Huang L, et al. Hrformer: high-resolution transformer for dense prediction[J]. arXiv preprint arXiv: 2110.09408, 2021.

[20] Vaswani A, Ramachandran P, Srinivas A, et al. Scaling local self-attention for parameter efficient visual backbones[C]//Proceed-

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),2021:12894-12904.

[21] Hu J,Shen L,Sun G. Squeeze and excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2018:7132-7141.

[22] Woo S,Park J,Lee J Y,et al. Cbam:convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision(ECCV),2018:3-19.

[23] Landman B,Xu Z,Igelsias J,et al. Miccai multi-atlas labeling beyond the cranial vault-workshop and challenge[C]//MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge,2015:5-12.

[24] Bernard O,Lalande A,Zotti C,et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis:is the problem solved? [J]. IEEE Transactions on Medical Imaging,2018,37(11):2514-2525.

[25] Rohlfing T. Image similarity and tissue overlaps as surrogates for image registration accuracy:widely used but unreliable[J]. IEEE Transactions on Medical Imaging,2011,31(2):153-163.

[26] Huttenlocher D P,Klanderman G A,Rucklidge W J. Comparing images using the Hausdorff distance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1993,15(9):850-863.

[27] Chen J,Wan Z,Zhang J,et al. Medical image segmentation and reconstruction of prostate tumor based on 3D AlexNet[J]. Computer Methods and Programs in Biomedicine,2021,200(3):1-8.

[28] Sudre C H,Li W,Vercauteren T,et al. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations[C]//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support:3rd International Workshop,2017:240-248.

[29] Song H,Kim M,Park D,et al. Learning from noisy labels with deep neural networks:a survey[J]. IEEE Transactions on Neural Networks and Learning Systems,2022,34(11):8135-8153.

[30] Isensee F,Jaeger P F,Kohl S A A,et al. nnU-Net:a self-configuring method for deep learning-based biomedical image segmentation[J]. Nature Methods,2021,18(2):203-211.

[31] Cao H,Wang Y,Chen J,et al. Swin-unet:unet-like pure transformer for medical image segmentation[C]//European Conference on Computer Vision(ECCV),2022:205-218.

[32] Xu G,Zhang X,He X,et al. Levit-unet:make faster encoders with transformer for medical image segmentation[C]//Chinese Conference on Pattern Recognition and Computer Vision(PRCV),2023:42-53.

[33] Zhou H Y,Guo J,Zhang Y,et al. nnFormer:interleaved transformer for volumetric segmentation[J]. arXiv preprint arXiv:2109.03201,2021.

[34] Huang X,Deng Z,Li D,et al. Missformer:an effective medical image segmentation transformer[J]. arXiv preprint arXiv:2109.07162,2021.

[35] Hatamizadeh A,Nath V,Tang Y,et al. Swin unetr:swin transformers for semantic segmentation of brain tumors in mri images[C]//International MICCAI Brainlesion Workshop,2021:272-284.