基于多尺度卷积调制的医学图像分割

周新民[1,2],熊智谋[3*],史长发[4,5],杨 健[3]

(1. 湖南工商大学人工智能与先进计算学院,湖南长沙 410205;2. 湘江实验室,湖南长沙 410205;
3. 湖南工商大学计算机学院,湖南长沙 410205;4. 湖南工商大学智能工程与智能制造学院,湖南长沙 410205;
5. 湖南工商大学长沙人工智能社会实验室,湖南长沙 410205)

摘 要: 当前,越来越多的医学图像分割模型都采用Transformer模型作为基础结构,然而,Transformer模型的计算复杂度与输入序列呈二次关系且需要大量的数据进行预训练才能取得较好的结果,在数据量不足的情况下无法发挥优势;此外,Transformer往往无法有效提取图像的局部信息. 相比于Transformer,卷积神经网络则能够很好地规避上述两个问题. 为了充分发挥卷积神经网络与Transformer的各自优势并进一步挖掘卷积神经网络的潜力,本文提出一个多尺度卷积调制网络模型(Multi-Scale Convolution Modulation Network,MSCMNet),该模型将视觉Transformer领域模型结构设计方法融入传统卷积网络. 采用卷积调制和多尺度特征提取策略,构建基于多尺度卷积调制机制的特征提取模块(Multi-Scale Convolution Modulation,MSCM). 并提出高效的patch组合与patch分解策略分别用于特征图的下采样以及上采样,进一步提升模型的表征能力. 在腹部多器官、心脏、皮肤癌以及细胞核四个不同类型以及不同规模的医学图像分割数据集上取得的mDice分别为0.805 7、0.923 3、0.923 9、0.854 8,以较低的运算量和参数量取得了最好的分割性能,为卷积神经网络以及Transformer在医学图像分割领域提供了一个新颖而高效的模型结构设计范式.

关键词: 医学图像分割;多尺度;卷积调制;Transformer

基金项目: 国家自然科学基金(No.72091515);国家社会科学基金(No.21BGL231)

中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2024)09-3159-13

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20231068

Medical Image Segmentation Based on Multi-Scale Convolution Modulation

ZHOU Xin-min[1,2], XIONG Zhi-mou[3*], SHI Chang-fa[4,5], YANG Jian[3]

(1. *School of Artificial Intelligence and Advanced Computing, Hunan University of Technology and Business, Changsha, Hunan* 410205, *China*; 2. *Xiangjiang Laboratory, Changsha, Hunan* 410205, *China*;
3. *School of Computer Science, Hunan University of Technology and Business, Changsha, Hunan* 410205, *China*;
4. *School of Intelligent Engineering and Intelligent Manufacturing, Hunan University of Technology and Business, Changsha, Hunan* 410205, *China*;
5. *Changsha Social Laboratory of Artificial Intelligence, Hunan University of Technology and Business, Changsha, Hunan* 410205, *China*)

Abstract: Currently, more and more medical image segmentation models are using Transformer as their basic structure. However, the computational complexity of the Transformer model is quadratic with respect to the input sequence, and it requires a large amount of data for pre-training in order to achieve good results. In situations where there is insufficient data, the Transformer´s advantages cannot be fully realized. Additionally, the Transformer often fails to effectively extract local information from images. In contrast, convolutional neural networks can effectively avoid these two problems. In order to fully leverage the strengths of both convolutional neural networks and Transformers and further explore the potential of convolutional neural networks, this paper proposes a multi-scale convolution modulation network (MSCMNet) model. This model incorporates the design methodology of visual Transformer models into traditional convolutional networks. By using convolution modulation and multi-scale feature extraction strategies, a feature extraction module based on multi-scale convolution modulation (MSCM) is constructed. Efficient patch combination and patch decomposition strategies are also pro-

收稿日期:2023-11-15;修回日期:2024-01-28;责任编辑:梅志强

*通讯作者:熊智谋

posed for downsampling and upsampling of feature maps, respectively, further enhancing the model's representation ability. The mDice scores obtained on four different types and sizes of medical image segmentation datasets - multiple organs in the abdomen, heart, skin cancer, and nucleus - are 0.805 7, 0.923 3, 0.923 9 and 0.854 8, respectively. With lower computational complexity and parameter count, MSCMNet achieves the best segmentation performance, providing a novel and efficient model structure design paradigm for convolutional neural networks and Transformers in the field of medical image segmentation.

Key words: medical image segmentation; multi-scale; convolutional modulation; Transformer

Foundation Item(s): National Natural Science Foundation of China (No.72091515); National Social Science Fund of China (No.21BGL231)

1 引言

医学图像分割是智慧医疗领域的重要组成部分，为医疗诊断、治疗和研究提供了关键的信息和支持，如精确的病灶定位和分割、自动化的病变检测和诊断、量化的疾病评估和监测、智能影像导航和手术辅助等[1]，具有重要的临床应用价值. 然而，由于医学影像本身的特点如低对比度、噪声、图像复杂性等使自动分割任务依然面临许多挑战，因此，开发准确、高效、鲁棒的医学图像分割算法，对于提高医学图像处理的准确性和效率具有重要的意义. 近年来，深度学习技术的不断发展和创新使其在智慧医疗领域的应用前景越来越广阔，为医疗系统的智能化和个性化提供了重要的支持，不仅提高了医疗效率和准确性，还为患者提供了更好的医疗体验和健康管理服务.

卷积神经网络[2](Convolutional Neural Network，CNN)是目前医学图像分割领域的主要方法，现有方法主要基于U-Net[3]结构. 近年来，随着医学图像分割性能要求的提升，针对U-Net结构也在不断地改进和扩展，比如编解码器的改进、外接特征金字塔[4]、构建块优化[5,6]等，U-Net++[7]、Attention U-Net[8]、SAU-Net[9](Self-Attention U-Net)等均基于U-Net进行改进并取得较为不错的性能. 由于医学图像通常是三维体积数据，传统的二维方法可能无法充分利用三维图像的空间相关性，因此同时有研究人员将U-Net扩展到3D医学图像分割领域，例如3D-UNet[10]和VNet[11]，3D U-Net的网络结构与2D U-Net相似，包括编码器和解码器部分. 3D医学图像分割方法的关键优势在于它能够处理三维图像数据的空间信息，在三个空间方向上进行特征提取和重建，更准确地捕捉目标在三维空间中的形状和结构，然而，3D医学图像分割方法也存在一些挑战，如计算资源需求较高、训练数据的获取和标注困难等. 此外，由于CNN感受野有限，只能对局部区域的依赖关系进行建模，缺乏全局上下文建模能力.

Transformer[12]最初提出时用于自然语言处理领域，它能够捕获输入序列之间的长距离依赖关系，可以有效解决CNN不能对全局上下文建模的问题. 随着ViT[13](Vision Transformer)和DETR[14](Detection Transformer)等模型的提出，越来越多的研究人员开始尝试利用Transformer的优势来解决计算机视觉领域中的各种问题. TransUNet[15]首次将Transformer引入医学图像分割领域，它首先利用卷积神经网络来提取低级特征，然后通过Transformer来模拟全局信息交互，结合跳跃连接，TransUNet在CT多器官分割任务中刷新了新的记录. 在TransUNet之后，为了提高对全局上下文建模的效率，同时保持对底层细节的强掌握，TransFuse[16]提出了一种新的并行分支架构，以并行的方式结合了Transformer和CNN. UTNet[17]则将自注意力整合到卷积神经网络中，在编码器和解码器中都应用了自注意力模块. 上述模型均同时结合CNN与Transformer而设计，Swin UNet[18]则在Swin Transformer[19]的基础之上提出一个完全基于Transformer的医学图像分割模型，该模型在U-Net的基础之上，将基础构建块替换为Swin Transformer，通过PatchMerging和PatchExpand分别进行下采样以及上采样. MISSFormer[20]在Swin UNet的基础上引入增强型Transformer块重新设计前馈网络，并提出一种带有增强型Transfomrer块的上下文桥接模块对分层Transformer编码器生成的多尺度特征的远程依赖关系和局部上下文进行建模. 然而基于Transformer的模型也有较明显的缺陷，首先，其计算复杂度与输入序列长度成二次关系；其次，Transformer通常需要大量的数据预训练才能够达到较好的结果，不同于自然图像，获取大量的医学图像数据是一件极为困难的事情，且医学图像标注有较高的专业要求. 此外，Transformer虽然能够捕获全局依赖关系，但是无法捕获局部细节信息，而对于医学图像来说，局部细节信息至关重要. 虽然Transformer具有上述缺陷，但视觉Transformer广泛使用的模型结构设计策略已被证明有益于提升卷积神经网络在视觉任务中的表现. 首先，Transformer将输入图像进行基于patch的嵌入表示，该过程会大幅降低模型后续的计算复杂度，减轻了Transformer自注意力模块带来的计算压力. 其次视觉Transformer广泛使用patch合并策略用于特征图的下采样，该方式优于卷积神经网络中传统的池化以及步幅为2的卷积下采样方式.

另外在Swin-UNet[18]医学图像分割模型中,引入patch扩展策略用于特征图的上采样,该方式也被证明优于卷积网络中的转置卷积以及插值. 此外,Transformer结构中为输入patch添加位置编码信息能够有效提升patch的特征表示能力.

基于Transformer结构存在的缺陷,部分研究转而站在Transformer模型结构的视角来重新探索卷积网络的潜力,代表性模型有VAN[21](Visual Attention Network)、SegNeXt[22]以及ConvNeXt[23]. VAN和SegNeXt二者的核心思想是一种卷积调制机制,旨在解决Transformer模型在语义分割任务中的性能瓶颈问题,与传统的卷积神经网络不同,卷积调制是一种将卷积层的输出与上下文信息相结合的技术,该机制可以更有效地捕获图像中的语义信息,并能够在保证模型性能的同时减少模型的计算复杂度,二者在自然图像分割任务中均取得出色的性能. 此外,VAN中提出了一种高效的大核卷积分解方法,通过将较大卷积核分解为相对简单且高效的多个组件,在实现卷积网络高效特征提取的同时大幅降低计算复杂度. SegNeXt则提出了一种多尺度模块作为卷积调制权重的生成器,并同时应用大核卷积的思想,与VAN不同的是,为了避免较大卷积核所带来的计算开销,作者将大核卷积分解为两个分别沿着宽度和高度方向的条形卷积,在自然图像分析中取得了较好的性能,但医学图像中往往更多的是块状组织结构,因此SegNeXt无法应对较为复杂的医学图像处理任务. ConvNeXt则在ResNet[24]模型的基础上,仿照Swin Transformer的结构进行逐步改进而得到,并发现在此过程中导致性能差异的几个关键组件,模型在精度和可扩展性方面可以与Transformer相当,同时保持了卷积网络的简单性和效率.

基于上述分析,本文充分利用Transformer的模型结构设计优势,将其深度融合到卷积神经网络结构当中. 同时采用卷积调制以及多尺度和大核卷积特征提取策略构建特征提取模块,提取更加丰富的语义信息,并设计更为高效的上采样和下采样策略,为卷积神经网络在医学图像分割领域探索一种全新的模型结构设计范式,以轻量级的卷积网络结构获取能够超越Transformer的性能. 本文的主要贡献如下:(1)引入卷积调制机制,同时采用大核卷积以及多尺度特征提取策略,构建了多尺度卷积调制模块,通过提取的多尺度信息指导调制权重的生成,该模块以较少的参数量和运算量实现了高效的特征提取,在有效避免Transformer计算代价高以及需要大量数据进行预训练的缺陷的同时具备超越Transformer结构的性能;(2)受现有视觉Transformer模型结构设计方法的启发,将其与卷积网络结构深度融合,并提出新颖而高效的patch组合(PatchCombining)以及patch分解(PatchDecomposing)策略分别用于特征图的下采样以及上采样,使模型具备更优的性能,进一步挖掘卷积神经网络的潜能;(3)提出了MSCMNet模型,并将其应用于医学图像分割领域,在四个不同类型以及不同规模的医学图像分割数据集上均取得最好的分割性能,为卷积网络与Transformer在医学图像分割领域提供了一个新颖而高效的模型结构设计范式.

2　MSCMNet模型设计

MSCMNet模型遵循U-Net标准的编码器和解码器结构,模型的结构概览如图1所示.

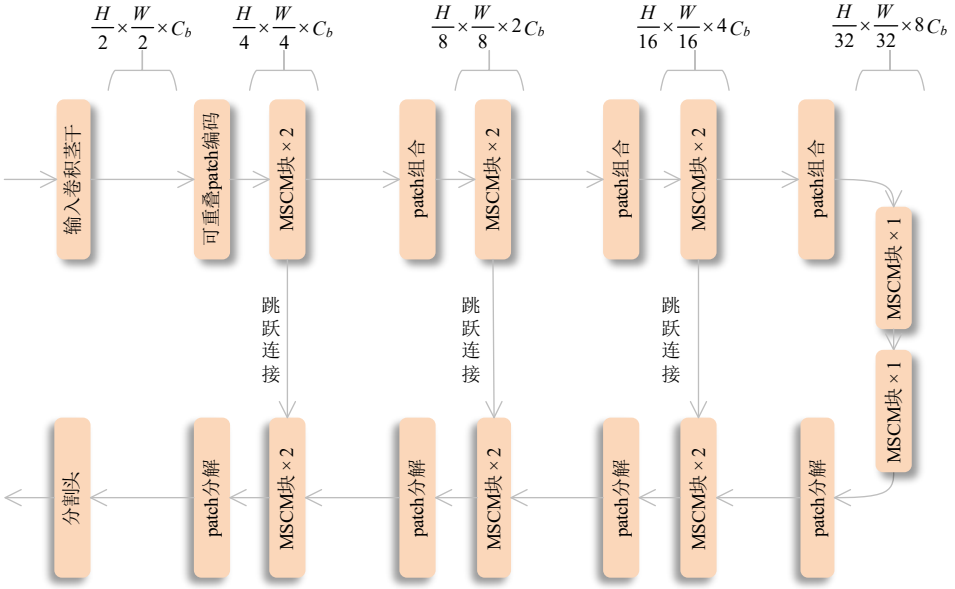

图1　模型结构图概览

相比UNet及其过往的改进模型，我们在编码器的输入新增卷积茎干模块，用于提取输入图像的全局结构化信息并降低分辨率，从而降低后续模块的计算复杂度，进一步提高模型的运行效率．在解码器的尾部新增分割头模块用于获取分割结果．在卷积茎干模块后紧接着对特征图进行可重叠的patch嵌入表示，基础构建块从U-Net的3×3卷积修改为MSCM模块，编码器共分为4个阶段，每个阶段先进行下采样，随后使用2个MSCM模块提取特征．除第一阶段使用patch嵌入表示降低分辨率之外，其它阶段均首先通过对patch进行组合实现下采样．在解码器中使用patch分解策略实现上采样．我们基于不同的通道数和基础通道数设计了不同规模的模型，模型规模对应的参数设置见表1.

表1 不同规模模型4个阶段通道数以及基础通道数设置

模型规模	通道数	基础通道数
MSCMNet-S	[32,64,128,256]	32
MSCMNet-B	[48,96,192,384]	48
MSCMNet-L	[96,192,384,768]	96

2.1 MSCM模块

MSCM模块综合借鉴了卷积调制和大核卷积的思想，并同时采用多尺度特征提取策略，综合利用不同尺度信息指导卷积调制权重生成过程，使生成的权重具备更好的鲁棒性，从而促使特征提取过程更为高效．并在每个模块的输入部分添加位置编码．具体来说，MSCM模块由四部分组成：位置编码模块、局部特征提取模块、多尺度卷积调制模块以及特征映射模块，如图2所示.

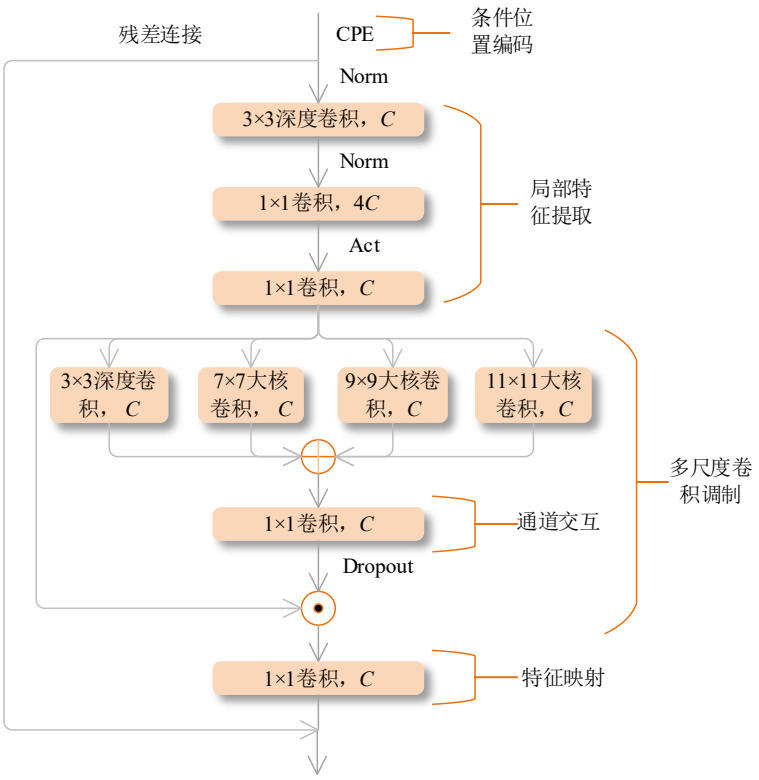

图2 MSCM模块

位置编码对于以patch划分特征图的Transformer结构来说可以显著提升其特征表示能力，比较常用的位置编码包括绝对位置编码、相对位置编码以及条件位置编码，绝对位置编码需要输入图像尺寸固定，不能适应任意分辨率的输入图像，相对位置编码依赖于参考点的选择，条件位置编码[25]（Conditional Position Encoding，CPE）简单高效，只需要一个卷积核大于或等于3、填充值为0的深度卷积即可，并且通过实验证明其具有较好的位置信息表示能力，本文在ConvNeXt[23]的基础之上，探索了将Transformer结构中的位置编码融入到卷积神经网络后对模型的影响.

局部特征提取模块主要用于提取局部区域特征，相较于SegNeXt[22]中的局部特征提取模块仅由一个简单的5×5的深度卷积构成，没有考虑通道维度之间的信息交互，本文模型借鉴ConvNeXt[23]中的特征提取方式，在提取局部特征的同时，更好的考虑特征图通道维

度之间的信息交互,更充分的挖掘图像的特征,有效提取局部区域信息.

多尺度卷积调制模块在局部特征提取模块的基础之上通过多个不同分支提取不同尺度区域局部信息.对于医学图像分割任务来说,病灶区域或器官常常呈现不同的大小,单一尺度无法同时应对大目标与小目标,通过多种不同尺度卷积核,可以有效覆盖不同病灶区域或器官,从而避免较大卷积核无法有效提取小目标与较小卷积核无法有效提取大目标的问题.卷积调制机制是一种受Transformer自注意力机制启发而设计的一种自调制机制,它通过自身提取的调制权重与特征值以逐元素相乘的方式实现特征图在整体空间位置上的重新加权,该机制同时具备卷积网络的轻量性与Transformer的高性能,医学图像往往对比度低、边界模糊,待分割目标与周边区域不易区分,通过卷积调制机制则能够有效关注重要区域,实现更精确的分割.具体来说,多尺度卷积调制模块由一个基础的$\mathrm{DW}_{3\times3}$分支以及多个不同尺度的大核卷积分支组成.近年来大核卷积已被证明对于密集预测任务有很好的效果,VAN[21]中提出了一种高效的大核卷积分解方法,它表明一个$K\times K(K>3)$的大核卷积可以分解为一个扩张率为d、核大小为$\lceil K/d\rceil\times\lceil K/d\rceil$的深度卷积以及一个核大小为$(2d-1)\times(2d-1)$的深度卷积和一个对通道进行自适应选择的点卷积.在本文模型中,我们将最后一个用于对通道进行自适应选择的点卷积抽离出来放在多尺度融合之后,以用于对不同尺度特征图之间的通道关系进行建模.这样,通过多尺度融合信息指导卷积调制权重的生成,进而获取更加准确的权重矩阵.最后将生成的卷积调制权重与输入特征图进行逐元素相乘,得到经过空间位置自调制的特征图.最后通过特征映射模块实现通道信息的自适应选择.调制权重的加权过程如以下公式所示:

$$\boldsymbol{f}=\mathrm{DW}_{(3\times3)}(\boldsymbol{x})+\sum_{i=1}^{n}\boldsymbol{L}_i \tag{1}$$

$$\boldsymbol{c}=\mathrm{PW}_{(1\times1)}(\boldsymbol{f}) \tag{2}$$

$$\boldsymbol{m}=\mathrm{Dropout}(\boldsymbol{c}) \tag{3}$$

$$\boldsymbol{y}=\boldsymbol{m}\odot\boldsymbol{f} \tag{4}$$

其中,n表示大核卷积分支的个数,$\boldsymbol{L}_i$表示第i个大核卷积分支,DW表示深度卷积,PW表示点卷积,$\odot$表示哈达玛乘积(Hadamard Product).需要注意的是,在单个MSCM模块的处理过程中,最终的输出通道数与输入通道数保持一致.总的来说,MSCM模块的整体计算过程如以下公式所示:

$$\boldsymbol{x}=\mathrm{CPE}(\boldsymbol{x})+\boldsymbol{x} \tag{5}$$

$$\boldsymbol{l}=\mathrm{LocalFeature}(\mathrm{norm}(\boldsymbol{x})) \tag{6}$$

$$\boldsymbol{s}=\mathrm{MSCM}(\boldsymbol{l}) \tag{7}$$

$$\boldsymbol{y}=\mathrm{Projection}(\boldsymbol{s})+\boldsymbol{x} \tag{8}$$

2.2　编码器

2.2.1　输入卷积茎干

视觉Transformer中采用patch划分方式的研究倾向于直接将输入图像进行patch嵌入表示,同时直接将输入图像分辨率降低4倍以减少模型的计算复杂度,造成模型无法有效提取全局特征表示,而全局特征表示对于后续网络模块的训练和推理至关重要,因为它包含了输入图像的全局结构化信息,这对于语义分割任务来说至关重要.为此,在MSCMNet模型中,输入图像首先经过卷积茎干模块,该模块由三个3×3卷积堆叠而成,通过堆叠扩大卷积核的有效感受野,更好的提取输入图像的全局结构化信息,中间一个3×3卷积同时用于将输入特征图的分辨率减半,从而降低后续模块的计算复杂度.卷积茎干模块的结构如图3所示.

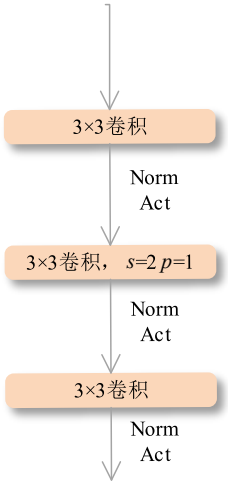

图3　卷积输入茎干

2.2.2　PatchCombining

CNN网络普遍使用池化或stride设置为2的卷积操作实现下采样,在视觉Transformer模型结构中则使用patch合并的方式.patch之间的合并方式会在极大程度上影响模型的泛化能力,这是一个非常重要且关键的问题,需要在设计模型结构时给予高度重视.在Swin transformer[19]中使用了一种PatchMerging合并方式用于下采样,具体来说,首先将输入patch划分为4个不同部分,随后在通道维连接,从而实现分辨率减半且通道数随之增加4倍,最后通过一个线性层将通道数减少为输入特征图通道数的2倍.然而这样的合并方式没有考虑被划分为相同部分的特征图中patch之间的信息交互,导致模型的泛化能力相对较弱.本文通过深入研究和实践,基于PatchMerging设

计了一种全新而又高效的patch组合方式(PatchCombining)实现下采样. 具体来说,PatchCombining包含两个部分:组内patch信息融合、组间通道信息交互,如图4所示.

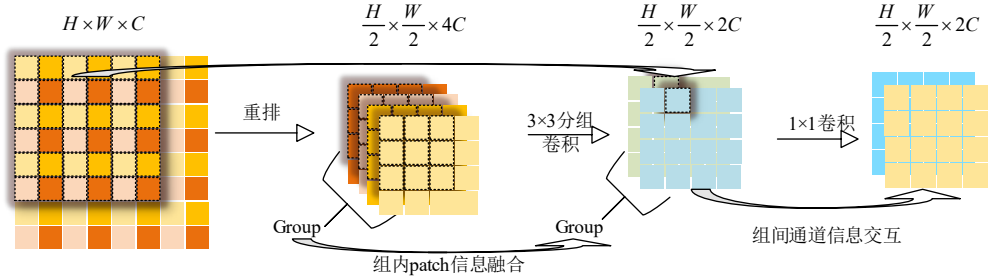

图4 单个通道特征图的PatchCombining实现过程

在将分辨率为$H\times W$、通道数为C的输入patch分为4个部分之后($H\times W\times C\rightarrow H/2\times W/2\times 4C$),我们首先将其进行分组,组数等于输入patch的通道数C,然后将每个分组通过一个核大小$K\geqslant(3\times 3)$的深度卷积映射到通道数为$2C$的patch表示空间(即组内变换为$H/2\times W/2\times 4\rightarrow H/2\times W/2\times 2$),通过这种方式,在组合过程中不仅可以实现被划分为相同部分的特征图中一个局部区域之间的信息交互,同时4个部分的相同局部区域之间的patch对应着原始输入之间一个较大的区域,也即最终组合得到的特征图中每个patch都是输入特征图中一个较大区域之间patch信息交互的结果(见图4中大箭头以及突出显示区域),从而使得到的patch表示包含更加丰富的语义信息,显著提升模型的泛化能力. 最后通过一个点卷积实现特征图通道之间的信息交互.

2.3 解码器

2.3.1 PatchDecomposing

在基于CNN的模型中,上采样的方式有转置卷积以及线性插值,在基于Transformer的模型中,Swin-UNet[18]引入了PatchExpand操作来实现上采样,具体来说,对于一个分辨率为$H\times W$、通道数为C的特征图,首先通过一个线性层将通道数增加2倍($H\times W\times C\rightarrow H\times W\times 2C$),然后通过重新排列操作将特征图的分辨率扩大2倍,通道数相较于原始输入减少2倍($H\times W\times 2C\rightarrow H\times W\times 2\times 2\times C/2\rightarrow 2H\times 2W\times C/2$). 实验也证明PatchExpand的做法要优于线性插值以及转置卷积. 同上述PatchCombining过程,我们对PatchExpand做进一步的优化,并将其命名为PatchDecomposing,如图5所示.

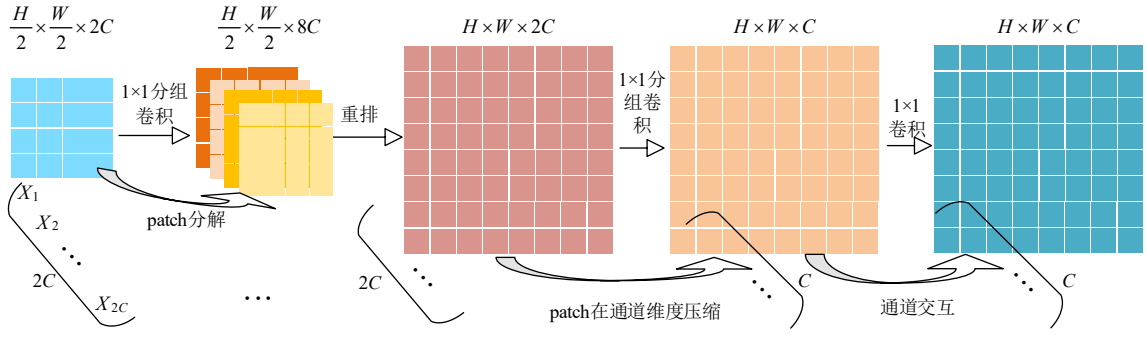

图5 PatchDecomposing实现过程

PatchDecomposing主要利用分组思想,首先将特征图分成C_b组,将每一组的特征图表示空间分解到4个不同的子空间($H\times W\times C\rightarrow H\times W\times 4C$),随后将获取的特征图进行重新排列,在此过程中实现分辨率加倍($H\times W\times 4C\rightarrow 2H\times 2W\times C$),然后以每$C_b$个通道为一组,将每组patch表示空间压缩到1/2表示空间,从而实现通道减半($2H\times 2W\times C\rightarrow 2H\times 2W\times C/2$). 最后通过一个点卷积对通道进行自适应选择. 具体的实现过程如以下公式所示:

$$\boldsymbol{x}_{\mathrm{f}}=\mathrm{DW}_{(1\times 1)}(\boldsymbol{x},\mathrm{groups}=C_b) \tag{9}$$

$$\boldsymbol{x}_{\mathrm{r}}=\mathrm{Restructure}(\boldsymbol{x}_{\mathrm{f}}) \tag{10}$$

$$\boldsymbol{x}_{\mathrm{c}}=\mathrm{DW}_{(1\times 1)}(\boldsymbol{x}_{\mathrm{r}},\mathrm{groups}=C_b) \tag{11}$$

$$\boldsymbol{y}=\mathrm{PW}_{(1\times 1)}(\boldsymbol{x}_{\mathrm{c}}) \tag{12}$$

如上述公式所示,在具体实现中,我们使用一个核大小1×1、组数为C_b的深度卷积实现多组特征图的分解过程,其中$C_b\left(C_b\leqslant C\right)$表示基础通道数(见表1),即每个PatchDecomposing中组数固定. 使用一个核大小1×

1、组数为 C_b 的深度卷积实现多组 patch 重建并压缩的过程,促使压缩后的特征表示更有助于分割任务. 最后,我们使用一个 1×1 的点卷积实现通道之间的关系建模. 需要注意的是模型中的最后一个 PatchDecomposing 操作在通道压缩的过程中不改变输入的通道数(即通道压缩过程变为普通映射:$2H\times2W\times C\rightarrow 2H\times 2W\times C$).

2.3.2　跳跃连接层

为了适应基于 patch 划分方式实现的卷积网络结构,本文的跳跃连接层与原始跳过连接略有不同,共包含两个部分,首先是一个特征连接部分,用于连接低层特征与高级特征,两者在通道维进行连接,这一部分也即原始的跳过连接,其次是一个特征映射部分,由一个点卷积实现,用于将连接后的特征图通道映射到较低维度,方便后续模块进行特征处理. 如式(13)所示:

$$\boldsymbol{y}=\mathrm{PW}_{(1\times1)}(\mathrm{Concat}(\boldsymbol{x},\boldsymbol{e})) \tag{13}$$

其中 $\mathrm{PW}_{(1\times1)}$ 表示点卷积,$\boldsymbol{x}$ 表示解码器输入特征图,$\boldsymbol{e}$ 表示编码器对应阶段的输出特征图.

2.3.3　分割头

在 Swin-UNet[18] 中,模型最后使用一个 4 倍的 PatchExpand 来将特征图分辨率直接扩大 4 倍,这种做法很有可能会降低模型的精度,为此,我们使用一个与卷积茎干相对应的分割头来逐渐的恢复特征图分辨率,避免可能出现的精度丢失,分割头的结构如图 6 所示,其中转置卷积操作将特征图分辨率扩大 2 倍.

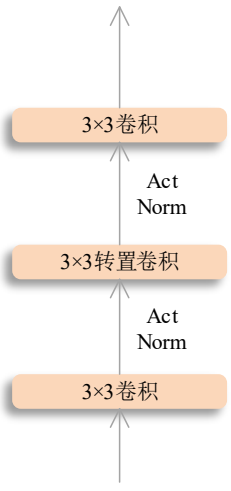

图 6　分割头

2.4　其它设计

2.4.1　添加 Dropout

Dropout 通常用于防止模型过拟合[26],但我们通过实验发现,将生成的调制权重通过 Dropout 随机舍弃部分,每次训练只更新部分权重值的做法能够提高调制权重的准确性,在避免模型过拟合的同时显著提高模型的泛化能力.

2.4.2　使用 LayerNorm 代替 BatchNorm

原始 Transformer 结构中使用 LayerNorm[27] 作为数据标准化方式,而在卷积神经网络中往往使用 BatchNorm[28],近些年基于 Transformer 的视觉模型中也广泛使用 LayerNorm 作为数据标准化方式,ConvNeXt[23] 中的一个卷积设计理念就是将 BatchNorm 替换为 LayerNorm,模型在原有的基础之上性能得到 0.1% 的提升,但由于深度学习模型参数初始化的影响,在 GPU 上的训练结果带有一定的随机性,这一实验结果不具备 LayerNorm 优于 BatchNorm 的验证能力,并且在 SegNeXt[22] 中,作者表明文中模型使用 BatchNorm 的结果要优于 LayerNorm. 基于此,本文在所有的数据集上都将 BatchNorm 替换为 LayerNorm,进一步探索使用 LayerNorm 替换 BatchNorm 对模型的影响,实验结果表明,对于 MSCMNet 模型来说,LayerNorm 要远远优于 BatchNorm.

3　实验结果与分析

3.1　数据集

我们选取了不同部位以及不同类型和不同规模的医学图像分割数据集来验证本文提出模型的泛化能力,包括腹部多器官分割、心脏分割、皮肤癌以及细胞核分割,各数据集的详情如见表 2.

AMOS22[29] (Abdominal Multi-Organ Segmentation22)数据集:来源于 MICCAI 2022 腹部多器官多模态分割挑战赛(Multi-Modality Abdominal Multi-Organ Segmentation Challenge 2022),共 15 个目标类别.

ACDC[30](Automated Cardiac Diagnosis Challenge):为自动心脏诊断挑战数据集,共包含 100 例从不同患者收集到的 MRI(Magnetic Resonance Imaging)扫描图像,每个图像标注了 3 个器官,分别为左心室(Left Ventricle,LV)、右心室(Right Ventricle,RV)以及心肌(Myoglobin,MYO).

ISIC2018[31,32](International Skin Imaging Collaboration2018)数据集:来自 MICCAI 2018 Workshop,主要目的在于使用计算机辅助自动诊断皮肤癌,对黑色素瘤检测的皮肤病变进行分析.

MoNuSeg[33,34](Multi-organ Nuclei Segmentation):是一个细胞核数据集,通过仔细标注在多家医院确诊的不同器官肿瘤患者的组织图像而获得. 该数据集是通过从 TCGA(The Cancer Genome Atlas)档案库下载以 40 倍放大率拍摄的 H&E 染色组织图像创建.

表2 数据集详情表

数据集	模态	部位	数量	数据划分	类别数	输入分辨率	特点
AMOS22	CT	腹部	300(41 430张切片)	训练集:120(16 361张切片) 验证集:60(8 430张切片) 测试集:120(16 639张切片)	15	512×512、 768×768	数量多、 类别多、 难度大
ACDC	MRI	心脏	100(1 902张切片)	训练集:70(1 304张切片) 验证集:10(182张切片) 测试集:20(416张切片)	3	224×224	数量少、 类别适中、 难度适中
ISIC2018	RGB	皮肤	3 694	训练集:2 594 验证集:100 测试集:1 000	2	512×512	数量少、 类别少、 较容易
MoNuSeg	数字显微组织图像	细胞核	44	训练集:24 验证集:6 测试集:14	2	1 000×1 000	数量很少、 边界密集、 难度适中

3.2 评估指标

本文使用医学图像分割的常用评估指标[4]来评估模型的性能,主要包括mDice和mIoU. mDice是通过在每个类上计算Dice相似系数之后计算平均得到,Dice相似系数常用于度量两个集合的相似度,计算方式如下:

$$\mathrm{Dice}=2\frac{|P\cap G|}{|P|+|G|}=2\frac{\mathrm{TP}}{\mathrm{FP}+2\mathrm{TP}+\mathrm{FN}} \tag{14}$$

其中,TP表示正类样本被成功预测为正类样本的个数,FP表示负类样本中被错误预测为正类样本的个数,FN表示正类样本中被错误预测为负类样本的个数,P表示预测值,G表示真实值,$|P\cap G|$表示预测值和真实值中相交的部分,$|P|$表示预测值中元素的个数,$|G|$表示真实标签中元素的个数. mIoU则是语义分割的标准度量,它通过在每个类上计算交并比(IoU)之后计算平均得到,计算方式如下:

$$\begin{aligned}\mathrm{mIoU}&=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}}\\&=\frac{1}{k+1}\sum_{i=0}^{k}\frac{\mathrm{TP}}{\mathrm{FN}+\mathrm{FP}+\mathrm{TP}}\end{aligned} \tag{15}$$

其中,$k+1$表示样本类别个数(包含一个背景类),p_{ii}表示预测正确的数量,p_{ij}和p_{ji}分别表示假正和假负的数量. 对于AMOS22以及ACDC数据集,使用mDice作为评估指标,并分别记录每个目标类别的Dice相似系数,对于ISIC2018以及MoNuSeg数据集,使用mDice以及mIoU作为评估指标.

3.3 实验设置

模型基于PaddlePaddle框架,使用Paddleseg套件构建分割模型训练方法,使用单个V100 16 GB GPU进行训练. 训练细节如表3所示,取训练过程中在验证集上性能最好的模型用于测试. 对于所有数据集的训练集部分,图像和标签输入到模型之前首先将其缩放到224×224,对于验证数据集以及测试数据集,首先将图像缩放到224×224,随后将模型的预测结果通过插值恢复到原始分辨率,最后与标签进行对比并计算相应的评估指标. 在模型的训练过程中,学习率逐步下降到0,初始训练阶段的1 500次迭代用于Warm Up,Warm Up阶段的初始学习率设置为0.00 000 1. 使用的激活函数均为GELU. Dropout操作的丢弃概率值设置为0.5.

表3 各数据集上的训练参数设置

数据集	迭代次数	批大小	优化器	学习率	权重衰减	数据增强策略	损失函数	评估指标
AMOS22	260 000	16	AdamW	0.000 6	0.01	随机旋转、 随机翻转、 随机亮度、 随机对比度	交叉熵损失 + Dice损失	mDice、各目标类别Dice
ACDC	120 000							
ISIC2018	80 000							mDice、mIoU
MoNuSeg	10 000	8						

3.4 对比分析

为了验证本文模型优于传统的CNN模型结构以及现有的卷积调制结构,同时为了表明本文模型与Transformer结构具有可匹敌的性能表现,分别选取了基于卷积神经网络、基于卷积调制以及基于Transformer的代表性网络作为对比模型,并分析其在4个不同类型以及不同规模医学图像分割数据集中的性能,实验结果如表4~7所示:

综合表4~7的实验数据可以看出本文提出的MSCMNet模型在4个数据集上均取得了最优的评估结果. 需要特别留意的是SegNeXt模型在MoNuSeg数据集上的表现非常差而在其它数据集上的表现较好,可能原

表4　AMOS22数据集上的测试性能

方法	UNet	UNet++	Swin-UNet	TransUNet	MISSFormer	SegNeXt-L	MSCMNet-S	MSCMNet-B	MSCMNet-L
Spleen	0.885 5	0.881 2	0.876 4	0.846 3	0.857 3	**0.931 4**	0.915 8	0.912 6	0.923 7
Right kidney	0.862 7	0.851 6	0.806 8	0.743 2	0.755 4	0.880 9	0.877 6	0.879 9	**0.891 3**
Left kidney	0.874 2	0.853 1	0.822 5	0.752 8	0.771	0.894 2	0.884 4	0.881 4	**0.907 9**
Gallbladder	0.746 7	0.700 5	0.661 9	0.0	0.584 7	0.7813	0.711 8	0.753 3	**0.782 5**
Esophagus	0.735 7	0.701 8	0.633 2	0.537 6	0.601 9	0.711 8	0.705 6	0.723	**0.749**
Liver	0.928 9	0.926 2	0.915	0.904 7	0.904 8	**0.947 5**	0.927 7	0.931 6	0.941 4
Stomach	0.836 1	0.820 8	0.792 5	0.720 5	0.732 9	0.862 4	0.816 3	0.821 6	**0.863 1**
Aorta	0.894 2	0.890 1	0.858 6	0.814 5	0.820 5	**0.899 7**	0.894 6	0.892 5	0.894 5
Inferior vena cava	0.803 2	0.787 9	0.728	0.671 4	0.687 2	0.791 6	0.792 2	0.791 5	**0.815 8**
Pancreas	0.739 3	0.720 3	0.661 3	0.585 5	0.632 3	0.746 5	0.705 6	0.736	**0.767 5**
Right adrenal gland	0.641 2	0.594 1	0.230 1	0.0	0.432 3	0.563 5	0.569 1	0.585 7	**0.631 7**
Left adrenal gland	0.589 5	0.507 7	0.000 3	0.0	0.376 5	0.509 6	0.519	0.540 7	**0.596 1**
Duodenum	0.643 5	0.579	0.532 8	0.399 1	0.459 3	0.646 4	0.615 9	0.627 8	**0.680 5**
Bladder	0.877 5	0.867	0.842 3	0.799	0.825 2	**0.891 8**	0.858	0.857 7	0.872 5
Prostate/uterus	0.733 1	0.733 1	0.706 2	0.576 1	0.672 1	**0.784 5**	0.736 5	0.765 9	0.768 5
mDice	0.786 1	0.761	0.671 2	0.556 7	0.674 2	0.789 5	0.768 7	0.780 1	**0.805 7**

注:粗体表示最优

表5　ACDC数据集上的测试性能

方法	UNet	UNet++	Swin-UNet	TransUNet	MISSFormer	SegNeXt-L	MSCMNet-S	MSCMNet-B	MSCMNet-L
RV	0.894	0.885 5	0.885 8	0.887 4	0.691 8	0.894 9	0.896 1	**0.907 2**	0.901 4
MYO	0.893 8	0.892 1	0.882 5	0.884 3	0.724 6	0.896	0.898 4	**0.902 1**	0.901 8
LV	0.956 7	0.958 2	0.953 9	0.954 3	0.881 6	0.957 3	0.959 6	0.960 6	**0.960 9**
mDice	0.914 8	0.911 9	0.907 4	0.908 7	0.766	0.916 1	0.91 8	**0.923 3**	0.921 4

注:粗体表示最优

表6　ISIC2018数据集上的测试性能

方法	UNet	UNet++	Swin-UNet	TransUNet	MISSFormer	SegNeXt-L	MSCMNet-S	MSCMNet-B	MSCMNet-L
mIoU	0.845 9	0.838 6	0.858 6	0.855 4	0.842 8	0.839 3	**0.860 4**	0.858 3	0.851
mDice	0.915 2	0.910 9	0.922 9	0.921 1	0.913 4	0.911 4	**0.923 9**	0.922 7	0.918 4

注:粗体表示最优

表7　MoNuSeg数据集上的测试性能

方法	UNet	UNet++	Swin-UNet	TransUNet	MISSFormer	SegNeXt-L	MSCMNet-S	MSCMNet-B	MSCMNet-L
mIoU	0.743 2	0.733 7	0.751 5	0.725 3	0.715 8	0.521 6	0.744 7	**0.756 3**	0.745 6
mDice	0.845 2	0.839 2	0.851 0	0.832 4	0.825 9	0.655 1	0.846 6	**0.854 8**	0.847 2

注:粗体表示最优

因是MoNuSeg数据集包含非常多的细胞核,而这些细胞核基本都呈现块状结构,SegNeXt使用的条状卷积在分割这些细胞核时很容易在核边界产生误差,而由于MoNuSeg数据集的核边界数量非常多,误差累积导致效果极差,此外结合图7的分割结果可以看出SegNeXt对目标的分割趋向于将块状结构分割为条状结构,从中可以看出,条状卷积不适用于包含很多块状结构的医学图像处理任务.

另外基于Transformer的模型在AMOS22数据集上的分割能力普遍较弱,而基于卷积以及卷积调制机制的模型则能够较好工作,一个重要原因就是Transformer无法有效提取局部细节信息,AMOS22数据集由于目标类别众多,多种器官之间对比度低,此时局部细节信息至关重要. 而除了本文所提模型,SegNeXt在其中2个数据集中均取得次优的结果,值得注意的是SegNeXt也基于卷积调制机制,充分表明卷积调制机制在医学图像分割领域具备很好的应用前景.

除了对模型的性能做对比之外,我们还对模型的运算量和参数量进行了分析,如表8所示,我们提出的MSCMNet-S和MSCMNet-B的参数量和运算量远远小于

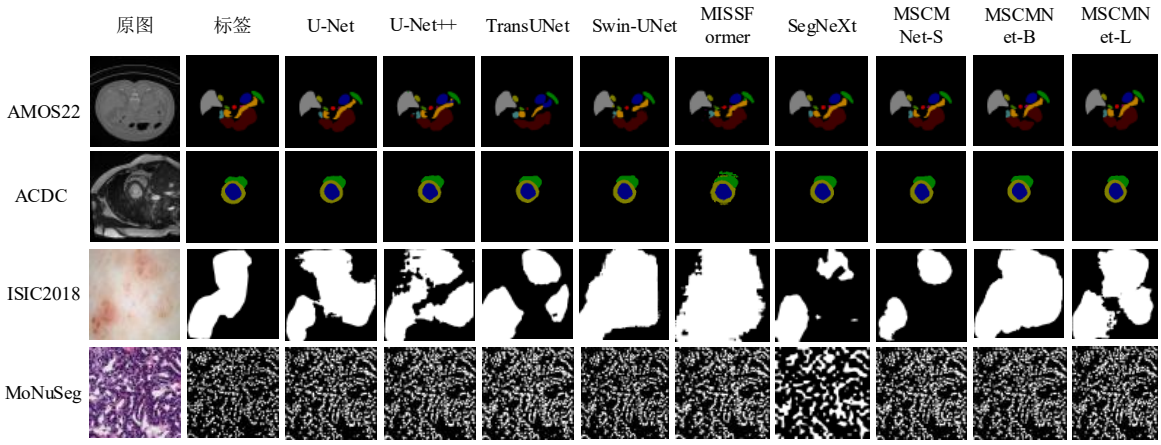

图7 模型分割结果对比

表8 模型的运算量以及参数量对比

模型	UNet	UNet++	Swin-UNet	TransUNet	MISSFormer	SegNeXt-L	MSCMNet-S	MSCMNet-B	MSCMNet-L
Flops/G	23.79	22.91	5.86	24.57	7.18	9.47	**1.54**	3.31	12.64
Params/M	13.41	8.37	27.15	105.13	35.45	45.11	**2.13**	4.65	18.01

其它模型(基于输入形状为(3,224,224)计算),充分证明了本文模型的高效性.

3.5 消融分析

为了验证多尺度卷积调制模块中各个组成部分的有效性和不可或缺性,以及PatchCombining下采样和PatchDecomposing上采样策略的有效性和模型中一些细节设计的重要性,我们从以下两个方面入手对MSCMNet模型进行消融分析:结构设计与细节设计. 为了方便对模型结构进行消融分析,我们基于MSCMNet-S设计了一个Baseline模型,该模型的基础构建块只包含MSCMNet-S中MSCM模块的局部特征提取以及特征映射部分. 在Baseline模型的基础之上,我们选择在4个数据集上以增量的方式分析MSCMNet中各个部分的有效性和不可或缺性,实验结果如表9所示.

表9 模型在4个数据集上的消融分析结果

方法	AMOS22	ACDC	ISIC2018	MoNuSeg
V1 (Baseline)	0.763 3	0.909 6	0.918 8	0.821 4
V2 (Baseline+CPE)	0.759 9	0.913 7	0.908 5	0.804 8
V3 (Baseline+MSCM)	0.742 6	0.911 4	0.918 8	0.838 5
V4 (Baseline+SSCM+CPE)	0.765 1	0.917 2	0.919 1	0.835
V5 (Baseline+MSCM+CPE)	**0.768 7**	**0.918**	**0.923 9**	**0.846 6**

在表9中,SSCM表示单一尺度卷积调制(只包含$DW_{3\times3}$分支,不包含大核卷积分支),MSCM表示多尺度卷积调制(同时包含$DW_{3\times3}$和3个大核卷积分支),通过实验我们可以看出:单一尺度的卷积调制(V4)与没有调制权重加持的模型(V2)在性能上均有较大提升,充分表明了卷积调制机制的高效性. 而多尺度的卷积调制加持的模型(V5)mDice则在4个数据集中均高出单一尺度卷积调制模型(V4),有效证明了多尺度所蕴含的丰富的语义信息能够指导生成更可靠的调制权重. 为了更进一步验证本文提出的多尺度卷积调制确实有效,而不是完全得益于所提取的多尺度信息,我们将哈达玛乘积修改为求和操作,构成单纯的多尺度卷积结构,并在4个数据集上进行对比实验,使用mDice评估模型,结果如表10所示.

通过表10的实验结果可以看出多尺度卷积调制结构在4个数据集上均明显优于单纯的多尺度结构,充分证明了卷积调制机制的有效性. 另外通过表9可以看出向模型的构建块添加条件位置编码(V5)相比于没有位置编码的模型(V3)在性能上有较为显著的提升,表明通过引入Transformer的位置编码能够有效提升模型的泛化能力. 而相比于Baseline(V1)模型,仅添加位置编码(V2)或者仅添加多尺度卷积调制模块(V3)并不能保证在所有数据集上均能优于Baseline,尤其是仅添加位置编码时除了在ACDC数据集上性能有所提升外,在另外3个数据集上均会导致性能有较大幅度的下降,而同时添加了位置编码与多尺度卷积调制模块的模型(V5)则能够确保在4个数据集上均能取得最优的性能,充分证明了MSCM构建块中各个部分的不可或缺性.

为了进一步挖掘卷积神经网络的潜力,提供一个较优的卷积网络模型结构设计方法,我们在上面分析了模

表10 多尺度卷积与多尺度卷积调制结构对比

方法	AMOS22	ACDC	ISIC2018	MoNuSeg
多尺度卷积	0.756 3	0.910 7	0.921 7	0.838 3
多尺度卷积调制	**0.768 7**	**0.91 8**	**0.923 9**	**0.846 6**

型结构设计的基础之上同时分析了不同标准化方式以及下采样和下采样方式以及Dropout等细节设计对MSCMNet模型性能的影响,使用mDice作为评估指标.

从表11的实验结果可以看出在MSCMNet模型中应用BatchNorm效果较不理想,甚至在AMOS22数据集上无法分割任何目标类别,而使用LayerNorm代替BatchNorm在4个数据集上均能够为模型带来很好的性能提升,表明LayerNorm不只在自然语言处理领域具备优势,在视觉任务中也能够发挥很好的作用,结合SegNeXt中BatchNorm优于LayerNorm的实验结果,我们可以看出在不同的模型中使用哪种标准化方式并没有统一的标准,需要根据具体的模型而定. 此外,与PatchMerging下采样方式相比,我们提出的PatchCombining不仅拥有更少的参数量,而且在4个数据集上均显著提升了模型性能,充分显示了PatchCombining的高效性,PatchDecomposing同样以更低的参数量在4个数据集上均取得了优于PatchExpand的性能. 另外通过表11的实验结果可以看出向生成的调制权重应用Dropout有助于提高模型的泛化能力,尤其在ISIC2018以及MoNuSeg数据集上使模型的性能显著提升,由此可以看出通过在训练过程中随机更新部分调制权重值,能够迫使模型学习更加鲁棒的调制权重.

表11　不同标准化、上采样、下采样以及Dropout对应的模型性能

方法	AMOS22	ACDC	ISIC2018	MoNuSeg	计算量/GFlops	Params/M
BatchNorm	—	0.318 2	0.839	0.690 8	1.535	2.13
PatchMerging	0.764 4	0.914 7	0.922 3	0.835 6	1.541	2.2
PatchExpand	0.767 5	0.909 4	0.916 2	0.843 5	1.539	2.27
No Dropout	0.760 6	0.915 2	0.915 7	0.830 3	1.535	2.13
LayerNorm+PatchCombining+PatchDecomposing+Dropout(p=0.5)	**0.768 7**	**0.918**	**0.923 9**	**0.846 6**	**1.535**	**2.13**

4　总结

针对视觉Transformer在医学图像分割领域运算代价高、需要预训练以及不能有效提取局部细节信息的问题,本文提出了一个基于多尺度卷积调制机制的医学图像分割方法MSCMNet,该方法采用了大核卷积策略,并结合多尺度以及自调制机制构造了一个多尺度卷积调制模块,有效规避了Transformer在医学图像分割任务中存在的缺陷,同时提出了高效的patch组合与patch分解策略用于特征图下采样以及上采样. 大量实验表明,MSCMNet模型在4种不同类型以及不同规模的医学图像分割任务中均以较低的运算量和参数量取得最优的性能表现. 然而MSCMNet在训练时需要更多的迭代次数才能够很好的拟合数据,后续将对模型训练时的收敛速度进一步优化.

参考文献

[1] 郑光远, 刘峡壁, 韩光辉. 医学影像计算机辅助检测与诊断系统综述[J]. 软件学报, 2018, 29(5): 1471-1514.
ZHENG G Y, LIU X B, HAN G H. Survey on medical image computer aided detection and diagnosis systems[J]. Journal of Software, 2018, 29(5): 1471-1514. (in Chinese)

[2] LE C Y, BOSER B, DENKER J S, et al. Handwritten digit recognition with a back-propagation network[C]//Advances in Neural Information Processing Systems 2. San Francisco: Morgan Kaufmann Publishers Inc, 1990: 396-404.

[3] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation [C]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Cham: Springer, 2015: 234-241.

[4] 殷晓航, 王永才, 李德英. 基于U-Net结构改进的医学影像分割技术综述[J]. 软件学报, 2021, 32(2): 519-550.
YIN X H, WANG Y C, LI D Y. Suvery of medical image segmentation technology based on U-net structure improvement[J]. Journal of Software, 2021, 32(2): 519-550. (in Chinese)

[5] 周涛, 霍兵强, 陆惠玲, 等. 残差神经网络及其在医学图像处理中的应用研究[J]. 电子学报, 2020, 48(7): 1436-1447.
ZHOU T, HUO B Q, LU H L, et al. Research on residual neural network and its application on medical image processing[J]. Acta Electronica Sinica, 2020, 48(7): 1436-1447. (in Chinese)

[6] 刘金平, 吴娟娟, 张荣, 等. 基于结构重参数化与多尺度深度监督的COVID-19胸部CT图像自动分割[J]. 电子学报, 2023, 51(5): 1163-1171.
LIU J P, WU J J, ZHANG R, et al. Toward automated segmentation of COVID-19 chest CT images based on structural reparameterization and multi-scale deep supervision [J]. Acta Electronica Sinica, 2023, 51(5): 1163-1171. (in Chinese)

[7] ZHOU Z W, RAHMAN SIDDIQUEE M M, TAJBAKHSH N, et al. UNet++: A nested U-Net architecture for medical image segmentation[C]//Deep Learning in Medical Image Analysis and Multimodal Learning for

Clinical Decision Support. Cham: Springer International Publishing, 2018: 3-11.

[8] OKTAY O, SCHLEMPER J, FOLGOC L L, et al. Attention U-Net: Learning where to look for the pancreas[EB/OL]. (2018-05-20) [2022-11-25]. http://arxiv. org/abs/1804.03999.

[9] 张淑军, 彭中, 李辉. SAU-Net:基于U-Net和自注意力机制的医学图像分割方法[J]. 电子学报, 2022, 50(10): 2433-2442.
ZHANG S J, PENG Z, LI H. SAU-net: Medical image segmentation method based on U-net and self-attention[J]. Acta Electronica Sinica, 2022, 50(10): 2433-2442. (in Chinese)

[10] ÇIÇEK Ö, ABDULKADIR A, LIENKAMP S S, et al. 3D U-Net: Learning dense volumetric segmentation from sparse annotation[C]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016. Cham: Springer, 2016: 424-432.

[11] MILLETARÌ F, NAVAB N, AHMADI S A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation[C]//2016 Fourth International Conference on 3D Vision (3DV). Stanford: IEEE, 2016: 565-571.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017: 6000-6010.

[13] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2021006-03) [2022-09-15]. http://arxiv.org/abs/2010.11929.

[14] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 213-229.

[15] CHEN J, LU Y, YU Q, et al. TransUNet: Transformers make strong encoders for medical image segmentation [EB/OL]. (2021-02-08) [2022-09-18]. http://arxiv.org/abs/2102.04306.

[16] ZHANG Y D, LIU H Y, HU Q. TransFuse: Fusing transformers and CNNs for medical image segmentation[C]// Medical Image Computing and Computer Assisted Intervention-MICCAI 2021. Cham: Springer, 2021: 14-24.

[17] GAO Y H, ZHOU M, METAXAS D N. UTNet: A hybrid transformer architecture for medical image segmentation [C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2021. Cham: Springer, 2021: 61-71.

[18] CAO H, WANG Y Y, CHEN J, et al. Swin-Unet: Unet-like pure transformer for medical image segmentation[EB/OL]. (2021-05-12) [2022-09-18]. http://arxiv. org/abs/2105.05537.

[19] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 9992-10002.

[20] Huang X H, Deng Z F, Li D D, et al. MISSFormer: An effective medical image segmentation transformer [EB/OL]. (2021-09-15) [2022-09-18]. http://arxiv.org/abs/2109.07162.

[21] GUO M H, LU C Z, LIU Z N, et al. Visual attention network[EB/OL]. (2022-07-11) [2022-11-13]. https://arxiv.org/abs/2202.09741v5.

[22] GUO M H, LU C Z, HOU Q, et al. SegNeXt: Rethinking convolutional attention design for semantic segmentation [EB/OL]. (2022-09-18) [2022-11-10]. https://arxiv. org/abs/2209.08575.

[23] LIU Z, MAO H Z, WU C Y, et al. A ConvNet for the 2020s[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 11966-11976.

[24] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778..

[25] CHU X X, TIAN Z, ZHANG B, et al. Conditional positional encodings for vision transformers[EB/OL]. (2023-02-13)[2023-07-13]. http://arxiv.org/abs/2102.10882.

[26] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[27] Ba J L, Kiros J R, Hinton G E. Layer Normalization[EB/OL]. (2016-07-21) [2023-09-07]. http://arxiv. org/abs/1607.06450.

[28] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. Lille, France: JMLR.org, 2015: 448-456.

[29] JI Y F, BAI H T, YANG J, et al. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation[EB/OL]. (2022-16-16) [2023-05-11].

http://arxiv.org/abs/2206.08023.
[30] 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.
[31] CODELLA N C F, GUTMAN D, CELEBI M E, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC) [C]//2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). Piscataway: IEEE, 2018: 168-172.
[32] TSCHANDL P, ROSENDAHL C, KITTLER H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions [J]. Scientific Data, 2018, 5: 180161.
[33] KUMAR N, VERMA R, ANAND D, et al. A multi-organ nucleus segmentation challenge[J]. IEEE Transactions on Medical Imaging, 2020, 39(5): 1380-1391.
[34] KUMAR N, VERMA R, SHARMA S, et al. A dataset and a technique for generalized nuclear segmentation for computational pathology[J]. IEEE Transactions on Medical Imaging, 2017, 36(7): 1550-1560.

作者简介

周新民　男,1977年5月出生于湖南省邵阳市.2010年5月博士毕业于同济大学计算机应用技术专业,2014年6月国防科技大学管理科学与工程博士后出站.现为湖南工商大学人工智能与先进计算学院副院长、教授、硕士生导师.主要研究方向为新型智慧城市、商务智能与大数据.
E-mail: zhouxinmin2699@163.com

熊智谋　男,1995年3月出生于河南省信阳市.湖南工商大学计算机学院软件工程专业学术硕士在读.主要研究方向为医学图像处理与分析、计算机视觉.
E-mail: Xzhimou@163.com

史长发　男,1985年2月出生于湖南省株洲市.2016年博士毕业于哈尔滨工业大学机械电子工程专业.现为湖南工商大学智能工程与智能制造学院副院长、副教授、硕士生导师.主要研究方向为医学图像处理与分析、深度学习.
E-mail: ivanhanks@yeah.net

杨　健　男,2000年7月出生于湖南省益阳市.湖南工商大学计算机学院电子信息专业硕士在读.主要研究方向为智慧医疗、计算机视觉.
E-mail: 1468554194@qq.com