DOI:10.3969/j.issn.1005-202X.2025.02.008

医学影像物理

MFMANet:一种融合多尺度特征的多重注意力医学图像分割网络

袁金丽,李博华,陈沐萱,蒋仁鼎,JUI SHANAZ SHARMIN,郭志涛
河北工业大学电子信息工程学院,天津 300401

【摘要】医学图像分割研究在推动高效且精确的自动化图像处理技术方面具有重要意义。然而,为解决医学图像中存在的器官组织形状差异大、边界模糊等导致图像分割结果不准确的问题,提出一种MFMANet新型网络,该网络以“U”型架构为基础,并集成了多尺度信息融合模块和多重注意力模块。具体而言,多尺度信息融合模块通过捕捉网络浅层中的多尺度信息,以弥补编码器和解码器特征之间的语义差距,从而提升了网络应对器官尺寸差异大问题的能力。同时,网络使用多重注意力方法,利用Swin Transformer作为网络深层编解码器,采用通道、空间注意力取代传统的跳跃连接,进而实现了特征信息的更精细提取,以应对边界模糊问题。通过在ACDC和Synapase这两个公共数据集上进行实验,结果显示,与MTUNet方法相比,该方法在骰子相似系数这一关键指标上取得了1.51%和1.29%的显著提升,充分证明了该方法在提高分割网络准确性方面的有效性。

【关键词】医学图像分割;多尺度信息融合;注意力机制
【中图分类号】R318;TP391 【文献标志码】A 【文章编号】1005-202X(2025)02-0190-09

MFMANet: a multi-attention medical image segmentation network fused with multi-scale features

YUAN Jinli, LI Bohua, CHEN Muxuan, JIANG Rending, JUI SHANAZ SHARMIN, GUO Zhitao
School of Electronic Information Engineering, Hebei University of Technology, Tianjin 300401, China

Abstract: The research on medical image segmentation is of great significance in advancing efficient and accurate automated image processing techniques. To address the problem of inaccurate segmentation results caused by significant variations in organ tissue shapes and blurred boundaries present in medical images, a novel network named MFMANet is proposed. Built upon a "U"-shaped architecture, the network integrates multi-scale information fusion modules and multi-attention modules. Specifically, multi-scale information modules capture multi-scale information in the shallow layers of the network to bridge the semantic gap between encoder and decoder features, thereby enhancing the network's ability to handle large variations in organ sizes. Regarding the issue of blurred boundaries, multi-attention mechanism utilizes Swin Transformer as the deep encoder-decoder network, employing channel and spatial attention instead of traditional skip connections to achieve finer feature extraction. Experimental results on the ACDC and Synapse public datasets show that the proposed method achieves improvements of 1.51% and 1.29% in Dice similarity coefficient as compared with MTUNet, fully demonstrating its effectiveness in enhancing segmentation network accuracy.

Keywords: medical image segmentation; multi-scale information fusion; attention mechanism

前 言

在临床诊断中,医学图像分割的任务是从通过X射线、计算机断层扫描、核磁共振成像及超声等方式得到的生物医学图像中精准勾勒出器官或病灶等待观察对象,以帮助医生对人体器官和组织中发生的病变进行定量分析和形态分析。传统的手动分割方法虽然准确,但其对专业知识有很强的依赖性,不仅耗时还可能存在标注不一致的问题。自动医学图像分割方法可减少人工干预,大幅度提高工作效率,然而其仍存在以下问题。首先,由于医学图像分割任务是在特定的局部进行分割,医学样本内部之间的差异很小,但样本之间的差异却十分显著。近期研究基于自注意力模块或其变体来改进模型,而这些

【收稿日期】2024-10-11
【基金项目】河北省教育厅重点项目(ZD2022115)
【作者简介】袁金丽,博士,副教授,研究方向:智能信息处理、计算机视觉、机器学习,E-mail: jinli_yuan@hebut.edu.cn
【通信作者】郭志涛,博士,教授,研究方向:射频识别、嵌入式系统、图像处理,E-mail: 2002089@hebut.edu.cn

方法在建模样本之间的关系还存在一定缺陷。其次，医学图像中的解剖结构通常具有复杂的形态和组织结构，导致难以准确地判断物体位置或形状，而利用多尺度信息是解决结构复杂性的关键策略之一。基于此，本文结合卷积和注意力机制，提出一种融合多尺度特征的多重注意力网络。其主要思想是，在网络浅层融合来自编码器的多尺度特征，最大限度地提高不同尺度特征信息利用率，从而捕获更复杂的通道依赖关系。在网络深层阶段，将通道、空间注意力引入跳跃连接中，协同网络深层的Transformer模块，对图像特征进行精细化建模，帮助模型更准确地定位边界，并减少边界模糊对分割的影响。

1 研究现状

UNet是常用的医学图像分割模型，是一种具有U形架构的编码器-解码器模型[1]，由于UNet的简单性和可扩展性优势，许多研究人员以它为基础提出了改进模型，并已被证明对许多不同的分割任务有效[2-5]。尽管UNet及其变体在医学图像分割中取得了很大的成就，但其在捕获远程依赖能力方面依旧受到限制。这种限制源于卷积操作的内在局部性，即卷积主要关注在有限的感受野中捕获局部模式和空间关系。因此，如何克服这种局限性，仍然是CNN架构领域最关键的挑战之一。

空间金字塔池化（Spatial Pyramid Pooling, SPP）和空洞空间卷积金字塔池化（Atrous Spatial Pyramid Pooling, ASPP）等结构已经证明，有效地结合多尺度特征可以显著提高感受野并提高模型性能[6-7]。为了充分利用多尺度上下文信息，Zhou等[2]和Huang等[5]引入了多级特征融合机制，相较于原始的UNet模型具有更深的网络结构，通过使用密集跳跃连接聚合不同语义的特征，可以捕获更复杂的特征和信息层次。Jha等[8]采用两个UNet叠加的方式以捕获更多语义信息，并在瓶颈层引入ASPP以有效提取高分辨率特征图，进一步提升模型性能。为了更好地获得全局信息以准确定位待分割区域，研究人员将ViT的自注意力机制（Self-Attenton, SA）引入到UNet[9-15]中。Transformer等注意力机制的加入协助模型能够在全局范围内理解图像的语义关系。这使得模型能够捕获图像中的长距离依赖关系，并在分割任务中更好地整合全局信息。Chen等[10]提出将UNet与Transformer相结合，利用Transformer对卷积神经网络的特征图谱进行标记，并将其作为输入序列，以获取整体上下文信息。在上采样过程中，解码器将其与高分辨率的CNN特征图相结合，从而实现准确的定位效果。这样的组合得到了出色的分割结果，并且成功将“卷积+Transformer”这种网络结构引入了医学图像分割领域。Cao等[12]在UNet的结构基础上，将Swin-Transformer模块作为基本模块，提出了一种纯Transformer的UNet网络。Hatamizadeh等[16]采用“收缩-扩展”模式，使用一堆Transformer模块组成UNTER，采用Transformer作为编码器部分，但在解码器中仍然使用UNet的卷积。Wang等[17]受到TransUnet的启发，基于encoder-decoder结构，提出了一种新的混合Transformer模块，用于同时进行样本内外部关系学习，在浅层阶段，它依赖卷积运算来提取特征，进入深层后，它首先运用局部-全局高斯权重自注意力模块来计算注意力，随后借助外部注意力模块，深入探索样本之间的关联。

Cai等[18]建立了一种多尺度机制，将中间层不同尺度的全局上下文信息直接聚合为最终特征表示，并利用额外的注意力机制来提高医学图像分割的预测精度。然而，一步级联可能会忽略大规模上采样过程中一些有价值的细节，仍然存在信息丢失的现象。Wang等[19]深入分析跳跃连接对分割工作的贡献之后，发现跳跃连接并不总是对分割有利，且简单的复制并不适合特征融合，因此提出了Ctrans模块来替代简单的跳跃连接实现更好的分割性能。这进一步证实了为更好地实现特征融合引入更合适的连接方式的必要性。Wu等[20]将高阶交互机制引入模型之中，并将多级和多尺度信息融合机制添加到跳跃连接路径中，以最大限度地提升不同阶段的特征信息的使用。Ruan等[21]提出通道注意桥模块和空间注意桥模块替代跳跃连接，对多级特征进行全局和局部融合，生成有效注意力图。Ates等[22]通过顺序捕获多尺度编码器特征之间的通道和空间依赖关系，来解决编解码器之间的语义差距。

2 研究方法

本文提出了一种融合多尺度信息的多重注意力医学图像分割网络模型（Fused Multi-Scale Feature Multi-Attention Net, MFMANet）。模型框架如图1所示。

延续之前医学图像分割任务工作，MFMANet建立在标准的“编码器-解码器”U型架构上。编解码器部分采用“C-C-C-T-T”的基础结构，其中，“C”是指两步卷积或反卷积，每一步卷积均包含“3×3卷积+批归一化+ReLU激活函数”的组合，以此来实现下采样或上采样中的通道扩展或挤压。“T”则代表包含Transformer的模块，为网络注入强大的全局信息处理能力，具体内容见2.2.1。网络在浅层使用卷积可

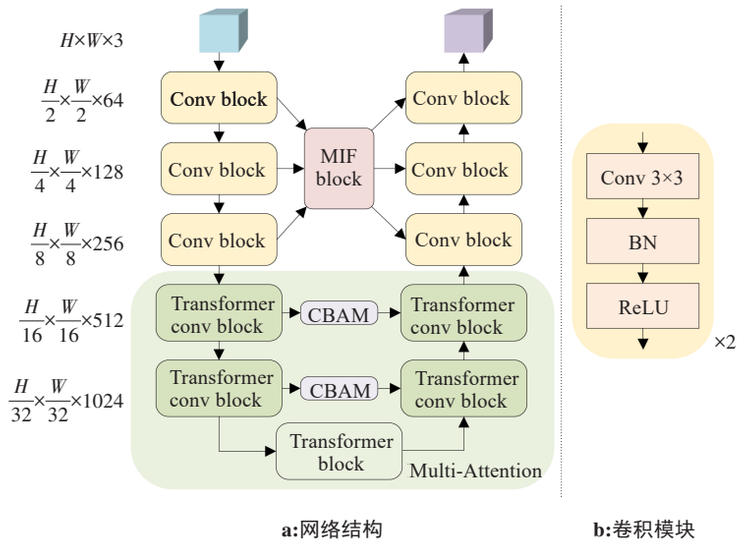

图1 MFMANet网络结构与卷积模块

Figure 1 MFMANet architecture and convolutional block

以更好地提取局部信息[23]，在深层则使用Swin Transformer提取全局特征[24]。卷积网络得益于归纳偏置先验，展现出强大的泛化能力和快速的收敛速度，而注意力层则拥有更强的建模能力，对大数据集更有益。通过将卷积和注意力层结合，MFMANet得以在医学图像分割任务中，充分提取医学图像的局部和全局信息，进而提升模型的分割性能。

针对前言部分提到的挑战，本文设计了多尺度信息融合（Multi-Information Fusion, MIF）模块和多重注意力（Multi-Attention）模块。MIF模块旨在抓取并融合不同层次的特征信息，通过密集连接不同语义尺度的通道特征，实现更全面的信息利用。而多重注意力模块的设计，则使网络能够更精准地聚焦于关键信息，显著提升分割结果的准确性。与基础UNet相比，MFMANet在充分利用多尺度信息、增强跳跃连接的有效性以及提高信息利用能力方面均展现出显著优势，从而能够进一步提高医学图像分割结果质量。

2.1 多尺度信息融合模块

多阶段和多尺度信息的获取在分割不同大小的目标及边缘信息提取中发挥着重要作用，这两者的有效融合已被证明是提高性能的关键[19]。MIF模块通过将扩张卷积层和通道注意力相结合，取代简单的跳跃连接，这一设计使网络能够融合不同尺度的特征，生成精细的通道注意图，保留更多有用特征信息，帮助网络更好地捕捉特征之间的关联性。MIF模块首先对浅层网络提取的信息进行逐元素叠加式融合，再利用扩张卷积层，使特征在不同尺度上同时卷积再聚合。这一步骤有效整合了多尺度信息，为后续处理提供了丰富的特征基础。随后使用通道注意力捕获局部跨通道交互，使网络通过协作学习的方式更加高效地融合多尺度通道特征，弥补特征之间的语义差距。MIF模块示意图如图2所示。

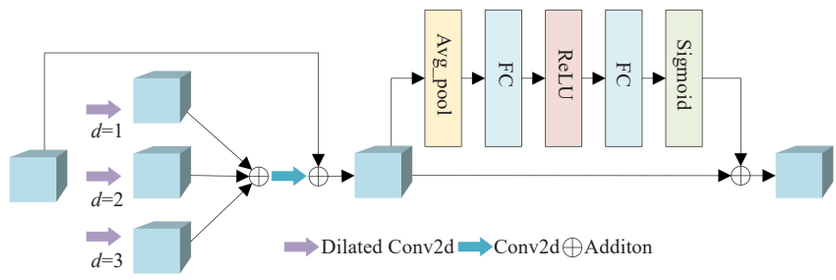

图2 多尺度信息融合模块

Figure 2 Multi-scale information fusion block

具体而言，特征图$X_{\text{input}} \in R^{H\times W\times C}$（$H$、W和$C$分别指图像高度、宽度和特征通道数）输入到网络之后，编码器浅层网络将捕获到不同层次的特征$X_1 \in R^{H\times W\times C}$，$X_2 \in R^{H/2\times W/2\times 2C}$，$X_3 \in R^{H/4\times W/4\times 4C}$。为了解决每个编码器块

的特征映射之间分辨率不匹配的问题，避免网络偏爱维度较大的特征图从而引起偏差的现象，模块采用卷积池化对低维度特征进行下采样，以及反卷积操作对高维度特征进行上采样的方法，统一特征图分辨率及维度。随后将这些经过处理之后的特征图X_i进行逐元素叠加，得到融合后的特征图X。其中，i=1, 2, 3，j=2。

$$X_i=\begin{cases}\text{Upsample}\left(X_i\right), & i>j\\ X_i, & i=j\\ \text{Downsample}\left(X_i\right), & i<j\end{cases} \tag{1}$$

$$X=X_1+X_2+X_3 \tag{2}$$

随后，特征图X被进一步送入扩张率分别为{1, 2, 3}的扩张卷积层，这些卷积层能够在不同尺寸上同时进行卷积操作，并将处理后的特征叠加起来得到新特征X'。这一步骤能够将相关性较强的特征聚集起来，同时弱化非关键特征的影响。为进一步优化特征，得到的特征X'还需要经过一个3×3的卷积，并与原始特征进行叠加。随后，叠加后的特征被送入通道注意力模块，该模块由平均池化、ReLU激活函数和Sigmoid函数组成。在扩张卷积层后加入通道注意力可根据每个通道的重要性动态调整特征图的权重，降低冗余信息，提高网络计算效率和泛化能力。具体过程如下：

$$X'=X+\text{conv}\left(\text{conv}_{(\text{rate}=1)}(X)+\text{conv}_{(\text{rate}=2)}(X)+\text{conv}_{(\text{rate}=3)}(X)\right) \tag{3}$$

$$Y=X+\text{Sigmoid}\left(\text{ReLU}\left(\text{Mean}\left(X'\right)\right)\right) \tag{4}$$

其中，$\text{conv}_{(\text{rate}=i)}$表示扩张率为$i$的扩张卷积，conv表示普通2D卷积。Mean表示全局平均池化，ReLU和Sigmoid分别对应两个激活函数，X'表示经过扩张卷积层之后得到的特征，Y表示MIF模块最终得到的特征。

最后将得到的特征Y还原到对应解码器的特征大小$Y_1\in R^{H\times W\times C}$，$Y_2\in R^{H/2\times W/2\times 2C}$，$Y_3\in R^{H/4\times W/4\times 4C}$，再与上层解码器的输出按元素相加。

2.2 多重注意力模块

多重注意力模块由网络深层Transformer conv模块和跳跃连接中的卷积块注意模块（Convolutional Block Attention Module, CBAM）构成[25]。在Transformer conv模块中，窗口注意力通过分层的注意力机制，能够同时捕获全局信息和局部信息，确保模型在理解数据时能够全面而深入地洞察数据内在结构。CBAM使用通道、空间混合注意力进一步优化特征提取过程，这一模块不仅增强了模型对关键特征的敏感度，还使其能够更好地捕获空间和通道之间的相关性，进而提升了特征表达的准确性和丰富性。结合以上两种注意力模块，能够使网络同时关注全局和局部特征，提取重要特征抑制无关特征，全面理解输入数据。这种结构有助于网络了解整体的病变结构及其与背景的关系，从而更准确地定位病变区域。

2.2.1 Transformer conv模块和瓶颈层

Transformer conv模块包括两层的Swin Transformer（以下简称SwinT）模块和一层两步卷积，每个SwinT模块都包含一个层归一化（LayerNorm, LN）和一个多头自注意力（Multi-head Self Attention, MSA）模块、一个残差连接和一个具有GELU非线性激活函数的两层MLP。MFMANet网络模型将基于窗口的多头自注意力（Window-based Multi-head Self Attention, W-MSA）和基于移位窗口的多头自注意力（Shifted Window-based Multi-head Self Attention, SW-MSA）分别应用于这两个连续的SwinT模块，构造如图3所示，窗口大小为7。这种方法能增强模型对远程依赖的捕获能力以及对全局、局部语义信息交互的能力。相比于简单的插值方法操作，本模型在Transformer模块之后使用卷积进行上/下采样，能够更好地捕获图像中的局部特征和结构信息，保留输入特征图的空间信息，从而产生更准确的采样结果。模型瓶颈层使用的Transformer模块仅包括两个连接的SwinT模块。在瓶颈层中，特征大小和分辨率保持不变，以确保特征表示的一致性。

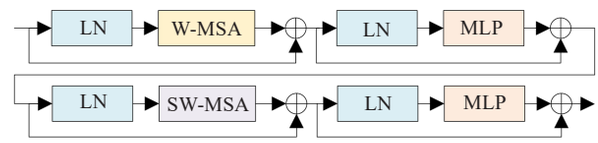

图3 Swin Transformer模块

Figure 3 Swin Transformer block

2.2.2 CBAM

为了能够更好地保留图像信息，该网络在深层Transformer编解码器的基础上，引入CBAM作为网络的跳跃连接。CBAM由通道注意力和空间注意力构成。通道注意力负责调整不同通道的重要性，以便网络更好地关注对分割任务有效的特征，提高特征表示的表现力和判别能力。通道注意力部分主要由全局平均池化、全局最大池化和多层感知机（Multilayer Perceptron, MLP）实现，池化用于获取通道的全局特征表示，通过MLP来学习每个通道权重以调整重要性。空间注意力则关注特征图中不同空间位置，使网络更好地捕获图像中不同区域的细节信息，提高特征的鲁棒性和区分度。空间注意力部分主要依赖卷积操作实现，根据输入特征图动态调整每个位置的权重，以实现空间注意力的

作用。具体实现过程如图4所示。计算步骤如下：

$$CA=X\otimes \mathrm{Sigmoid}\left(\mathrm{MLP}\left(\mathrm{Max}\left(X\right)\right)+\mathrm{MLP}\left(\mathrm{Mean}\left(X\right)\right)\right) \tag{5}$$

$$SA=CA\otimes \mathrm{Sigmoid}\left(\mathrm{conv}\left(\mathrm{concat}\begin{pmatrix}\mathrm{MAX}\left(CA\right),\\ \mathrm{Mean}\left(CA\right)\end{pmatrix}\right)\right) \tag{6}$$

其中，CA表示通道注意力特征图，SA表示空间注意力特征图。⊗表示逐元素相乘，Max表示全局最大池化，Mean表示全局平均池化。concat表示按通道拼接，Sigmoid对应Sigmoid激活函数。

这种混合通道、空间注意力结构的加入可促进通道和空间维度中注意权重的动态分布，同时也保证了较低的计算复杂度，避免除Transformer外更大的计算开销。

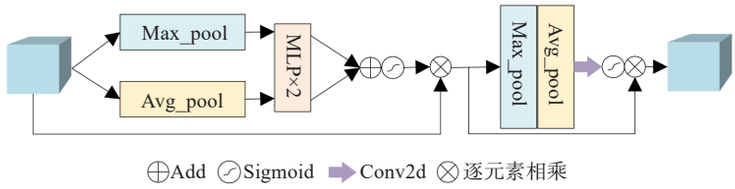

图4 卷积块注意力模块

Figure 4 Convolutional block attention module

3 分析与讨论

3.1 数据集和评价指标

Synapse是一个公共腹部多器官分割数据集，包含30例对比增强腹部临床CT病例，总计3 779张轴向切片。每个CT扫描都包涵85~198个512×512像素的切片，并对每个图像进行了8种腹部器官或组织（主动脉、胆囊、脾脏、左肾、右肾、肝脏、胰腺、脾脏和胃）的标注。按照文献[10]的划分，其中18个病例（共2 212个轴向切片）用于训练，12个案例（共849个轴向切片）用于测试。

ACDC是一个由100次检查组成的公共心脏MRI数据集。此MRI数据集由一系列覆盖心脏基底到左心室顶端的短周期案例组成。对于每个病例的短轴切片，有两种不同的模式：心脏舒张末期帧（ED）和心脏收缩末期帧（ES），相应的标签包括左心室（LV）、右心室（RV）和心肌（MYO）。数据集分为70个训练样本（共1 930个轴向切片）、10个验证样本和20个测试样本。

3.2 实验设置

本文使用PyTorch框架来构建MFMANet网络模型，并在Nvidia GTX 2080Ti GPU（11 GB）上对其进行训练和实验。具体来说，实验将批量大小设置为12，并使用学习率为$1e^{-4}$的Adam优化器。在进行实验之前需要将所有图像输入大小调整为224×224，实验所使用的数据增强方法包括随机旋转和随机翻转。除此之外，模型并没有使用预训练的方法。

3.2.1 评估指标　为了评估模型在图像分割中的性能，实验使用骰子相似系数（Dice Similarity Coefficient, DSC）和95%豪斯多夫距离（95% Hausdorff Distance, HD95）评估分割精度，DSC衡量预测结果和标签内部填充相似性，HD95衡量边界相似性。

$$\mathrm{DSC}=\frac{2\mathrm{TP}}{\mathrm{FP}+2\mathrm{TP}+\mathrm{FN}} \tag{7}$$

其中，TP表示实际为真且被预测为真的样本，FP表示实际为假但预测为真的样本，FN表示实际为真且预测为假的样本。

豪斯多夫距离（HD）如下式所示：

$$\mathrm{HD}=\max\left\{d_{XY},d_{YX}\right\}=\max\begin{Bmatrix}\max\limits_{x\in X}\min\limits_{y\in Y}d\left\{x,y\right\},\\ \max\limits_{y\in Y}\min\limits_{x\in X}d\left\{x,y\right\}\end{Bmatrix} \tag{8}$$

其中，X表示预测结果，Y表示真实标签，d表示X和Y之间的距离，d_{XY}表示对X中的每个点x到距离此点x最近的Y中点y之间的距离$\|x-y\|$。首先，对距离进行排序，选取最大值作为d_{XY}，d_{YX}同理，取这两者中最大值作为HD，在实际计算时，为排除可能由离群点引起的不合理距离，采用排名前95%的距离值，记为HD95，以确保整体数值的稳定。

3.2.2 损失函数　实验采用的损失函数为交叉熵-Dice混合损失函数，见式（9）~（11）。交叉熵损失函数被用作像素级别的标签预测与真实标签之间的差异度量，衡量两个概率分布之间的相似性。Dice损失函数不关注每个像素的分类，而是关注预测结果与真实结果的重叠情况，这使得它对于类别不平衡的图像分割任务效果更好，尤其是在背景像素占主导地位的情况下。以上两个损失函数混合使用可以综合考虑模型的分类准确性和相似性，从而更好地训练模型。

$$L_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}(y \times \log(p) + (1-y) \times \log(1-p)) \quad (9)$$

$$L_{\mathrm{Dice}} = 1 - \mathrm{Dice} = 1 - \frac{2\mathrm{TP}}{\mathrm{FP} + 2\mathrm{TP} + \mathrm{FN}} \quad (10)$$

$$L = \alpha \times L_{\mathrm{BCE}} + \beta \times L_{\mathrm{Dice}} \quad (11)$$

其中，L_{BCE}表示交叉熵损失，L_{Dice}表示Dice损失，L表示最终的损失函数，y表示期望的输出，p表示实际的输出，α和β分别为两个损失函数的权重。通过调整这两个权重的值，可以在一定程度上平衡交叉熵损失函数和Dice损失函数在最终损失值上的影响，本实验设定α、β都为0.5。

3.3 消融实验

为了研究网络组成部分是否有助于医学图像分割，本文在ACDC数据集上选用不同的网络架构进行实验。

3.3.1 MFMANet网络组成消融实验　表1数据结果表明，首先，UNet+SwinT的组合要优于R50_UNet和Swin-UNet这两个基线模型，这种卷积和Transformer组合实现了全局信息和局部信息的有效融合，使得模型能够更好地理解输入数据的整体结构和局部细节。其次，MIF模块和多重注意力模块的加入进一步提高了模型的分割结果，充分展现了多尺度特征信息的整合及注意力机制的恰当运用在医学图像分割任务中的重要性。最后，由以上模块组合成的MFMANet网络达到了最佳的分割结果，证明了该模型在医学图像分割任务上的有效性。

表1 网络组成消融实验（DSC, %）

Table 1 Ablation study of network composition (DSC, %)

网络结构	平均值	右心室	心肌	左心室
R50_UNet	87.55	87.10	80.63	94.92
Swin-UNet	90.00	88.55	85.62	95.83
UNet+SwinT	91.42	88.73	89.68	95.85
UNet+SwinT+MIF模块	91.53	89.19	89.86	95.55
UNet+SwinT+卷积块注意模块	91.68	89.60	89.58	95.86
MFMANet	91.94	90.05	89.93	95.85

3.3.2 MIF模块消融实验　MIF模块由扩张卷积层和通道注意力组成，为了深入探究扩张卷积层和通道注意力在特征提取中的增益效果，实验将这两个模块分别独立地加入网络浅层。由表2的结果可知，当两个模块以“扩张卷积层-通道注意力”的顺序组成MIF模块时，DSC指标达到最高，分割结果最好。这是因为MIF模块不仅能够考虑不同尺度下的特征信息，还能动态调整特征图中不同通道的重要性，从而获得更全面、更丰富的特征表示。这样的设计有助于网络获取区域边缘和角落细节信息，使预测更加完整。

3.3.3 多重注意力模块消融实验　为了验证不同注意力机制对网络的增益效果，本文将不同注意力加入到网络跳跃连接的不同位置。多重注意力机制消融实验结果由表3所示。结果显示，当编解码器中Transformer模块的跳跃连接上分别加入混合注意力时（按通道注意力、空间注意力的顺序加入其中），分割性能达到最高。这证明多重注意力机制有助于模型了解整体待分割结构及其与背景的关系，能够更准确地定位病变区域，协助模型更好地捕获分割图像的全局信息。除此之外，本文还尝试将编码器第4、5层得到的特征叠加之后输入混合注意力，再将得到的特征输入对应的解码器中。这种方法得到的DSC指标并不是最佳，这表明对于低分辨率特征，简单叠加这种融合方式虽然能够提升模型的分割结果，但并不是融合特征的最佳选择。

3.4 实验结果分析

3.4.1 ACDC数据集实验结果分析　与TransUNet、SwinUNet、MTUNet等几种流行的模型架构相比，本文提出的网络实现了91.94%的DSC，达到了最好的分割指标结果，相比于MTUNet这种同样基础架构的方法提升了1.51%，同时也比最新的方法TransCASCADE[26]提高了0.31%。由表4可知，MFMANet对左心室和右心室的分割效果也优于其他方法。图5显示了部分分割方法的可视化结果，红色区域代表左心室，蓝色区域代表右心室，绿色区域则代表心肌，如图所示，与其他基于Transformer的方

表2 多尺度信息融合模块消融实验（DSC, %）

Table 2 Ablation study of multi-scale information fusion block (DSC, %)

模块内容	扩张卷积层	通道注意力	平均值	右心室	心肌	左心室
扩张卷积层	√	×	91.76	89.26	90.02	95.98
通道注意力	×	√	91.38	88.56	89.84	95.75
MIF模块	√	√	91.94	90.05	89.93	95.85

表3 多重注意力模块消融实验(DSC, %)

Table 3 Ablation study of multi-attention block (DSC, %)

注意力组合方式	SC4	SC5	平均值	右心室	心肌	左心室
通道注意力	√	√	91.70	88.91	90.15	96.04
空间注意力	√	√	91.80	89.33	90.12	95.95
卷积块注意力模块	√	×	91.83	89.73	89.94	95.82
卷积块注意力模块	×	√	91.27	88.06	89.97	95.79
两个SC共用一个模块	√	√	91.65	89.15	90.14	95.66
卷积块注意力模块	√	√	91.94	90.05	89.93	95.85

SC表示跳跃连接,SC4、SC5分别表示网络中跳跃连接第4层和第5层

法相比,本文方法对右心室的边界分割更加准确,这一结果证明了本文方法在理解结构形态和组织特征方面具有更优异的性能。

表4 ACDC数据集实验结果(DSC, %)

Table 4 Experimental results on ACDC dataset (DSC, %)

模型	平均值	右心室	心肌	左心室
R50+UNet	87.55	87.10	80.63	94.92
R50+AttnUNet	86.75	87.58	79.20	93.47
TransUNet	89.71	88.86	84.53	95.73
SwinUNet	90.00	88.55	85.62	95.83
MTUNet	90.43	86.64	89.04	95.62
PVT-CASCADE[26]	91.46	88.90	89.97	95.50
TransCASCADE[26]	91.63	89.14	90.25	95.50
MFMANet	91.94	90.05	89.93	95.85

3.4.2 Synapse数据集实验结果分析　如表5所示,本文提出的MFMANet获得了最好的DSC结果(79.88%)。特别是胆囊、胰腺等这些边界不明显、形状变化大导致难以分割的结构,相较于MTUNet分别获得了6.33%和2.16%的提升。测试集部分分割结果可视化如图6所示,在案例1中,MFMANet在肝脏(黄色)上突出了正确的显著区域,其像素准确性和边界分割精度方面优于其他方法。在案例2中,MFMANet在胃部(亮蓝色)、胰腺(红色)、肝脏(黄色)的内部填充和边界轮廓上分割更准确。不同的方法比较和可视化结果证明了MIF模块和多重注意力的有效性,本方法使得模型能够在保留详细形状信息的同时,生成更精细的分割结果。与此同时,由表5中的HD95指标结果可知,尽管DSC指标表现良好,但模型在HD95指标方面仍存在改进空间,这可能是因为腹部器官存在更加复杂的器官边界或者相互重叠的现象。

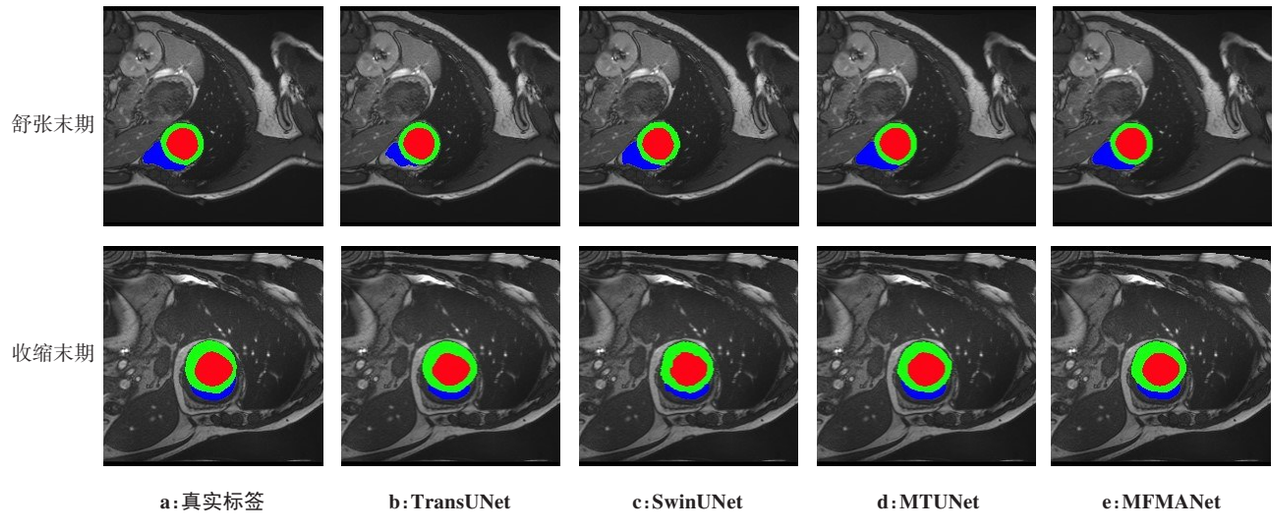

图5 ACDC数据集可视化结果

Figure 5 Visualization results on ACDC dataset

表5 Synapse数据集实验结果

Table 5 Experimental results on Synapse dataset

模型	平均HD95/mm	DSC/%								
		主动脉	胆囊	左肾	右肾	肝脏	胰腺	脾脏	胃	平均值
R50+UNet	36.87	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92	74.68
R50+AttenUNet	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95	75.57
TransUNet	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62	77.48
SwinUNet	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60	79.13
MTUNet	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81	78.59
MFMANet	28.70	88.39	71.32	83.02	79.55	93.80	61.62	86.93	74.44	79.88

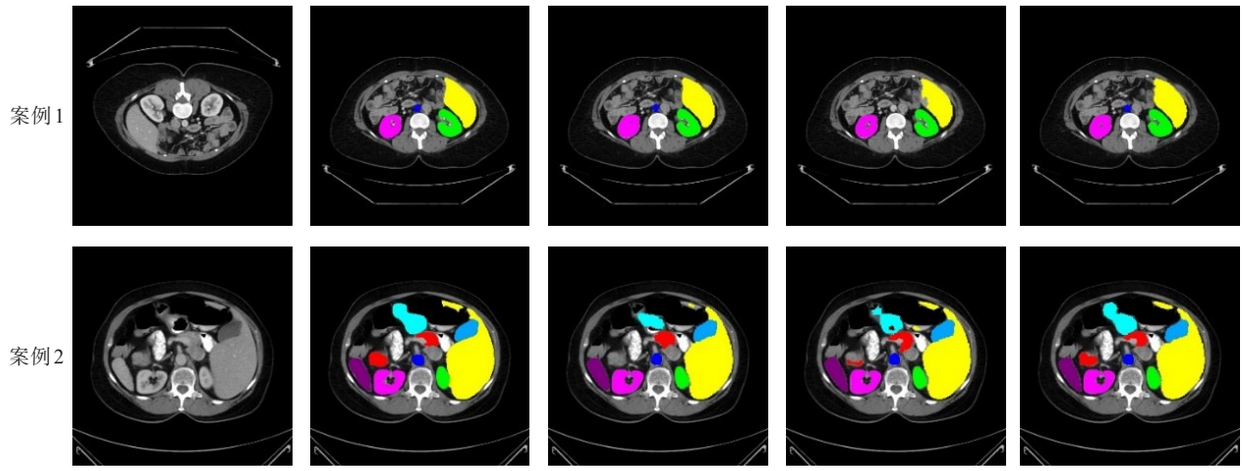

图6 Synapse数据集可视化结果

Figure 6 Visualization results on Synapse dataset

4 总 结

为解决医学图像分割中器官组织形状变化差异大和成像边界模糊等问题，本文提出的MIF模块增强了对多尺度特征信息的利用，采用多重注意力协助网络捕获局部-全局信息，成功缩小了编码器和解码器之间的语义差距，表现出良好的分割性能。在ACDC数据集和Synapse腹部多器官数据集上与其他现有模型进行比较，本文提出的MFMANet具有出色的分割性能和泛化能力，进一步证明利用多尺度信息和同时关注信息对象和重要区域，可以有效应对器官形态变化和成像边界模糊等挑战，获得更好的分割性能。未来，研究将密切关注领域最新进展，继续优化模型，进一步增强对多尺度信息的利用，提高HD95的指标结果，以进一步提升网络在边界区域的表现效果。

【参考文献】

[1] 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 International Publishing, 2015: 234-241.

[2] Zhou ZW, Siddiquee MMR, Tajbakhsh N, et al. UNet++: redesigning skip connections to exploit multiscale features in image segmentation[J]. IEEE Trans Med Imaging, 2020, 39(6): 1856-1867.

[3] Oktay O, Schlemper J, Le Folgoc L, et al. Attention U-net: learning where to look for the pancreas[EB/OL]. (2018-05-20). https://arxiv.org/abs/1804.03999.

[4] Milletari F, Navab N, Ahmadi SA. V-net: fully convolutional neural networks for volumetric medical image segmentation[C]//2016 Fourth International Conference on 3D Vision (3DV). Piscataway, NJ, USA: IEEE, 2016: 565-571.

[5] Huang HM, Lin LF, Tong RF, et al. UNet 3+: a full-scale connected UNet for medical image segmentation[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ, USA: IEEE, 2020: 1055-1059.

[6] He KM, Zhang XY, Ren SQ, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Trans Pattern Anal Mach Intell, 2015, 37(9): 1904-1916.

[7] Chen LC, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-12-05). https://arxiv.org/abs/1706.05587.

[8] Jha D, Riegler MA, Johansen D, et al. DoubleU-net: a deep convolutional neural network for medical image segmentation[C]// 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). Piscataway, NJ, USA: IEEE, 2020: 558-564.
[9] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. (2021-06-03). https://arxiv.org/abs/2010.11929.
[10] Chen JN, Lu YY, Yu QH, et al. TransUNet: transformers make strong encoders for medical image segmentation[EB/OL]. (2021-02-08). https://arxiv.org/abs/2102.04306.
[11] Wang WX, Chen C, Ding M, et al. TransBTS: multimodal brain tumor segmentation using transformer[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2021. Cham: Springer International Publishing, 2021: 109-119.
[12] Cao H, Wang YY, Chen J, et al. Swin-unet: unet-like pure transformer for medical image segmentation[C]//Computer Vision-ECCV 2022 Workshops. Cham: Springer Nature Switzerland, 2023: 205-218.
[13] Lin AL, Chen BZ, Xu JY, et al. DS-TransUNet: dual Swin transformer U-net for medical image segmentation[J]. IEEE Trans Instrum Meas, 2022, 71: 1-15.
[14] Zhang YD, Liu HY, Hu Q. TransFuse: fusing transformers and CNNs for medical image segmentation[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2021. Cham: Springer International Publishing, 2021: 14-24.
[15] Gao YH, Zhou M, Metaxas DN. UTNet: a hybrid transformer architecture for medical image segmentation[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2021. Cham: Springer International Publishing, 2021: 61-71.
[16] Hatamizadeh A, Tang YC, Nath V, et al. UNETR: transformers for 3D medical image segmentation[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway, NJ, USA: IEEE, 2022: 1748-1758.
[17] Wang HY, Xie SA, Lin LF, et al. Mixed transformer U-net for medical image segmentation[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ, USA: IEEE, 2022: 2390-2394.
[18] Cai YT, Wang Y. MA-Unet: an improved version of Unet based on multi-scale and attention mechanism for medical image segmentation[C]//Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021). Bellingham, WA, USA: SPIE, 2022: 121670X.
[19] Wang HN, Cao P, Wang JQ, et al. UCTransNet: rethinking the skip connections in U-net from a channel-wise perspective with transformer[C]//Thirty-Sixth AAAI Conference on Artificial Intelligence, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, the Twelveth Symposium on Educational Advances in Artificial Intelligence. Palo Alto, CA, USA: AAAI Press, 2022: 2441-2449.
[20] Wu RK, Liang PC, Huang X, et al. MHorUNet: high-order spatial interaction UNet for skin lesion segmentation[J]. Biomed Signal Process Control, 2024, 88, Part B: 105517.
[21] Ruan JC, Xiang SC, Xie MY, et al. MALUNet: a multi-attention and light-weight UNet for skin lesion segmentation[C]//2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Piscataway, NJ, USA: IEEE, 2022: 1150-1156.
[22] Ates GC, Mohan P, Celik E. Dual cross-attention for medical image segmentation[J]. Eng Appl Artif Intell, 2023, 126, Part D: 107139.
[23] Dai ZH, Liu HX, Le QV, et al. CoAtNet: marrying convolution and attention for all data sizes[C]//Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc., 2021: 3965-3977.
[24] Liu Z, Lin YT, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 9992-10002.
[25] Woo S, Park J, Lee JY, et al. CBAM: convolutional block attention module [C]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.
[26] Rahman MM, Marculescu R. Medical image segmentation *via* cascaded attention decoding[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway, NJ, USA: IEEE, 2023: 6211-6220.

（编辑：薛泽玲）