基于双分支特征交互和跨样本特征融合的胸腹部医学图像分割

王康，房乐鑫，李振宇，郝兴伟*，张彩明

(山东大学软件学院 济南 250101)
(hxw@sdu.edu.cn)

摘 要: 胸腹部医学图像分割在临床医学早期病变的筛查与诊断中具有不可替代的重要意义。针对胸腹部医学图像通常存在伪影干扰、边界模糊等因素，现有模型易产生分割错误的问题，通过构建双分支特征交互机制并联合跨样本特征融合，提出一种语义分割网络 MQUNet。首先采用基于 CNN-Transformer 的双分支架构提取当前图像的局部特征与全局特征，尽可能全面地捕捉器官间的局部细节和空间关联，并新增设一个基于改进的激活函数的负向关注分支融合这 2 种特征，同时兼顾特征间的相似性和差异性，避免语义信息的丢失；然后根据训练样本的类别标签建立记忆队列存储解剖先验特征，并通过跨样本特征检索与加权融合将训练样本中的稳定解剖先验动态加入当前分割过程，有效地校正由伪影和模糊情况导致的特征偏移。在公开的心脏肿瘤（ACDC）和腹部多器官（Synapse）数据集上进行实验的结果表明，Dice 系数指标分别达到 92.53%和 85.17%，优于文中对比的分割方法，MQUNet 在胸腹部医学图像分割任务中展现出显著的优势。

关键词: 胸腹部图像；医学图像分割；双分支架构；非对称交叉注意力；动态记忆队列
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2025-00248

Thoracic and Abdominal Medical Image Segmentation Based on Dual-Branch Feature Interaction and Cross-Sample Feature Fusion

Wang Kang, Fang Lexin, Li Zhenyu, Hao Xingwei*, and Zhang Caiming
(*School of Software, Shandong University, Jinan* 250101)

Abstract: Thoracic and abdominal medical image segmentation plays an irreplaceable role in the screening and diagnosis of early lesions in clinical medicine. In view of the fact that the thoracic and abdominal medical images usually have factors such as artifact interference and blurred boundaries, and existing models are prone to segmentation errors, a semantic segmentation network MQUNet is proposed by constructing a dual-branch feature interaction mechanism and combining cross-sample feature fusion. Firstly, the dual-branch architecture based on CNN-Transformer is adopted to extract the local and global features of the current image, capturing the local details and spatial correlations between organs as comprehensively as possible. It also integrates the two features by adding a negative attention branch based on an improved activation function, while taking into account both similarities and differences between features, avoiding the loss of semantic information. In addition, a memory queue is established based on the category labels of the training samples to store anatomical prior features. Through cross sample feature retrieval and weighted fusion, stable anatomical prior features from the training samples are dynamically added to the current segmentation process, effectively correcting feature shifts caused by artifacts and blurring. The experimental results on the publicly available cardiac cancer

收稿日期：2025-07-04；修回日期：2025-11-20。**基金项目**：国家自然科学基金（U22A2023，U24A20219）。**王康**（2000—），男，硕士，主要研究方向为计算机视觉、图像分割；**房乐鑫**（1999—），女，博士，主要研究方向为计算机视觉；**李振宇**（2001—），男，硕士，主要研究方向为计算机视觉；**郝兴伟**（1966—），男，博士，教授，CCF 会员，论文通信作者，主要研究方向为计算机软件；**张彩明**（1955—），男，博士，教授，博士生导师，CCF 高级会员，主要研究方向为 CAGD&CG、图像处理、计算金融。

(ACDC) and abdominal multi-organ (Synapse) datasets show that the Dice coefficient index reaches 92.53% and 85.17% respectively, which is superior to the segmentation methods compared in the paper. MQUNet demonstrates significant advantages in the task of thoracic and abdominal medical image segmentation.

Key words: thoracoabdominal images; medical image segmentation; dual-branch architecture; asymmetric cross-attention; dynamic memory queues

胸腹部医学图像分割作为精准医疗的核心技术，在胸腹部疾病的临床诊断与治疗规划中发挥着不可替代的作用。随着深度学习技术的发展，自动化分割方法显著地提升了早期病变检测与器官定量分析的效率，成为实现个性化诊疗的重要基础。然而，胸腹部解剖结构的复杂特性给分割任务带来了多重挑战。例如，心脏搏动产生的运动伪影导致心肌与纵隔结构在 MRI 影像中出现局部模糊或重影，干扰心腔边界与冠状动脉的精准勾画；纵隔脂肪、心肌等低对比度区域的边界模糊性，加之浸润性病变的细微影像学表现，也严重地制约分割精度。当前，基于深度学习的解决方案虽然取得了部分进展，但是在处理复杂病理改变与解剖变异共存的真实临床数据时，仍然面临误分割率升高的挑战。

在医学图像分割领域，局部感受野的渐进式特征提取机制是卷积神经网络（convolutional neural network，CNN）[1-5]的核心优势，即通过逐层堆叠的卷积核捕获从边缘、纹理到器官轮廓的局部空间关联，并结合下采样操作构建多尺度特征金字塔，使其在边界清晰的器官分割中展现出高精度和强鲁棒性。然而，CNN 的归纳偏置与局部性先验在面对胸腹部复杂解剖场景时暴露了明显的缺陷：一方面，卷积核的有限感受野难以建模远距离空间依赖，对胸腹部器官的分割易受局部伪影干扰；另一方面，局部卷积操作对低对比度边界的敏感性不足，并且其依赖局部梯度信息的特性易导致边界判定偏移或断裂，难以表征模糊的边界区域。这些局限性表明，单纯依赖局部卷积运算难以实现胸腹部多器官分割的临床级精度要求。随着 Transformer 模型在自然语言处理领域的突破性进展，其基于自注意力机制的长程依赖建模能力为医学图像分割提供了新的解决路径。与 CNN 的局部卷积运算不同，Transformer 通过全局注意力权重建立像素间的语义关联，显著地增强了对器官整体分割的适应能力。这类方法[6-9]的核心优势是其突破了卷积核的局部性限制，能够从全图尺度解析器官形态学特征和病理结构。但是，自注意力机制对图像细节的敏感性不足，容易忽略低对比度区域的微小结节或薄层解剖结构，如何综合利用两者的优势互补，提高系统的鲁棒性，以及在实际应用中提高图像分割的准确率，是值得深入研究的重要方向。然而，现有的特征融合模块[10-11]仅专注于 2 个特征之间的相似性而忽略差异性，融合后的特征未能充分地保留 CNN 的局部细节刻画能力和 Transformer 的全局语义一致性。因此，实现局部-全局特征优势互补的难点是构建同时具备相似性和差异性感知能力的特征融合机制。

现有基于深度学习的胸腹部器官分割模型[12-14]将每个训练样本视为独立同分布数据，通过端到端映射直接学习从图像到分割掩模的静态转换函数，这种范式忽视了跨样本间的潜在语义关联性，在图像伪影与模糊边界问题上受到局限。在单样本框架下，由于现有的模型仅依据当前受伪影污染的局部特征进行决策，无法对比区分解剖结构的固有特征和由伪影诱导的异常模式，因此易受伪影干扰而产生分割错误；另一方面，现有的方法仅利用单一样本内的多尺度上下文进行边界推断，由于缺乏对跨样本特征关联的显式建模机制，因此难以辨别边界模糊区域。如何结合解剖先验和跨样本特征消除伪影干扰，实现模糊区域的准确分割，依然是该领域的核心挑战。

本文提出一种基于 CNN-Transformer 混合架构的医学图像分割网络 MQUNet。首先通过双分支协同机制联合 CNN 的局部细节建模能力与视觉 Transformer 的长程依赖捕获特性，并通过非对称交叉注意力融合（asymmetric cross-attention fusion，ACAF）模块在常规交叉注意力路径外增设负向关注分支实现全局-局部语义的全面融合，充分地结合两者的互补优势。进一步，基于特征空间聚类构建多个类别的动态记忆队列（dynamic memory queue，DMQ）来存储解剖先验，并通过跨样本相似性检索与加权融合，将训练图像的解剖先验注入当前分割过程，解决胸腹部图像分割中的伪影干扰与边界模糊问题。

1 相关工作

1.1 CNN

CNN 凭借其局部感受野与平移不变性的固有优势，成为医学图像分割的核心技术，其通过层级卷积操作有效地捕获从边缘纹理到器官形态的局部解剖特征。自全卷积网络(fully convolution network，FCN) [1]开创端到端图像分割范式以来，U-Net[2]凭借其对称的编码器-解码器结构和跳跃连接机制，迅速成为医学图像分割的基准框架。U-Net 通过局部感受野逐级捕获多尺度特征，在器官边界清晰、形态相对固定的场景中展现出显著的优势。后续的研究中，ResUNet[3]通过残差连接增强深层特征稳定性，解决了梯度消失的问题；Attention U-Net[4]聚焦关键解剖区域，通过注意力门控机制动态地筛选关键特征；3D U-Net[5]通过体素级空间建模，突破了 2D 切片限制，具有构建 3D 空间建模能力；nnU-Net[15]通过自适应框架实现数据驱动的拓扑配置；EMCAD[12]通过构建高效解码器完成细粒度特征重建。这些模型通过对 CNN 的改进，进一步提升其对小样本数据与复杂解剖结构的适应性，推动了肿瘤分割、多器官勾画等临床任务的智能化进程。

1.2 视觉 Transformer

随着 Transformer 在自然语言处理领域的突破性进展，其基于自注意力机制的长程依赖建模能力逐渐被引入医学图像分析领域。在医学图像分割领域中，引入 Vision Transformer 与 U-Net 架构融合，如 TransUNet[16]通过将 CNN 提取的局部特征与 Transformer 编码的全局语义相结合，在小目标任务中实现了更精准的边界定位；Swin UNETR[6]利用分层移位窗口机制降低计算复杂度，在全身多器官分割中展现出对解剖结构空间关系的强建模能力；nnFormer[7]通过动态稀疏注意力机制优化心脏 MRI 序列的时空关联建模；UNesT[8]通过器官感知的位置编码提升腹部多器官分割的解剖合理性；SUnet[13]通过高效空间缩减注意力与多注意力特征融合机制，结合增强注意力门模块的残差设计，在腹部及胸部多器官分割中实现了高精度与低参数量的平衡。这些方法在需要复杂语义关联或跨区域交互的任务中表现出超越卷积模型的性能，其全局上下文感知特性显著地缓解了局部特征的局限性。

1.3 记忆网络

记忆网络通过显式存储原型特征增强模型表征能力，在无监督异常检测中展现出独特的优势。MemAE[17]中首次将可学习的记忆矩阵嵌入自编码器框架，通过编码器提取的特征与记忆槽进行相似性检索，重构正常模式的特征基元以抑制异常响应；然而，该记忆矩阵作为可训练参数存在过拟合风险，可能导致正常模式表征的多样性受限。针对这一问题，Park 等[18]提出非可学习的静态记忆模块，通过输入特征动态地更新记忆槽，缓解了梯度优化与模式固化的矛盾；但其采用基于余弦相似度的硬性匹配策略，限制了跨模态异常特征的泛化能力。在内存效率优化方面，Lv 等[19]设计动态原型单元，采用滑动平均机制实时更新紧凑的原型向量，显著地降低了传统记忆矩阵的空间复杂度；但是，特征压缩过程可能损失细粒度异常线索的判别性。Xiang 等[20]提出 SQUID，克服了记忆矩阵的局限性，并提出一种有效而高效的记忆队列。在医学图像分割中，记忆机制可以被重新定位为解剖先验的知识库，通过记忆队列动态地存储器官的典型形态特征，使模型在面对图像伪影或边界模糊区域时能够检索相似解剖模式并进行特征校准，以提升分割的鲁棒性。

2 MQUNet

2.1 整体框架

本文提出 MQUNet，通过双分支特征融合与动态记忆先验的协同设计，解决胸腹部医学图像分割中的图像伪影和边界模糊难题。MQUNet 采用 U-Net 的编码器-解码器架构，并通过跳跃连接传递语义信息，其整体框架如图 1 所示。该网络包含 3 个核心组件：（1）并行的 CNN-Transformer 双分支编码器，CNN 分支和 Transformer 分支分别提取局部细节和全局语义特征；（2）ACAF 模块，实现跨分支特征交互；（3）DMQ 模块，存储并更新解剖先验以增强模型泛化能力。

给定输入图像 $I \in \mathbb{R}^{H\times W\times D}$，$I$ 分别进入 CNN 和 Transformer 这 2 个分支。其中，CNN 分支包含 4 层，每层有 2 个连续的卷积，每个卷积都伴随着一个 Batch归一化函数和一个 ReLU激活函数；Transformer 分支包含 4 层，层内的结构沿用 SUnet[13]的配置，注意力头数为 8；为了取得最佳效果，2 个分支的层数和参数设置均由实验确定。

在 CNN 分支，首先将输入图像 I 划分为 $h\times w$ 个互不重叠的 Patch，以便在后续的 DMQ 模块中进行跨样本特征的交互；然后通过多层级联的卷积层和下采样层逐步构建多尺度局部特征 $\left\{\boldsymbol{F}_{\mathrm{c}}^{l}\right\}_{l=1}^{L}$。

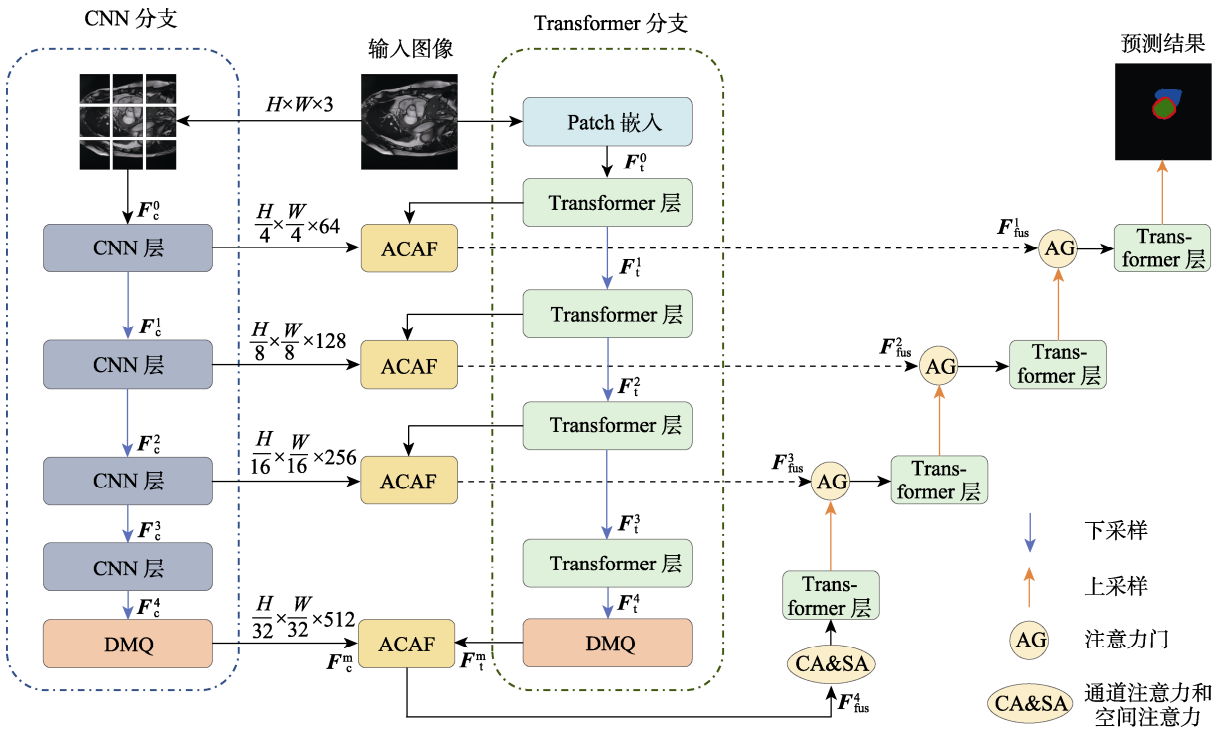


图 1　MQUNet 的整体框架

Transformer 分支将图像划分为 Patch 序列，经过多个自注意力层编码并下采样，得到多尺度全局上下文特征 $\{\boldsymbol{F}_{\mathrm{t}}^{l}\}_{l=1}^{L}$。在双分支相同尺度的每层的级联位置，由 ACAF 模块融合 $\boldsymbol{F}_{\mathrm{c}}^{l}$ 和 $\boldsymbol{F}_{\mathrm{t}}^{l}$ 生成增强特征 $\{\boldsymbol{F}_{\mathrm{fus}}^{l}\}$。双分支最后一层的结果 $\boldsymbol{F}_{\mathrm{c}}^{l}$ 和 $\boldsymbol{F}_{\mathrm{t}}^{l}$ 都会进入 DMQ 模块，该模块在训练过程中持续更新，采用跨样本特征检索为当前分割样本注入解剖先验知识，得到增强后的特征 $\boldsymbol{F}_{\mathrm{c}}^{\mathrm{m}}$ 和 $\boldsymbol{F}_{\mathrm{t}}^{\mathrm{m}}$，再由 ACAF 模块融合后进入解码器。解码器通过上采样与跳跃连接逐步恢复空间分辨率，最终输出分割掩模。

2.2　ACAF 模块

单独使用 CNN 难以对器官间的长程解剖依赖进行建模，容易受到局部伪影干扰；而纯 Transformer 网络对高频细节敏感性不足且计算复杂。为了克服单一构架的局限性，同时利用双分支的优点，本文提出 ACAF 模块。现有的融合模块往往只关注特征之间的相似性，如熊岚堃等[10]使用结合轴向增强的融合模块对 CNN 和 Transformer 进行融合，虽然保持了语义信息的一致性，但是这种融合方式往往忽略了特征之间的差异性信息，而这些差异性信息也是有效信息。ACAF 模块采用双向交互机制，在保留 CNN 局部细节和 Transformer 全局语义的基础上，利用正向-逆向注意力强化特征的相似性和差异性，实现复杂解剖结构的高精度分割，解决了现有特征融合模块仅专注 2 个特征之间的相似性而忽略差异性问题。

对于 CNN-Transformer 双分支中同一层级的特征 $\boldsymbol{F}_{\mathrm{c}}^{l}$ 和 $\boldsymbol{F}_{\mathrm{t}}^{l}$，ACAF 模块将它们投入到 2 个路径中，如图 2 所示。其中，一个路径将 CNN 分支的特征 $\boldsymbol{F}_{\mathrm{c}}^{l}$ 投影作为查询矩阵，Transformer 分支的特征 $\boldsymbol{F}_{\mathrm{t}}^{l}$ 投影作为键矩阵和值矩阵，实现非对称交叉注意力的计算，该过程称为 C-T 路径。与之对应，还存在一条由 $\boldsymbol{F}_{\mathrm{t}}^{l}$ 投影作为查询，$\boldsymbol{F}_{\mathrm{c}}^{l}$ 投影作为键和值，进行非对称交叉注意力计算的 T-C 路径。下面将以 C-T 路径为例，详细介绍每条路径的过程。

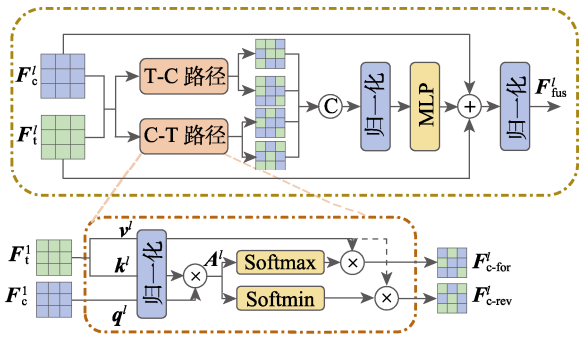


图 2　ACAF 模块示意图

传统的特征融合方法（如基于 Softmax 的注意力）倾向于强化 2 个分支的共识性特征，但可能会损失那些独特且具有判别性的特征。在医学图像

中，这些非共识区域往往对应着分割难点，如模糊边界或微小病变。负相关学习理论[21]通过鼓励模型间的差异性提升集成模型的泛化能力。受到该思想的启发，尽管经典负相关学习作用于独立模型之间，但是在同一网络内部，不同结构分支（如 CNN 与 Transformer）生成的特征间同样存在着宝贵的差异性信息，这些差异性信息可以显著地增强特征融合的互补性和多样性。因此，本文引入了 Softmin 函数，通过放大注意力得分中的最小值构建了一条显式的负相关注意力路径，以提取到不同特征之间的差异性信息。Softmin 函数表示为

$$\mathrm{Softmin}\left(z_i\right)=\exp\left(-z_i\right)\Big/\sum_{j=1}^{N}\exp\left(-z_j\right)。$$

其中，z_i 表示向量 $\boldsymbol{z}$ 的第 i 个元素，$i\in(1,2,\cdots,N)$。与常规交叉注意力不同，在进行注意力权重计算时，非对称交叉注意力不仅使用 Softmax 函数，还使用改进后的 Softmin 函数。其中，Softmax 建模正向关联的相似性，Softmin 建模逆向关联的差异性，两者协同使用，通过正向-逆向注意力机制强化特征，平衡模型对显著特征与细节特征的敏感性，提升了特征融合的效果。

如图 2 所示，在 C-T 路径中，CNN 分支的特征 $\boldsymbol{F}_{\mathrm{c}}^{l}$ 由投影得到查询矩阵 $\boldsymbol{q}_{\mathrm{c}}^{l}$，Transformer 分支的特征 $\boldsymbol{F}_{\mathrm{t}}^{l}$ 由投影得到键矩阵 $\boldsymbol{k}_{\mathrm{c}}^{l}$ 和值矩阵 $\boldsymbol{v}_{\mathrm{c}}^{l}$。经过归一化之后，$\boldsymbol{q}_{\mathrm{c}}^{l}$ 与 $\boldsymbol{k}_{\mathrm{c}}^{l}$ 进行点积注意力计算得到交叉注意力矩阵 $\boldsymbol{A}_{\mathrm{c}}^{l}$。$\boldsymbol{A}_{\mathrm{c}}^{l}$ 分别由 Softmax 函数和 Softmin 函数激活增强，提取到 2 个特征之间的相似性和差异性信息。根据激活后的 $\boldsymbol{A}_{\mathrm{c}}^{l}$ 对 $\boldsymbol{v}_{\mathrm{c}}^{l}$ 进行加权求和，获得 C-T 路径的正向注意力增强特征 $\boldsymbol{F}_{\text{c-for}}^{l}$ 与逆向注意力增强特征 $\boldsymbol{F}_{\text{c-rev}}^{l}$。这个过程可以表示为

$$\boldsymbol{q}_{\mathrm{c}}^{l}=\boldsymbol{F}_{\mathrm{c}}^{l}\boldsymbol{W}_{\mathrm{q}},$$

$$\left[\boldsymbol{k}_{\mathrm{c}}^{l},\boldsymbol{v}_{\mathrm{c}}^{l}\right]=\boldsymbol{F}_{\mathrm{t}}^{l}\boldsymbol{W}_{\mathrm{kv}},$$

$$\boldsymbol{A}_{\mathrm{c}}^{l}=\boldsymbol{q}_{\mathrm{c}}^{l}\cdot\left(\boldsymbol{k}_{\mathrm{c}}^{l}\right)^{\mathrm{T}},$$

$$\boldsymbol{F}_{\text{c-for}}^{l}=\mathrm{Softmax}\left(\frac{\boldsymbol{A}_{\mathrm{c}}^{l}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{v}_{\mathrm{c}}^{l},$$

$$\boldsymbol{F}_{\text{c-rev}}^{l}=\mathrm{Softmin}\left(\frac{\boldsymbol{A}_{\mathrm{c}}^{l}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{v}_{\mathrm{c}}^{l}。$$

同样地，由 $\boldsymbol{F}_{\mathrm{t}}^{l}$ 投影作为查询，$\boldsymbol{F}_{\mathrm{c}}^{l}$ 投影作为键和值，进行交叉注意力计算的 T-C 路径，也会产生 2 个特征 $\boldsymbol{F}_{\text{t-for}}^{l}$ 和 $\boldsymbol{F}_{\text{t-rev}}^{l}$。将这 4 个特征拼接起来，经过多层感知机（multilayer perceptron，MLP）进行通道压缩和非线性映射，并加上 $\boldsymbol{F}_{\mathrm{c}}^{l}$ 与 $\boldsymbol{F}_{\mathrm{t}}^{l}$ 的跳跃连接，得到的特征就是 $\boldsymbol{F}_{\mathrm{c}}^{l}$ 与 $\boldsymbol{F}_{\mathrm{t}}^{l}$ 经过 ACAF 的融合特征 $\boldsymbol{F}_{\mathrm{fus}}^{l}$。该过程表示为

$$\boldsymbol{F}_{\mathrm{fus}}^{l}=\mathrm{MLP}\left(\mathrm{Concat}\left(\boldsymbol{F}_{\text{c-for}}^{l},\boldsymbol{F}_{\text{c-rev}}^{l},\boldsymbol{F}_{\text{t-for}}^{l},\boldsymbol{F}_{\text{t-rev}}^{l}\right)\right)+\boldsymbol{F}_{\mathrm{c}}^{l}+\boldsymbol{F}_{\mathrm{t}}^{l}。$$

其中，Concat 表示通道维度的拼接；MLP 包括 3 层，第 1 层的输入维度是 $4c$，输出维度是 $8c$，第 2 层的输入维度是 $8c$，输出维度是 $4c$，第 3 层的输入维度是 $4c$，输出维度是 c，c 表示 $\boldsymbol{F}_{\mathrm{c}}^{l}$ 与 $\boldsymbol{F}_{\mathrm{t}}^{l}$ 的通道数，每层使用 ReLU 作为激活函数。

将非最后一层 ACAF 模块的融合结果作为多尺度上下文信息，经过跳跃连接传递到解码器中；最后一层 ACAF 模块的融合结果直接成为解码器的输入。

2.3 DMQ 模块

融合跨样本特征时，应有效地利用解剖先验知识。首先采用 *K*-means 算法对训练图像数据集基于其解剖结构进行聚类，然后构建记忆队列模块。该模块由若干个子队列组成，每个子队列中存储同一聚类图像的图像特征。这种子队列存储方式，不仅可以聚焦更具代表性的解剖特征，避免特征混淆，而且无需在全局记忆库中检索特征，降低了时间复杂度。

在训练集图像输入到分割模型之前，首先加载所有图像并将其展平为一维向量存储于列表中；值得注意的是，在网络训练过程中，由于不需要重复这个操作，因此不会增加网络训练时间。然后使用 *K*-means 算法（设聚类数为 N）对列表中的一维向量进行无监督聚类，生成每幅图像对应的类别标签。最后根据聚类结果将图像分组，同一组内的图像具有相似的解剖结构，同组的图像具有相同的类别标签，为后续分割任务中的记忆队列提供划分依据。

图 3 展示了 DMQ 模块的流程，图中以 CNN 分支的记忆队列为例，其中数字代表的操作分别是：①当前特征查询记忆队列；②按照余弦相似度匹配历史特征；③当前特征与历史特征进行融合；④当前特征存入记忆队列。这些操作是按时间顺序排列的。具体来说，在训练的过程中，图像 I 在 CNN 和 Transformer 这 2 个分支的最后一层的输出 $\boldsymbol{F}_{\mathrm{c}}^{L}$ 与 $\boldsymbol{F}_{\mathrm{t}}^{L}$ 进入 DMQ，其中包含 N 个子队列，若图像 I 的类别标签为 n，则会进入第 n 个子队列，$n\in(1,N)$。以 CNN 分支最后一层的结果 $\boldsymbol{F}_{\mathrm{c}}^{L}$ 为例，

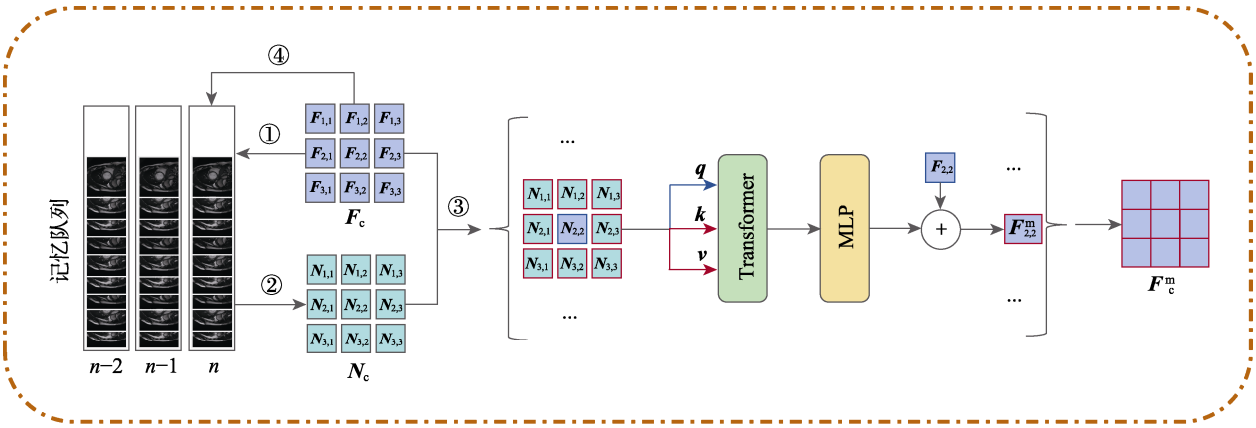


图 3　DMQ 模块的流程

按照第 2.1 节所述，CNN 分支从开始就将输入图像 I 划分为 $h\times w$ 个 Patch；经过 CNN 分支特征提取后得到的 $\boldsymbol{F}_{\mathrm{c}}^{L}$ 也同样具有 $h\times w$ 个 Patch，记为 $\boldsymbol{F}_{\mathrm{c}\{(1,1),\cdots,(h,w)\}}$。第 n 个子队列中存储训练集中之前进入分割框架的类别标签为 n 的图像特征，它们也具有相同的 Patch 结构。利用余弦相似性，从所有历史图像特征的 Patch 中匹配到与 $\boldsymbol{F}_{\mathrm{c}\{(1,1),\cdots,(h,w)\}}$ 每个对应位置 Patch 最相似的 Patch 特征，即用 $\boldsymbol{F}_{\mathrm{c}(i,j)}$ 匹配所有历史图像特征中下标为 (i,j) 的 Patch 特征，其中，$i\in(1,h)$，$j\in(1,w)$，得到的所有 Patch 特征组合为特征 $\boldsymbol{N}_{\mathrm{c}\{(1,1),\cdots,(h,w)\}}$。该过程表示为

$$\boldsymbol{N}_{\mathrm{c}\{(1,1),\cdots,(h,w)\}}=\mathrm{DMQ}\left(C\left(\boldsymbol{F}_{\mathrm{c}\{(1,1),\cdots,(h,w)\}}\right)\right)。$$

其中，C 表示计算余弦相似度。

然后，用 Transformer 块聚合当前特征 $\boldsymbol{F}_{\mathrm{c}}$ 与历史特征 $\boldsymbol{N}_{\mathrm{c}}$。对于 $\boldsymbol{F}_{\mathrm{c}}$ 中的每一个 Patch 特征 $\boldsymbol{F}_{\mathrm{c}(i,j)}$，选择以其为中心的 8 个相邻 Patch 在 $\boldsymbol{N}_{\mathrm{c}}$ 中的对应 Patch 组成 $\boldsymbol{N}_{\mathrm{c}\{(i-1,j-1),\cdots,(i+1,j+1)\}}$。以 $\boldsymbol{F}_{\mathrm{c}(i,j)}$ 作为查询矩阵，$\boldsymbol{N}_{\mathrm{c}\{(i-1,j-1),\cdots,(i+1,j+1)\}}$ 作为键矩阵和值矩阵，输入到 Transformer 块中得到结果与 $\boldsymbol{F}_{\mathrm{c}(i,j)}$ 进行残差连接后，记为 $\boldsymbol{F}_{\mathrm{c}(i,j)}^{\mathrm{m}}$，所有增强后的 Patch 特征 $\boldsymbol{F}_{\mathrm{c}(i,j)}^{\mathrm{m}}$ 组合为 $\boldsymbol{F}_{\mathrm{c}\{(1,1),\cdots,(h,w)\}}^{\mathrm{m}}$。这个过程可以表示为

$$\boldsymbol{q}_{\mathrm{c}}^{\mathrm{m}}=\boldsymbol{F}_{\mathrm{c}(i,j)}\boldsymbol{W}_{\mathrm{q}}，$$

$$\left[\boldsymbol{k}_{\mathrm{c}}^{\mathrm{m}},\boldsymbol{v}_{\mathrm{c}}^{\mathrm{m}}\right]=\boldsymbol{N}_{\mathrm{c}\{(i-1,j-1),\cdots,(i+1,j+1)\}}\boldsymbol{W}_{\mathrm{kv}}，$$

$$\boldsymbol{F}_{\mathrm{c}(i,j)}^{\mathrm{m}}=\mathrm{MLP}\left(\mathrm{Softmax}\left(\frac{\boldsymbol{q}_{\mathrm{c}}^{l}\cdot\left(\boldsymbol{k}_{\mathrm{c}}^{l}\right)^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{v}_{\mathrm{c}}^{l}\right)+\boldsymbol{F}_{\mathrm{c}(i,j)}，$$

$$\boldsymbol{F}_{\mathrm{c}\{(1,1),\cdots,(h,w)\}}^{\mathrm{m}}=\left\{\boldsymbol{F}_{\mathrm{c}(1,1)}^{\mathrm{m}},\cdots,\boldsymbol{F}_{\mathrm{c}(i,j)}^{\mathrm{m}},\cdots,\boldsymbol{F}_{\mathrm{c}(h,w)}^{\mathrm{m}}\right\}。$$

最后，图像 I 增强前的特征 $\boldsymbol{F}_{\mathrm{c}\{(1,1),\cdots,(h,w)\}}$ 被更新存储到当前子队列中，作为历史图像特征，供后续输入的图像特征进行跨样本特征融合。

与 CNN 分支相同，Transformer 分支也包含一个 DMQ。Transformer 分支中，图像 I 的特征 $\boldsymbol{F}_{\mathrm{t}}^{L}$ 被划分为 $\boldsymbol{F}_{\mathrm{t}\{(1,1),\cdots,(h,w)\}}$。经过类似的记忆队列模块后，$\boldsymbol{F}_{\mathrm{t}\{(1,1),\cdots,(h,w)\}}$ 转变为 $\boldsymbol{F}_{\mathrm{t}\{(1,1),\cdots,(h,w)\}}^{\mathrm{m}}$。$\boldsymbol{F}_{\mathrm{c}\{(1,1),\cdots,(h,w)\}}^{\mathrm{m}}$ 和 $\boldsymbol{F}_{\mathrm{t}\{(1,1),\cdots,(h,w)\}}^{\mathrm{m}}$ 经过 ACAF 模块融合，得到的结果输入到解码器中。

2.4　损失函数

本文设计的医学图像分割模型中使用骰子（Dice）损失函数和交叉熵（cross entropy，CE）损失函数。Dice 损失函数的表达式为

$$L_{\mathrm{Dice}}=1-\frac{2\left|X\cap Y\right|}{\left|X\right|+\left|Y\right|}。$$

其中，X 表示模型预测结果的像素集合，Y 表示真实标签中的像素集合。CE 损失函数的表达式为

$$L_{\mathrm{CE}}=-\sum_{i=1}^{N}y_i\ln\left(p_i\right)。$$

其中，N 表示类别数，y_i 表示样本 i 的真实标签，p_i 表示样本 i 的预测概率。模型的总体损失函数可以表示为

$$L=\lambda_1\times L_{\mathrm{Dice}}+\lambda_2\times L_{\mathrm{CE}}。$$

其中，λ_1 与 λ_2 是可设置的超参数，本文实验中设置为 λ_1=0.6，λ_2 =0.4。

3　实验及结果分析

本文通过多组实验，讨论和评估 MQUNet 的性能，证明其在医学图像分割任务中的有效性和优

越性。

3.1 数据集

本文在自动心脏诊断挑战数据集（ACDC）[22]和突触多器官分割数据集（Synapse）[16]这 2 个公开数据集上进行实验，分别对应胸部和腹部的医学图像分割。

ACDC 数据集源自 MICCAI 2017 挑战赛，广泛应用于心脏磁共振影像分析。该数据集包含 100 名不同患者的心脏 MRI 扫描数据，并标注了左心室（LV）、右心室（RV）和心肌（Myo）的精确分割标签；所有 3D 扫描数据均被转换为 2D 切片进行处理；按照 TransUNet，划分为 70 个训练用例、10 个验证用例和 20 个测试用例。

Synapse 数据集广泛应用于腹部多器官 CT 图像分割。该数据集包含 30 例腹部 CT 扫描病例（约 3 779 幅轴向切片），每例标注 8 个腹部关键器官的精确分割掩模，包括主动脉、胆囊、左肾、右肾、肝脏、胰腺、脾脏和胃；CT 图像分辨率范围为 512×512 像素，切片厚度 1.5~5.0 mm，涵盖多种成像条件；根据官方方法划分为 18 例训练集和 12 例测试集。

3.2 评价指标

为了定量地评估 MQUNet 的性能，使用 Dice 相似性系数（Dice similarity coefficient，DSC）、95% 豪斯多夫距离（Hausdorff distance，HD95）和交并比（intersection over union，IoU）这 3 个评价指标。它们分别定义为

$$\mathrm{DSC}=\frac{2|X\cap Y|}{|X|+|Y|},$$

$$\mathrm{HD95}(X,Y)=\max\left\{\sup_{x\in X}\inf_{y\in Y}d(x,y),\sup_{y\in Y}\inf_{x\in X}d(x,y)\right\},$$

$$\mathrm{IoU}=\frac{X\cap Y}{X\cup Y}。$$

其中，$d(\cdot)$表示 2 个点集之间的距离范式。

DSC 值越高，说明分割精度越好；HD95 值越低，说明预测结果与真实标签越吻合；IoU 值越高，说明预测结果和真实标签重叠区域越大。

3.3 实现设置

本文所有实验基于 Pytorch[23]框架，并在具有 48 GB 显存的 NVIDIA L40 GPU 上进行训练；使用 AdamW 优化器，学习率和权重衰减设置为10^{-4}。

在 ACDC 数据集的实验中，将 batch size 设置为 16，最大迭代次数设置为 150，输入图像大小和 Patch 大小分别设置为 224×224 和 16；使用随机翻转和旋转增强数据。

在 Synapse 数据集的实验中，batch size 设置为 24，最大迭代次数设置为 150，输入图像大小和 Patch 大小分别设置为 224×224 和 16；使用随机翻转和旋转增强数据。

在本文的实验设置下，模型的参数量为 69M，FLOPs 值为 19G。其中，DMQ 模块的参数量为 4.5M，FLOPs 值为 0.6G，该模块参数量固定，只在训练阶段替换历史特征，不参与梯度更新，存储占用和计算开销显著低于模型的总体消耗。

3.4 对比实验

在 2 个数据集上，将 MQUNet 与 10 种医学图像分割方法进行广泛的对比实验。

3.4.1 对比方法

本文实验中，10 种医学图像分割方法如下：

（1）U-Net[2]。一种对称编码器-解码器架构，通过跳跃连接实现多尺度特征融合，奠定了医学图像分割的基础范式。

（2）Attention U-Net（简称为 Att-Unet）[4]。在 U-Net 的跳跃连接中引入注意力门控机制，自适应地筛选空间显著特征，抑制无关背景干扰。

（3）R50-ViT[16]。构建了 ResNet50 和 Vision Transformer 的混合编码器，通过 CNN 局部感知与 Transformer 全局建模的互补性增强特征表达能力。

（4）TransUNet[16]。首次将 Transformer 作为 U-Net 的编码器，利用自注意力机制建立全局上下文依赖，提升对不规则病灶的建模能力。

（5）Swin-Unet[24]。基于 Swin Transformer 构建 U 形架构，通过移位窗口机制实现分层跨尺度全局推理，平衡计算效率与感受野的范围。

（6）PVTCASCADE[25]。金字塔视觉 Transformer 级联结构，通过渐进式多尺度特征精修策略实现细粒度病灶边界分割。

（7）MISSFormer[26]。多尺度交互式 Transformer，在编解码器间建立双向跨层注意力，实现解剖结构与局部细节的协同优化。

（8）TransCASCADE[25]。基于可变形 Transformer 的级联架构，通过动态感受野调整与迭代预测修正应对医学图像中的尺度敏感性问题。

（9）SUnet[13]。引入一个高效的空间缩减注意力模块，不仅可以更好地提取图像特征，还可以减少整体模型参数，减轻过拟合。

（10）EMCAD[12]。一种高效的多尺度卷积注意力解码器，实现了医学图像分割中的多阶段特征聚合和细化。

3.4.2　在 ACDC 数据集上的实验

在 ACDC 数据集上，将 MQUNet 与 10 种对比方法的 DSC 进行实验，结果如表 1 所示。其中，LV、RV 和 Myo 均使用 DSC 指标进行评价。可以看出：（1）与原始的 U-Net 相比，MQUNet 在 DSC 指标上提升 5.69%。（2）与其他方法相比，MQUNet 取得了最先进的性能，其中，与 TransCASCADE 相比，DSC 指标提升 0.98%；与 SUnet 相比，DSC 指标提升 0.31%；与 EMCAD 相比，的 DSC 指标提升 0.45%。这些实验结果充分验证 MQUNet 的卓越性能，反映了其双分支协同建模机制能够同时增强局部边界敏感性与全局伪影抗干扰能力。通过局部信息和全局信息全面建模，以及跨样本特征融合，MQUNet 实现了比对比医学图像分割方法更优的性能。

在 ACDC 数据集上，4 种方法的可视化结果如图 4 所示。可以看出，在心肌与心腔交界处，U-Net 和 MISSFormer 均出现断裂或模糊，而 MQUNet 保持了连续且清晰的边界；在受运动伪影影响的区域（如右心室边缘），U-Net、MISSFormer 和 SUnet 均出现不同程度的分割异常，而 MQUNet 通过跨样本特征融合有效地校正异常模式，避免了误分割；对于左心室乳头肌等细微结构，MQUNet 也展现出更强的识别能力，说明其双分支融合机制在局部-全局特征平衡方面的优势。实验结果表明，MQUNet 在各方面表现更优。

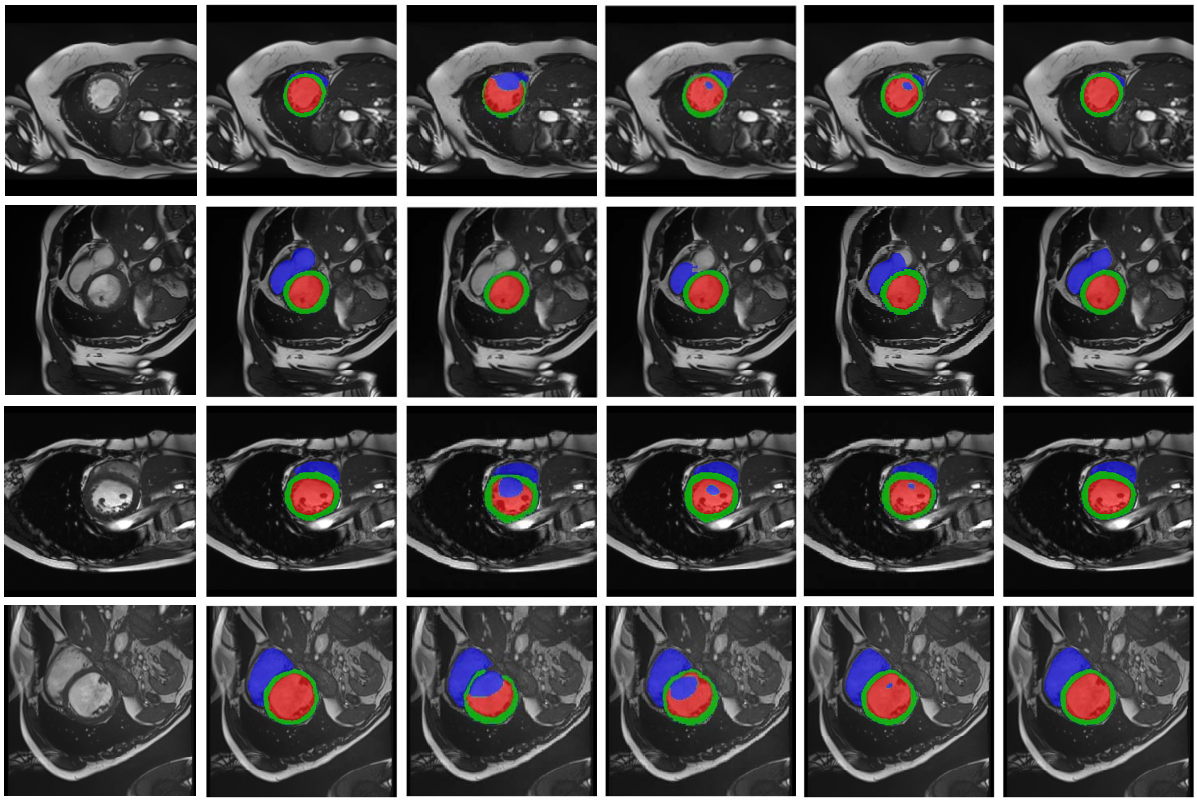
3.4.3　在 Synapse 数据集上的实验

在 Synapse 数据集上，将 MQUNet 与 10 种对比方法的性能进行实验，结果如表 2 所示。其中，主动脉（Aorta）、胆囊（GB）、左肾（KL）、右肾（KR）、

表 1　在 ACDC 数据集上 11 种方法 DSC 指标对比　%

方法	DSC ↑	不同标签 DSC ↑		
		RV	Myo	LV
U-Net[2]	87.55	87.10	80.63	94.92
Att-Unet[4]	86.75	87.58	79.20	93.47
R50-ViT[16]	87.57	86.07	81.88	94.75
TransUNet[16]	89.71	88.86	84.53	95.73
Swin-Unet[24]	90.00	88.55	85.62	95.83
PVTCASCADE[25]	91.46	88.90	89.97	95.50
MISSFormer[26]	90.86	89.55	88.04	94.99
TransCASCADE[25]	91.63	89.14	90.25	95.50
SUnet[13]	92.25	90.69	89.95	96.09
EMCAD[12]	92.12	90.65	89.68	96.02
MQUNet	**92.53**	**91.15**	**90.27**	**96.17**

注：粗体表示最优值。



a. 输入图像　b. 真实标签　c. U-Net[2]　d. MISSFormer[26]　e. SUnet[13]　f. MQUNet

图 4　在 ACDC 数据集上 4 种方法的可视化结果

表 2 在 Synapse 数据集上 11 种方法的性能对比 %

方法	DSC ↑	HD95 ↓	IoU ↑	不同标签 DSC ↑							
				Aorta	GB	KL	KR	Liver	PC	SP	SM
U-Net[2]	76.85	39.70	59.39	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
Att-Unet[4]	77.77	36.02	61.38	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
R50-ViT[16]	71.29	32.87		73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUNet[16]	77.48	31.69	67.32	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet[24]	79.13	21.55	66.88	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
PVTCASCADE[25]	81.06	20.23	70.88	83.01	70.59	82.23	80.37	94.08	64.43	90.10	83.60
MISSFormer[26]	81.96	18.20		86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
TransCASCADE[25]	82.68	17.34	73.48	86.63	68.48	87.66	84.56	94.43	65.33	90.79	83.50
SUnet[13]	84.29	19.46		87.29	73.92	86.85	83.44	**95.35**	69.70	**92.54**	85.24
EMCAD[12]	83.63	15.68	74.65	88.14	68.87	88.08	84.10	95.26	68.51	92.17	83.92
MQUNet	**85.17**	**14.88**	**75.83**	89.08	**74.17**	**90.01**	**85.29**	94.65	**70.44**	92.37	**85.31**

注：粗体表示最优值。

肝脏（Liver）、胰腺（PC）、脾脏（SP）和胃（SM）这 8 种标签均使用 DSC 指标进行评价。可以看出，与对比方法相比，MQUNet 的 DSC、HD95 和 IoU 这 3 种指标表现出更加优异的性能；与 TransCASCADE 相比，DSC 指标提升 3.01%，HD95 指标降低 14.19%，IoU 指标提升 3.19%；与 SUnet 相比，DSC 指标提升 1.04%，HD95 指标降低 23.54%；与 EMCAD 相比，DSC 指标提升 1.84%，HD95 指标降低 5.10%，IoU 指标提升 1.58%。实验结果表明，尤其是 HD95 指标的降低，充分地说明双分支的多尺度特征融合精准定位器官边缘，结合记忆队列的动态边界校正，大幅降低了分割轮廓的拓扑错误，表现出 MQUNet 在腹部 CT 图像分类任务中的显著优势。

3.5 消融实验

本文通过消融实验，系统地评估了 MQUNet 中各个组件的贡献，包括跨样本融合的记忆队列机制、基于相似特征聚类的子队列划分策略，以及基于正向-逆向注意力的相似性和差异性特征融合机制对模型性能的影响。

为了全面地评估各组件对模型性能的贡献，设置了涵盖 6 种情况的消融实验：

（1）Baseline。基线模型网络架构的编码器仅包含 CNN-Transformer 这 2 个分支，不使用 ACAF 和 DMQ 模块。

（2）Model1。作为 Baseline 的扩展，添加记忆队列（memory queue，MQ）来充分利用跨样本特征信息，但是不使用子队列划分机制（简称为 Sub）。

（3）Model2。在 Model1 的基础上，基于特征聚类将 MQ 划分为 Sub。

（4）Model3。在 Baseline 的基础上添加 ACAF 模块，促进双分支相似性和差异性特征融合。

（5）Model4。在 Model1 的基础上，添加 ACAF 模块。

（6）MQUNet。所有组件协作，构建完整的 MQUNet。

在 ACDC 数据集上，消融实验结果如表 3 所示。

表 3 在 ACDC 数据集上的消融实验结果

模型	Baseline	MQ	Sub	ACAF	DSC ↑
Baseline	√				91.67
Model1	√	√			92.24
Model2	√	√	√		92.42
Model3	√			√	92.13
Model4	√	√		√	92.35
MQUnet	√	√	√	√	92.53

可以看出：（1）Model1 仅使用未经特征聚类增强的 MQ，与 Baseline 相比，性能已有显著提升，表明 DMQ 能够有效地融合不同样本之间的特征信息，充分利用解剖先验，显著地提升分割精度。在不进行 Sub 的情况下，DMQ 存储的图像特征包含全部聚类特征。但是，检索特征时容易造成特征的混淆。（2）Model2 通过对样本划分类别，构建多个具有相似特征的 Sub，进一步提升了模型性能，说明该机制能够有效地增强 DMQ 的代表性，聚焦于更具典型性的解剖特征。（3）Model3 和 Model4 引入 ACAF 模块，性能比 Baseline 和 Model1 均有提升，表明，ACAF 模块能够充分地融合双分支的局部细节和全局语义，同时兼顾特征的相似性和差

异性。（4）MQUNet 中，DSC 指标显著提升，DMQ 融合跨样本特征解剖先验，特征聚类 Sub 增强特征代表性，ACAF 模块融合双分支局部与全局特征，这些组件协同作用，使得 MQUNet 能够在胸腹部医学图像数据下实现更高的分割精度。

3.6　参数敏感性实验

本文在 ACDC 数据集上进行参数敏感性实验，探讨特征聚类的类别数 N 对 MQUNet 性能的影响。

进行特征聚类时，设置的 N 将会影响后续记忆队列的划分，其 N 值越大，子队列的数目就越多，图像总数固定，每个子队列中存储的特征数量就越少。ACDC 数据集中主要包含 RV、Myo 和 LV 这 3 种解剖结构，在病理状态下，这些结构呈现出的典型形态变异（如扩张、萎缩、肥厚、变薄等），使心脏整体形态大致划分为若干模式，包括正常心脏、扩张型心肌病、肥厚型心肌病、心肌梗死、RV 异常等。基于不同解剖结构的不同影像表现（如表现为 LV 扩张与表现为双心室扩张的扩张型心肌病是 2 种不同的模式），可识别出 5~10 种主要的解剖-病理组合模式。因此，本文为子队列数量 N 选取了多个候选值：5、6、7、8、9、10，以期覆盖主要的临床特征。

在 ACDC 数据集上，参数敏感性实验结果如图 5 所示。

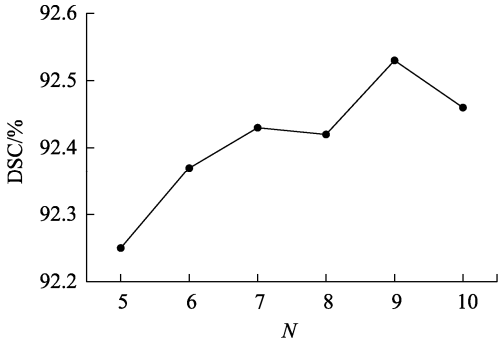


图 5　在 ACDC 数据集上的参数敏感性实验结果

可以看出，N 对于分割精度有显著影响，当 N 值过小时，分割精度下降，这是因为过少的子队列需要同时表征多种病理-解剖组合模式，内部特征分布离散度过高，破坏了同类解剖特征的语义一致性；反之，当 N 值过大时同样导致精度下降，过多的子队列虽然能够学习更细粒度的局部特征模式，但是每个子队列覆盖的样本数量显著减少，容易引发局部过拟合，损害模型的泛化能力。实验结果表明，最佳取值是 $N = 9$，此时子队列数量能够有效地覆盖上述解剖-病理组合模式，既避免因聚类过粗导致的模式混杂问题，又规避了因聚类过细引发的特征稀疏和过拟合风险，在解剖相似性和特征多样性之间取得平衡，取得了最佳的分割效果。

4　结　语

为了解决胸腹部医学图像因伪影干扰与边界模糊导致的特征偏移问题，本文提出一种基于双分支特征交互和跨样本特征融合的语义分割网络 MQUNet。通过构建 CNN 与 Transformer 双分支架构，实现了局部细节特征与全局空间关联建模的协同提取；并结合 ACAF 模块，有效地避免了多尺度语义信息的丢失；进一步，设计了 DMQ 模块，通过跨样本解剖特征的聚类存储和加权检索机制，将训练数据中的稳定先验动态地注入当前分割过程，显著地提升了模型应对伪影和模糊情况时的泛化能力。实验结果表明，MQUNet 在胸腹部复杂解剖场景下展现出优异的分割精度和鲁棒性。未来，将探索记忆队列的轻量化压缩策略，并研究其在多模态影像联合分割中的迁移机制，推动临床复杂病例的智能化诊疗应用。

参考文献（References）：

[1] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3431-3440

[2] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C] //Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2015: 234-241

[3] Diakogiannis F I, Waldner F, Caccetta P, *et al*. ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 162: 94-114

[4] Oktay O, Schlemper J, Folgoc L L, *et al*. Attention U-Net: learning where to look for the pancreas[OL]. [2025-07-04]. https://arxiv.org/abs/1804.03999

[5] Çiçek Ö, Abdulkadir A, Lienkamp S S, *et al*. 3D U-Net: learning dense volumetric segmentation from sparse annotation[C] //Proceedings of the 19th International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2016: 424-432

[6] Hatamizadeh A, Nath V, Tang Y C, *et al*. Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images[C] //Proceedings of the 7th International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Heidelberg: Springer, 2021: 272-284

[7] Zhou H Y, Guo J S, Zhang Y H, *et al*. nnFormer: volumetric medical image segmentation via a 3D transformer[J]. IEEE Transactions on Image Processing, 2023, 32: 4036-4045

[8] Yu X, Yang Q, Zhou Y C, *et al*. UNesT: local spatial representation learning with hierarchical transformer for efficient medical segmentation[J]. Medical Image Analysis, 2023, 90: Article No.102939

[9] Lu Ling, Qi Weimin. Spine CT image segmentation based on Transformer[J]. Journal of Image and Graphics, 2023, 28(11): 36 18-3628(in Chinese)
(卢玲, 漆为民. 基于Transformer的脊椎CT图像分割[J]. 中国图象图形学报, 2023, 28(11): 3618-3628)

[10] Xiong Lankun, Zhang Guimei, Liu Huiqun, *et al*. Combination of axial enhanced transformer and CNN network for medical image segmentation[OL]. [2025-07-04]. https://www.jcad.cn/article/doi/10.3724/SP.J.1089.2024-00356(in Chinese)
(熊岚堃, 张桂梅, 刘晖群, 等. 结合轴向增强Transformer与CNN双编码的医学图像分割方法[OL]. [2025-07-04]. https://www.jcad.cn/article/doi/10.3724/SP.J.1089.2024-00356)

[11] Shi Songlin, Pazilai M. Left ventricular segmentation algorithm for echocardiography combining transformer and CNN[J]. Journal of Computer-Aided Design & Computer Graphics, 2024, 36(9): 1418-1426(in Chinese)
(史松林, 帕孜来·马合木提. 结合Transformer和CNN的超声心动图左心室分割算法[J]. 计算机辅助设计与图形学学报, 2024, 36(9): 1418-1426)

[12] Rahman M M, Munir M, Marculescu R. EMCAD: efficient multi-scale convolutional attention decoding for medical image segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 11769-11779

[13] Li X S, Qin X, Huang C L, *et al*. SUnet: a multi-organ segmentation network based on multiple attention[J]. Computers in Biology and Medicine, 2023, 167: Article No.107596

[14] Jiang Ting, Li Xiaoning. Segmentation of abdominal CT and cardiac MR images with multi scale visual attention[J]. Journal of Image and Graphics, 2024, 29(1): 268-279(in Chinese)
(蒋婷, 李晓宁. 采用多尺度视觉注意力分割腹部CT和心脏MR图像[J]. 中国图象图形学报, 2024, 29(1): 268-279)

[15] 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

[16] Chen J N, Lu Y Y, Yu Q H, *et al*. TransUNet: transformers make strong encoders for medical image segmentation[OL]. [2025-07-04]. https://arxiv.org/abs/2102.04306

[17] Gong D, Liu L Q, Le V, *et al*. Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 1705-1714

[18] Park H, Noh J, Ham B. Learning memory-guided normality for anomaly detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 14360-14369

[19] Lv H, Chen C, Cui Z, *et al*. Learning normal dynamics in videos with meta prototype network[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 15420-15429

[20] Xiang T G, Zhang Y X, Lu Y Y, *et al*. SQUID: deep feature in-painting for unsupervised anomaly detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 23890-23901

[21] Liu Y, Yao X. Ensemble learning via negative correlation[J]. Neural Networks, 1999, 12(10): 1399-1404

[22] 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

[23] Paszke A, Gross S, Massa F, *et al*. PyTorch: an imperative style, high-performance deep learning library[C] //Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2019: Article No.721

[24] Cao H, Wang Y Y, Chen J, *et al*. Swin-Unet: unet-like pure transformer for medical image segmentation[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2022: 205-218

[25] Rahman M M, Marculescu R. Medical image segmentation via cascaded attention decoding[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 6211-6220

[26] Huang X H, Deng Z F, Li D D, *et al*. MISSFormer: an effective transformer for 2D medical image segmentation[J]. IEEE Transactions on Medical Imaging, 2023, 42(5): 1484-1494