计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(6):232−241 [doi: 10.15888/j.cnki.csa.009528]

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

多级特征交互 Transformer 的多器官图像分割①

武书磊[1], 张方红[2], 杨　有[2], 刘学文[2]

[1](重庆师范大学 计算机与信息科学学院, 重庆 401331)
[2](重庆师范大学 重庆国家应用数学中心, 重庆 401331)
通信作者: 杨　有, E-mail: 20130958@cqnu.edu.cn

摘　要: 多器官医学图像分割有助于医生做出临床诊断. 针对 CNN 提取全局特征能力弱, Transformer 提取局部特征能力弱, 以及 Transformer 具有二次方计算复杂度的问题, 提出了用于多器官医学图像分割的多级特征交互 Transformer 模型. 所提模型采用 CNN 提取局部特征, 局部特征经 Swin Transformer 输出全局特征; 通过下采样分别产生多级局部和全局特征, 每级局部和全局特征经过交互并增强; 每级增强后的特征经多级特征融合模块进行交叉融合; 再次融合后的特征经过上采样和分割头输出分割掩码. 所提模型在 Synapse 和 ACDC 数据集上进行实验, 平均 DSC 和平均 HD95 系数值为 80.16% 和 19.20 mm, 均优于 LGNet 和 RFE-UNet 等代表性模型. 该模型对多器官医学图像分割是有效的.

关键词: 多器官医学图像分割; 多级特征交互; Transformer; 卷积神经网络 (CNN); 语义分割; 深度学习

引用格式: 武书磊,张方红,杨有,刘学文.多级特征交互 Transformer 的多器官图像分割.计算机系统应用,2024,33(6):232–241. http://www.c-s-a.org.cn/1003-3254/9528.html

Multi-level Feature Interaction Transformer for Multi-organ Image Segmentation

WU Shu-Lei[1], ZHANG Fang-Hong[2], YANG You[2], LIU Xue-Wen[2]

[1](College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)
[2](National Center for Applied Mathematics in Chongqing, Chongqing Normal University, Chongqing 401331, China)

Abstract: Clinical diagnoses can be facilitated through the utilization of multi-organ medical image segmentation. This study proposes a multi-level feature interaction Transformer model to address the issues of weak global feature extraction capability in CNN, weak local feature extraction capability in Transformer, and the quadratic computational complexity problem of Transformer for multi-organ medical image segmentation. The proposed model employs CNN for extracting local features, which are then transformed into global features through Swin Transformer. Multi-level local and global features are generated through down-sampling, and each level of local and global features undergo interaction and enhancement. After the enhancement at each level, the features are cross-fused by multi-level feature fusion modules. The features, once again fused, pass through up-sampling and segmentation heads to produce segmentation masks. The proposed model is experimented on the Synapse and ACDC datasets, achieving average dice similarity coefficient (DSC) and average 95th percentile Hausdorff distance (HD95) values of 80.16% and 19.20 mm, respectively. These results outperform representative models such as LGNet and RFE-UNet. The proposed model is effective for multi-organ medical image segmentation.

Key words: multi-organ medical image segmentation; multi-level feature interaction; Transformer; convolutional neural network (CNN); semantic segmentation; deep learning

① 基金项目: 重庆市教委科技研究项目 (KJZD202200504); 重庆市自然科学基金创新发展联合基金 (市教委)(CSTB2023NSCQ-LZX0142); 重庆市高等教育教学改革研究项目 (232062)
收稿时间: 2023-12-25; 修改时间: 2024-01-23; 采用时间: 2024-01-29; csa 在线出版时间: 2024-04-19
CNKI 网络首发时间: 2024-04-23

1 引言

多器官医学图像的自动分割在计算机辅助诊断中有着重要作用[1]. 它通过提取器官区域的信息, 辅助医生做出适当的诊断. 自 21 世纪来, 随着医学成像领域的不断发展, 由此产生了大量的医学图像, 从中手动分割器官会耗费大量时间, 且分割同一器官时, 不同医生也会产生判断偏差[2,3]. 因此, 在医学图像中, 多器官分割仍然是一项具有挑战性的任务.

CNN 已被广泛应用于医学图像分割领域[4]. Ronneberger 等人[5]提出的 U-Net 采用端到端的网络形式, 编码器采用卷积层和池化层, 分层提取图像的局部特征. 解码器采用上采样操作, 将特征图恢复到原始图像的尺寸. 为提升医学图像分割任务的精度, U-Net 通过跳跃连接, 将编码器的特征图与解码器的特征图相融合, 从而保留更多的细节信息. U-Net 中的这种跳跃连接, 大幅度减少带标签的训练集. 后续的 U-Net 变种模型, 如 Zhou 等人[6]提出的 U-Net++, 通过加入密集连接方式重新设计编码器-解码器结构. Huang 等人[7]提出的 UNet 3+, 通过全尺度跳跃连接, 将来自不同尺度特征图的低级细节与高级语义相结合的一种编码器-解码器结构. 这些模型通过捕获目标特征以及位置信息对目标区域进行分割, 在医学图像分割领域取得了一定的效果, 但由于感受野限制和平移不变性所带来固有的局限性, 导致 CNN 在全局特征能力的建模和位置信息的识别是存在缺陷的. 在一些复杂的场景中, 受限的感受野不足以捕获全局特征信息, 平移不变性会使得模型对物体位置信息的提取能力有所下降, 从而导致分割性能下降[8,9].

Dosovitskiy 等人[10]提出的视觉 Transformer (vision Transformer, ViT) 克服了 CNN 的缺点. ViT 借助独特的自注意力机制 (self-attention, SA), 可以对输入位置与其他位置相关性进行建模, 使 Transformer 能够轻松捕捉不同位置之间的长距离依赖关系, 有助于全局特征信息的提取[11,12]. Wang 等人[13]在 Transformer 中引入 CNN 中金字塔结构, 以处理不同尺度的特征. 与传统的 ViT 不同, 金字塔结构允许 PVT 在不同尺度上执行特征提取和处理, 从而提高多尺度目标分割的性能. 虽然 ViT 缓解了 CNN 中感受野限制和平移不变性的局限性, 但 ViT 在捕获局部特征表示上表现不佳, 将 Transformer 应用于特定情况下的医学图像分割会面临新的挑战[14,15]. 为达到与 CNN 相似性能, 需要大量数据进行训练. 此外, 由于 SA 需要将每个位置与其他位置计算相似度分数, 所带来的计算复杂度是二次方[16]. 在医学图像分割任务中, 面对高维的医学图像数据时, 会降低模型的处理速度.

Liu 等人[17]提出的 Swin Transformer 降低了 ViT 的二次方计算复杂度. 作者采用一种移动窗口划分方法, 将 SA 限制在一个局部的窗口中, 捕获局部特征的同时降低自注意力的计算复杂度. 尽管 Swin Transformer 有能力在对全局上下文表示进行建模的同时, 提高计算效率, 但对局部特征能力的提取仍然受到限制. Chen 等人[18]提出的 TransUNet 缓解 Transformer 在医学图像中的弱局部表示. TransUNet 是首个将 Transformer 应用于医学图像分割领域的 U 型网络, 编码器采用 CNN 和 Transformer 提取局部和全局特征, 解码器通过上采样与高分辨率的局部特征相融合, 以达到精准定位的效果. 然而, TransUNet 的缺点是参数量大和计算效率低.

针对 CNN 和 Transformer 的上述缺点, 设计多级特征交互融合 Transformer 模型. 该模型首先通过 CNN 提取分级局部特征, 首级局部特征经 Swin Transformer 输出多级全局特征, 随后将同级的全局和局部特征进行加性融合, 融合后的特征再经过全局和局部特征增强 (global and local feature enhancement, GLFE) 模块, 在通道上做深度卷积降低运算参数量, 其次在 X 方向和 Y 方向进行聚合增强特征表示, 再采用逐点卷积对特征进行升维. 多个 GLFE 模块的输出采用 Shen 等人[19]提出的高效注意力 (efficient attention, EA) 和 Lin 等人[20]提出的交叉注意力 (cross attention, CA) 的双注意力机制组成的多级融合 (multi-level feature fusion, MLFF) 模块进行特征交叉融合. 最后, 通过上采样操作将交叉融合后的特征映射成原始图像尺寸, 通过分割头输出分割掩码.

- 设计 GLFE 模块, 使模型可以同时捕获跨通道信息以及长距离位置信息. 仿真实验表明, 加入 GLFE 模块在平均 DSC 和平均 HD95 系数上提升 0.58% 和降低 9.01 mm, GLFE 模块可以提升模型的定位能力.
- 设计 MLFF 模块, 使模型在降低计算复杂度的同时计算多级特征向量之间的注意力权重. 仿真实验表明, 添加 MLFF 模块的模型在平均 DSC 和平均 HD95 系数上提升 2.67% 和降低 8.53 mm, MLFF 模块可以提

升模型识别目标的能力和定位能力.

● 提出一种多级特征交互 Transformer 的多器官医学图像分割模型. 实验结果表明所提模型在 Synapse 和 ACDC 数据集上与代表性模型相比, 在平均 DSC 系数和平均 HD95 系数上具有竞争性优势.

2 所提模型结构

所提模型的网络结构如图 1 所示, 其中浅紫色和深蓝色块分别表示 CNN 和 Swin Transformer 层级. 首先编码器采用多级特征提取结构提取局部和全局特征信息, 通过 GLFE 模块将上述同级特征交互后再增强特征表示. 其次使用 MLFF 模块计算不同级特征之间的注意力权重, 捕获上下文语义信息. 最后, 将 MLFF 的特征映射经过级联上采样器和分割头, 输出分割掩码.

2.1 多级特征提取结构

如图 1 所示, 多级特征提取结构由 CNN 和 Swin Transformer 组成. CNN 和 Swin Transformer 分别包括 3 个不同的层级, 通过增大感受野的方式对图像进行分级特征提取.

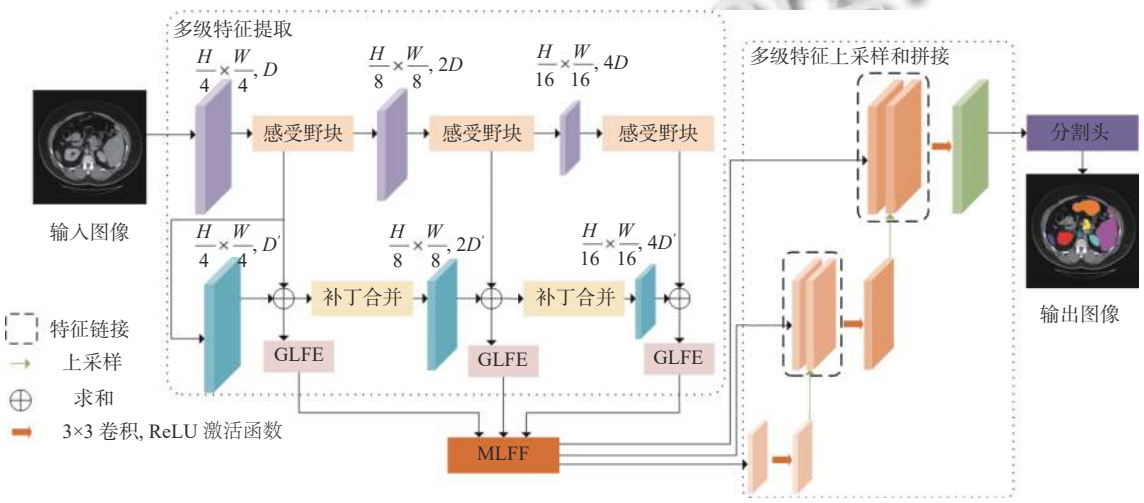

图 1 整体网络框架图

首先使用不同分辨率的 CNN 作为多级特征提取器, 以提取不同层级的特征图. 给定一个具有空间维度 H 和 W 以及通道数 C 的输入图像 $x\in R^{H\times W\times C}$, x 首先被送入由 3 级 CNN 组成的特征提取模块中. 在每一层 CNN 后, 通过 Liu 等人[21]提出的感受野块 (receptive field block, RFB) 再加强模型的特征提取能力. 再通过 1×1 的卷积, 将 RFB 的输出连接到同一级的 Swin Transformer 进行特征加性融合, 完善 Swin Transformer 捕获的局部信息.

在首层局部特征的基础上使用 Swin Transformer[17] 输出多级全局特征. ViT[10]主要由 L 层多头自注意力 (multi-head self-attention, MSA) 和多层感知机 (multi-layer perceptron, MLP) 组成, 在 MSA 和 MLP 之前采用层归一化 (LayerNorm, LN) 操作. 与 ViT 不同的是, Swin Tranformer 模块采用滑动窗口机制构建窗口多头自注意力 (window based MSA, W-MSA) 和滑动窗口多头自注意力 (shifted window based MSA, SW-MSA). 在 W-MSA 中, 在 $(W/4)\times(H/4)$ 的补丁基础上采用 $M\times M$ 大小的窗口重新划分补丁, 每个窗口只在内部进行相似计算. SW-MSA 利用窗口向右下移动的方式构建新的窗口, 从而计算不同窗口之间的相似计算, 其计算公式如下:

$$\hat{z}^{l}=\text{W-MSA}(\text{LN}(z^{l-1}))+z^{l-1} \tag{1}$$

$$z^{l}=\text{MLP}(\text{LN}(\hat{z}^{l}))+\hat{z}^{l} \tag{2}$$

$$\hat{z}^{l+1}=\text{SW-MSA}(\text{LN}(z^{l}))+z^{l} \tag{3}$$

$$z^{l+1}=\text{MLP}(\text{LN}(\hat{z}^{l+1}))+\hat{z}^{l+1} \tag{4}$$

其中, $\hat{z}^{l}$ 表示 W-MSA 的输出, z^{l} 表示 l 层的 MLP 的输出, $\hat{z}^{l+1}$ 表示 $l+1$ 层的 SW-MSA 的输出, z^{l+1} 表示第 $l+1$ 层的 MLP 的输出.

2.2 全局和局部特征增强模块

常见的特征融合操作有拼接, 相加, 相乘操作, 但在某些条件下会带来一系列问题. 例如, 相加操作是将

两个相同维度的特征进行叠加, 通过增加描述图像的特征信息量, 提高模型的表示能力. 同时会使得不正确的特征信息被加入融合后的特征中, 导致一些重要特征的丢失, 从而降低模型的分割性能.

GLFE 模块如图 2 所示, 实现降低卷积计算的参数量的同时整合空间位置信息. 先用深度卷积对特征进行降维, 降维后的特征在 X 方向和 Y 方向通过全局平均池化压缩成一维向量, 最后使用逐点卷积对特征进行升维. 压缩成一维向量的计算公式如下:

$$z_c^h(h)=\frac{1}{W}\sum_{0\leqslant i\leqslant W}x_c(h,i) \tag{5}$$

$$z_c^w(w)=\frac{1}{H}\sum_{0\leqslant j\leqslant H}x_c(j,w) \tag{6}$$

其中, $z_c^h(h)$表示为高度为h的第c个通道的输出. $z_c^w(w)$表示为宽度为w的第c个通道的输出. 通过将输入特征沿着水平方向和垂直方向聚合特征, 一方面捕捉沿着空间方向的长期依赖关系, 另一方面保存沿着另一个空间方向的精确位置信息. 这有助于模型学习到感兴趣的目标区域, 从而提高模型的分割性能.

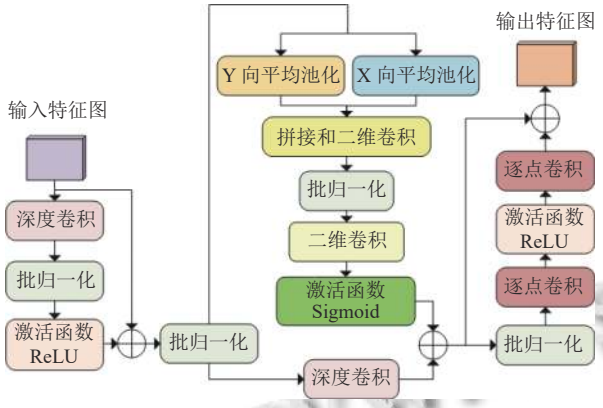

图 2 全局和局部特征增强模块

2.3 多级融合模块

不同级别的特征对应不同的语义信息, 如何将这些语义信息进行关联, 是模型捕获目标区域信息的关键. 针对 GLFE 输出的多级特征信息, 设计多级融合模块, 使模型同时捕获输入特征映射的空间重要性和关联不同特征映射之间的语义内容.

图 3 所示为 MLFF 模块. 增强后的多级特征作为输入, 采用 EA 和 CA 的双注意力结构捕获输入特征映的空间重要性的同时计算注意力权重. 标准的 SA 的计算复杂度是O(N^2), 限制了 Transformer 对高分辨率图像的适用性, EA 提出一种计算 SA 机制的有效方法, 计算公式如下所示:

$$\mathrm{EA}(Q,K,V)=\rho_q(Q)(\rho_k(K)^{\mathrm{T}}V) \tag{7}$$

其中, ρ_q和ρ_k表示归一化函数. 这种计算方式不同于 SA, 将$K\in R^{n\times d_k}$看作全局特征表示, 和V相乘后得到全局上下文向量, Q看作全局上下文向量的系数. 这种全局特征表示代表整个输入特征的语义信息, 而不是点与点之间的相似性. 这种计算方式极大地降低注意力机制的计算复杂性, 采用有效注意力可以捕获输入特征的空间重要性. CA 不同于 SA 的是, 在 CA 中, 查询矩阵来自一个输入序列, 键和值矩阵来自另一个输入序列. MLFF 模块的输入为 3 个不同级的特征图向量, 采用全局平均池化为每一个特征图向量分配一个对应的 CLS 令牌. 计算公式如下所示:

$$CLS^l=\mathrm{GAP}(\mathrm{LN}(P^l)) \tag{8}$$

其中, CLS^l为第l层通道维度的 CLS 令牌, P^l为当前层特征映射. 通过添加位置编码为每个CLS^l和P^l添加可学习的位置信息. 再使用 SA 融合每个级别的特征. 先交换融合不同级别的CLS^l和P^l, 具体来说, CLS^l和P^{l+1}拼接, CLS^{l+1}和P^{l+2}拼接, CLS^{l+2}和P^l拼接. 通过 SA, 最后在映射投影到各自的维度, 与当前层特征映射连接输出. 通过与其他层级CLS^l融合计算的方式, 实现跨层级共享信息. 计算公式如下:

$$z^l=\left[f^l(CLS^l),P^{l+1}\right] \tag{9}$$

$$p^l=f^l(CLS^l)+\mathrm{EA}(\mathrm{LN}(z^l)) \tag{10}$$

$$p^{l'}=f^{l'}(p^l)+\mathrm{CA}(\mathrm{LN}(p^l)) \tag{11}$$

$$Z^l=\left[P^l,g^{l'}(p^{l'})\right] \tag{12}$$

其中, z^l表示连接后的特征, [·]表示连接操作, P^{l+1}表示$l+1$层的特征映射, $f^l(\cdot)$表示将l层的CLS^l维度映射到P^l维数上, p^l表示为 EA 的特征映射, $p^{l'}$表示为 CA 的特征映射, $g^{l'}(\cdot)$表示将l层的$p^{l'}$维度映射到CLS^l维数上, $Z^{l+2}\in R^{(HW/P^2)\times 4D}$, $Z^{l+1}\in R^{(HW/(P/2)^2)\times 2D}$, $Z^l\in R^{(HW/(P/4)^2)\times D}$. 将$CLS^l$的映射和$P^{l+1}$作为键和值矩阵, 计算注意力的查询矩阵. 由于只查询 CLS 令牌, 所以交叉注意力机制为线性复杂度.

2.4 级联上采样器

MLFF 模块的输出为 3 个特征映射Z^l, Z^{l+1}, Z^{l+2},

采用连续卷积融合上采样解码器将 3 个特征映射结合为统一的特征掩码. 首先通过3×3的卷积运算, 双线性插值上采样, 分组归一化操作和激活函数 ReLU 构成的级联上采样器将Z^l映射到Z^{l+1}尺寸, 再与Z^{l+1}进行特征连接. 同理, 将连接后的特征映射至Z^{l+2}大小, 再与Z^{l+2}进行特征连接. 再将融合后的特征经过卷积块, 得到统一的$H\times W$特征图. 将获得的特征图输入进分割头中的3×3卷积, 生成最终的分割掩码.

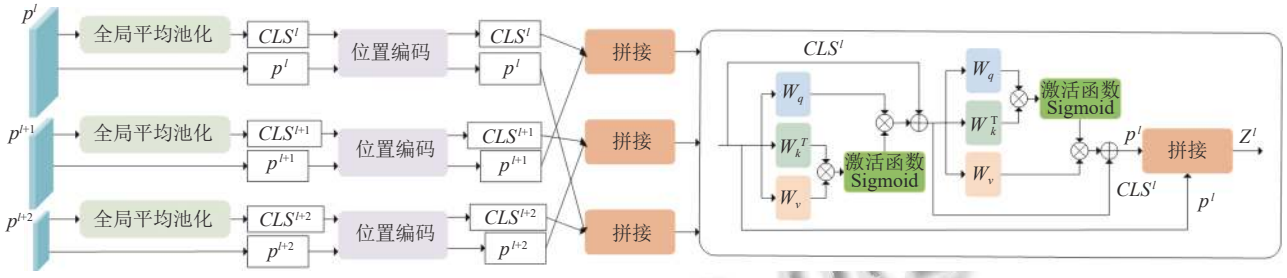

图 3 多级特征融合模块

3 实验

3.1 数据集和评价指标

Synapse: Synapse 数据集由 30 例病人共 3 779 张腹部轴向临床 CT 图像构成. CT 体积由 85–198 张切片组成. 为保证实验的一致性, 与 TransUNet[18]的数据集划分设置相同, 18 个病例用于模型的训练, 12 个病例用于模型的测试.

ACDC: ACDC 数据集由 100 个心脏磁共振图像序列构成. 每个序列包含收缩末期和舒张末期, 包括左心室, 右心室和心肌等心脏结构. 为保证实验的一致性, 与 TransUNet[18]的数据集划分设置相同, 70 样本用于训练, 10 个样本用于验证和 20 个样本用于测试.

为了分析所提模型在 Synapse 和 ACDC 数据集上的实验结果. 研究中使用平均骰子相似系数 (Dice similarity coefficient, DSC) 和平均 95% 的豪斯多夫距离 (Hausdorff distance, HD95) 作为评估指标.

DSC: *Dice* 相似系数是一种用于比较两个样本相似度的指标. 在医学图像分割领域, 它的取值范围为 0–1, 其中 1 表示两个样本完全一致, 0 表示两个样本完全不同. 在医学图像分割中, *Dice* 相似系数用于评估预测结果与真实结果之间的重叠程度, 即两者之间的交集与并集的比值. 较高的 *Dice* 相似系数通常意味着较好的分割效果. 计算公式如下:

$$Dice(P,T)=\frac{2|P\cap T|}{|P|+|T|} \tag{13}$$

其中, P表示模型的分割结果图像, T表示真实标签图像.

HD95: HD 是一种用于比较两个集合之间相似度的度量方法, 其度量单位是 mm. 在医学图像分割领域, HD 被广泛用于评估预测分割结果与真实分割结果之间的相似度, 特别是用于评估分割结果的边界或轮廓的准确性. 计算过程如式 (14) 所示. HD95 是通过 HD 结果值乘以 95%, 目的是消除离群值的一个非常小的子集的影响.

$$H(X,Y)=\max\{h(X,Y),h(Y,X)\} \tag{14}$$

其中, $H(X,Y)$表示双向 HD95, $h(X,Y)$表示为集合X到集合Y的单向 HD95, $h(Y,X)$表示为集合Y到集合X的单向 HD95.

3.2 实验配置

所有实验均在 Linux 操作系统上进行, CPU 配置为 i7-12600KF, 显卡配置为 RTX 3060. 深度学习框架为 PyTorch 1.10 和 Python 3.7, CUDA 版本为 11.2. 训练分割模型时, 通过翻转和旋转数据集中的样本, 改善数据多样性; 从 ImageNet 上获得 CNN 和 Swin Transformer 模块的预训练权重, 使用该权重对所提模型进行初始化; 模型输入图像大小为 224×224, 训练期间的批量大小和学习率分别为 8 和 0.01; 所提模型使用 SGD 算法进行优化, 其中动量设置为 0.9, 权重衰减设置为 0.000 1.

3.3 对比实验

为了验证所提模型在多器官医学图像分割任务上的有效性. 与近两年较新的医学图像分割网络 (如 MT-Unet, LGNet 和 RFE-UNet) 进行比较. 所有实验都在统一的基准和评价指标上进行, 不同的模型在两个数据集上的实验结果如表 1 和表 2 所示.

表 1 所提模型在 Synapse 数据集上的实验结果

网络模型	平均		DSC (%)							
	DSC (%)↑	HD95 (mm)↓	主动脉	胆囊	左肾	右肾	肝脏	胰腺	脾脏	胃
V-Net[22]	68.81	—	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR[23]	69.77	—	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
U-Net[5]	76.85	39.70	89.07	**69.72**	77.77	68.60	93.43	53.98	86.67	75.58
AttenUNet[24]	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
TransUNet[18]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet[25]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
TransClaw[26]	78.09	26.38	85.87	61.38	**84.83**	79.36	94.28	57.65	87.74	73.55
MT-Unet[27]	78.28	32.07	86.95	64.01	82.18	77.606	94.30	60.88	88.16	72.19
CTC-Net[28]	78.41	22.52	86.46	63.53	83.71	80.79	93.78	59.73	86.87	72.39
TransDeepLab[29]	80.16	21.25	86.04	69.16	84.08	79.88	93.53	**61.19**	89.00	78.40
RFE-UNet[30]	79.77	21.75	87.32	65.40	84.18	**81.92**	**94.34**	59.02	89.56	76.45
LGNet[31]	80.15	21.21	88.06	66.48	83.98	81.57	94.00	59.99	90.90	76.21
所提模型	**80.17**	**19.20**	87.47	68.62	82.69	76.94	94.11	60.26	**91.65**	**79.62**

表 2 所提模型在 ACDC 数据集上的实验结果 (%)

网络模型	平均DSC ↑	左心室	心肌	右心室
R50 UNet[18]	87.60	84.62	84.52	93.68
R50 Attn UNet[18]	86.90	83.27	84.33	93.53
ViT-CUP[18]	83.41	80.93	78.12	91.17
R50 ViT[18]	86.19	82.51	83.01	93.05
TransUNet[18]	89.71	86.67	87.27	95.18
Swin-Unet[25]	88.07	85.77	84.42	94.03
MT-Unet[27]	90.43	86.64	**89.04**	95.62
HiFormer[32]	89.20	86.43	86.50	94.69
TransUnet+[33]	90.47	**89.13**	87.96	94.31
DAE-Former[16]	84.42	81.99	79.77	91.50
所提模型	**91.07**	88.86	88.74	**95.61**

表 1 为所提模型在 Synapse 数据集上实验结果, 加粗字体表示最佳结果. 所提模型在平均 DSC 系数上到达最高的 80.17%, 在平均 HD95 系数上降低为 19.20 mm. 所提模型相比于基于 CNN 的 U-Net, 在平均 DSC 系数上提升 3.32%. 在平均 HD95 系数上降低 20.50 mm. 相比于基于 Transformer 的 Swin-Unet, 在平均 DSC 系数上提升 1.04%, 在平均 HD95 系数上降低 2.35 mm. 相比于基于 CNN-Transformer 的 TransUNet, 平均 DSC 系数上分别提升 2.69%, 在平均 HD95 系数上降低 12.49 mm. 实验结果表明, 所提模型在这两个评估指标上都表现出优秀的性能, 特别是在脾脏分割上优于其他模型, 而在胰腺和胃的分割上也优于大多数模型.

表 2 为所提模型在 ACDC 数据集上的实验结果, 表中加粗字体表示最佳结果. 所提模型在平均 DSC 系数上取得最高的值为 91.07%. 对比基于 Transformer 的 Swin-Unet, 在平均 DSC 系数提升 3%, 对比基于 CNN-Transformer 混合所提模型的 TransUNet 和 HiFormer, 平均 DSC 系数分别提升 1.36% 和 1.87%. 实验结果表明, 所提模型对心脏数据集的分割精度更高.

3.4 可视化结果及分析

图 4 为部分多器官分割实例的定性结果, 图中红色矩形框中的分割效果差异更大. 首先, 与基于混合架构的模型相比, 基于 Transformer 的模型在捕捉局部细节方面较弱. 在病例 2 中, Swin-Unet 在分割胃和胰腺方面表现不佳, 主要因为基于 Transformer 的模型在建模过程中强调上下文信息, 而忽略局部细节的重要性. 因此, 基于 Transformer 的模型会丢失局部信息, 进而影响分割性能. 其次, 基于混合架构的模型直接将局部特征与全局特征相结合, 这会引入信息冗余或信息丢失, 导致模型分割效果不好, 如 TransUNet 对胃和胰腺的分割是残缺的. 最后, 考虑到器官的复杂性和多样性, 局部特征和全局特征的融合对于优化分割性能至关重要. 与其他模型相比, 所提模型显示出优秀的分割结果. 首先, 通过 GLFE 模块增强特征表示. 其次, MLFF 模块在允许不同分辨率的局部和全局特征融合. 特别是, 该模型在精确分割复杂结构方面表现出色, 从而产生更精确的分割结果. 在 4 个案例中, 所提模型在胃、脾脏和胰腺分割上具有优势, 与表 1 中的结果相符合.

图 5 为部分心脏分割实例的定性结果, 图中红色矩形框中的分割效果差异更大. 首先基于混合架构的模型, 在有效融合局部和全局特征方面面临挑战. 例如, 在病例 1 中, TransUNet 对于收缩末期的右心室分割是不足的. HiFormer 缓解了这个问题, 它通过 DLF 模块巧妙地融合局部和全局特性. 其次基于 Transformer 的模型, 由于局部特征相关性建模的不足, 往往会出现分割错误. 在

案例 2 中, DAE-Former 对于收缩末期的右心室分割是残缺的. Swin-Unet 和 MT-Unet 对收缩末期的右心室分割不够准确. 最后所提模型在分割舒张末期和收缩末期的心肌方面表现优秀, 与表 2 中的结果一致.

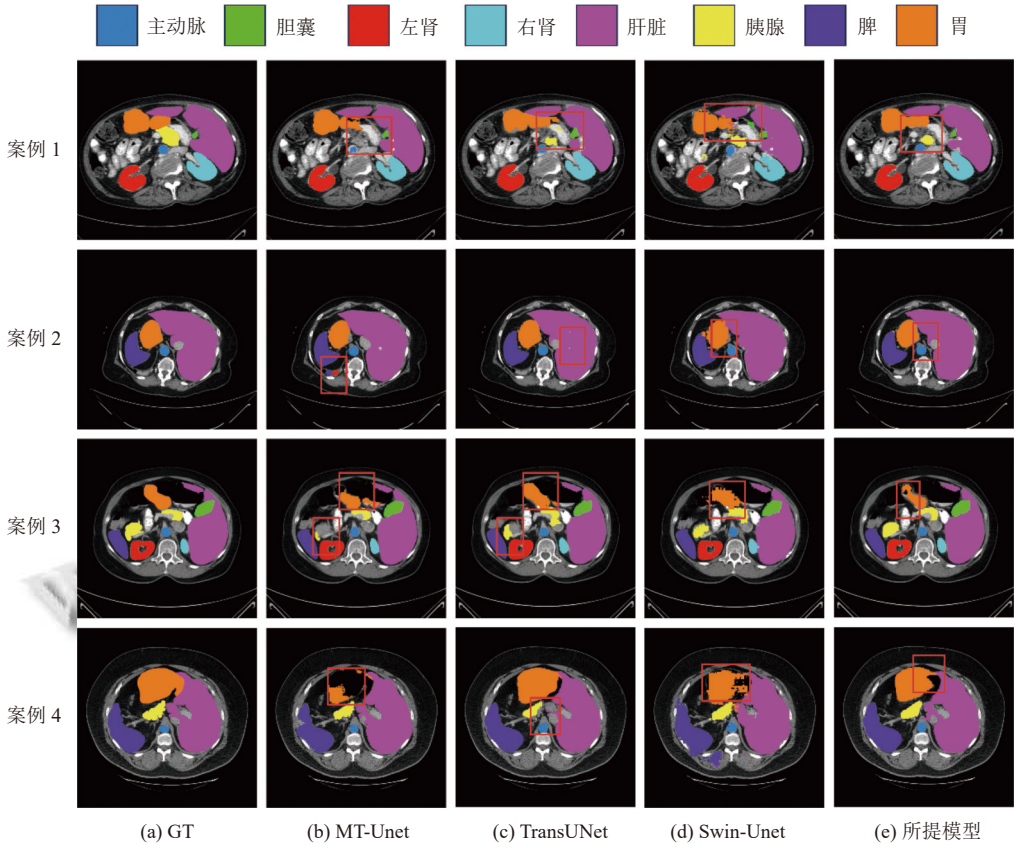

图 4 Synapse 数据集上可视化结果比较

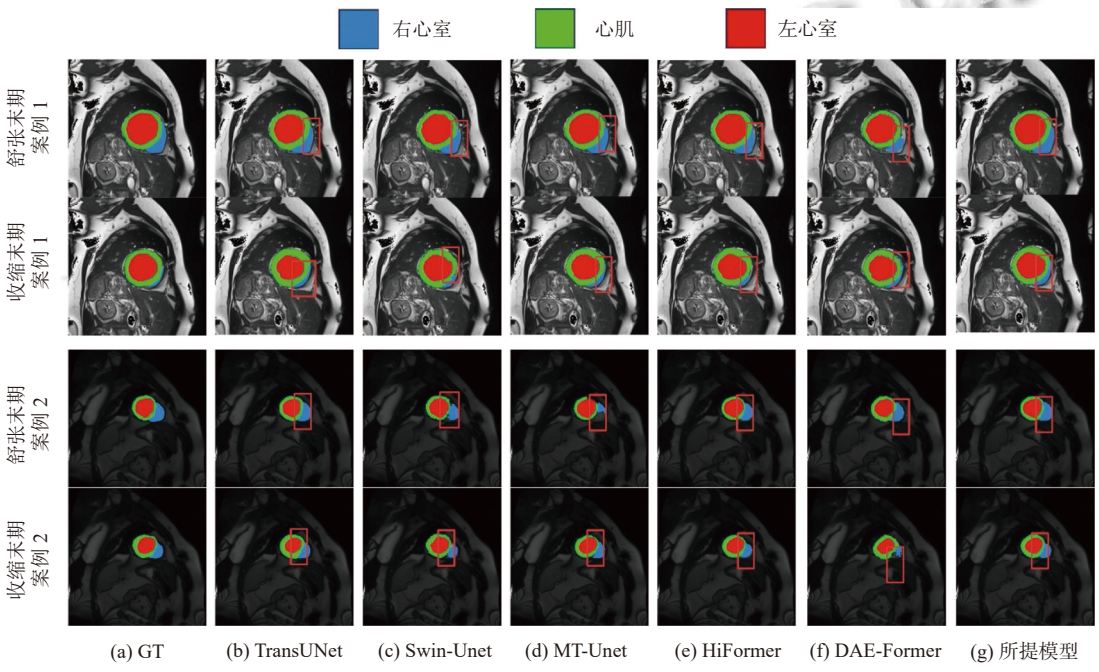

图 5 ACDC 数据集上的可视化结果比较

3.5 消融实验

为了验证所设计的 MLFF 模块和 GLFE 模块的有效性, 在 Synapse 和 ACDC 数据集上对所提模型中的 MLFF 模块和 GLFE 模块进行消融实验.

GLFE 模块对模型的分割是有贡献的. 在模型训练相同条件下, 实验结果如表 3 和表 4 所示. 表 3 中, 在使用 GLFE 模块比不用 GLFE 模块导致平均 DSC 系数增加 1.71%, 平均 HD95 系数下降 9.01 mm. 表 4 中, 添加 GLFE 模块比不添加 GLFE 模块使左心室, 心肌和右心室在 DSC 系数上分别提升 1.57%, 0.39% 和 0.01%, 总体提升 0.66%. 融合后的局部特征和全局特征, 通过 GLFE 模块, 帮助网络在局部特征和全局特征关注更为重要的信息, 从而避免信息的冗余和丢失. 实验结果表明, GLFE 模块在平均 DSC 和 HD95 系数上均有提升, 可以提高模型对目标边缘区域的分割性能, 验证了 GLFE 模块对模型是有贡献的.

表 3 GLFE 模块在 Synapse 数据集上的影响

GLFE	MLFF	平均	
		DSC (%)↑	HD95 (mm)↓
×	√	79.59	28.21
√	√	80.17	19.20

表 4 GLFE 模块在 ACDC 数据集上的影响 (%)

GLFE	MLFF	平均DSC ↑	左心室	心肌	右心室
×	√	90.41	87.29	88.35	95.60
√	√	91.07	88.86	88.74	95.61

MLFF 模块对模型的分割也是有贡献的. 在模型训练相同条件下, 实验结果如表 5 和表 6 所示. 表 5 中, MLFF 模块在平均 DSC 系数上提升 2.67%, 在 HD95 系数上降低 8.53 mm. 表 6 中, 添加 MLFF 模块相较于不添加 MLFF 模块在平均 DSC 系数上, 使左心室提升 2.37%, 心肌提升 2.81%, 右心室提升 1.32%, 总体提升 2.17%. MLFF 模块通过将 3 个不同层次的特征表示进行关联, 从而提高模型的分割性能. 实验结果表明, 所设计的 MLFF 模块在两个数据集上对平均 DSC 和 HD95 系数均有提升, 验证了 MLFF 模块对模型是有贡献的.

表 5 MLFF 模块在 Synapse 数据集上的影响

GLFE	MLFF	平均	
		DSC (%)↑	HD95 (mm)↓
√	×	77.50	27.73
√	√	80.17	19.20

表 6 MLFF 模块在 ACDC 数据集上的影响 (%)

GLFE	MLFF	平均DSC ↑	左心室	心肌	右心室
√	×	88.90	86.49	85.93	94.29
√	√	91.07	88.86	88.74	95.61

3.6 计算复杂度分析

所提模型与不同医学图像分割模型参数量的比较如表 7 所示. 提出的模型在平均 DSC 系数上优于其他模型, 在参数量上优于大部分模型. 实验结果表明, 与基于 CNN 和基于 Transformer 的模型相比, 所提模型不论是在评估指标或模型计算力和参数量上, 具有一定的竞争优势.

表 7 模型计算力和参数比较

模型网络	参数量 (M)	GFLOPs	平均	
			DSC (%)↑	HD95 (mm)↓
TransUNet[18]	110.15	24.73	77.48	31.69
Swin-Unet[25]	41.34	8.73	79.13	21.55
MT-Unet[27]	79.08	44.80	78.28	32.07
所提模型	35.94	12.98	80.17	19.20

4 结论与展望

针对多器官医学图像分割任务, 提出一种多级特征交互融合 Transformer 模型. 所提模型通过 CNN 和 Swin Transformer 分别提取局部特征和全局特征; 通过 GLFE 模块在降低卷积计算的参数量的同时捕捉融合后局部特征和全局特征的长期依赖关系和精确的位置信息; 通过 MLFF 模块捕获输入特征映射的空间重要性和关联不同级特征映射之间的语义内容, 从而得到精细化融合后的特征. 实验结果表明, 所提模型在 Synapse 和 ACDC 数据集上的分割性能优于基于 Transformer 和基于 CNN 的模型. 所提模型有助于提高医生的工作效率, 但模型的泛化能力仍待提高.

参考文献

1 Bao H, Zhu YQ, Li Q. Hybrid-scale contextual fusion network for medical image segmentation. Computers in Biology and Medicine, 2023, 152: 106439. [doi: 10.1016/j.compbiomed.2022.106439]

2 Dhamija T, Gupta A, Gupta S, *et al*. Semantic segmentation in medical images through transfused convolution and Transformer networks. Applied Intelligence, 2023, 53(1): 1132–1148. [doi: 10.1007/s10489-022-03642-w]

3 Zhang YC, Jiao RS, Liao QC, *et al*. Uncertainty-guided mutual consistency learning for semi-supervised medical

image segmentation. Artificial Intelligence in Medicine, 2023, 138: 102476. [doi: 10.1016/j.artmed.2022.102476]

4 徐光宪, 冯春, 马飞. 基于 UNet 的医学图像分割综述. 计算机科学与探索, 2023, 17(8): 1776–1792. [doi: 10.3778/j.issn.1673-9418.2301044]

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

6 Zhou ZW, Rahman Siddiquee MM, Tajbakhsh N, *et al*. UNet++: A nested U-Net architecture for medical image segmentation. Proceedings of the 4th International Workshop on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Granada: Springer, 2018. 3–11. [doi: 10.1007/978-3-030-00889-5_1]

7 Huang HM, Lin LF, Tong RF, *et al*. UNet 3+: A full-scale connected UNet for medical image segmentation. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 1055–1059.

8 Ranjbarzadeh R, Tataei Sarshar N, Jafarzadeh Ghoushchi S, *et al*. MRFE-CNN: Multi-route feature extraction model for breast tumor segmentation in Mammograms using a convolutional neural network. Annals of Operations Research, 2023, 328(1): 1021–1042. [doi: 10.1007/s10479-022-04755-8]

9 Nirthika R, Manivannan S, Ramanan A, *et al*. Pooling in convolutional neural networks for medical image analysis: A survey and an empirical study. Neural Computing and Applications, 2022, 34(7): 5321–5347. [doi: 10.1007/s00521-022-06953-8]

10 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.

11 Jiang Y, Liang J, Cheng TT, *et al*. MTPA_Unet: Multi-scale Transformer-position attention retinal vessel segmentation network joint Transformer and CNN. Sensors, 2022, 22(12): 4592. [doi: 10.3390/s22124592]

12 Wu YL, Wang GL, Wang ZY, *et al*. DI-Unet: Dimensional interaction self-attention for medical image segmentation. Biomedical Signal Processing and Control, 2022, 78: 103896. [doi: 10.1016/j.bspc.2022.103896]

13 Wang WH, Xie EZ, Li X, *et al*. Pyramid vision Transformer: A versatile backbone for dense prediction without convolutions. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 548–558.

14 石磊, 籍庆余, 陈清威, 等. 视觉 Transformer 在医学图像分析中的应用研究综述. 计算机工程与应用, 2023, 59(8): 41–55. [doi: 10.3778/j.issn.1002-8331.2206-0022]

15 傅励瑶, 尹梦晓, 杨锋. 基于 Transformer 的 U 型医学图像分割网络综述. 计算机应用, 2023, 43(5): 1584–1595.

16 Azad R, Arimond R, Aghdam EK, *et al*. Dae-former: Dual attention-guided efficient Transformer for medical image segmentation. Proceedings of the 6th International Workshop on Predictive Intelligence in Medicine. Vancouver: Springer, 2023. 83–95.

17 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10002.

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

19 Shen ZR, Zhang MY, Zhao HY, *et al*. Efficient attention: Attention with linear complexities. Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021. 3530–3538.

20 Lin HZ, Cheng X, Wu XY, *et al*. CAT: Cross attention in vision Transformer. Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME). Taipei: IEEE, 2022. 1–6.

21 Liu ST, Huang D. Receptive field block net for accurate and fast object detection. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 404–419.

22 Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. Proceedings of the 4th International Conference on 3D Vision (3DV). Stanford: IEEE, 2016. 565–571.

23 Fu SH, Lu YY, Wang Y, *et al*. Domain adaptive relational reasoning for 3D multi-organ segmentation. Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention. Lima: Springer, 2020. 656–666.

24 Schlemper J, Oktay O, Schaap M, *et al*. Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis, 2019, 53: 197–207. [doi:

10.1016/j.media.2019.01.012]

25 Cao H, Wang YY, Chen J, *et al*. Swin-Unet: Unet-like pure Transformer for medical image segmentation. Proceedings of the 2023 European Conference on Computer Vision. Tel Aviv: Springer, 2023. 205–218.

26 Yao C, Hu MH, Li QL, *et al*. TransClaw U-Net: Claw U-Net with Transformers for medical image segmentation. Proceedings of the 5th International Conference on Information Communication and Signal Processing (ICICSP). Shenzhen: IEEE, 2022. 280–284.

27 Wang HY, Xie SA, Lin LF, *et al*. Mixed Transformer U-Net for medical image segmentation. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022. 2390–2394.

28 Yuan FN, Zhang ZX, Fang ZJ. An effective CNN and Transformer complementary network for medical image segmentation. Pattern Recognition, 2023, 136: 109228. [doi: 10.1016/j.patcog.2022.109228]

29 Azad R, Heidari M, Shariatnia M, *et al*. TransDeepLab: Convolution-free Transformer-based DeepLab v3+ for medical image segmentation. Proceedings of the 5th International Workshop on Predictive Intelligence in Medicine. Singapore: Springer, 2022. 91–102.

30 Zhong XX, Xu LH, Li CQ, *et al*. RFE-UNet: Remote feature exploration with local learning for medical image segmentation. Sensors, 2023, 23(13): 6228. [doi: 10.3390/s23136228]

31 Xu GP, Zhang X, Liao WT, *et al*. LGNet: Local and global representation learning for fast biomedical image segmentation. Journal of Innovative Optical Health Sciences, 2023, 16(4): 2243001. [doi: 10.1142/S1793545822430015]

32 Heidari M, Kazerouni A, Soltany M, *et al*. HiFormer: Hierarchical multi-scale representations using Transformers for medical image segmentation. Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2023. 6191–6201.

33 Liu YH, Wang H, Chen ZG, *et al*. TransUNet+: Redesigning the skip connection to enhance features in medical image segmentation. Knowledge-based Systems, 2022, 256: 109859. [doi: 10.1016/j.knosys.2022.109859]

(校对责编: 孙君艳)