计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(5):64−76 [doi: 10.15888/j.cnki.csa.009852] [CSTR: 32024.14.csa.009852]

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

多注意力融合的 TransUNet 医学影像分割模型①

赵　亮, 赵雨祺, 金海波

(辽宁工程技术大学 软件学院, 葫芦岛 125105)
通信作者: 赵雨祺, E-mail: 18341854479@163.com

摘　要: 精确识别组织器官和病变区域是医学影像分析中最重要的任务之一. 在现有的医学影像语义分割研究中, 基于 U-Net 结构的模型占据了主导地位. TransUNet 结合了 CNN 和 Transformer 的优势, 弥补了两者在捕捉长程依赖和提取局部特征方面的不足, 但在提取和复原特征的位置时仍不够准确. 针对此问题, 提出了一种多注意力融合机制的医学影像分割模型 MAF-TransUNet. 该模型首先在 Transformer 层之前增加一个多注意力融合模块 (MAF) 来增强位置信息的表达; 然后在跳跃连接中再次结合多注意模块 (MAF) 使位置信息能够有效地传递到解码器一侧; 最后在解码阶段使用深度卷积注意力模块 (DCA) 保留更多的空间信息. 实验结果显示, MAF-TransUNet 相较 TransUNet 在 Synapse 多器官分割数据集和 ACDC 自动心脏诊断数据集上的 *Dice* 系数分别提升了 3.54% 和 0.88%.
关键词: 医学影像分割; Transformer; TransUNet; 注意力机制

引用格式: 赵亮,赵雨祺,金海波.多注意力融合的 TransUNet 医学影像分割模型.计算机系统应用,2025,34(5):64–76. http://www.c-s-a.org.cn/1003-3254/9852.html

TransUNet Medical Image Segmentation Model with Multi-attention Fusion

ZHAO Liang, ZHAO Yu-Qi, JIN Hai-Bo

(Software College, Liaoning Technical University, Huludao 125105, China)

Abstract: Accurate identification of tissues, organs, and lesion regions is one of the most important tasks in medical image analysis. Models based on the U-Net structure dominate the existing research on semantic segmentation of medical images. Combining the advantages of CNN and Transformer, TransUNet has superiority in capturing long-range dependencies and extracting local features, but it is still not accurate enough in extracting and recovering the locations of features. To address this problem, a medical image segmentation model MAF-TransUNet with a multi-attention fusion mechanism is proposed. The model first adds a multi-attention fusion module (MAF) before the Transformer layer to enhance the representation of location information. Then it combines the MAF again in the hopping connection so that the location information can be efficiently transmitted to the decoder side. Finally, the deep convolutional attention module (DCA) is used in the decoding stage to retain more spatial information. The experimental results show that MAF-TransUNet improves the *Dice* coefficients on the Synapse multi-organ segmentation dataset and ACDC automated cardiac diagnostic dataset by 3.54% and 0.88%, respectively, compared with TransUNet.
Key words: medical image segmentation; Transformer; TransUNet; attention mechanism

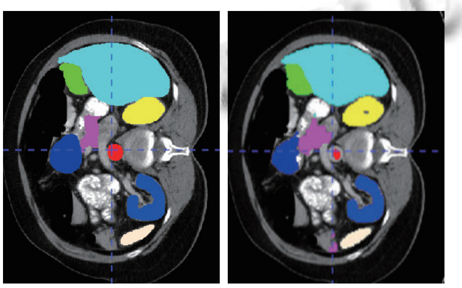
随着民众对医疗服务需求的不断提升, 医生在诊断过程中所需分析的医学影像数量也日益增多, 这不仅加重了医生的工作负担, 还在一定程度上增加了误诊的风险. 基于人工智能的医学影像分析能够迅速完

① 基金项目: 国家自然科学基金 (62173171)
收稿时间: 2024-10-22; 修改时间: 2024-11-12; 采用时间: 2024-12-11; csa 在线出版时间: 2025-03-24
CNKI 网络首发时间: 2025-03-25

成对医学影像的初步处理, 为医生的诊断提供辅助信息, 不但可以提高医生的工作效率, 还能间接提升诊断质量. 因此, 医学影像分析技术的进步对于提高医疗质量有积极的促进作用.

TransUNet 结合了 U-Net[1]和 Transformer[2]的优点, 在医学影像分割领域取得了较好的效果, 但仍存在一些问题: 首先, TransUNet 将提取图像局部特征有优势的 CNN 与提取全局信息特征有优势的 Transformer 相结合, 起到了互补的作用, 但在对位置信息尤其重要的分割问题上, 并没有针对位置信息的提取进行加强; 其次, 在 U-Net 模型中, 尽管上采样和下采样之间使用跳跃连接降低了信息的损耗, 但也没有针对空间信息的增强; 最后, 虽然 U 型结构在提取特征和保留全局信息方面非常强大, 但在上采样过程中仍然会有位置信息损失的结构性问题, 需要有针对性地进行弥补. 综上所述, 在图像分割过程中减少特征信息的损失, 尤其是减少在上下采样过程中损失的位置信息, 能有效提高医学图像的分割精度.

图 1 所示是 Synapse 多器官分割数据集中图像的分割结果, 图 1(a) 为人工标注的结果, 图 1(b) 为模型预测的结果. 显而易见, TransUNet 在关键特征的辨认和位置定位方面与人工标注仍存在差距. 具体来说, 导致模型的预测结果在定位方面存在偏差的原因可能有两种: 一是特征提取不准确. 对像素所属的器官做了错误的分类; 二是特征定位不准确, 在解码器上采样的过程中, 对特征位置的还原产生了偏差. 这种偏差会影响模型的分割精度, 特别是在需要高精度特征位置识别的任务中, 这一问题显得尤为突出.



(a) Ground truth (b) TransUNet

图 1 深度学习结果与人工结果的比较

针对以上问题, 本文提出了一种多注意力融合的 TransUNet 医学影像分割模型 (TransUNet medical image segmentation model with multi-attention fusion, MAF-TransUNet). 在 TransUNet 模型的编码阶段、解码阶段及跳跃连接中引入新模块, 以提高特征位置信息在各个阶段的准确性.

本文的主要工作如下.

1) 提出了一种多注意力融合编码机制, 运用位置注意力模块和通道-空间注意力混合模块来连接 CNN 和 Transformer, 增强其位置信息特征提取能力.

2) 在跳跃连接中引入多注意力融合模块, 在信息传递的过程中对位置信息做有针对性地加强.

3) 在解码阶段增加了深度卷积注意力模块, 该模块将深度卷积和卷积注意力进行结合, 能够在上采样过程中保留更多的空间信息.

将本文提出的方法在 Synapse 多器官分割数据集和 ACDC 自动心脏诊断数据集上进行验证, 实验结果显示, MAF-TransUNet 较其他基准模型性能均有一定提升. 其中在 Synapse 数据集上相较于 Swin-Unet 模型, 胰腺部位分割精度提高了 7.67%, 更准确地预测了多器官的边缘细节. 在 ACDC 数据集上相较于 TransUNet 模型, 心肌部位分割精度增加了 3.66%.

1 相关工作

2015 年, Ronneberger 等人[3]提出一种 U 型结构用于医学影像分割; 2016 年, Diakogiannis 等人[4]使用 ResNet 网络残差模块[5]代替 U-Net 中的普通卷积, 不仅提高了分割效果, 还降低了 1/4 的参数量; Zhou 等人[6]提出的 UNet++由深度不同的 U-Net 组成, 在解码器中引入更密集的跳跃连接能更有效地整合多层次特征信息, 并提升模型对图像细节和结构的感知能力; Li 等人[7]将密集连接与 U-Net 结合, 提升了特征的重用性和表征能力, 有助于加强网络泛化和小目标检测能力; Alom 等人[8]通过结合 U-Net、残差网络和 R-CNN (region-based convolutional neural network) 的优势, 提升了模型捕获多尺度信息的能力; Valanarasu 等人[9]使用一个过完备的卷积架构将输入图像投影到一个更高的维度, 限制感受野在网络深层增加, 从而改善模型对细节的捕捉能力; Huang 等人[10]利用全尺度的跳跃连接把来自不同尺度特征中的高级语义与低级语义直接结合, 利用深度监督学习来增强多尺度特征的层次表示. 上述网络模型都在一定程度上提高了图像分割的精度, 但都仅使用了 CNN 来提取特征, 虽然 CNN 可以通过卷积层和下采样层的堆叠来扩大感受野并增强局部信息交互, 但这种方式只能部分弥补卷积核在提取全局

信息能力上的不足, 而且会增加模型复杂度导致模型过拟合[11].

一些工作[12,13]尝试利用注意力机制等方法来建模长程依赖. Wang 等人[14]设计了一个非局部算子, 通过在多个中间卷积层中插入非局部算子可以扩展模型对于全局信息的感知范围, 从而提升模型的表征能力和泛化性能; Schlemper 等人[15]在编码器-解码器网络中引入注意力门控机制, 使得网络能更好地控制信息的传递和整合, 提高了网络的学习能力、表征能力和鲁棒性. 有部分研究人员尝试使用 Transformer 嵌入全局自注意力. Transformer 最初由 Vaswani 等人[16]提出并用于机器翻译, 目前已成为许多自然语言处理 (NLP) 任务的主流方法. 为了使 Transformer 适用于计算机视觉任务, Parmar 等人[17]只在每个查询像素的局部邻域中应用自注意力, 而非全局应用, 这能够在保证模型性能的同时提升计算效率和内存使用效率; Child 等人[18]提出稀疏变换, 采用可伸缩的近似值来实现全局自注意力, 降低了计算成本和内存消耗, 使模型更具有可扩展性和高效性; Liu 等人[19]提出的 Swin Transformer 既保留了局部自注意力的计算效率, 又通过移位窗口机制捕获了全局特征; Carion 等人[20]提出的 DETR 模型利用 Transformer 设计构建了首个完全端到端的目标检测模型, 为后续研究和应用提供了新思路; Lin 等人[21]在跳跃连接中加入时序信息融合模块 (temporal information fusion, TIF), 通过自注意机制有效建立不同尺度特征间的全局依赖关系来获取多尺度特征. 以上注意力机制主要关注跨区域特征之间的相互联系, 但在一定程度上可能忽视了局部区域的独立特征. 为确保局部特征的完整性, 还需设计专门的位置信息整合模块, 来增强模型对位置信息特征的捕捉和利用.

Chen 等人[22]结合了 Transformer 和 U-Net 在医学图像分割领域的优势, 提出 TransUNet 模型. 一方面, TransUNet 将 Transformer 和 CNN 结合, 提升了获取全局信息的能力. 另一方面, 解码器对编码特征进行上采样, 后与高分辨率 CNN 特征结合, 提高了特征的定位精度; Cao 等人[23]用 Swin-Transformer 块替换 U-Net 中的每个卷积模块, 提高了注意力机制对图像信息的处理能力; Yang 等人[24]提出的 AA-TransUNet 利用卷积块注意力模型 (convolutional block attention module, CBAM)[25]和深度可分离卷积 (depthwise separable convolution, DSC)[26]进一步优化 TransUNet, 使上采样过程中保留了更多的空间信息; Sun 等人[27]在解码器和编码器中使用双注意力块 (dual attention block, DA-Block), 充分利用全局和局部特征提高对医学图像的分割性能. 因此, 基于卷积的 U-Net 虽然在局部细节处理上表现出色, 但全局建模能力相对不足. 而引入自注意力机制的 U-Net, 通过增强全局信息的捕捉与特征表达能力, 能更精准地进行特征识别和位置定位.

Wu 等人[28]通过改进的相对位置编码提升了 ViT 的性能和计算效率; Gao 等人[29]提出了一种基于 U-Net 的增强特征提取网络, 运用位置编码和交叉注意力机制, 精细识别每个跳跃连接的特征, 从而在解码过程中减少特征噪声, 提高边界的位置精度; Su 等人[30]将旋转位置编码结合到 Transformer 模型中, 旨在更有效地整合位置信息, 以提高模型对长文本的处理能力.

2 方法

2.1 网络结构概述

本文提出一种应用在医学图像分割领域的 MAF-TransUNet 网络模型, 如图 2(a). MAF-TransUNet 由编码器、跳跃连接和解码器 3 部分构成. 编码阶段首先使用 CNN 和多注意力融合模块 (multi-attention fusion, MAF-Block) 来提取图像特征. 随后, 这些特征被送入 Transformer 结构, 以便学习图像的全局特征并完成重建. 跳跃连接部分使用多注意力融合模块来增强特征提取能力. 解码阶段由新构建的深度卷积注意力模块 (deep convolutional attention, DCA) 与双线性上采样模块组成, 用来恢复更精细的位置信息. MAF-TransUNet 通过整合传统卷积、Transformer 层、深度卷积注意力模块以及多注意力融合模块来实现增强图像分割性能的目标.

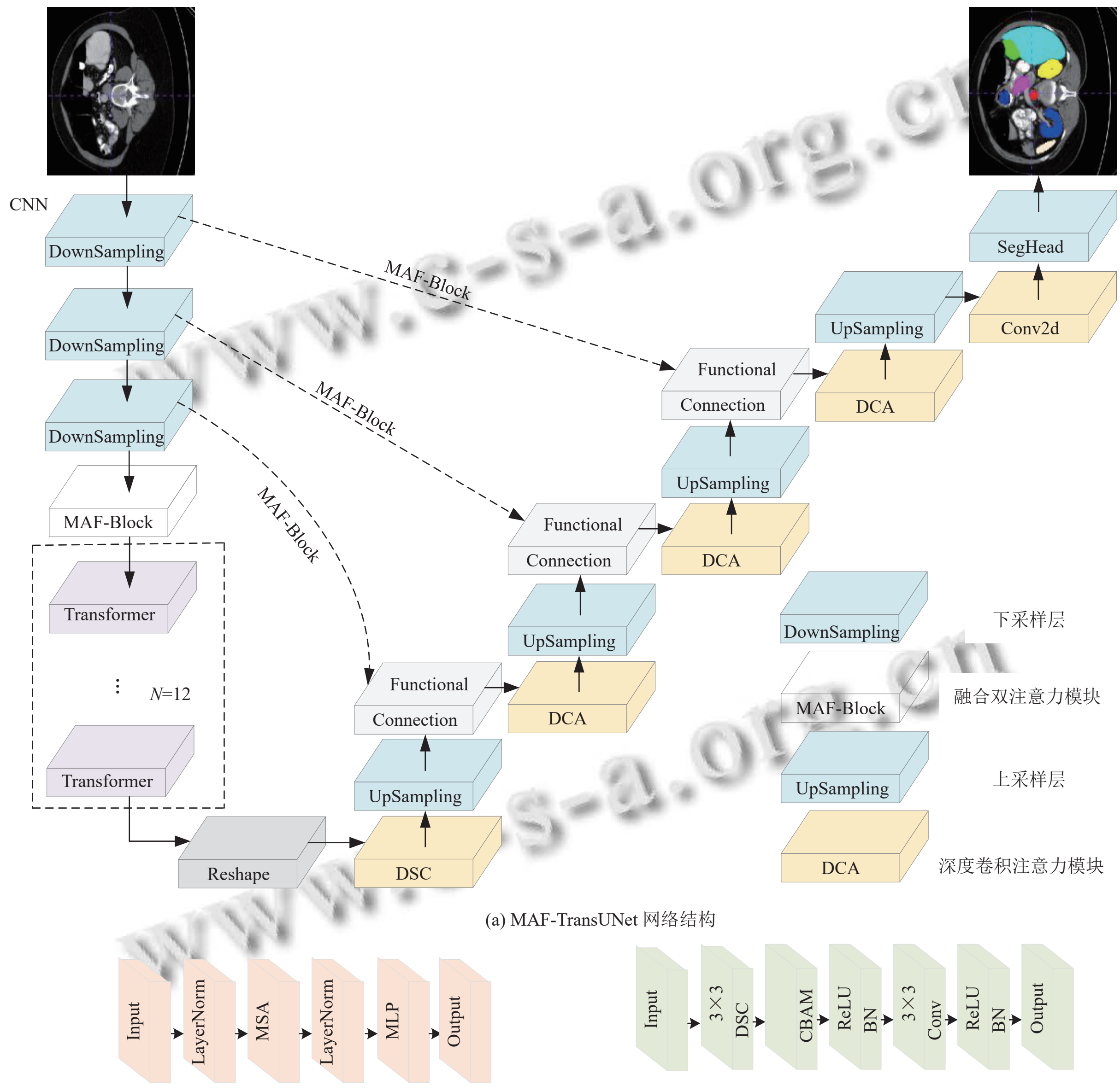
2.2 模型编码

编码器由 CNN 层、MAF-Block 和 Transformer 层组成. 首先, 输入的图片进入到 CNN 层进行局部特征提取, 先按照顺序进入到 CNN 中, 此过程中图像维度依次为 (256, 56, 56)、(512, 28, 28)、(1024, 14, 14). 然后, 通过 MAF-Block 产生一个特征图作为输出, 接着对输出进行图像块嵌入, 最后将嵌入后的一维向量输入到 12 层 Transformer 结构中 (如图 2(b)). 其中 Transformer 由多头自注意力 (multi-head self-attention, MSA) 层和多层感知器 (multilayer perceptron, MLP) 层组成,

每个 MSA 和 MLP 都通过归一化层来保持模型稳定性和性能.

2.2.1 多注意力融合模块

尽管 Transformer 在特征提取方面表现出色, 但作为处理文字信息的结构, 它忽视了部分通道和空间的特征信息. 为更有效地提取和复原图像特征, 本文在编码器和跳跃连接中引入多注意力融合模块 (MAF-Block), 该模块将 DA-Block 中通道注意力替换为 CBAM 模块. CBAM 包括通道注意力模块和空间注意力模块, 搭配位置注意力模块 (position attention block, PAM) 使用, 同时关注通道和空间特征, 不仅能够增强位置信息的表达, 还能获取更全面、更精确的边缘特征集合.



(a) MAF-TransUNet 网络结构

(b) Transformer

(c) 深度卷积注意力模块 (DCA)

图 2 MAF-TransUNet 网络结构

如图 3 所示, MAF-Block 由两个主要部分组成: 一个是 PAM 模块, 专门用于捕获输入特征中的位置信息; 另一个是 CBAM 模块, 主要提取特征中通道和空间的信息. 因此, 本文将 MAF-Block 集成到编码器和跳跃连接中, 来获取更准确的位置信息特征, 以增强模型的分割能力. MAF-Block 把相同的特征用不同的特征提取模块进行处理, 将这两种不同类型的输出特征结合后输出结果.

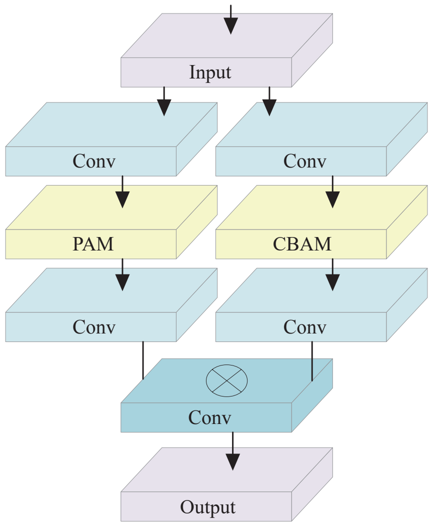


图 3　多注意力融合模块

通过卷积将输入特征的通道数减少为原来的 1/16, 以生成γ^1, 这简化了 *PAM* 的特征提取; 通过 *PAM* 特征提取及卷积操作, 得到$\gamma^{1'}$.

$$\gamma^1 = Conv(input) \tag{1}$$

$$\gamma^{1'} = Conv(PAM(\gamma^1)) \tag{2}$$

如图 4 所示, *PAM* 可以捕获特征图中任意两个位置之间的空间依赖关系, 通过所有位置特征的加权和来更新指定特征. 因此, *PAM* 具有较强的空间特征提取能力. *PAM* 中权重由两个位置之间的特征相似性确定, 输入采用局部特征, 表示为$A \in R^{C\times H\times W}$($C$表示通道, H表示高度, W表示宽度). 最初将A送入卷积层, 产生 3 个新的特征图, 即B、C和D, 每个特征图的大小为$R^{C\times H\times W}$.

接下来, 将B和C整形为$R^{C\times N}$, 其中$N = H\times W$表示像素的数量. 在C和B的转置之间执行矩阵乘法, 然后使用 Softmax 层来计算空间注意力图$F \in R^{N\times N}$:

$$F_{cb} = \frac{\mathrm{e}^{(B_b\times C_c)}}{\sum\limits_{b=1}^{N}\mathrm{e}^{(B_b\times C_c)}} \tag{3}$$

其中, F_{cb}测量第b个位置对第c个位置的影响. 将矩阵D重建为$R^{C\times N}$. 在D和F的转置之间执行矩阵乘法并将结果重建, 设为$R^{C\times H\times W}$. 最后, 将其乘以参数α, 并对特征A执行元素求和运算, 以获得最终输出$E_c \in R^{C\times H\times W}$:

$$E_c = \alpha\sum_{b=1}^{N}(F_{cb}D_b) + A_c \tag{4}$$

E_c是所有位置特征和原始特征的加权和, 因此它能够在空间注意力图基础上聚合上下文信息. 既确保了位置信息特征的有效提取, 又保留了更多的全局上下文信息.

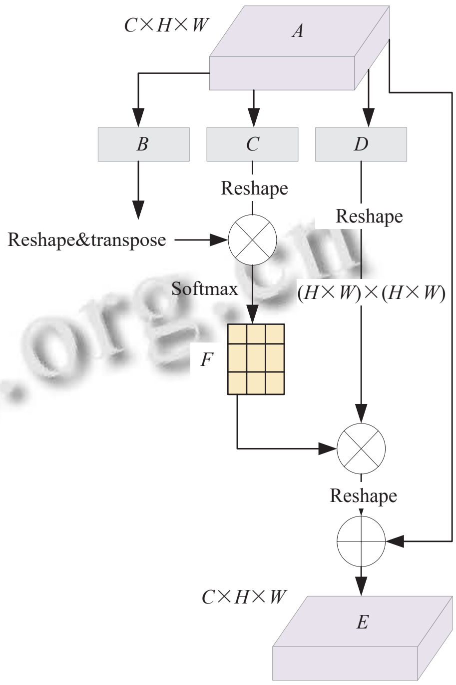


图 4　PAM 结构图

CBAM 模块是一种前馈卷积神经网络的注意力模块, 包含两个主要部分: 通道注意力模块和空间注意力模块. 这两个模块协同工作, 来保留更多的空间位置信息特征. 通道注意力模块负责分析不同通道之间的依赖关系, 以便更好地捕捉图像的语义信息. 空间注意力模块则关注特征图上不同位置间的关系, 从而更有效地捕捉图像的局部细节. 具体流程如图 5 所示. CBAM 生成注意力过程可描述为: 输入特征$A \in R^{C\times H\times W}$, 首先按通道进行全局最大池化和平均池化, 将池化后结果送入共享全连接层得到 Maxout 及 Avgout 后相加, 使用Sigmoid生成通道注意力$M_c \in R^{C\times 1\times 1}$, 再将通道注意力与输入元素相乘获得特征图$A'$; 其次将$A'$按照空间进行全局最大池化和平均池化, 将池化结果拼接进行卷积操作, 后通过 Sigmoid 生成空间注意力$M_s \in R^{1\times H\times W}$, 最终将其与$A'$按元素相乘输出特征图$A''$.

$$A' = M_c(A)\times A' \tag{5}$$

$$A'' = M_s(A')\times A' \tag{6}$$

2.3　模型解码

解码器先采用深度可分离卷积、级联上采样、深度卷积注意力这 3 个阶段产生维度为 (64, 112, 112) 的

特征图. 之后, 通过双线性上采样及 3×3 卷积、归一化和 ReLU, 将特征图尺寸放大至 (16, 224, 224). 最后, 利用分割头 (SegHead) 处理特征图, 得到最终分割结果.

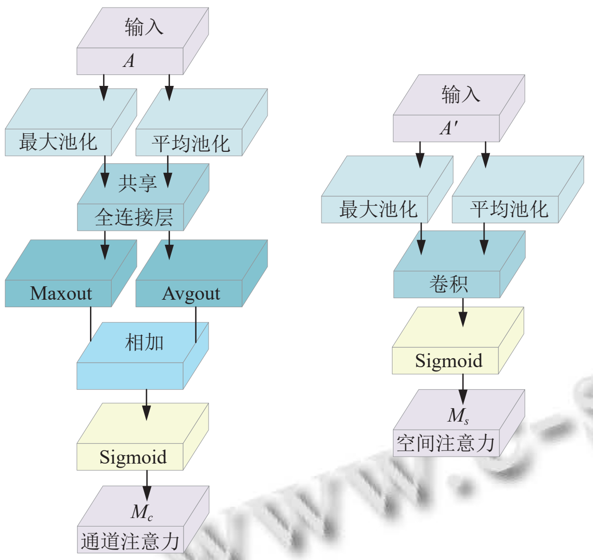


图 5　CBAM 中通道注意力和空间注意力结构图

2.3.1　深度卷积注意力模块

输入特征在深度卷积注意力模块 (DCA) 内依次进行以下操作, 如图 2(c) 所示, 深度可分离卷积、CBAM、ReLU、BatchNorm、普通卷积、ReLU 和 BatchNorm. DCA 中两个关键元素: 深度可分离卷积和 CBAM. 受 AA-TransUNet 启示, 两者搭配使用构建的 DCA 模块不仅提升了模型的计算效率, 还增强了对于位置信息特征的提取, 使模型处理医学图像时能够提供更准确的分割结果.

深度可分离卷积如图 6 所示, 该过程可以被看作两个步骤的组合: 先进行逐通道卷积, 后进行逐点卷积. 在逐通道卷积中, 每个卷积操作只处理一个通道, 也就是说, 输入数据的一个通道只被一个卷积核处理, 这样处理后的特征通道数与输入通道数相同. 逐通道卷积操作减少了计算量, 但它在特征提取时无法利用不同通道间的相关性, 逐点卷积则解决了这个问题.

逐点卷积是一个 1×1 的卷积操作, 能够保留空间信息, 因为其在处理数据时不会改变数据的空间维度. 主要有以下几个特点.

1) 逐点卷积关注输入特征的每个单独点, 而不是像传统卷积那样处理多个邻近点. 这就意味着逐点卷积可以独立处理每个像素点, 而不会丢失其在原始空间中的位置信息.

2) 由于其卷积核大小为 1×1, 逐点卷积不会像较大卷积核那样降低特征的空间分辨率, 因此输出特征的空间尺寸与输入特征相同.

3) 逐点卷积可以在不同的通道之间进行混合, 而不会影响每个通道内的空间结构, 这使得网络能够在保留空间信息的同时学习跨通道的特征.

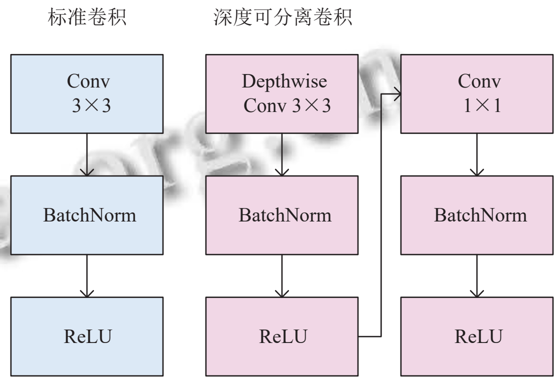


图 6　标准卷积和深度可分离卷积结构图

DCA 先通过深度可分离卷积提取和精简特征, 然后使用 CBAM 对特征进行加权, 进一步强化有效特征. 这种组合方式不仅提高了模型的计算效率, 还显著提高了特征位置信息提取的准确度. CBAM 在解码路径中被放置在深度可分离卷积后, 使得模型能够执行通道级和空间级注意力, 从而更好地探索特征在通道间和空间维度上的关系.

3　损失函数

本文将交叉熵损失函数和 *Dice* 损失函数结合应用在 MAF-TransUNet 模型训练中. 交叉熵损失函数主要运用于分类任务, 用于衡量预测的结果与人工标注之间的差异. 在计算机视觉领域, 它用来预测图像类别, 实现像素级的前景和背景分类. 在医学图像分割中, 它能帮助模型更准确地区分类别来减少误分割现象, 并提高模型的分割精度.

$$L_{\text{CrossEntropyLoss}} = -\sum_{x}(p(x_i)\ln q(x_i)) \tag{7}$$

Dice 损失函数常用于图像分割任务, 尤其在处理类别间样本不均衡时表现出色, 如医学图像分割中的病灶检测. 该损失函数基于 *Dice* 系数, 衡量预测的结果和人工标注之间的相似性, 从而帮助模型更好地保留目标区域的边缘信息.

$$L_{\text{DiceLoss}} = 1 - 2 \times \frac{\sum_i P_i G_i}{\sum_i P_i + \sum_i G_i} \tag{8}$$

其中, P_i表示第i个样本属于某一类别的概率, G_i表示第i个样本的人工标注. 交叉熵-*Dice* 组合损失函数定义如下所示, 其中λ_1和λ_2分别代表交叉熵损失函数和 *Dice* 损失函数的权重因子. 通过间隔为 0.1 的一系列消融实验, 对比分析实验结果中的各项评价指标, 发现最优权重为$\lambda_1 = \lambda_2 = 0.5$.

$$L_{\text{Loss}} = \lambda_1 \times L_{\text{CrossEntropyLoss}} + \lambda_2 \times L_{\text{DiceLoss}} \tag{9}$$

4 实验结果与分析

4.1 数据集

实验采用 Synapse[31]多器官分割数据集和 ACDC[32]自动心脏诊断数据集. Synapse 多器官分割数据集包含 30 例腹部电子计算机断层扫描 (computed tomography, CT), 共含有 3 779 张轴向对比增强的腹部临床 CT 图像, 每个 CT 扫描中含有 85–198 张 512×512 像素的切片, 其空间分辨率为 0.54×0.98×2.5–0.54×0.98×5.0 mm^3. Synapse 多器官分割数据集包含 8 个腹部器官分别为主动脉、胆囊、脾脏、左肾、右肾、肝脏、胰腺、脾脏和胃, 整个数据集被划分为训练集和测试集, 分别包含 18 个训练样本和 12 个测试样本.

ACDC 自动心脏诊断数据集包含 100 个心脏磁共振图像序列, 每个序列都包含收缩末期和舒张末期的图像. 这些图像捕捉了左心室、右心室和心肌等心脏结构的详细信息.

4.2 实验设置

所有实验均应用简单的数据扩充, 例如随机、旋转和翻转. 数据集的预处理分为 4 步: 将医学图像数据从原始格式转为多维数组格式; 把像素值限制在[−125, 275]范围内剪切图片; 归一化数据至[0, 1]; 从三维数据中提取二维切片. 对比基于纯 Transformer 的编码器, MAF-TransUNet 采用具有 12 层 Transformer 层结构. 对比混合编码器设计, MAF-TransUNet 结合了 ResNet-50 和 ViT. 本文所有的 Transformer 主干和 ResNet-50 (R50) 都经过了 ImageNet 的预训练. 输入分辨率和补丁大小设定为 224×224 和 16. 因此, 本文在级联上采样金字塔中连续级联 4 个 2 倍上采样块, 才能达到完整的分辨率. 模型采用交叉熵和 *Dice* 组合损失函数, 采用 SGD 优化器进行训练, 学习率为 0.01, 优化算法的最大迭代次数为 30000, 动量为 0.9, 权重衰减为 0.0001. Synapse 多器官分割数据集和 ACDC 自动心脏诊断数据集默认的批处理大小分别为 24 和 16, 神经网络训练的最大时期数分别为 150 和 100. 实验使用了 PyTorch 1.13.0+cu116 作为深度学习框架, Python 3.10.9 作为开发环境, 运行在 12th Gen Intel(R) Core(TM) i5-12400F 2.50 GHz 处理器上, GPU 选择 GeForce RTX 3090, 显存大小为 24 GB.

4.3 评价指标

本文使用平均 *Dice* 相似系数和平均 Hausdorff 距离作为评价指标. *Dice* 系数对分割的内部匹配度较为敏感, 主要衡量预测区域和真实区域的重叠程度. 而 Hausdorff 对分割结果的边界匹配度较为敏感, 特别关注了预测边界和真实边界之间的最大差异, 即使是少量的边界不匹配也可能导致较大的 Hausdorff 值.

Dice 系数一般用于计算两个样本的相似度, 值的范围 0–1, 越接近 1 分割效果越好, 越接近 0 分割效果越差. *P* 代表真实标注, *T* 代表预测结果.

$$Dice(P,T) = \frac{2|P \cap T|}{(|P| + |T|)} \tag{10}$$

Hausdorff 距离 (HD) 是两组点集之间相似程度的一种量度, 对分割出的边界比较敏感, 相似程度越高, 值越小. *P* 代表真实标注, *T* 代表预测结果.

$$\begin{aligned} d_{\text{H}}(P,T) &= \max|d_{PT}, d_{TP}| \\ &= \max\{\max_{p\in P}\min_{t\in T} d(p,t), \max_{t\in T}\min_{p\in P} d|p,t|\} \end{aligned} \tag{11}$$

4.4 对比实验

在 Synapse 多器官分割数据集上, 将所提出的分割方法与其他模型进行比较.

1) R50 U-Net[3]提出了一种 U 型结构来进行图像的语义分割.

2) R50-AttnUNet[14]在多个中间卷积层中插入非局部算子.

3) R50-ViT[33]将图像分割成多个小块并利用 Transformer 处理.

4) TransUNet[22]在 U-Net 模型基础上引入了混合编码器, 将 CNN 和 Transformer 相结合.

5) Swin-Unet[23]用 Swin-Transformer 块替换 U-Net 中每个卷积块.

6) AA-TransUNet[24]在 TransUNet 模型上配备了 CBAM 和 DSC.

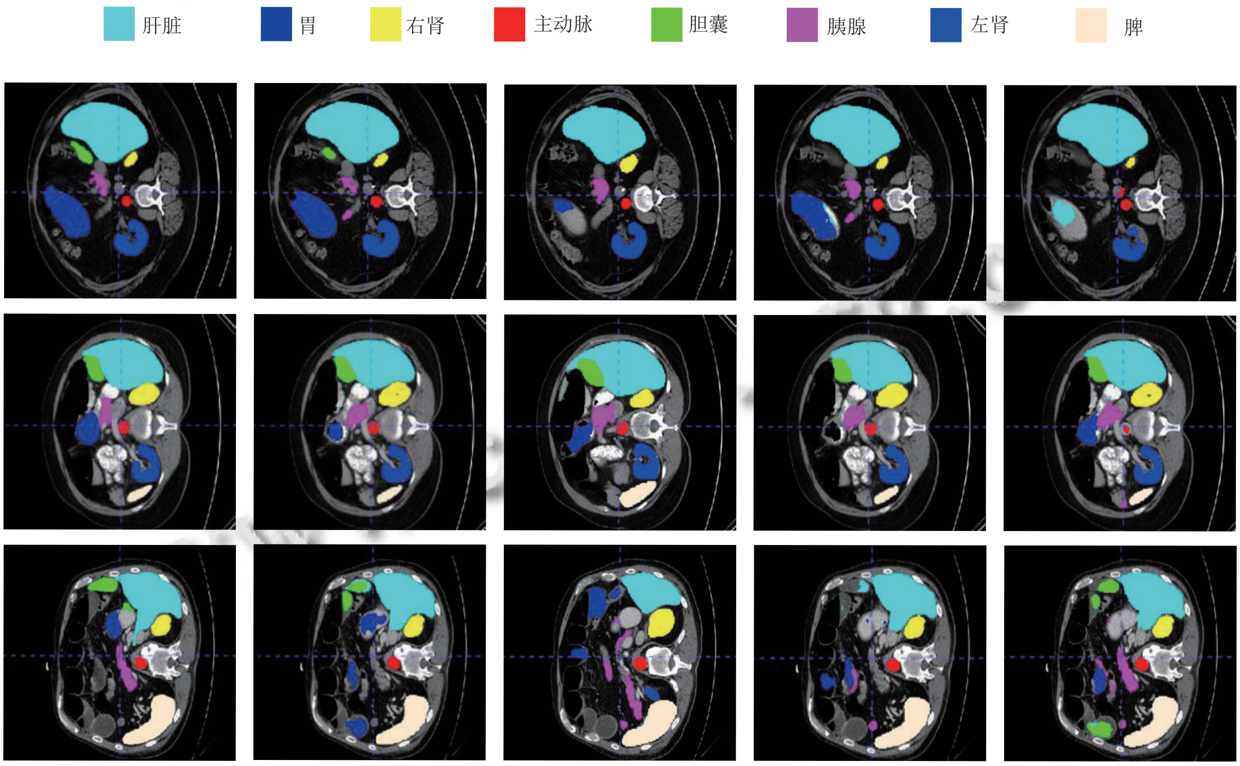
表 1 显示 MAF-TransUNet 相较于先前技术 *Dice* 相似系数提升范围为 1.89%–9.73% 不等. 将 Transformers 与 CNN 结合, 即 R50-ViT 产生比 U-Net 和 R50-AttnUNet 这种完全以 CNN 作为编码器差的结果. 当 U-Net 结构与跳跃连接相结合时, TransUNet 比 R50-ViT 和 R50-AttnUNet 分别提高了 6.19% 和 1.91%. AA-TransUNet 在 TransUNet 基础上配备 CBAM 和 DSC 反而降低了 1.23%, 说明该结构并不一定适用于医学图像分割. 在此基础上进行尝试得出的 MAF-TransUNet 相较应用 Transformer 的 Swin-Unet 则提高了 1.89%. MAF-TransUNet 在 TransUNe 基础上融合 MAF-Block、替换重建卷积及加入 DCA 后比原始数据提升了 3.54%. 展现了 MAF-TransUNet 具有较强的学习高级语义特征及低级细节的能力, 这在医学图像分割中至关重要. 对于平均 Hausdorff 距离也可以看到类似的趋势, 进一步证明了 MAF-TransUNet 相对于其他模型的优势.

表 1　在 Synapse 数据集上的比较 (%)

方法	*Dice*↑	HD↓	主动脉	胆囊	肾 (左)	肾 (右)	肝脏	胰腺	脾	胃
R50 U-Net[3]	76.85	39.70	85.54	64.77	77.77	68.60	93.43	53.98	86.67	75.58
R50-AttnUNet[14]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
R50-ViT[33]	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUNet[22]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet[23]	79.13	**21.55**	85.47	66.53	**83.28**	**79.61**	94.29	56.58	**90.66**	76.60
AA-TransUNet[24]	76.25	35.89	77.28	64.54	79.55	76.83	92.64	53.21	88.93	77.02
MAF-TransUNet	**81.02**	27.06	**88.22**	**67.73**	82.24	78.55	**94.62**	**64.25**	90..00	**82.53**

注: 最佳结果加粗表示

通过可视化对不同方法在 Synapse 多器官分割数据集上分割进行定性比较, 结果如图 7 所示.



(a) Ground truth　(b) MAF-TransUNet　(c) AA-TransUNet　(d) TransUNet　(e) R50-ViT

图 7　在 Synapse 数据集上对不同方法进行可视化

根据图 7 可得到结论: R50-ViT 和 AA-TransUNet 在分割器官时更容易出现过度或不足的情况 (例如在第 3 行中, R50-ViT 过度分割了胆囊, 而 AA-TransUNet 则欠分割). 相比而言, 基于 CNN 与 Transformer 的 U 型结构, 例如 AA-TransUNet 或 TransUNet, 在编码全局上下文及保留高级特征方面具有一定优势, 但未对不

同部位位置信息的提取进行有效增强.

MAF-TransUNet 相较于其他方法, 其预测结果中假阳性较少, 这表明它在抑制错误预测方面更为有效. 这意味着 MAF-TransUNet 对各部位位置信息提取更加准确, 在降低模型产生错误预测的能力上比其他方法更胜一筹.

从图 7 中还可以看出, 与 TransUNet 相比, MAF-TransUNet 在边界和形状方面的预测更为精细 (例如, 在第 1 行中胰腺的预测表现得更为准确). 在第 2 行中, MAF-TransUNet 更准确地预测了左右肾脏的边缘细节. 此外, MAF-TransUNet 检测到的范围明显超过了 TransUNet. 这些观察结果表明, MAF-TransUNet 保留更多特征位置信息的同时能够进行更精细的分割. 这得益于 MAF-TransUNet 能够同时利用高级全局上下文信息和低级细节的优势, 并通过多注意力融合机制准确提取特征位置信息.

在 ACDC 自动心脏诊断数据集上, 将所提分割方法与其他模型进行比较: 1) U-Net; 2) R50-AttnUNet; 3) R50-ViT; 4) TransUNet; 5) Swin-Unet; 6) AA-TransUNet. 表 2 显示所提出的 MAF-TransUNet 相对于先前技术提升了 0.59%–3.84%, 相较于 TransUNet 和 AA-TransUNet 分别提升了 0.88%、2.60%. 实验结果证明 MAF-TransUNet 在 ACDC 数据集上分割效果更好.

表 2 在 ACDC 数据集上的比较 (%)

方法	*Dice*↑	右心室	心肌	左心室
U-Net[3]	88.28	86.08	86.04	92.72
R50-AttnUNet[14]	86.75	87.58	79.20	93.47
R50-ViT[33]	87.57	86.07	81.88	94.75
TransUNet[22]	89.71	**88.86**	84.53	95.73
Swin-Unet[23]	90.00	88.55	85.62	**95.83**
AA-TransUNet[24]	87.99	87.83	82.46	93.68
MAF-TransUNet	**90.59**	88.34	**88.19**	95.23

注: 最佳结果加粗表示

如图 8 所示, 在 ACDC 数据集上对不同方法进行可视化分析, 其中 (a) 为人工标注的结果, (b)–(e) 分别是 MAF-TransUNet、AA-TransUNet、TransUNet、R50-ViT 的分割结果, 可以看出以上方法都能将器官的大概轮廓分割出. 但是, R50-ViT 右心室部位分割不足; TransUNet 心肌与左心室交界处分割不充分; AA-TransUNet 对各部位整体定位不够细致. 相较于其他模型, 本文提出的 MAF-TransUNet 通过 MAF-Block 和 DCA 模块将通道信息和空间信息相融合, 通过充分提取特征位置信息来增强对器官的整体定位及边缘分割的准确性, 使分割的结果与人工标注的结果更加相似.

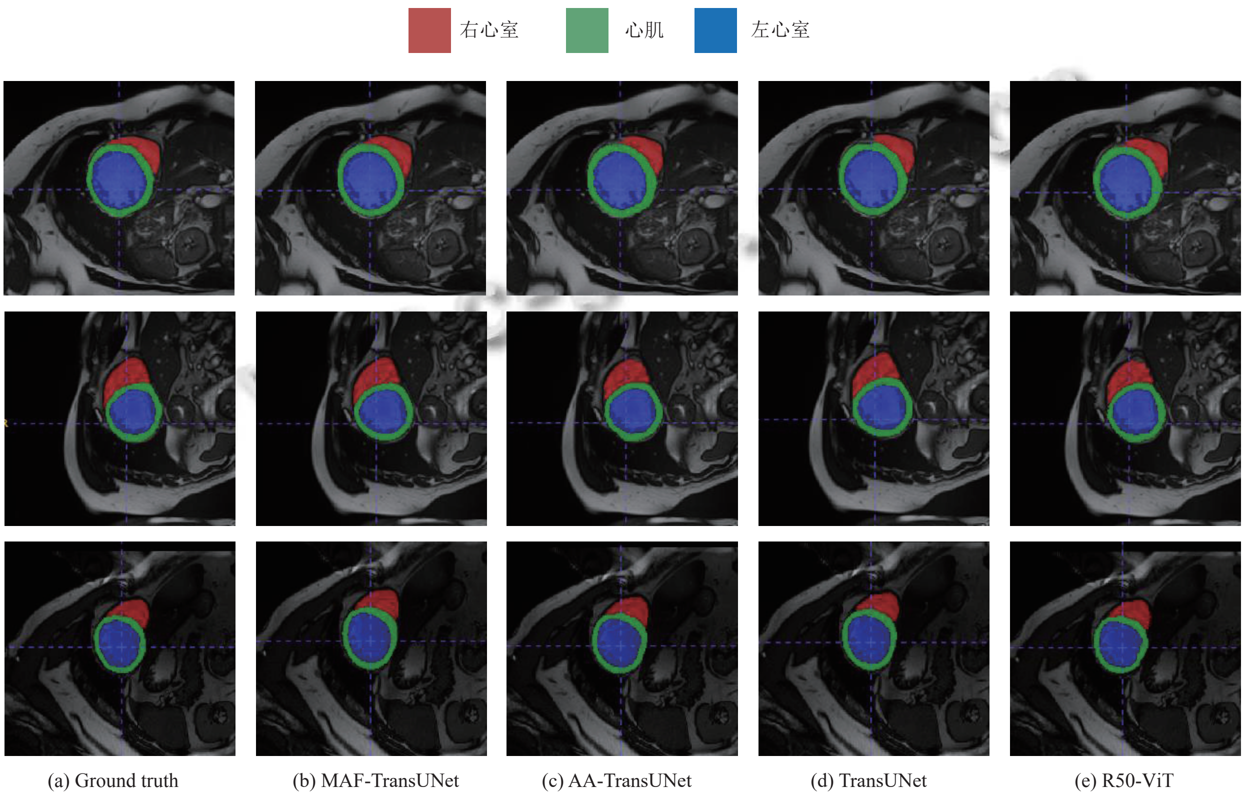


图 8 在 ACDC 数据集上对不同方法进行可视化

MAF-TransUNet 模型在 Synapse 数据集中的损失函数曲线如图 9 所示, 随着迭代次数增加, 损失率逐渐降低, 并在约第 67 轮时趋于稳定, 达到平稳的最优状态.

图 9 中 Smoothed 为平滑损失, 即对原始损失曲线进行平滑处理后的当前损失值, 有助于观察整体趋势; Value 为未经过平滑处理的原始损失值, 通常波动较大; Step 表示当前训练迭代步数, 本次训练已进行 27 750 次迭代; Relative 则为相对时间, 表示训练到当前步数所消耗的时间, 本次训练已运行 2.757 h.

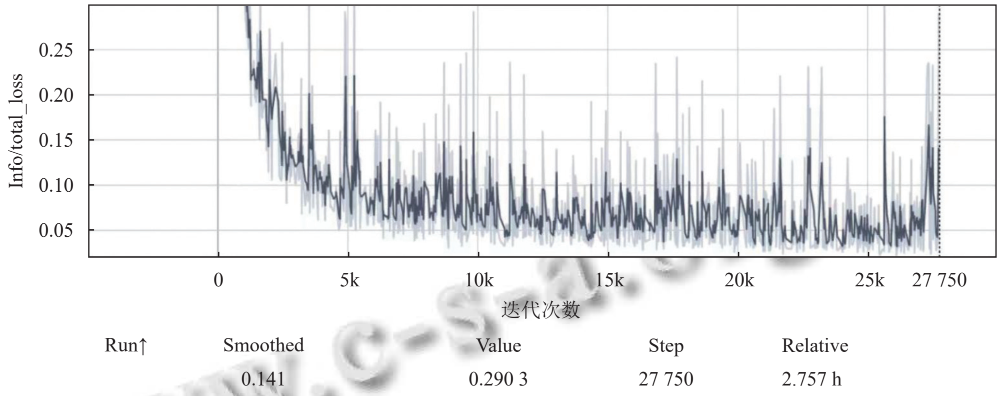


图 9 MAF-TransUNet 模型训练过程中损失函数曲线

4.5 消融实验

为了探究不同因素对网络性能的影响, 本文在 Synapse 数据集上进行了一系列实验. 这些实验主要关注以下几方面: 一方面是输入不同图像分辨率及在不同位置使用深度可分离卷积对分割性能的影响; 另一方面是在不同位置使用 MAF-Block 及加入不同模块对分割性能的影响.

实验结果表明, 如表 3 所示, AA-TransUNet[24]在 Synapse 数据集上的表现比 TransUNet 下降了 1.23%. 这一结果表明 AA-TransUNet 并不适用于医学图像分割, 但可以尝试其他方法进行改进; DA-TransUNet[27]以 TransUNet 为基础, 在编码器和跳跃连接中使用 DA-Block, 提高了 U 形网结构的特征提取能力, 还增强了编码器与解码器之间的特征传递, *Dice* 系数相较于基础模型提高了 2.32%. 然而, DA-TransUNet 相较于 MAF-TransUNet 的 *Dice* 系数降低了 1.22%. 是因为 DA-Block 在增强特征融合的过程中, 难以做到全局和局部信息之间的平衡精确统一. MAF-TransUNet 使用 MAF-Block 能够增强模型对细节和整体结构的捕捉能力并提升位置信息表达的准确性, 同时运用 DCA 模块探索通道和空间维度之间的关系, 保留了更多的空间信息.

MAF-TransUNet 的默认输入分辨率是 224×224. 如表 4 所示, 提供了在 512×512 高分辨率下训练 MAF-TransUNet 的结果. 当使用 512×512 作为输入时, 保持相同的 patch 大小, 使 Transformer 的序列长度增大. 正如 Chen 等人[22]所指出的, 增加有效序列长度可以显著提高模型性能. 对于 MAF-TransUNet, 将分辨率从 224×224 更改为 512×512, 性能提高了 2.20%, 但代价是计算成本大幅增加. 因此, 考虑到计算成本, 所有实验比较均采用默认的 224×224 分辨率, 以展示 MAF-TransUNet 的有效性.

表 3 不同模型对分割性能的影响 (%)

方法	*Dice*↑
TransUNet	77.48
AA-TransUNet	76.25
DA-TransUNet	79.80
MAF-TransUNet	81.02

表 4 输入不同图像分辨率对分割性能的影响 (%)

分辨率	*Dice*↑	主动脉	胆囊	肾 (左)	肾 (右)	肝脏	胰腺	脾	胃
224	81.02	88.22	67.73	82.24	78.55	94.62	64.25	90.00	82.53
512	83.22	90.63	71.94	86.43	82.85	95.57	64.32	90.23	83.77

表 5 所示为不同位置使用深度可分离卷积对分割性能的影响. DecoderCup 模块下使用深度可分离卷积 (DecoderCup-DSC, DC-D)、DecoderBlock 模块均用深度可分离卷积 (DecoderBlock-DSC, DB-D)、DecoderCup

模块与 DecoderBlock 模块下 Conv1 使用深度可分离卷积 (DecoderCup-DecoderBlock_Cov1, DC-DBC1)、DecoderCup 模块与 DecoderBlock 模块下 Conv2 使用深度可分离卷积 (DecoderCup-DecoderBlock_Cov2, DC-DBC2)、DecoderCup 模块与 DecoderBlock 模块全部使用深度可分离卷积 (DecoderCup-DecoderBlock, DC-DB).

具体结果如表 5 所示, DC-DBC1 效果最好. 深度可分离卷积在某些位置的引入可能有助于模型捕获局部特征, 特别是对于图像中细节信息的提取, 这是因为深度可分离卷积在处理细粒度特征上表现良好. 但有时也可能带来信息丢失和空间连续性的问题.

表 5 不同位置使用深度可分离卷积对分割性能的影响 (%)

方法	*Dice*↑
DC-D	69.41
DB-D	69.28
DC-DBC1	78.76
DC-DBC2	70.02
DC-DB	74.43

为了探索在编码层和跳跃连接中引入 MAF-Block 是否有助于提升模型的分割能力. 实验结果如表 6 所示, 编码器具有 MAF-Block 的模型, 与未添加 MAF-Block 的情况相比, 显著地提高了模型的性能. 其中, 运用 MAF-Block 来连接 CNN 和 Transformer 大大提高了系统的效率, 增强了位置信息特征提取能力. 通过为解码器提供更精细的位置信息特征, 可以减少在上采样过程中特征丢失的风险, 降低过拟合的可能性, 并保留更多的空间信息.可以得出结论: 在跳跃连接中添加 MAF-Block 并在 CNN 和 Transformer 间添加 MAF-Block 能够更有效地提取并复原特征的位置信息并增强 Transformer 学习特征的能力.

表 6 不同区域使用 MAF-Blcok 对分割性能的影响 (%)

方法	编码器中运用MAF	跳跃连接中运用MAF	*Dice*↑
TransUNet	—	—	77.48
TransUNet	√	—	79.21
TransUNet	—	√	77.98
TransUNet	√	√	80.25

如表 7 所示, 不同模块的加入对分割性能的影响. Synapse 数据集上的实验结果显示, 在 TransUNet 中加入 DSC 使性能降低了 8.07%, 之所以还使用 DSC, 是因为在上述实验中探讨了不同位置加入深度可分离卷积对分割性能的影响. 在深度卷积注意力模块中包含了深度可分离卷积操作使模型性能达到最高. DSC 的加入可以充分缓解模型中注意力冲突, 一定程度上平衡模型的复杂性, 避免引入过多复杂的结构, 同时提高了模型的可解释性和分割性能. 除探讨过的 DSC 外, 单独加入 MAF-Block 和 DCA 结果分别提高 2.32% 和 0.85%. 说明 MAF-Block 在位置和通道两个方面的特征提取方面效果突出, 能够获取更详细、更准确的特征集, 对于 DCA 则是在降低模型复杂度的同时能够关注通道间的特征重要性和空间位置的特征关联性来增强模型性能. 对于 MAF-Block、DSC 及 DCA 的不同组合, 将性能提升范围从 0.15% 到 1.55% 不等. 相对于单独加入略显降低, 相反将三者同时使用, 模型性能竟然提高至 81.02%, 比 TransUNet 性能增强了 3.54%. 三者结合可以同时在编码器、跳跃连接和解码器中增强位置信息的特征提取, 每一部分的缺失都会造成位置信息提取不足.

表 7 不同模块的加入对分割性能的影响 (%)

方法	*Dice*↑
TransUNet	77.48
TransUNet+MAF-Block	80.25
TransUNet+DSC	69.41
TransUNet+DCA	78.33
TransUNet+DCA+MAF-Block	77.63
TransUNet+MAF-Block+DSC	79.03
TransUNet+DCA+DSC	78.21
MAF-TransUNet	81.02

本文测试了不同层跳跃连接使用 MAF-Block 对分割性能的影响, 如表 8 所示. 可以看出, 在向编码层添加 MAF 块的基础上, 在跳跃连接的每一层都加入 MAF-Block 效果最好. 传统的跳跃连接只传递来自编码层的特征, 而不考虑所传递特征的质量. 通过实验可以看出, 传递给解码层更准确的位置信息特征可以更大程度上提高模型的特征学习能力和图像分割能力.

表 8 不同层跳跃连接使用 MAF-Block 对分割性能的影响 (%)

方法	1层	2层	3层	*Dice*↑
MAF-TransUNet	—	—	—	78.21
MAF-TransUNet	√	—	—	79.54
MAF-TransUNet	—	√	—	79.15
MAF-TransUNet	—	—	√	80.23
MAF-TransUNet	√	√	√	81.02

5 结论与展望

针对 TransUNet 等 U 型结构的模型在特征提取与

位置重建过程中, 模型在精度上仍存在一定偏差的问题. 本文提出 MAF-TransUNet 模型, 该模型通过引入 MAF 模块, 除了为编码层提供更准确的位置信息, 还减少了传统 U 形结构跳跃连接中的位置信息损耗, 增强了模型对位置信息的提取能力. 解码阶段加入的深度卷积注意力模块, 有效结合了不同特征通道中的空间位置信息, 提高了模型的分割精度. MAF-TransUNet 在 Synapse 数据集和 ACDC 数据集上表现出的性能提升, 证实了交叉注意力结构可以有效增强特定信息的表达能力. 未来计划引入多尺度可变形位置编码来进一步提高位置信息特征提取的准确性.

参考文献

1 周涛, 董雅丽, 霍兵强, 等. U-Net 网络医学图像分割应用综述. 中国图象图形学报, 2021, 26(9): 2058–2077. [doi: 10.11834/jig.200704]

2 Nam H, Ha JW, Kim J. Dual attention networks for multimodal reasoning and matching. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 299–307.

3 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.

4 Diakogiannis FI, Waldner F, Caccetta P, *et al*. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 162: 94–114. [doi: 10.1016/j.isprsjprs.2020.01.013]

5 Xie SN, Girshick R, Dollár P, *et al*. Aggregated residual transformations for deep neural networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1492–1500.

6 Zhou ZW, Rahman Siddiquee M, 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.

7 Li XM, Chen H, Qi XJ, *et al*. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Transactions on Medical Imaging, 2018, 37(12): 2663–2674. [doi: 10.1109/TMI.2018.2845918]

8 Alom Z, Hasan M, Yakopcic C, *et al*. Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. arXiv:1802.06955, 2018.

9 Valanarasu JMJ, Sindagi VA, Hacihaliloglu I, *et al*. KiU-net: Towards accurate segmentation of biomedical images using over-complete representations. Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention. Lima: Springer, 2020. 363–373.

10 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. Barcelona: IEEE, 2020. 1055–1059.

11 Lin M, Chen Q, Yan SC. Network in network. arXiv:1312.4400, 2013.

12 Huang ZL, Wang XG, Huang LC, *et al*. CCNet: Criss-cross attention for semantic segmentation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 603–612.

13 Hou QB, Zhou DQ, Feng JS. Coordinate attention for efficient mobile network design. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13713–13722.

14 Wang XL, Girshick R, Gupta A, *et al*. Non-local neural networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7794–7803.

15 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]

16 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

17 Parmar N, Vaswani A, Uszkoreit J, *et al*. Image Transformer. Proceedings of the 35th International Conference on Machine Learning. Stockholmsmässan: PMLR, 2018. 4055–4064.

18 Child R, Gray S, Radford A, *et al*. Generating long sequences with sparse Transformers. arXiv:1904.10509, 2019.

19 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. 10012–10022.

20 Carion N, Massa F, Synnaeve G, *et al*. End-to-end object detection with Transformers. Proceedings of the 16th

European Conference on Computer Vision. Glasgow: Springer, 2020. 213–229.

21 Lin AL, Chen BZ, Xu JY, *et al*. DS-TransUNet: Dual swin Transformer U-Net for medical image segmentation. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 4005615.

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

23 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.

24 Yang YM, Mehrkanoon S. AA-TransUNet: Attention augmented TransUNet for nowcasting tasks. Proceedings of the 2022 International Joint Conference on Neural Networks. Padua: IEEE, 2022. 1–8.

25 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19.

26 Chollet F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1251–1258.

27 Sun GQ, Pan YZ, Kong WK, *et al*. DA-TransUNet: Integrating spatial and channel dual attention with Transformer U-Net for medical image segmentation. Frontiers in Bioengineering and Biotechnology, 2024, 12: 1398237. [doi: 10.3389/fbioe.2024.1398237]

28 Wu K, Peng HW, Chen MH, *et al*. Rethinking and improving relative position encoding for vision Transformer. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 10033–10041.

29 Gao Y, Che XJ, Xu H, *et al*. An enhanced feature extraction network for medical image segmentation. Applied Sciences, 2023, 13(12): 6977. [doi: 10.3390/app13126977]

30 Su JL, Ahmed M, Lu Y, *et al*. RoFormer: Enhanced Transformer with rotary position embedding. Neurocomputing, 2024, 568: 127063. [doi: 10.1016/j.neucom.2023.127063]

31 Landman B, Xu Z, Igelsias J, *et al*. Miccai multi-atlas labeling beyond the cranial vault—Workshop and challenge. Proceedings of the 2015 MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. 2015. 12.

32 Baumgartner CF, Koch LM, Pollefeys M, *et al*. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. Proceedings of the 8th International Workshop on Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. Quebec City: Springer, 2018. 111–119.

33 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.

(校对责编: 张重毅)