

基于多尺度自注意力 Transformer 的医学图像分割方法

金晶, 胡楚笛, 陈刚

(武汉大学国家网络安全学院空天信息安全与可信计算教育部重点实验室, 湖北 武汉 430072)

摘要: Transformer 模型凭借其出色的全局信息捕获能力和强大的表示能力, 被广泛应用于医学图像分割领域, 并取得了显著成效。然而, 这些方法在对图像进行序列化时, 会将图像分割成固定大小的块, 仅提取单一尺度的全局特征, 在一定程度上割裂了图像的语义特征, 最终导致分割精度不佳。为此, 提出一种多尺度自注意力 Transformer 架构(MultiFormer)。首先, 采用连续卷积和下采样模块对图像进行处理; 然后, 使用多尺度卷积投影模块来替换原来的 1×1 投影模块; 最后, 在自注意力模块生成的特征图中引入可变形卷积。与传统 Transformer 图像序列化过程相比, 这种连续卷积在产生相同分辨率的特征时, 有效增大了感受野, 保留了二维(2D)图像的空间相关性, 避免固定位置和大小图像块带来的语义信息损失。同时, 多尺度卷积投影模块利用 4 种不同大小的卷积核捕捉图像中的上下文信息, 并通过通道拼接实现多尺度特征融合, 反映了不同尺度下局部与局部间的注意力, 而不仅限于单一尺度, 实现了模型对不同尺度语义信息的聚合能力, 进一步减轻了语义割裂问题。此外, 可变形卷积通过引入一个额外的卷积层来学习并生成偏移场, 使得卷积核能够灵活调整其形状, 以适应图像中形态多变的病变或器官, 提升模型对复杂医学图像的处理能力。将该模块分别插入到 SETR、TransUNet、TransFuse 这 3 个网络结构中, 并在 ACDC 心脏数据集和 ISIC 2018 皮肤病变数据集上进行实验, 结果表明, Dice 系数分别提升了 3.63、1.06、2.30 个百分点和 1.22、2.31、3.01 百分点。MultiFormer 具有即插即用性, 能够方便地集成到各种下游医学图像分析任务中。

关键词: 医学图像分割; Transformer; 注意力机制; 多尺度卷积投影; 可变形卷积

源代码链接: <https://github.com/vicky716/MultiFormer>

中图分类号: TP391

文献标志码: A

DOI: 10.19678/j.issn.1000-3428.0070385

Medical Image Segmentation Method Based on Multi-Scale Self-Attention Transformer

JIN Jing, HU Chudi, CHEN Gang

(Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education,
School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, Hubei, China)

【Abstract】 The Transformer model, owing to its excellent global information capture and powerful representation abilities, has been widely used in medical image segmentation and has achieved remarkable results. However, when serializing images, these methods divide them into fixed-sized blocks and extract only global features of a single scale, which, to some extent, fragments the semantic features of the images, ultimately leading to poor segmentation accuracy. To address this issue, this paper proposes a Multi-scale self-attention Transformer architecture (MultiFormer). First, continuous convolution and downsampling modules are used to process images. Then, the original 1×1 projection module is replaced with a multi-scale convolution projection module. Finally, deformable convolution is introduced into the feature maps generated by the self-attention module. Compared to the traditional Transformer image serialization process, this continuous convolution effectively enlarges the receptive field while generating features of the same resolution, retains the spatial correlation of 2-Dimensional (2D) images, and avoids the loss of semantic information caused by fixed-position and fixed-size image blocks. Meanwhile, the multi-scale convolution projection module captures contextual information in the images using four convolution kernels of different sizes and achieves multi-scale feature fusion through channel concatenation, reflecting attention between local regions at different scales, rather than being limited to a single scale, which enables the model to aggregate semantic information at different scales and further alleviates the problem of semantic fragmentation. Deformable convolution introduces an additional convolution layer to learn and generate an offset field, allowing the convolution kernel to flexibly adjust its shape to adapt to morphologically diverse lesions or organs in the images and

基金项目: 国家自然科学基金(U1936107)。

作者简介: 金晶(CCF 学生会员), 女, 硕士研究生, 主研方向为医学图像处理、网络空间安全、深度学习; 胡楚笛, 博士研究生; 陈刚(通信作者), 教授、博士。

收稿日期: 2024-09-18

修回日期: 2025-01-19

E-mail: chenzuolin@whu.edu.cn

enhancing the ability of the model to process complex medical images. This module is inserted into three network structures—SETR, TransUNet, and TransFuse—and experiments are conducted on the ACDC heart and ISIC 2018 skin lesion datasets. The results show that the Dice coefficient improves by 3.63, 1.06, and 2.30 percentage points and by 1.22, 2.31, and 3.01 percentage points, respectively. MultiFormer is a plug-and-play tool and can be easily integrated into various downstream medical image analysis tasks.

【Key words】 medical image segmentation; Transformer; attention mechanism; multi-scale convolutional projection; deformable convolution

0 引言

近年来,Transformer^[1]在自然语言处理(NLP)领域取得了突破性的进展,它逐渐成为序列建模任务的主流方法。受此启发,一些研究成功地将Transformer应用于计算机视觉领域,在图像分类^[2]、目标检测^[3]、语义分割^[4]等方面取得了良好的性能。与NLP类似,视觉Transformer(ViT)^[5]通常将输入图像划分为一系列固定大小的patch(例如 16×16),并通过多头自注意力机制建模不同patch之间的上下文关系。与卷积神经网络(CNN)相比,Transformer凭借其强大的长距离依赖关系捕捉能力和丰富的语义信息提取能力,为视觉任务带来了全新的视角。

尽管Transformer在视觉任务中已经取得了一些成果,但在医学图像领域应用Transformer仍有许多挑战。例如,在大规模训练下,ViT显示出比CNN更好的效果,但由于缺乏归纳偏置,因此ViT在小规模数据上的泛化性能存在一定挑战,考虑到疾病亚型、数据模态、采集困难、隐私问题和高昂的注释成本,基于图像的任务往往只有少量可用的样本,这使得大规模训练或预训练变得困难。DeiT^[6]提出了一种有效的知识蒸馏训练方案,以克服ViT需要大量数据学习的困难。Swin Transformer^[7]通过引入窗口注意力机制,将自注意力机制的计算限制在不相交的局部窗口内,并允许窗口之间进行交互,从而提升了模型的效率和泛化能力。同时,由于医学图像的分割目标形态差异大,如患者之间的体型差异(胖瘦、高矮、年龄层差异)以及病灶在大小、形状和位置上的显著变化,因此,有必要融合多尺度语义特征,以帮助勾画详细的边界。现有的多尺度方法多采用多分辨率输入策略,如MG-Trans^[8]通过不同分辨率尺度的输入提取图像特征,分别送入Transformer模块处理,MSPT^[9]将不同分辨率的图像通过Transformer模块后进行特征融合。

通过分析上述工作可知,现有的多尺度方法处理主要集中在Transformer模块的输入或输出阶

段,依赖多分辨率输入或后期特征融合策略,而未深入探索Transformer内部对多尺度信息的建模。这些策略虽然有效,但是在隐式假设下使用强制的固定大小的patch划分可能会带来2个问题:

1)图像中局部结构的坍塌。固定大小的patch总是很难捕捉到完整的、与对象相关的局部结构,因为对象在不同的图像中具有不同的尺度^[10]。

2)不同图像之间的语义不一致。同一对象在不同图像中可能呈现出不同的视觉属性,例如尺度变化和旋转,采用固定的分割方法可能无法准确捕获这些变化,从而导致捕获的信息在不同图像之间出现差异。这些固定的patch可能会破坏语义信息,导致性能下降。

为了解决由固定patch划分引起的语义割裂问题,本文提出了一种新型Transformer架构——MultiFormer。MultiFormer是一个全卷积的多尺度Transformer结构:首先,通过尺度感知配置的连续卷积来替代传统的patch划分过程,有效缓解patch划分带来的语义割裂问题;其次,设计多尺度卷积投影模块,通过在查询(query)、键(key)和值(value)向量生成过程中捕获不同感受野范围的邻域信息,实现Transformer自注意力机制内部对多尺度信息的建模,反映不同尺度下局部与局部间的注意力,而不仅限于单一尺度,从而增强局部与全局特征的空间表达能力,进一步减轻语义割裂问题;最后,在聚合信息后引入可变形卷积,以增强其对分割边界复杂形态变化的灵活适应能力。MultiFormer作为一种全卷积即插即用的模块,可以轻松集成于各种Transformer架构中,有效地克服Transformer在小数据集上的限制。本文的主要工作如下:

1)提出一种全卷积的多尺度Transformer架构,通过尺度感知配置的连续卷积替代传统固定大小的patch划分方式,显著增大感受野并保留二维(2D)图像的空间相关性,缓解patch划分带来的语义割裂问题。

2)设计多尺度卷积投影模块,在Transformer自注意力机制中引入多感受野范围的邻域信息,提

高不同尺度下局部特征的上下文建模能力。

3)在特征聚合阶段引入可变形卷积模块,用于适应和处理不规则边界的形态变化,以提高分割边界的精确性。

1 相关工作

1.1 医学图像分割

医学图像分割是医学影像分析中的关键任务,广泛应用于器官结构识别、病灶检测、疾病诊断等场景。传统的医学图像分割主要依赖于基于规则或特定特征的算法,包括基于阈值的分割^[11]、区域增长^[12]、边缘检测^[13]等方法。但是,这些方法难以适应复杂的医学图像特征。随着深度学习技术的发展,CNN成为医学图像分割领域的主流方法^[14],U-Net^[15]及其变种通过编码器-解码器结构显著提升了分割性能。然而,CNN因其感受野有限,难以捕获全局上下文信息,在处理大尺度变化和复杂目标边界时表现受限。这些问题推动了 Transformer 在医学图像分割任务中的探索。

1.2 基于 Transformer 的医学图像分割方法

Transformer 通过自注意力机制在 NLP 和计算机视觉中取得了显著成功,并逐步应用于医学图像分割任务。与 CNN 相比,Transformer 的优势在于其强大的全局信息建模能力。然而,直接应用纯 Transformer 架构会带来计算复杂度高、局部特征建模能力不足的问题。

现有基于 Transformer 的医学图像分割方法主要分为纯 Transformer 架构^[4]和 CNN-Transformer 混合架构^[16-18]两类。SETR^[4]是将纯 Transformer 引入医学图像分割的早期工作之一,它通过固定大小的 patch 划分输入图像,利用多层 Transformer 实现全局语义信息的建模,但这种固定划分割裂了图像的空间相关性,尤其在处理局部边界细节时表现不足。此外,SETR^[4]的高计算复杂度也限制了其在实际场景中的应用。为克服纯 Transformer 的局限性,近年来,研究人员提出了结合 CNN 的混合架构。TransUNet^[19]结合了 U-Net 的编码器结构,通过 CNN 提取局部特征,再利用 Transformer 捕获全局信息,实现了全局与局部特征的融合。然而,由于 Transformer 仍依赖于 patch 划分机制,TransUNet 对局部与全局特征的深度融合仍有局限。Swin-Unet^[20]通过滑动窗口机制改进了空间相关性丢失问题,但是窗口之间的上下文交互仍然受限,难以完全弥补全局建模能力的缺失。

尽管融合 Transformer 的 U-Net 方法在医学图像分割中表现出良好性能,但对混合架构中 Transformer 模块效果的研究仍然存在较大的空白,且在保留图像空间相关性、多尺度特征建模和适应复杂目标形态变化方面仍存在局限。

1.3 多尺度 Transformer 在医学图像中的应用

在医学图像处理任务中,目标区域通常具有显著的尺度变化特性,因此,如何有效捕获和整合多尺度信息是提升分割性能的关键。近年来,研究者尝试将多尺度设计与 Transformer 架构相结合,以增强模型对不同尺度特征的表达能力。

现有工作提出了多种方法来整合不同尺度的特征。MG-Trans^[8]通过在不同分辨率下提取特征,分别输入多个 Transformer 模块中以建模语义信息,并在 Transformer 的输出阶段进行特征融合。MSPT^[9]则设计了特征融合模块,将不同分辨率的图像通过 Transformer 整体框架后,在输出阶段统一结合多尺度语义特征,从而提升分割性能。MMT^[21]采用 U-Net 结构,通过下采样与 Transformer 模块交替设计,对不同尺度的特征进行逐层聚合,捕获复杂目标的多层次信息。这些方法充分验证了在不同分辨率下处理特征并通过特定融合策略提升分割性能的有效性。

然而,上述研究大多在 Transformer 模块的输入或输出阶段进行多尺度处理,依赖于多分辨率输入或后期特征融合策略,而未深入探讨 Transformer 自注意力机制内部对多尺度信息的直接建模。因此,本文在 Transformer 的自注意力模块中引入多尺度卷积操作,直接在生成查询、键和值向量的过程中嵌入不同感受野的上下文信息。这种设计不仅在特征生成阶段实现了多尺度信息的高效融合,还减轻了语义割裂问题。

2 方法设计

本文提出一种多尺度的 Transformer 架构 MultiFormer,该架构由 3 个关键模块组成,即尺度感知卷积序列化模块、多尺度自注意力模块和前馈神经网络(FFN)模块,如图 1 所示(彩色效果见《计算机工程》官网 HTML 版,下同),其中 MCP 为多尺度卷积投影模块。尺度感知卷积序列化模块用于替换 ViT^[5]中的 patch embedding,在不依赖额外位置嵌入的情况下很好地保留了局部特征和位置信息。多尺度自注意力模块是 MultiFormer 的核心组件,通过多尺度卷积生成查询、键和值,以实现多层次的上下文信息捕获,并进一步通过自注意力机

制进行特征交互,实现全局和局部空间特征表达。在此基础上,该模块进一步引入可变形卷积,在注意力矩阵后动态调整特征,以适应形态复杂的目标区

域,提高对边界区域和局部结构变化的建模能力。前馈神经网络模块通过非线性变换和映射,增强特征的表达能力,提高模型性能。

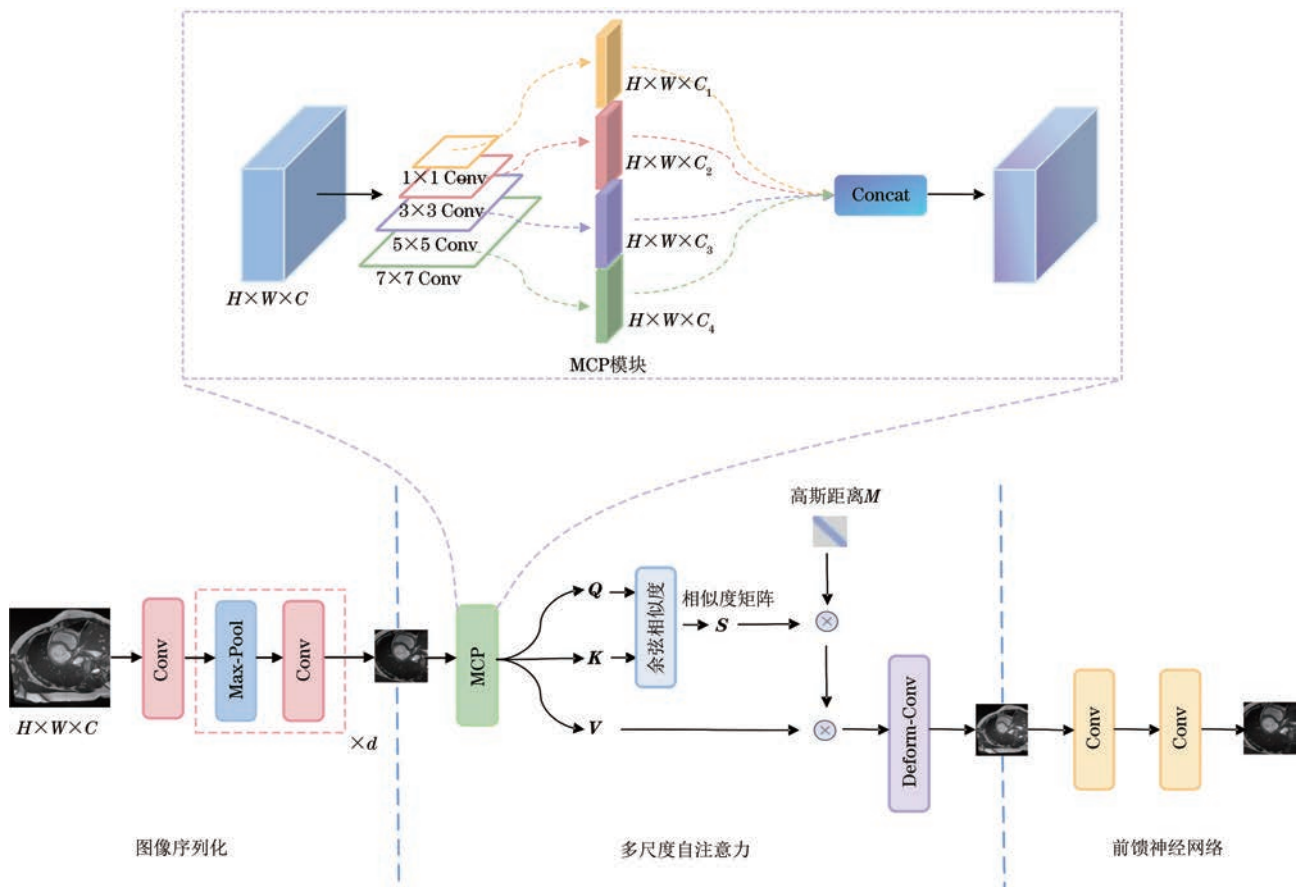


图 1 MultiFormer 的结构
Fig. 1 Structure of MultiFormer

2.1 尺度感知卷积序列化

给定一个数据集 $\{X, Y\}$, 其中, X 是模型的输入图像, Y 是相应的语义或二进制分割图。对于每个图像 $X_i \in \mathbb{R}^{H \times W \times C}$, 其中, H 和 W 是图像的空间分辨率, $C \in \{1, \dots, N\}$ 是输入通道的数量。

在传统的 ViT 中, 将输入图像分为固定大小的 patch, patch 的大小为 S , 则每张图像会生成 T 个 patch, 其中, $T = H/S \times W/S$, 即输入序列长度为 T , 每个 patch 维度为 K , 其中, $K = S \times S \times C$ 。然后, 经过一个线性投影层, 通过线性投影层之后的维度依然为 $T \times K$, 即共有 T 个 token, 每个 token 的维度是 K 。这里还需要加上一个特殊字符 cls, 因此最终的维度是 $(T+1) \times K$ 。通过 patch 嵌入(patch embedding)将一个视觉问题转化为一个序列到序列的问题。由于 ViT 是一种基于自注意力机制(Self-Attention)的模型, 没有明确的位置信息, 因此需要引入位置编码来帮助模型理解输入序列的相对位置。位置编码是通过将可学习的一维(1D)位

置嵌入(learnable 1D position embedding)与输入序列的嵌入进行相加(sum)而得到的, 这个可学习的 1D 位置嵌入是一个独立于输入数据的参数, 用来表示序列中每个位置的信息。加入位置编码信息之后, 维度依然是 $(T+1) \times K$ 。

在本文方法中, 通过尺度感知卷积序列化模块来实现 token 化的功能。所谓尺度感知, 就是使输入适合于通道维度上的 Transformer, 并且根据不同 patch 大小, 自适应地调整输入大小。对于输入 $X_i \in \mathbb{R}^{H \times W \times C}$, 首先应用核大小为 3 的卷积, 然后进行批归一化和 ReLU 来捕获局部特征, 使网络更好地学习和表示数据的特征, 提高模型性能和泛化能力。由于 ViT 中 patch 的大小为 S , 通过 $d = \lfloor \log S \rfloor$ 次下采样操作产生相同的分辨率, 输入 X_i 通过下采样操作变为 $X_{im} \in \mathbb{R}^{\frac{H}{2^d} \times \frac{W}{2^d} \times C}$ 。本文方法在进行 token 化时直接对 2D 输入进行操作, 无须像传统 Transformer 一样通过 patch 嵌入将 2D 输入转换为 1D 序列, 这种设计使得该方法在不需要位

置嵌入的情况下,更好地保留了图像中的位置信息。

2.2 多尺度自注意力

细胞核、器官、组织等生物医学图像中需要分割的对象通常大小、形状各异,例如,皮肤病变图像如图 2 所示,不同图像中的皮肤病变具有显著性差异。因此,分割模型需要在各种尺度和类型下具有鲁棒性。

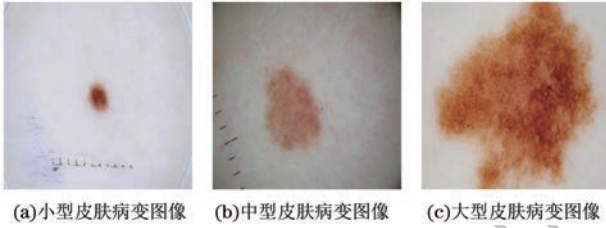


图 2 不同大小和形状的皮肤病变图像

Fig. 2 Images of skin lesions of different sizes and shapes

现有的大多数模型采用线性投影进行多头自注意力(MHSA)计算,这导致 Transformer 失去空间上下文。MultiFormer 用 MCP 替代原来的自注意力中的线性投影,形成多尺度卷积投影层。通过使用 4 个不同尺度的平行卷积,然后将结果按通道进行拼接,形成最终的输出,如图 1 所示。核大小为 3、5、7 的卷积核可以有效地捕获多尺度信息,而核大小为 1 的卷积核实际上是一个可学习的通道尺度因子。具体来说,对于输入 $\mathbf{X}_{im} \in \mathbb{R}^{H' \times W' \times C}$,其中, $H' = H/2^d, W' = W/2^d, x_{i,j}$ 为 $\mathbf{X}_{im}$ 中的每个像素, $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ 通过式(1)进行计算:

$$\mathbf{Q}/\mathbf{K}/\mathbf{V}_{i,j} = \text{Concat} \left(\sum_{l=\frac{f-1}{2}}^{\frac{f-1}{2}} \sum_{g=\frac{f-1}{2}}^{\frac{f-1}{2}} \mathbf{E}_{\frac{f+1}{2}+l, \frac{f+1}{2}+g}^{q/k/v} x_{i+l, j+g} \right) \quad (1)$$

$f=1, 3, 5, 7$

式中: f 为卷积核大小; $\mathbf{E}^{q/k/v} \in \mathbb{R}^{C_t \times C_t \times C_m \times C_m}$ 是可学习的投影矩阵; C_t 为输入通道的数量,对应 ViT 中 $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ 的嵌入维度,将 $f \times f$ 邻域内相邻像素的特征整合为 $\mathbf{X}_{i,j}$; C_m 为输出通道的数量。对于不同大小的卷积核,分别设置其输出通道的数量,按通道拼接得到最终的输出通道数量 C_{out} ,其对应于 ViT 中线性投影的输出通道数。 C_{out} 可表示如下:

$$C_{out} = \alpha_1 C_t + \alpha_2 C_t + \alpha_3 C_t + \alpha_4 C_t \quad (2)$$

$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$

将输出的结果分别通过 1×1 的卷积进行多尺度信息融合。

在计算 $\mathbf{Q}$ 和 $\mathbf{K}$ 的相似度时,本文没有采用传统 Transformer 中的点乘方法,而是采用

ConvFormer^[22]中的余弦相似度方法,得到相似度矩阵 $\mathbf{S}^{i,j} \in \mathbb{R}^{H' \times W'}$:

$$\mathbf{S}_{p,q}^{i,j} = \frac{\sum_{l=0}^{C_t} \mathbf{Q}_{i,j}^l \mathbf{M}_{p,q}^l}{\sqrt{\sum_{l=0}^{C_t} \mathbf{Q}_{i,j}^{2l}} \sqrt{\sum_{l=0}^{C_t} \mathbf{M}_{p,q}^{2l}}}, \mathbf{S}_{p,q}^{i,j} \in [-1, 1] \quad (3)$$

$\mathbf{S}_{p,q}^{i,j}$ 对应于 ViT 自注意力中的权重矩阵。这种方法的优点在于,它减少了相似度矩阵中零元素的出现,从而提供了更为丰富的信息。为了动态地调整相似度矩阵 $\mathbf{S}$,通过引入一个可学习的高斯距离矩阵 $\mathbf{M}$,将 $\mathbf{S}$ 与 $\mathbf{M}$ 相乘,从而根据距离的高斯分布来调整相似度矩阵 $\mathbf{S}_{p,q}^{i,j}$:

$$\mathbf{M}_{p,q}^{i,j} = e^{-\frac{(i-p)^2/H'^2 + (j-q)^2/W'^2}{2(\theta \times \alpha)^2}} \quad (4)$$

$$\mathbf{S}_{p,q}^{i,j} = \mathbf{S}_{p,q}^{i,j} \times \mathbf{M}_{p,q}^{i,j} \quad (5)$$

式中: $\theta \in (0, 1)$ 是一个可学习的网络参数; α 是一个控制感受野大小的超参数。最后,将 $\mathbf{S}$ 和 $\mathbf{V}$ 矩阵相乘,并通过一个可变形卷积(Deform-Conv)^[23],得到最终的自注意力层的输出结果。

可变形卷积结构如图 3 所示。可变形卷积能够通过整数偏移量来对采样网格进行动态调整,以实现自由变形。在可变形卷积中,通过一个额外的卷积层,从深层特征图中捕捉和学习这些变形,从而创建一个偏移场,通过双线性插值方法计算不在图像网格上的偏移处的像素值。基于特征本身学习变形会导致自适应卷积核。这种灵活的卷积核形状能够针对病变或器官的特征进行自适应调整,从而显著增强它们在图像中的表示能力,并有效提升物体边界的识别精度与清晰度。

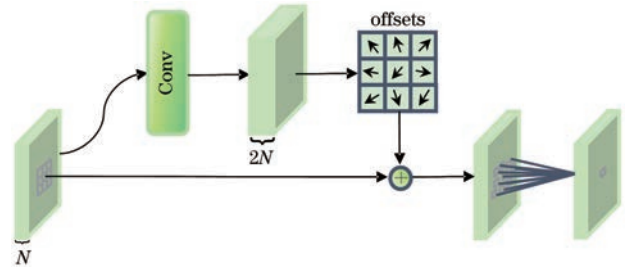


图 3 可变形卷积的结构

Fig. 3 The structure of deformable convolution

2.3 前馈神经网络

在前馈神经网络中,传统的线性投影被 1×1 的卷积所替代。首先,自注意力的输出通过一个 1×1 的卷积被映射到高维空间;然后,通过另一个 1×1 的卷积,高维空间的表示被映射回原始的空间维度。

3 实验分析

3.1 数据集

1) ACDC (Automated Cardiac Diagnosis Challenge)^[24]。ACDC 是用于自动心脏诊断挑战的公开可用数据集。ACDC 挑战赛收集了 100 例 MRI 扫描图像,与之前方法一致^[19-20],随机选择其中 70 例案例用于模型训练,10 例用于验证,20 例用于测试。在提供的数据中,共标记了左心室(LV)、右心室(RV)和心肌(MYO)3 个器官。

2) ISIC 2018^[25]。ISIC 2018 是用于皮肤病变分割的公开可用数据集,包含 2 594 张不同类型的皮肤病变图像。与之前方法^[16-17]一致,随机选择其中 2 074 张图像用于训练,520 张图像用于验证和测试,并对这一过程重复 5 次,最终取平均值作为实验结果。

3.2 评价指标

1) Dice 系数(E_{Dice})。Dice 系数主要用来计算预测值和真实值之间的差异,被广泛地运用在图像分割任务中,取值范围为 0~1。Dice 系数越高,表示分割结果越准确。给定预测结果 X 以及真实标签 Y ,Dice 系数定义如下:

$$E_{\text{Dice}} = \frac{2|X \cap Y|}{|X| + |Y|} \quad (6)$$

2) Hausdorff 距离(D_H)。Dice 系数对 mask 的内部填充比较敏感,而 Hausdorff 距离对分割出的边界比较敏感,它表示预测分割区域边界与真实区域边界之间的最大距离,其值越小,代表预测边界分割误差越小,质量越好。设 X 和 Y 是度量空间 M 的 2 个真子集,则集合之间的 Hausdorff 距离定义为:

$$D_H(X, Y) = \max\{d_{XY}, d_{YX}\} = \max\{\max_{x \in X} \min_{y \in Y} d(x, y), \max_{y \in Y} \min_{x \in X} d(y, x)\} \quad (7)$$

在计算出式(7)后,HD 95 就是最后的值乘以 95%。

3.3 实验细节

MultiFormer 基于 PyTorch 1. 8. 1 和 CUDA 10. 2 实现,并使用单个 NVIDIA Tesla V100 GPU 进行训练。模型输入的分辨率为 256×256 像素,所有模型均由 Adam 优化器训练,学习率为 0. 000 1,批量大小为 4,训练 400 轮。模型在训练过程中未使用预训练数据集或预训练机制,而是从头开始训练。最终依据验证集 Dice 系数的最高表现选择最优模型参数,用于测试集性能评估。在 ACDC 心脏数据集和 ISIC 2018 皮肤病变数据集中均使用数据增强方法,包括随机旋转、缩放、对比度增强和伽马变换。

3.4 实验结果

MultiFormer 可以作为即插即用模块,直接替换基线模型中的传统 Transformer 块。为了评估 MultiFormer 的有效性,选择了 3 种具有代表性的先进方法作为基线和主干: SETR^[4] 作为纯 Transformer 架构的代表, TransUNet^[19] 和 TransFuse^[18] 代表 CNN-Transformer 混合架构。这些模型能够全面验证 MultiFormer 在不同主干架构中的适用性和性能提升效果。此外,在上述基于 Transformer 的基线中,还与 CNN 式的 Transformer 模块 ConvFormer^[22] 进行比较,以验证本文多尺度可变形 Transformer 的有效性。为了确保实验的公平性,所有方法均采用与对应主干方法相同的配置。

3.4.1 ACDC 数据集

在 ACDC 数据集上,对基于 Transformer 的 3 种基线方法 SETR^[4]、TransUNet^[19]、TransFuse^[18] 以及替换这 3 种主干方法上 Transformer 模块的 ConvFormer^[22]、MultiFormer 模型进行对比实验,不同网络模型分割结果比较如表 1 所示,最优结果加粗标注。

表 1 在 ACDC 数据集上的实验结果对比

Table 1 Comparison of experimental results on the ACDC dataset

网络模型	Dice 系数 ↑ / %				HD 95 ↓ / mm
	右心室	心肌	左心室	平均	平均
SETR	84. 06	84. 25	93. 55	87. 20	16. 38
SETR+ConvFormer	88. 93	87. 54	95. 08	90. 52	14. 05
SETR+MultiFormer	89. 15	88. 05	95. 28	90. 83	13. 64
TransUNet	89. 59	87. 81	94. 99	90. 80	14. 58
TransUNet+ConvFormer	89. 38	89. 08	95. 64	91. 36	13. 75
TransUNet+MultiFormer	90. 28	89. 49	95. 80	91. 86	13. 13
TransFuse	87. 85	85. 73	93. 73	89. 10	14. 98
TransFuse+ConvFormer	89. 33	88. 57	95. 03	91. 01	13. 98
TransFuse+MultiFormer	89. 90	88. 51	95. 78	91. 40	13. 55

从表 1 可以观察到,在 ACDC 数据集中,MultiFormer 模块的引入显著提升了 3 种基线模型的性能。具体而言,SETR 模型在采用 MultiFormer 后,Dice 系数平均提升了 3.63 百分点,同时 Hausdorff 距离平均降低了 2.74 mm;TransUNet 模型在 Dice 系数上平均提高了 1.06 百分点,Hausdorff 距离则平均降低了 1.45 mm;TransFuse 模型也展现出了相似的效果,Dice 系数平均增长了 2.30 百分点,Hausdorff 距离则平均下降了 1.43 mm。这些结果表明,MultiFormer 模块在 3 种基线模型上均实现了一致且显著的性能改进。

与最新研究 ConvFormer 相比,MultiFormer 在 3 个基线模型上均展现出了更优的性能。具体来说,在 SETR、TransUNet 和 TransFuse 中,Dice 系数分别提高了 0.31、0.50 和 0.39 百分点,而 Hausdorff 距离则分别降低了 0.41、0.62 和 0.43 mm。这一对比进一步验证了 MultiFormer 模块在医学图像分割任务中的优越性和有效性。

对比模型的可视化分割结果如图 4 所示,以 SETR 和 TransUNet 这 2 个基线方法为例,红框突出显示了使用 MultiFormer 相较于其他方法性能更为出色。

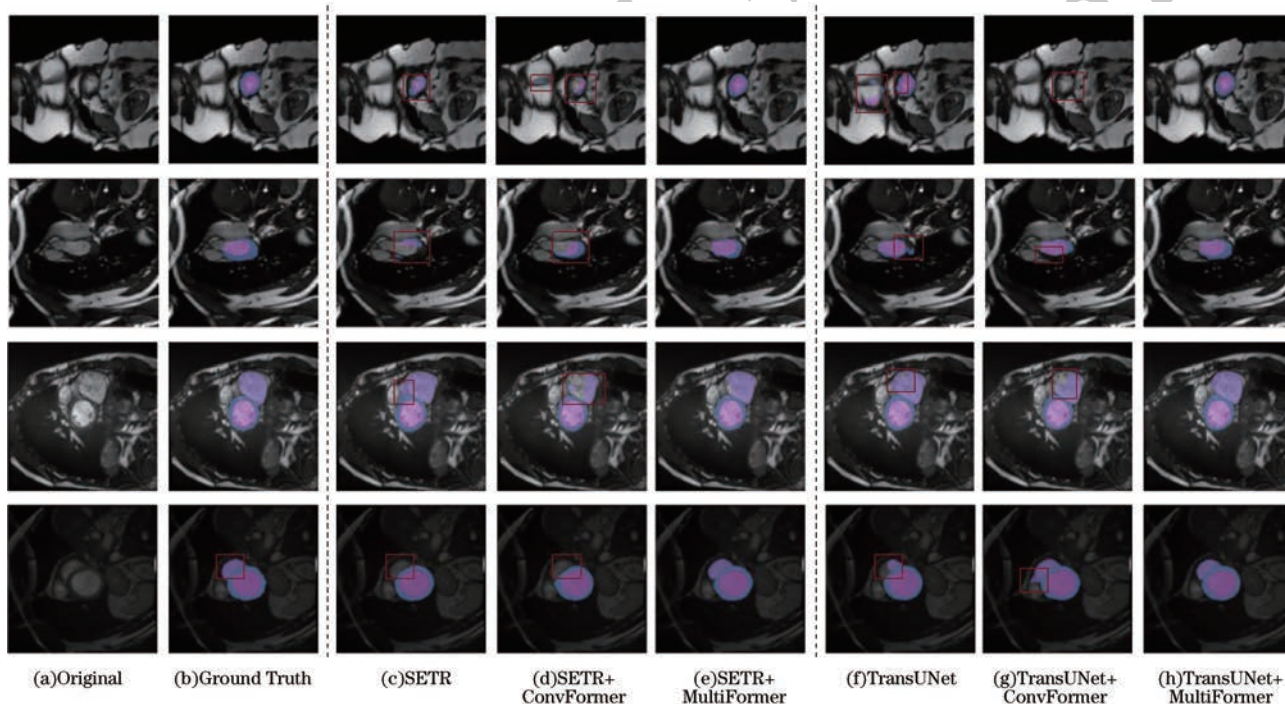


图 4 在 ACDC 数据集上的可视化对比

Fig.4 Visual comparison on the ACDC dataset

为进一步分析 MultiFormer 的作用,图 5 展示了引入 MultiFormer 后对基线方法的改进效果,其中,图 5(a)和图 5(b)为图 5(f)中红框区域的放大图,用于更加直观地展示注意力计算过程。图 5(f)中的紫色区域表示一个完整的器官,而传统 Transformer 方法(以 SETR 为例)在进行 patch 划分时,容易将完整的器官区域分割到不同的 patch 中,如图 5(a)所示,导致语义割裂。此外,基线方法中的注意力计算局限于单个固定大小的窗口[图 5(a)绿色窗口和黄色窗口],仅在单一尺度下捕获像素与像素之间的注意力。较小的窗口会导致语义割裂,而较大的窗口可能引入冗余信息,导致建模不精确。针对这些问题,本文 MultiFormer 通过以下方式有效缓解语义割裂问题:首先,使用连续卷积代替固定大小的 patch 划分,避免语义信息的丢失,

同时保留 2D 图像的空间上下文关系;其次,在注意力计算中,通过多个不同尺度的窗口[图 5(b)]来计算局部与局部区域之间的注意力,实现自注意力机制内部对多尺度信息的捕获,以适应不同大小的目标。从注意力图[图 5(c)和图 5(d)]中可以发现,模型显著提升了不同尺度下局部区域之间的注意力表达能力,缓解了固定 patch 划分导致的语义割裂问题。与基线方法相比,MultiFormer 生成的分割结果更接近真实值,并且产生了更精确的边界。

3.4.2 ISIC 2018 数据集

本文将 MultiFormer 网络模型与其他对比模型应用于 ISIC 2018 数据集,实验结果如表 2 所示。在 SETR、TransUNet、TransFuse 这 3 种基线方法上,通过引入 MultiFormer 模块,可以观察到,Dice

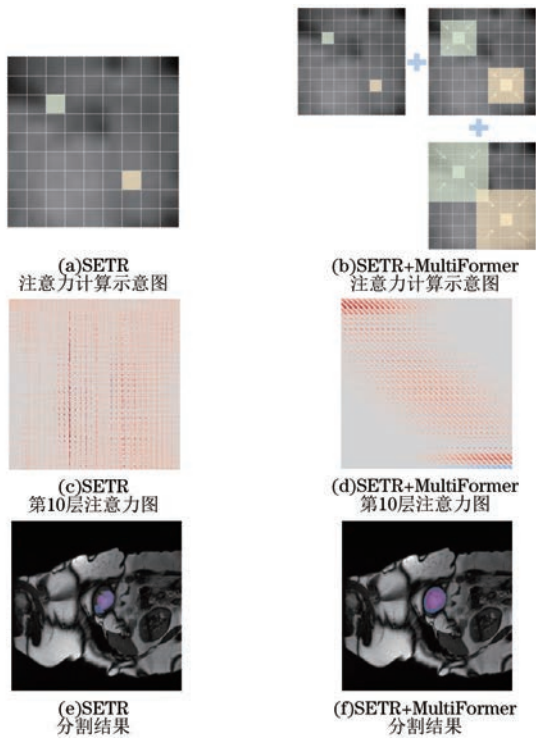


图 5 基线模型引入 MultiFormer 前后的对比分析

Fig.5 Comparative analysis of the baseline model before and after the introduction of MultiFormer

系数分别提升了 1.22、2.31 和 3.01 百分点,而 Hausdorff 距离则相应降低了 1.52、4.13 和 2.98 mm。MultiFormer 模块在各项评估指标中始

终优于其他方法,该模块不仅能够聚合多尺度语义信息,减轻语义割裂问题,还能灵活适应病变或器官的复杂边界形态,从而显著提升医学图像分割任务的精确度,并展现了其即插即用的广泛适用性。

表 2 在 ISIC 2018 数据集上的实验结果对比

Table 2 Comparison of experimental results on the ISIC 2018 dataset

网络模型	Dice 系数 ↑ / %	HD 95 ↓ / mm
SETR	88.13	24.60
SETR+ConvFormer	88.73	24.03
SETR+MultiFormer	89.35	23.08
TransUNet	86.26	27.66
TransUNet+ConvFormer	88.46	24.47
TransUNet+MultiFormer	88.57	23.53
TransFuse	86.28	25.08
TransFuse+ConvFormer	88.55	23.50
TransFuse+MultiFormer	89.29	22.10

对比模型的可视化分割结果如图 6 所示,以 SETR 和 TransUNet 这 2 个基线方法为例,MultiFormer 方法在捕捉病变复杂轮廓方面表现更为出色,更好地遵循了病变的实际边界。

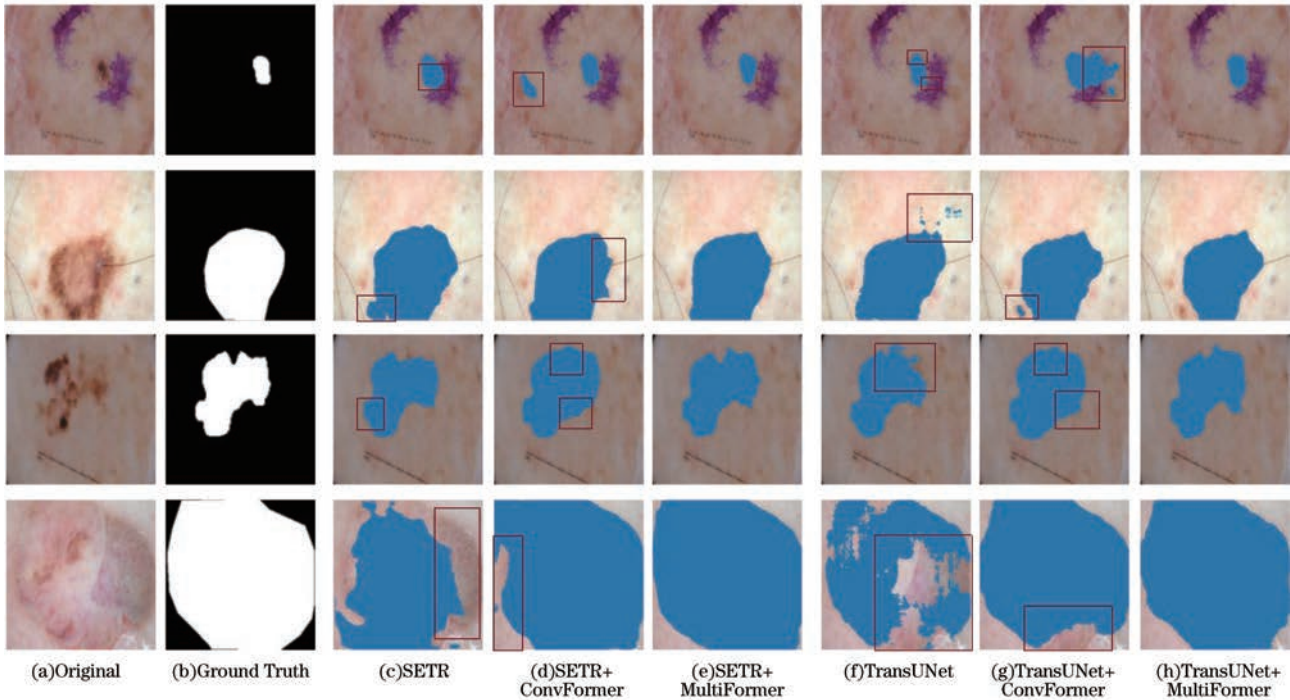


图 6 在 ISIC 2018 数据集上的可视化对比

Fig.6 Visual comparison on the ISIC 2018 dataset

3.5 消融实验及最佳参数选择实验

为了验证 MultiFormer 的有效性,在 ACDC 心脏

数据集和 ISIC 2018 皮肤病变数据集上进行消融实验,选取 SETR、TransUNet、TransFuse 这 3 种基线方

法,将 MultiFormer 模块分别集成到这些网络结构中。实验结果如表 3 所示,MultiFormer 在所有数据

集上的性能均优于原始基线方法,表明了多尺度卷积投影和可变形卷积(De-Conv)2 个模块组合的有效性。

表 3 消融实验结果

Table 3 Ablation experiment results

网络模型	ACDC		ISIC 2018	
	Dice 系数 $\uparrow$ /%	HD 95 $\downarrow$ /mm	Dice 系数 $\uparrow$ /%	HD 95 $\downarrow$ /mm
SETR	87.20	16.38	88.13	24.60
SETR+ MCP	90.74	13.64	89.26	23.20
SETR+De-Conv	90.69	13.72	89.13	23.42
SETR+MCP+De-Conv	90.83	13.51	89.35	23.08
TransUNet	90.80	14.58	86.26	27.66
TransUNet+MCP	91.73	13.34	88.48	23.93
TransUNet+De-Conv	91.68	13.45	88.45	24.12
TransUNet+MCP+De-Conv	91.86	13.13	88.57	23.53
TransFuse	89.10	14.98	86.28	25.08
TransFuse+MCP	91.80	13.67	89.06	22.21
TransFuse+De-Conv	91.14	13.82	88.93	22.42
TransFuse+MCP+De-Conv	91.40	13.55	89.29	22.10

为验证多尺度卷积投影层的有效性,首先对卷积核大小的选择进行实验。以 TransUNet 为基线,在 ACDC 数据集上评估 MultiFormer 中多尺度卷积投影层的性能,实验结果如表 4 所示, f 表示卷积核的大小。从表 4 可以看出,独立使用更大的卷积核虽然扩大了感受野,但未能显著提升分割性能。相反,当使用多尺度卷积时,核大小为 1、3、5、7 的卷积组合分割效果明显优于其他单个或多个组合,这表明了该卷积核组合的有效性。实验结果进一步揭示了多尺度卷积投影层在 Transformer 架构中的关键作用,不仅通过扩大感受野提升了模型的全局感知能力,还通过捕获不同尺度的特征信息增强了模型的适应性。

表 4 ACDC 数据集上卷积核大小选择实验结果

Table 4 The results of the experiment on selecting convolution kernel sizes on the ACDC dataset

卷积核大小	Dice 系数 $\uparrow$ /%	HD 95 $\downarrow$ /mm
Baseline	90.80	14.58
$f=1$	91.52	13.45
$f=3$	91.37	13.75
$f=\{1,3\}$	91.53	13.17
$f=\{3,5\}$	91.45	14.23
$f=\{1,3,5\}$	91.54	13.43
$f=\{1,3,5,7\}$	91.86	13.13

对于不同大小的卷积核,对其输出通道的权重进行消融实验,以 TransUNet 为基线,在 ACDC 数据集上对卷积核输出通道权重的选择进行评估,实验结果如表 5 所示,其中, α_1 、 α_2 、 α_3 、 α_4 分别代表卷

积核大小为 1、3、5、7 的输出通道在按通道拼接后对应的通道权重。

表 5 ACDC 数据集上卷积核输出通道权重选择实验结果

Table 5 The results of the experiment on selecting the output channel weights of convolutional kernels on the ACDC dataset

输出通道权重 ($\alpha_1, \alpha_2, \alpha_3, \alpha_4$)	Dice 系数 $\uparrow$ /%	HD 95 $\downarrow$ /mm
Baseline	90.80	14.58
(1/2, 1/4, 1/8, 1/8)	91.65	13.38
(1/4, 1/2, 1/8, 1/8)	91.55	13.51
(1/4, 1/4, 1/4, 1/4)	91.86	13.13

通过对比分析发现,与单独将核大小为 1 或 3 的卷积核的通道权重设置为 1/2 相比,当 4 种不同大小的卷积核通道权重均等分配为 1/4 时,模型取得了最佳的分割性能。实验结果表明,合理分配卷积核通道权重可以显著提高模型性能,平衡多尺度特征信息在提升医学图像分割效果中具有一定作用。

4 分析讨论

本文对 MultiFormer 模型的设计进行深入探讨,特别是在卷积设置、高斯距离图以及多尺度卷积投影操作等方面进行合理选择。首先,对于多尺度投影层中卷积的设置,为了减轻模型计算复杂度和减少参数量,尝试使用膨胀卷积实现核大小为 5、7 的卷积感受野、使用深度卷积实现 4 个尺度的卷积感受野以及使用核大小为 3 的小卷积核序列等价代替大卷积核的方法来替换普通卷积,在保持感受野一致的同时降低计算成本。以 TransUNet 为基线,

本文在 ACDC 数据集上进行了实验评估。如表 6 所示,使用普通卷积的方法取得了最优的分割效果,这证明了本文设计的多尺度卷积结构能够有效提取并融合不同尺度的特征信息,从而显著提升模型的

分割性能。使用深度卷积的方法实验效果次优,也展示了其在多尺度特征融合方面的潜力。这些实验结果为后续进一步优化 MultiFormer 模型提供了方向。

表 6 ACDC 数据集上卷积方式选择实验结果

Table 6 The results of the convolution mode selection experiment on the ACDC dataset

卷积设置	Dice 系数 $\uparrow$ / %				HD 95 $\downarrow$ / mm
	右心室	心肌	左心室	平均	平均
膨胀卷积	89.16	89.10	95.73	90.31	13.48
深度卷积	89.90	89.07	95.65	91.54	13.45
小卷积核序列	88.81	88.70	95.46	90.99	14.67
普通卷积	90.28	89.49	95.80	91.86	13.13

此外,由于高斯距离图 M 也能够影响感受野的大小,本文探讨使用高斯距离图 M 是否对模型性能有影响;在多尺度卷积投影方面,由于在 Transformer 中是通过将 Query 和 Key 进行相似度计算得到注意力分数,再与 Value 进行加权求和,将原始 Value 值映射到一个新的空间,因此,本文也对在 V 中是否使用多尺度卷积投影进行探讨。在参数设置方面,对于

高斯距离图 M 的使用选择,用 0 表示未使用 M ,用 1 表示使用 M 。在参数 V 的选择中,用 0 表示 V 使用单个卷积投影,用 1 表示使用多尺度卷积投影。其他参数设置相同,实验结果如表 7 所示。值得注意的是,在右心室分割任务中,未使用 M 且 V 使用单个卷积投影取得了较好的结果,而使用 M 且 V 使用多尺度卷积投影在其他方面表现出一致的性能提升。

表 7 ACDC 数据集上参数选择实验结果

Table 7 The results of parameter selection experiment on the ACDC dataset

参数选择		Dice 系数 $\uparrow$ / %				HD 95 $\downarrow$ / mm
M	V	右心室	心肌	左心室	平均	平均
0	0	89.91	89.24	95.52	91.55	13.66
0	1	89.28	89.02	95.59	91.30	13.53
1	0	89.66	89.05	95.60	91.44	13.30
1	1	89.50	89.45	95.72	91.56	13.28

5 结束语

本文提出一种多尺度自注意力 Transformer 架构 MultiFormer,旨在应对医学图像分割领域中的关键挑战。MultiFormer 通过尺度感知配置的连续卷积生成 patch 的方式保留图像空间相关性,同时通过多尺度卷积投影和可变形卷积实现多尺度特征聚合与对边界形态变化的灵活建模能力。实验结果表明,MultiFormer 在 ACDC 和 ISIC 2018 医学图像分割数据集上取得了显著的性能提升,尤其在处理形态多变的目标时表现出明显优势,这证明了 MultiFormer 在医学图像分割任务中的有效性和实用性。

然而,本文仍存在一些局限性需要进一步研究。例如,MultiFormer 的多尺度设计和卷积投影机制虽然提升了分割性能,但可能带来额外的计算开销。未来研究可以从以下 2 个方面展开:首先,探索轻量化设计以降低计算复杂度,使模型适用于资源受限

的场景;其次,将模型扩展至多模态任务或更复杂的医学影像分割场景,以验证其在更广泛领域中的应用潜力。

参考文献

- [1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. [2024-08-05]. <https://arxiv.org/abs/1706.03762?ref=andrealrosa.org>.
- [2] 石争浩,李成建,周亮,等. Transformer 驱动的图像分类研究进展[J]. 中国图象图形学报, 2023, 28(9): 2661-2692. SHI Z H, LI C J, ZHOU L, et al. Survey on Transformer for image classification[J]. Journal of Image and Graphics, 2023, 28(9): 2661-2692. (in Chinese)
- [3] 付苗苗,邓森磊,张德贤. 基于深度学习和 Transformer 的目标检测算法[J]. 计算机工程与应用, 2023, 59(1): 37-48. FU M M, DENG M L, ZHANG D X. Object detection algorithms based on deep learning and Transformer[J]. Computer Engineering and Applications, 2023, 59(1): 37-48. (in Chinese)
- [4] ZHENG S X, LU J C, ZHAO H S, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2021: 6877-

- 6886.
- [5] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. [2024-08-05]. <https://arxiv.org/abs/2010.11929>.
 - [6] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[EB/OL]. [2024-08-05]. <https://arxiv.org/abs/2012.12877>.
 - [7] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: hierarchical vision Transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Washington D. C., USA: IEEE Press, 2022: 9992-10002.
 - [8] SHI J B, TANG L F, GAO Z Y, et al. MG-Trans: Multi-scale Graph Transformer with information bottleneck for whole slide image classification[J]. IEEE Transactions on Medical Imaging, 2023, 42(12): 3871-3883.
 - [9] DING S S, WANG J, LI J C, et al. Multi-scale prototypical transformer for whole slide image classification[EB/OL]. [2024-08-05]. <https://arxiv.org/abs/2307.02308>.
 - [10] HAN K, XIAO A, WU E, et al. Transformer in Transformer[EB/OL]. [2024-08-05]. <https://arxiv.org/abs/2103.00112>.
 - [11] OTSU N. A threshold selection method from gray-level histograms[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1): 62-66.
 - [12] ADAMS R, BISCHOF L. Seeded region growing[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(6): 641-647.
 - [13] 张博. 基于边缘检测的细胞图像分割方法研究与实现[D]. 武汉: 武汉理工大学, 2006.
ZHANG B. Research and implement of cell image segmentation based on edge detect[D]. Wuhan: Wuhan University of Technology, 2006. (in Chinese)
 - [14] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Pattern Analysis and Machine Intelligence. Washington D. C., USA: IEEE Press, 2016: 640-651.
 - [15] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[EB/OL]. [2024-08-05]. <https://arxiv.org/abs/1505.04597>.
 - [16] LIN A L, CHEN B Z, XU J Y, et al. DS-TransUNet: Dual Swin Transformer U-Net for medical image segmentation[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 4005615.
 - [17] CHEN B Z, LIU Y S, ZHANG Z, et al. TransAttUnet: multi-level attention-guided U-Net with Transformer for medical image segmentation[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 8(1): 55-68.
 - [18] ZHANG Y D, LIU H Y, HU Q. TransFuse: fusing transformers and CNNs for medical image segmentation[EB/OL]. [2024-08-05]. <https://arxiv.org/abs/2102.08005>.
 - [19] CHEN J N, LU Y Y, YU Q H, et al. TransUNet: transformers make strong encoders for medical image segmentation[EB/OL]. [2024-08-05]. <https://arxiv.org/abs/2102.04306>.
 - [20] CAO H, WANG Y Y, CHEN J, et al. Swin-Unet: Unet-like pure Transformer for medical image segmentation[EB/OL]. [2024-08-05]. <https://arxiv.org/pdf/2105.05537>.
 - [21] LIU J, PASUMARTHI S, DUFFY B, et al. One model to synthesize them all: multi-contrast multi-scale Transformer for missing data imputation[J]. IEEE Transactions on Medical Imaging, 2023, 42(9): 2577-2591.
 - [22] ZHANG Y, LIU H, HU Q, et al. ConvFormer: plug-and-play CNN-style Transformers for improving medical image segmentation[EB/OL]. [2024-08-05]. https://link.springer.com/chapter/10.1007/978-3-031-43901-8_61.
 - [23] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision (ICCV). Washington D. C., USA: IEEE Press, 2017: 764-773.
 - [24] ACDC challenge[EB/OL]. [2024-08-05]. <https://www.creatis.insa-lyon.fr/Challenge/acdc/index.html>.
 - [25] ISIC 2018: skin lesion analysis towards melanoma detection[EB/OL]. [2024-08-05]. <https://challenge2018.isic-archive.com/>.

文字编辑 吴云芳
栏目编辑 赖玉玲