计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(4):141−148 [doi: 10.15888/j.cnki.csa.009010]

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

基于 CNN 与 Transformer 的医学图像分割①

王金祥[1,2], 付立军[1,2,3], 尹鹏滨[4], 李　旭[5]

[1](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[2](中国科学院大学, 北京 100049)
[3](山东大学 大数据技术与认知智能实验室, 济南 250100)
[4](中国人民解放军总医院, 北京 100039)
[5](中科智禾数字科技 (北京) 有限公司, 北京 101499)
通信作者: 付立军, E-mail: fu_lijun@ucas.ac.cn

摘　要: 医学图像对疾病的诊断、治疗和评估均有所帮助, 准确分割医学图像中的器官对于辅助医生的诊断具有重要的实际意义. 由于医学图像中各器官部位与周围组织的图像对比度低, 不同器官的边缘和形状也会存在很大差异, 从而增加了分割的难度. 针对这些问题, 本文提出了一种基于卷积神经网络和 Transformer 的医学图像语义分割网络, 有效提高了医学图像语义分割的精度. 特征提取部分使用 ResNet-50 网络结构, 在特征提取后使用 Transformer 模块来扩大感受野. 在上采样过程中加入多个跳跃连接层, 充分利用各阶段的特征提取信息, 来恢复至与输入图像相近的分辨率. 在胃肠道医学图像分割数据集上的实验结果证明本文的方法可以有效分割医学图像中的器官组织, 提升分割准确率.

关键词: 深度学习; 语义分割; 医学图像; U-Net; Transformer; 卷积神经网络 (CNN)

引用格式: 王金祥,付立军,尹鹏滨,李旭.基于 CNN 与 Transformer 的医学图像分割.计算机系统应用,2023,32(4):141–148. http://www.c-s-a.org.cn/1003-3254/9010.html

Medical Image Segmentation Based on CNN and Transformer

WANG Jin-Xiang[1,2], FU Li-Jun[1,2,3], YIN Peng-Bin[4], LI Xu[5]

[1](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](Laboratory of Big Data and Artificial Intelligence Technology, Shandong University, Jinan 250100, China)
[4](Chinese PLA General Hospital, Beijing 100039, China)
[5](Zhongke Zhihe Digital Technology (Beijing) Co. Ltd., Beijing 101499, China)

Abstract: Medical images are helpful for the diagnosis, treatment, and evaluation of diseases. Accurate segmentation of organs in medical images is of great practical significance to assist doctors in diagnosis. Due to the low contrast between organ parts and surrounding tissues in medical images, the edges and shapes of different organs are very different, which increases the difficulty of segmentation. To solve these problems, this study proposes a semantic segmentation network for medical images based on a convolutional neural network and Transformer, which effectively improves the accuracy of semantic segmentation of medical images. The feature extraction part uses a ResNet-50 network structure, and a Transformer module is employed to expand the receptive field after feature extraction. In the process of up-sampling, multiple skip connection layers are added, and the feature extraction information of each stage is fully utilized to make the resolution close to that of input images. The experimental results on the segmentation dataset of gastrointestinal medical images prove that the proposed method can effectively segment organs and tissues in medical images and improve the segmentation accuracy.

Key words: deep learning; semantic segmentation; medical image; U-Net; Transformer; convolutional neural network (CNN)

① 基金项目: 科技部重点研发计划 (2019YFC0840705)
收稿时间: 2022-08-23; 修改时间: 2022-09-27; 采用时间: 2022-09-30; csa 在线出版时间: 2022-12-23
CNKI 网络首发时间: 2022-12-27

1 引言

随着机器学习与深度学习的发展, 计算机视觉技术在医学图像分析中得到了广泛的应用. 图像分割是医学图像分析的重要组成部分, 特别是在计算机辅助诊断与图像引导的临床手术中发挥着重要作用[1].

现有的医学图像分割方法主要基于 U 形结构的全卷积神经网络 (FCNN). 独特的 U 型结构 U-Net[2], 由一个带有跳跃连接的对称编码器-解码器组成. 在编码器中, 采用一系列卷积层和连续下采样层来提取具有较大感受野的深度特征. 然后, 解码器将提取的深度特征上采样进行像素级语义预测, 在此过程中将不同尺度的高分辨率特征图信息通过跳跃连接融合, 来减少下采样造成的空间信息损失. U-Net 凭借其优秀的结构设计, 在各种医学影像的任务中取得了优异的成绩. 跟随着这一技术路线, 许多算法如 3D U-Net[3]、Res-UNet[4]、U-Net++[5] 和 UNet3+[6] 已被创造出来, 用于各种医学格式的图像分割. 这些基于 FCNN 的方法在心脏分割、器官分割和病变部位分割中的出色表现证明了 CNN 具有较强的提取特征的学习能力.

目前, 尽管基于 CNN 的方法在医学图像分割领域取得了优秀的表现, 但仍然不能完全满足医学实际应用中对分割精度的严格要求. 在医学图像分析中, 图像分割仍然是一个具有挑战性的任务. 由于卷积运算的固有局域性, 基于 CNN 的方法很难学习到全局语义信息和像素相对距离较远的语义信息[1]. 一些研究试图通过使用空洞卷积、自注意力机制[7] 和图像金字塔[8] 来解决这个问题. 然而, 这些方法在建立较远的依赖关系方面仍然有局限性. 最近, 受到 Transformer 在自然语言处理 (NLP) 领域取得巨大成功的启发[9], 研究人员试图将 Transformer 引入视觉领域[10]. 文献 [11] 提出了使用 Vision Transformer (ViT) 来进行图像识别任务. 用带有位置信息的二维图像块序列作为输入, 在大数据集上进行预训练, ViT 的性能与基于 CNN 的方法相当. 此外, 文献 [12] 还提出了 data-efficient image Transformer (DeiT), 使得 Transformer 可以在中等规模的数据集上训练, 并且结合知识蒸馏方法可以得到一个更稳健的 Transformer. 文献 [13] 设计了一个分层的 Swin Transformer, 将 Swin Transformer 作为视觉特征提取骨干网络, 在图像分类、目标检测和语义分割方面取得了明显的进步. ViT、DeiT 和 Swin Transformer 在图像识别方面的成功应用表现出了 Transformer 在视觉领域的应用潜力.

本文利用 Transformer 的优势进行 2D 医学图像分割, 模型是将 Transformer、ResNet 和 U-Net++思想于一体的语义分割网络. 与其他网络结构相似, 网络由编码器、解码器和跳跃连接组成. 编码器基于 ResNet-50 和 Vision Transformer 构建的, 输入的医学图像经过多层 CNN 特征提取结构后, 被分割成不重叠的图像块, 将分割的图像块组合成图像块序列输入到基于 Transformer 的编码器学习全局特征. 然后进入基于 CNN 的解码层对提取的多层次特征进行上采样, 并通过跳跃连接与下采样过程中的多尺度特征融合, 恢复特征图到近似输入图像的分辨率尺寸, 最后进行像素级的分割预测. 在 UW 胃肠道图像分割数据集上的实验表明, 该方法具有优秀的分割精度和泛化能力. 具体而言, 模型的贡献可以概括为: (1) 基于 CNN 和 Transformer, 构建了一个具有多层级跳跃连接的非对称编码器-解码器体系结构. 在编码器中实现了从局部到全局的自注意力机制. 在解码器中, 将全局特征上采样到输入分辨率, 进行像素级的分割预测; (2) 充分结合卷积神经网络 CNN 和 Transformer 结构的优势对医学图像进行分割预测; (3) 利用多层级跳跃连接结构, 充分利用各层级特征信息, 最终构造了一个基于 CNN 和 Transformer 的带有跳跃连接的 U 型编码器-解码器网络结构.

2 相关工作

早期的医学图像分割方法主要是基于轮廓和传统的基于机器学习的算法. 随着深度学习卷积神经网络的发展, U-Net 神经网络被提出用于医学图像分割. 由于其简单的 U 形结构和优越的性能, 各种类似 U-Net 的方法不断涌现, 如 Res-UNet、Dense-UNet[14]、U-Net++和 UNet3+. 并将其引入 3D 医学图像分割领域, 如 3D-UNet 和 V-Net[15]. 目前, 基于 CNN 的方法凭借其强大的表现能力在医学图像分割领域取得了巨大成功.

众所周知, 上下文信息是提升语义分割性能的关键因素, 而感受野的大小就基本决定了网络可以利用多少有用的信息. 理论上, 通过堆叠足够深的卷积层, 网络的感受野便能够覆盖到输入图像的全局领域, 然而 He 等人[16] 发现, 随着模型层数的增加, 识别精度不会提升反而会下降. Transformer 最初是为机器翻译任务而提出的[9]. 在自然语言处理领域, 基于 Transformer 的方法在各种任务中均取得了最优秀的成绩[17]. 在

Transformer 的成功引领下, 研究人员引入了一种开创性的方法 Vision Transformer (ViT) , 它实现了惊人的速度与精度均衡的图像识别任务. Transformer 不像卷积操作那样有固定且有限的感受野, 它的核心 self-attention 可以直接提取整体的感受野. 也正是这样丰富的感受野, 以及它在各项挑战赛中优异的性能, 吸引了很多图像领域的研究学者. 值得一提的是 Swin-UNet[18], 使用纯 Vision Transformer 结构来进行提取特征. 基于 U-Net 网络结构, Swin-UNet 实现了医学图像分割的优秀性能, 但 Swin-UNet 网络的分割结果相较于 CNN 方法的分割结果表现在分割的边缘较为粗糙.

近年来, 研究人员试图在 CNN 中引入自注意力机制来提高网络模型的性能. 文献 [7] 在 U 型结构中加入了带有附加注意门的跳跃连接, 进行医学图像分割. 然而, 这仍然是基于 CNN 的方法. 目前, 人们正在努力将 CNN 和 Transformer 结合起来, 打破 CNN 在医学图像分割中的主导地位[19]. 文献 [1] 将 Transformer 与 CNN 相结合, 构成一个用于二维医学图像分割的强编码器. 与文献 [1] 的工作相似, 文献 [20] 和文献 [21] 利用 Transformer 和 CNN 的互补性来提高模型的分割性能. 在本文中, 尝试进一步融合 CNN 与 Vision Transformer 模块, 构建一个具有多层级跳跃连接的 U 型编码器-解码器架构, 用于医学图像分割, 从而为 Transformer 在医学图像领域的发展进一步做出探索.

3 医学图像分割网络模型

给定一个分辨率为$H\times W$, 通道数为C的图像$X\in R^{H\times W\times C}$. 任务的目标是预测大小为$H\times W$的相应的像素级标签掩码图片. 最常见的方法是直接训练 CNN 网络, 首先将图像编码为高层次特征, 然后解码至完整图像分辨率. 与上述方法有所不同, 此网络通过在编码器中使用 Transformer 来引入自注意力机制, 接下来将详细介绍网络各部分的结构.

3.1 网络结构概述

对于图像分割, 最直接的上采样解决方法是简单地将编码的特征表示$Z_L\in R^{\frac{HW}{P^2}\times D}$上采样到全分辨率来预测输出, 式中$P$表示 Transformer 模块分割的图像块尺寸, D表示通道数. 但是为了恢复各个图像块的空间顺序, 首先要将编码特征序列的大小从$\frac{HW}{P^2}$重构为$\frac{H}{P}\times\frac{W}{P}$. 首先使用$1\times1$的卷积将特征通道大小减少到分割目标类别的数量, 然后将特征图直接使用双线性插值上采样到输入分辨率$H\times W$来预测最终的分割结果. 尽管 Transformer 与 CNN 上采样相结合已经获得了不错的性能, 但如上所述, 这种策略并不是 Transformer 在分割中的最佳使用, 因为$\frac{H}{P}\times\frac{W}{P}$通常比原始图像分辨率$H\times W$小得多, 因此不可避免地会导致低层次细节信息的丢失, 例如器官的形状和边界. 因此, 为了弥补这种信息损失, 采用级联跳跃上采样的方法来进一步提升准确率.

网络模型的整体结构如图 1 所示. 网络并没有使用纯 CNN 或者纯 Transformer 作为特征提取器, 而是采用了 CNN-Transformer 混合网络模型. 网络由编码器、解码器和跳跃连接组成. 对于编码器, 首先将图片输入 CNN 结构进行局部特征提取, CNN 特征提取部分使用 ResNet-50 网络结构. 为了将输入信息转化为序列编码, Transformer 部分首先将 CNN 下采样部分的特征图像块进行编码转换. 转换后的图像块序列经过 12 个 Transformer 模块, 来学习全局特征. 受到 U-Net++ 网络模型的启发, 设计了一个多层级的跳跃连接结构. 解码器部分由图像块序列扩展层、CNN 上采样层和跳跃连接模块组成. 通过跳跃连接将提取的上下文特征与编码器的多尺度特征相融合, 来弥补下采样造成的信息损失. 解码器部分首先将 Transformer 结构输出的结果进行位置恢复, 之后利用级联的上采样器将相邻维度的特征图多次进行 2 倍分辨率上采样. 最后利用一个扩展层进行 4 倍上采样, 将特征图恢复到与输入相近的分辨率$H\times W$, 然后对这些上采样的特征使用线性映射层得到像素级的分割预测结果, 将在下面详细阐述每个模块的构造及参数.

3.2 ResNet-50 编码器

首先对输入的图片使用 CNN 卷积神经网络进行下采样局部特征提取, 这部分使用残差网络 ResNet-50 作为特征编码, 共分为 3 个下采样阶段, 详细 CNN 下采样模块参数见表 1. 其中每一个模块由多个 bottleneck 结构重复叠加组成. Bottleneck 部分先通过一个 1×1 的卷积来减少通道数, 使得中间卷积的通道数缩小 4 倍, 中间的普通 3×3 卷积做完卷积后输出的通道数等于输入通道数, 第 3 个 1×1 的卷积用于恢复通道数, 使得输出通道数等于输入通道数. 这两个 1×1 的卷积有效减少了卷积的参数个数和计算量, bottleneck 的结构如图 2.

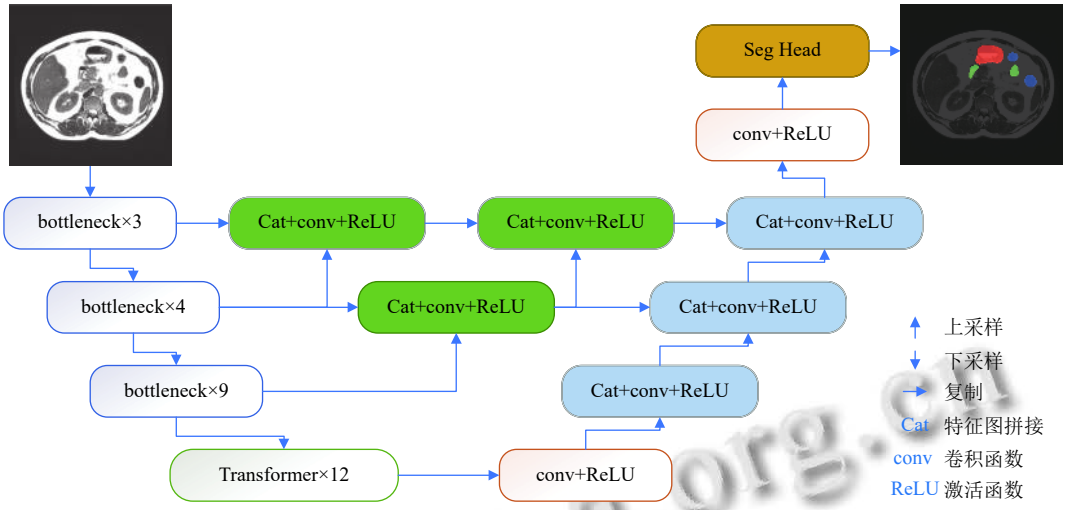


图 1 网络基本架构

表 1 ResNet-50 结构

卷积部分	输出尺寸	50层
卷积层	128×128	$7\times7, 64, \text{stride}2$
池化层	64×64	$3\times3, \text{max pool}, \text{stride}2$
第1阶段	64×64	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
第2阶段	32×32	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$
第3阶段	16×16	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times9$

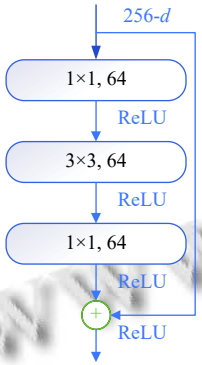


图 2 Bottleneck 结构

图 2 中的符号 d 表示特征图的通道数, 256-d 即表示输入的特征图通道数有 256 个.

3.3 Transformer 编码器

首先对输入的特征图x进行位置标记, 将其切分展开为一个 2D 图像块$\left\{x_p^i \in R^{P^2\cdot C} | i=1,\cdots,N\right\}$, 其中每个图像块的大小为$P\times P$, 图像块的数量为$N=\dfrac{HW}{P^2}$, 图像块的数量也是输入图像块序列长度.

使用可训练的线性映射将向量化的图像块x_p映射到一个 D 维编码空间. 为了对图像块的空间信息进行编码, 将特定的位置信息加入到图像块序列中来保留其位置信息, 具体方法如式 (1):

$$z_0=\left[x_p^1E;x_p^2E;\cdots;x_p^NE\right]+E_{\text{pos}} \tag{1}$$

其中, $E\in R^{(P^2\cdot C)\times D}$表示图像块的编码映射, $E_{\text{pos}}\in R^{N\times D}$表示位置信息编码.

Transformer 编码结构由 L 层多头自注意力机制 (MSA) 和多层感知机 (MLP) 组成. 其中, 多头自注意力机制即含有多个分支的 self-attention 模块, 每个分支即表示一个头, 而多层感知机即隐藏层和输出层都是全连接层的神经网络. 编码结构中第ℓ层的输出如式 (2) 和式 (3) 表示:

$$z_\ell'=MSA(LN(z_{\ell-1}))+z_{\ell-1} \tag{2}$$

$$z_\ell=MLP\left(LN\left(z_\ell'\right)\right)+z_\ell' \tag{3}$$

其中, $LN(\cdot)$表示层标准化操作, z_ℓ是编码的图像表示, Transformer 模块的结构如图 3 所示.

3.4 模型上采样

解码部分引入了一种级联上采样器 (CUP), 它由多个上采样及多层跳跃连接组合来解码隐藏特征, 来得到最终的分割掩码图. 在将隐藏特征$Z_L\in R^{\frac{HW}{P^2}\times D}$序列重构为$\dfrac{H}{P}\times\dfrac{W}{P}\times D$的后, 通过级联多个上采样块和跳跃连接来实现上采样, 每个上采样块依次由特征图拼接、卷积函数和 ReLU 激活函数组成, 以达到从$\dfrac{H}{P}\times\dfrac{W}{P}$恢复到$H\times W$的完整分辨率.

与 U-Net 相似, 跳跃式连接用于将编码器的多尺度特征与上采样特征进行融合. 将浅层特征和深层特征拼接在一起, 来减少下采样造成的空间信息损失. 但是第 1 个阶段下采样仅提取了图像的浅层特征, 直接与上采样的最后一层特征融合会丢失很多信息, 所以加入了密集的跳跃连接, 将 U-Net 网络架构的中心填满, 使水平方向上的每一尺度之间都有连接. 上采样的方法有很多, 本文中使用双线性插值上采样方法. 将两个特征图进行拼接后使用 1×1 卷积进行通道数减半操作, 这个操作的目的是使得特征图的通道数与拼接前的通道数相同, 具体操作如图 4 所示.

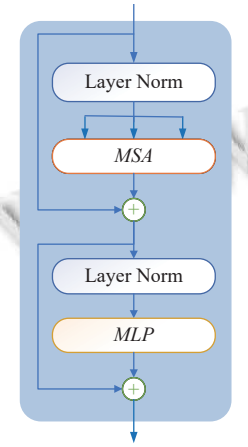


图 3 Transformer 编码结构

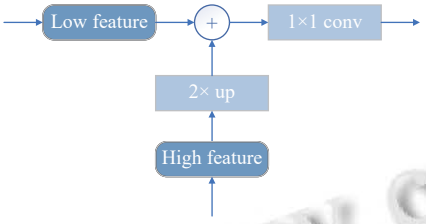


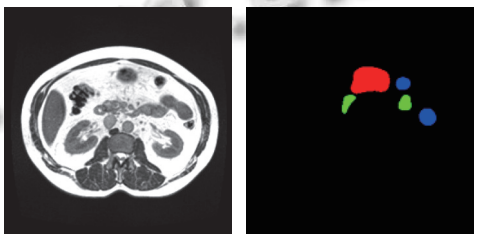
图 4 特征拼接结构

4 实验分析

实验中使用 UW-Madison GI Tract Image Segmentation 胃肠道医学图像分割数据集 (https://www.kaggle.com/competitions/uw-madison-gi-tract-image-segmentation/data) 对本文算法进行详细的评估, 并且和近几年优秀的医学图像分割模型进行比较, 最后对所提出的模型的多个模块部分进行消融实验.

4.1 数据集介绍

实验使用的 UW-Madison GI Tract Image Segmentation 胃肠道医学图像数据集来源于 kaggle 的胃肠道图像分割比赛. 如图 5(a) 所示, 原始数据集的图片存在肉眼完全看不到的情况, 肉眼看不到并不是图片中不存在器官, 肉眼是否可见与图片中有无器官并无直接关联, 通过将图像的全局亮度调高, 可以看到图像的各个部位. 数据集共包含 85 个病例的 CT 扫描图片, 每个病人包含多个时期阶段的扫描结果, 每个时期阶段包含 144 张图片, 图片尺寸不统一, 大多为266×266像素, 每个像素对应物理尺寸为 3 mm, 数据集共计包含 38 496 张图片. 在以下实验中将此数据集按 8:1:1 的比例分为训练集、验证集和测试集.



(a) 数据集图片 (b) 标签图片

图 5 数据集样本

此数据集官方所提供的图像为 16 位灰度的 PNG 格式, 标签以 RLE 编码的掩码形式提供, 标签文件不可将其直接作为网络模型的输入. 因此, 首先需要将标签文件转换为适合网络模型输入的格式, 实验中将标签文件转换为与图像格式相同的 PNG 格式. 如图 5(b), 此为转换为 PNG 格式后的标签图片.

4.2 实验环境及参数设置

实验环境使用的深度学习框架为 PyTorch 1.8.2, 开发环境为 Python 3.8.13, 操作系统为 Ubuntu 18.04.3 LTS, GPU 选用的是一张 GeForce RTX 3090, 显存为 24 GB. 以下所有的实验均使用此实验环境进行.

实验详细参数设置如下: 首先将原始图片转换为同一尺寸, 使网络模型的输入大小为128×128, 批大小为 160, 迭代次数为 100, 损失函数使用交叉熵损失函数, Transformer 模块的图像块尺寸设置为 16, 使用 SGD 优化且初始学习率为 0.01, 动量为 0.9, 权重衰减系数为 0.000 1, 如式 (4) 所示:

$$W \leftarrow W - \eta \frac{\partial L}{\partial W} \tag{4}$$

其中, W表示需要更新的权重参数, 损失函数关于W的梯度记为$\frac{\partial L}{\partial W}$, η表示学习率.

4.3 评价指标

为了评估图像语义分割方法的性能, 需要从语义分割

的准确性来进行衡量算法模型的性能, 性能评价指标使用 Dice 相似系数 (Dice similarity coefficient, DSC) 和 Hausdorff_95 (HD95). 其中, Dice 系数对分割的内部填充比较敏感, 而 Hausdorff_95 对分割出的边界比较敏感.

Dice 相似系数 (DSC) 是一种集合相似度度量函数, 通常用于计算两个样本的相似度, 取值范围为 [0, 1], 分割结果最好时值为 1, 最差时值为 0, 如式 (5) 所示:

$$S = \frac{2|X \cap Y|}{|X| + |Y|} \tag{5}$$

其中, X表示真实的分割标签, Y表示网络模型所输出的预测的结果, $|X \cap Y|$表示X和Y之间的交集; $|X|$和$|Y|$分别表示X和Y的元素个数. 其中分子的系数 2 是因为分母存在重复计算X和Y之间的共同元素的原因.

HD95 计算两个集合之间的距离, 值越小, 代表两个集合的相似度越高, 故数值越小越好, 如式 (6) 所示. 最后的值乘以 95%, 目的是为了消除离群值的一个非常小的子集的影响.

$$d_H(X,Y) = \max\left\{\max_{x\in X}\min_{y\in Y}(x,y), \max_{y\in Y}\min_{x\in X}(x,y)\right\} \tag{6}$$

其中, X表示真实的分割标签, Y表示网络模型所输出的预测结果, max(·)表示取最大值, min(·)表示取最小值.

4.4 与其他网络模型的对比实验

表 2 和表 3 展示了所提出的网络模型在 UW-Madison GI Tract Image Segmentation 胃肠道医学图像分割数据集上与目前各类型优秀方法的比较. 所比较的方法包括 R50 UNet、R50 Att-UNet 、U-Net++、TransUNet 和 Swin-UNet. 实验结果表明, 所提出的方法获得了最佳的分割精度, 其分割精度 DSC 达到 84.54%.

在分割结果的 DSC 上可以看出 U-Net++网络的分割结果准确率高于除本文模型之外的其他对比方法, R50 UNet 的分割结果准确率低于其他对比方法. 但是, 通过观察图 6 可以发现, U-Net++的分割结果似乎看起来是最差的, 而 R50 UNet 的结果看起来似乎比其他对比方法要好. 根据上述评价指标 DSC 可以推断, 这是因为 U-Net++产生了欠分割, 导致预测的结果与真实值的和变小, 导致 DSC 分母偏小, 进而导致准确率相对高一些. 同理, R50 UNet 是产生了过分割, 导致预测的结果与真实值的和偏大, 分母偏大, 而分子中的交集并未增大, 所以 DSC 准确率偏低. 而本文的方法的分割效果则处于二者之间, 所以准确率与分割效果均优于上述两个网络. 同时, 本文模型的参数量与 TransUNet 网络参数量相当, 但网络 DSC 准确率远高于 TransUNet, 且在 HD95 评价指标中此方法的数值均小于对比模型, 进一步证明了本文模型的有效性.

表 2 测试集对比结果 (DSC) (%)

方法	平均DSC	大肠	小肠	胃
R50 UNet	76.62	70.49	78.73	80.65
R50 Att-UNet	77.51	71.28	79.43	81.82
TransUNet	77.16	71.10	78.47	81.90
Swin-UNet	81.44	74.59	83.79	85.95
U-Net++	82.38	76.22	86.39	84.51
本文模型	**84.54**	**79.88**	**87.19**	**86.54**

表 3 测试集对比结果 (HD95) (mm)

方法	平均HD95	大肠	小肠	胃
R50 UNet	4.34	6.36	5.01	4.35
R50 Att-UNet	6.31	8.51	8.53	**1.88**
TransUNet	5.29	7.50	6.28	2.08
Swin-UNet	6.47	8.95	8.47	1.99
U-Net++	7.67	11.90	8.94	2.16
本文模型	**3.86**	**5.30**	**4.67**	4.59

4.5 各个模块的消融实验

为了探究不同因素对模型性能的影响, 又继续在 UW-Madison GI Tract Image Segmentation 胃肠道医学图像分割数据集上进行了各个模块的消融实验. 分别对模型的输入图像分辨率、模型中不同数量的 Transformer 层以及 Transformer 层中图像块的不同尺寸和序列长度进行详细实验对比.

网络模型的默认输入分辨率为128×128, 在实验中也对高分辨率256×256进行了实验, 详细实验结果见表 4. 在使用256×256作为模型的输入时, Transformer 部分保持相同的图像块尺寸, 这导致了 Transformer 的序列长度增大了 4 倍. 结果表明, 增加的序列长度提升了模型的稳健性. 分辨率从128×128增大到256×256使模型的准确率提升了 1.59%, 但这也伴随着更大的计算时间成本. 因此, 考虑到计算时间成本, 本文所有的实验都采用默认的128×128来验证模型的有效性.

表 4 不同输入分辨率的消融实验

分辨率	平均DSC (%)	大肠 (%)	小肠 (%)	胃 (%)
128×128	79.13	73.38	81.18	82.83
256×256	**82.35**	**78.17**	**85.74**	**83.15**

为了验证模型的有效性, 继续对不同数量的 Transformer 层进行了消融实验, 共进行了 3 种不同数量 Transformer 层的影响, 分别为 2 层、8 层、和 12 层, 详细结果见表 5. 实验过程中除 Transformer 层的数量变化外, 其他模型参数均保持不变. 实验结果表明, 随着 Transformer 层数的增加, 模型的准确率也随之增加, 进一步验证了 Transformer 结构的有效性.

表 5 Transformer 层数量的消融实验

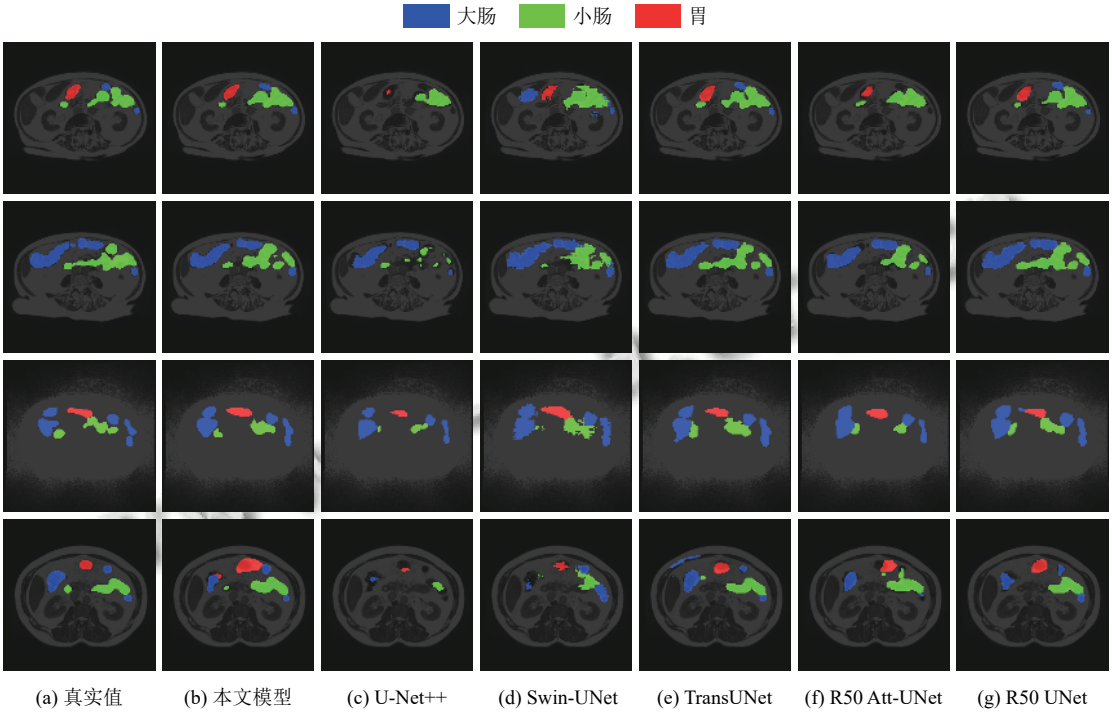
层数	平均DSC (%)	大肠 (%)	小肠 (%)	胃 (%)
2	81.73	76.11	84.64	84.45
8	82.71	76.08	86.72	85.33
12	**84.54**	**79.88**	**87.19**	**86.54**

实验中又进一步研究了 Transformer 层中的图像块尺寸 (patch size) 对模型效果的影响. 结果汇总在表 6 中. 从实验结果可以观察到, 图像块的尺寸越小, 分割性能越好. 可以注意到, Transformer 的序列长度与图像块大小成反比, 比如图像块大小为 16 的序列长度为 64, 而图像块大小为 32 的序列长度仅为 16. 为了更长的序列长度, 通过减小图像块的大小, 模型表现出了更强大的分割准确率, 这是因为 Transformer 对每个元素之间的更复杂的依赖关系进行了编码学习, 使得分割准确度更高. 参考 ViT 中的设置, 本文中的其他相关实验均使用 16×16 的默认图像块尺寸.

如图 6 所示, 在 UW-Madison GI Tract Image Segmentation 胃肠道医学图像分割数据集上进行了定性比较. 可以看出: (1) 基于纯 CNN 的方法 U-Net++、R50 UNet 和 R50 Att-UNet 更容易对器官进行过分割或欠分割. 而基于纯 Transformer 的 Swin-UNet 的分割结果相比于其他网络模型的分割结果相对边缘粗糙. 这表明结合 CNN 与 Transformer 的模型, 如 TransUNet 和本文的模型对全局上下文进行特征学习和语义分割的能力更强. (2) 第 1 行的结果表明, 与其他方法相比, 本文提出的模型预测的假阳性更少, 这表明在抑制噪声方面, 此方法比其他方法更有优势. (3) 在第 3 行中, 正确预测了大肠和小肠的位置, 而 TransUNet 则出现了过分割现象, U-Net++则出现了欠分割现象. 这些对比结果表明, 本文所提出的网络模型能够进行更精细的分割各器官, 并保留详细的边缘形状信息. 这进一步证明了本文模型方法的有效性.

表 6 图像块尺寸及序列长度消融实验

尺寸	序列长度	平均DSC (%)	大肠 (%)	小肠 (%)	胃 (%)
32	16	82.89	76.84	86.98	84.85
16	64	83.38	77.22	87.39	85.51
8	256	**83.73**	**77.64**	**87.86**	**85.69**



(a) 真实值 (b) 本文模型 (c) U-Net++ (d) Swin-UNet (e) TransUNet (f) R50 Att-UNet (g) R50 UNet

图 6 对不同方法进行定性比较

5 结论与展望

本文提出将 CNN 与 Transformer 并结合多层级跳跃连接结构应用于医学图像分割. 不仅通过将图像特征作为序列来编码学习全局信息, 而且结合 U 形结构和多层级跳跃连接结构来融合多尺度特征, 这种设计充分利用了 CNN 各层次的特征信息. 本文网络模型通过 ResNet-50 结构对图像进行 CNN 提取特征, 然后利用 Vision Transformer 结构提取全局空间特征, 加强器

官各部分内在的关联性信息提取. 在解码器中通过融合不同尺度的器官特征, 使网络模型保留更多的细节和边缘信息. 下一步的研究工作将进一步改进 Transformer 模块, 进一步提升医学图像分割的准确率.

参考文献

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

2 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. [doi: 10.1007/978-3-319-24574-4_28]

3 Çiçek Ö, Abdulkadir A, Lienkamp SS, *et al*. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. Proceedings of the 19th International Conference on Medical Image Computing and Computer-assisted Intervention. Athens: Springer, 2016. 424–432. [doi: 10.1007/978-3-319-46723-8_49]

4 Xiao X, Lian S, Luo ZM, *et al*. Weighted Res-UNet for high-quality retina vessel segmentation. Proceedings of the 9th International Conference on Information Technology in Medicine and Education. Hangzhou: IEEE, 2019. 327–331. [doi: 10.1109/ITME.2018.00080]

5 Zhou ZW, Rahman Siddiquee MM, Tajbakhsh N, *et al*. U-Net++: 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]

6 Huang HM, Lin LF, Tong RF, *et al*. UNet3+: A full-scale connected UNet for medical image segmentation. Proceedings of the 2020 International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020. 1055–1059. [doi: 10.1109/ICASSP40776.2020.9053405]

7 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]

8 Zhao HS, Shi JP, Qi XJ, *et al*. Pyramid scene parsing network. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2019. 6230–6239. [doi: 10.1109/CVPR.2017.660]

9 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.

10 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. [doi: 10.1007/978-3-030-58452-8_13]

11 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16×16 words: Transformers for image recognition at scale. Proceeding of the 9th International Conference on Learning Representations. OpenReview.net, 2021. 1–21.

12 Touvron H, Cord M, Douze M, *et al*. Training data-efficient image transformers & distillation through attention. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 10347–10357.

13 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 (ICCV). Montreal: IEEE, 2021. 9992–10002. [doi: 10.1109/ICCV48922.2021.00986]

14 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]

15 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. [doi: 10.1109/3DV.2016.79]

16 He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

17 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186.

18 Cao H, Wang YY, Chen J, *et al*. Swin-UNet: UNet-like pure transformer for medical image segmentation. arXiv:2105.05537, 2021.

19 Hatamizadeh A, Tang YC, Nath V, *et al*. UNETR: Transformers for 3D medical image segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2022. 1748–1758. [doi: 10.1109/WACV51458.2022.00181]

20 Valanarasu JMJ, Oza P, Hacihaliloglu I, *et al*. Medical transformer: Gated axial-attention for medical image segmentation. Proceedings of the 24th International Conference on Medical Image Computing and Computer-assisted Intervention. Strasbourg: Springer, 2021. 36–46. [doi: 10.1007/978-3-030-87193-2_4]

21 Zhang YD, Liu HY, Hu Q. TransFuse: Fusing transformers and CNNs for medical image segmentation. Proceedings of the 24th International Conference on Medical Image Computing and Computer-assisted Intervention. Strasbourg: Springer, 2021. 14–24. [doi: 10.1007/978-3-030-87193-2_2]

(校对责编: 孙君艳)