聚知刊出版社
JZK publishing

基于 Transformer 在腹部多器官图像分割的研究进展

康运成

湖南工业大学生物与医学工程学院，湖南省株洲市，412007；

摘要：腹部多器官图像精准分割在现代临床影像检查、精准诊断和治疗规划中意义至关重要。CNN 架构中存在的局部感受野和固有归纳偏置局限，限制其对图像中远程依赖关系的有效建模。近年来，Transformer 架构依赖其对全局信息的捕获能力，有助于建模长距离的依赖关系并挖掘语义信息，在生物医学图像分割领域展示出卓越的性能和巨大潜力。在此，对 Transformer 架构的组成及其在腹部多器官图像分割中的应用进行了全面综述，并对 Transformer 模型在分割任务中存在的局限不足进行了概括总结，最后对其未来发展趋势及优化路径进行了探讨展望。

关键词：深度学习；Transformer；腹部多器官图像分割；卷积神经网络

DOI：10.69979/3029-2808.25.12.049

引言

图像分割是指将图像划分为若干区域或片段，以实现对感兴趣对象或区域的识别与提取的过程[1]。在医学图像分析中，分割技术被视为基础且关键的环节，其核心目标是从 CT 或 MRI 等医学成像中，准确定位并分割出特定器官或病变区域的像素[2][3]。其中，腹部多器官的精准分割由于涉及器官种类多、尺度差异大、边界模糊等复杂因素，一直是医学图像分割领域的研究重点和难点。

传统的医学图像分割方法多依赖于人工设计的图像处理技术与数学模型，如阈值法、图割法、区域生长法等，或依靠医生手动提取图像特征。这类方法往往依赖于人为设定的特征、阈值或初始种子，分割结果容易受到噪声干扰和图像质量波动的影响，鲁棒性不足，适应性差，难以推广到复杂或多样化的器官结构，尤其在大规模图像处理任务中，效率与精度常常难以兼顾。此外，手工特征提取过程不仅需要操作者具备高度的专业知识和丰富经验，而且容易受到主观判断的影响，限制了自动化程度与临床可扩展性。

近年来，随着深度学习的发展，基于卷积神经网络(CNN)的模型在医学图像分割中取得了显著进展，尤以 U-Net 及其变体为代表。然而，CNN 固有的局部感受野限制了其对图像中远程依赖关系和全局语义信息的建模能力，需通过堆叠更深的网络层次来扩展感受野，但这也带来了特征冗余、梯度消失和计算资源增加等问题。

为解决这一瓶颈，研究者将最初应用于自然语言处理的 Transformer[4]引入计算机视觉任务，凭借其多头自注意力机制，能够高效建模长距离依赖，捕捉图像中的全局上下文关系，展现出在语义理解与结构建模方面的独特优势。2020 年，Vision Transformer(VIT)[5]首次将纯 Transformer 架构引入图像分类任务，随后研究者开始探索其在医学图像分割中的潜力。例如，Swin-UNet[6]通过引入分层 Transformer 模块和滑动窗口注意力机制，在腹部多器官分割中展现出良好的边缘结构还原能力；MISSFormer[7]通过上下文桥接模块增强多尺度语义信息的表达效果，有效提升了对器官间结构差异的感知能力等。

针对医学影像数据稀缺、标注困难等问题，越来越多的研究采用 CNN 与 Transformer 融合的混合架构，以结合 CNN 在局部特征提取方面的高效性与 Transformer 在建模全局依赖关系方面的优越性，从而提升分割性能与模型泛化能力。为了更深层次地探讨 Transformer 深度学习模型在腹部多器官图像分割、分类等任务发挥的潜力，本文将介绍其技术演进、代表性模型、应用成效与存在的挑战，探讨 Transformer 在未来腹部多器官图像分割中的发展潜力。

1 Transformer 概述

Transformer 架构主要由多头自注意力机制、前馈神经网络层、残差连接以及层归一化等关键模块组成。其中，自注意力机制是该架构的核心。其基本思想是将输入序列中的每个位置表示为查询(Query, Q)，通过计算 Q 与序列中各位置对应的键(Key, K)之间的相似度，

进而加权求和对应的值(Value, V)，以捕捉序列内部的空间依赖关系与语义联系。

多头自注意力机制通过并行地使用多个独立的注意力头，对输入进行多维度的特征关注，从而在保持较高计算效率的同时，提升模型对不同语义信息的表达能力。具体而言，注意力机制的计算过程包括以下步骤：首先对查询矩阵 Q 与键矩阵 K 进行点积操作，并根据维度缩放因子进行归一化处理；随后通过 softmax 函数获取注意力权重；最后将该权重与值矩阵 V 相乘，得到最终的注意力输出。

在编码器和解码器结构中，Transformer 为每个位置的隐藏状态引入独立的前馈网络，该模块通常由两个全连接层和一个非线性激活函数构成，目的是提升模型的非线性表达能力与建模深度。

为了缓解深层网络中常见的梯度消失问题，Transformer 引入了残差连接机制，它能够将输入信息直接传递至后续模块，增强了梯度流动性，有助于模型的稳定训练。此外，层归一化作为标准化技术之一，有效抑制了训练过程中因特征分布变化带来的不稳定现象，从而提升了模型在处理长序列和复杂任务时的鲁棒性与泛化能力。

2 Transformer 模型在腹部多器官图像分割中的应用

2.1 Swin Transformer 在腹部多器官图像分割中的应用

为提升计算效率并克服ViT在小规模数据集上的表现不足，Swin Transformer[6]应运而生。该模型采用窗口多头自注意力(W-MSA)机制，在局部窗口内计算注意力，显著降低了高分辨率图像处理时的计算负担；同时引入滑动窗口多头自注意力(SW-MSA)，通过窗口间的位移操作实现跨区域信息交互，从而在保持效率的同时增强了特征的上下文建模能力。此外，Swin Transformer 还引入分层特征图结构，通过补丁合并操作逐步缩小空间维度，以构建多尺度特征表示，更好地适配图像中不同尺度的视觉对象。这些结构改进使 Swin Transformer 相较于传统 ViT 更加高效，并在各类图像分割任务中表现出更优性能。

在腹部多器官图像分割任务中，Swin Transformer 展现出良好的适用性。由于腹部器官形态复杂、尺度差异显著且边界模糊，传统卷积模型难以准确建模其空间结构关系。Swin Transformer 通过其强大的局部-全局建模能力和分层特征表达机制，有效提升了对小器官(如胆囊)与边缘模糊区域的分割精度。

2.2 TransUNet 在腹部多器官图像分割中的应用

尽管 ViT 拥有较强的全局建模能力，Swin Transformer 在计算效率上也有所提升，但这两者在捕捉空间细节与建模远程依赖方面仍存在一定局限。为此，Chen 等人[8]提出了 TransUNet，首次将 Transformer 融入 U-Net 架构，应用于医学图像分割任务。该模型通过 ViT 编码器提取图像的全局语义信息，同时借助 U-Net 的跳跃连接机制保留浅层的空间细节，实现了局部与全局特征的有机融合。TransUNet 在多个分割任务中表现出较高的准确性，尤其在处理形态复杂、边界模糊的目标区域时效果显著。在腹部多器官图像分割中，TransUNet 在 Synapse、BTCV 等数据集上获得了优于传统 CNN 模型的性能，特别是在肝脏、肾脏等大器官的分割上展现出较强的稳定性。TransUNet 为将 Transformer 引入医学图像分割提供了有效范式，在腹部多器官任务中具有较强的参考价值。

2.3 Swin-Unet 在腹部多器官图像分割中的应用

Swin-Unet[9]是首个完整基于 Transformer 构建的 U 形网络结构，其采用 Swin Transformer 作为编码器核心，利用窗口注意力(W-MSA)与滑动窗口注意力(SW-MSA)机制进行特征提取，不仅保留了 Transformer 在建模全局依赖方面的优势，同时有效控制了计算资源消耗。该模型结合了 U-Net 的跳跃连接设计，使浅层局部细节信息得以传递与融合，进一步增强了分割边界的准确性。与 TransUNet 相比，Swin-Unet 采用分层特征提取策略，并通过 Patch Expanding 操作在解码阶段逐步恢复图像分辨率，在分割精度与计算效率之间取得了更优的平衡，尤其适用于高分辨率医学图像的处理。在腹部多器官图像分割任务中，Swin-Unet 在 Synapse、BTCV 等腹部 CT 数据集上表现出良好的稳定性和泛化能力，对于肝脏、脾脏等大器官的分割精度显著提升。同时，由于其多尺度建模能力和局部注意力机制，小器官如胰腺、胆囊等的识别效果也较传统 CNN 网络有所改善。不过，该模型仍存在一定局限，例如对于极小目标仍依赖细粒度解码策略，且模型结构复杂，对硬件资源有一定要求。总体而言，Swin-Unet 在融合全局建模与局部特征提取方面

提供了一种高效路径，已成为腹部多器官图像分割研究中的关键代表模型之一。

3 总结与展望

本文系统梳理了 Transformer 及其衍生架构与经典网络融合在医学图像分割中的研究进展。首先概述了医学影像分割的临床价值以及传统分割方法在精度和效率上的不足，进而详细分析了 Transformer 的核心组件，包括自注意力机制、多头注意力等，并着重讨论了 Swin Transformer、TransUNet 和 Swin-Unet 等代表性融合模型。在此基础上，进一步聚焦 Transformer 在腹部多器官图像分割中的实际应用与性能表现。尽管现有研究显著提升了分割准确度与模型表征能力，该领域仍面临高昂计算资源需求以及标签稀缺条件下的模型适应等关键挑战，尚存在诸多待深入探索的方向：

（1）算法优化与模型轻量化：尽管 Transformer 模型在医学图像分割任务中表现出卓越的性能，但其高昂的计算成本与内存占用仍是阻碍临床实际部署的主要瓶颈。未来的研究方向可集中于开发更高效的模型结构，例如采用轻量化 Transformer 设计(如参数共享机制和稀疏注意力)、引入动态位置编码方案以及嵌入局部感受野增强模块，从而显著降低计算资源消耗并提升推理速度。此外，融合模型剪枝、量化技术等压缩策略，可在维持高分割精度的同时，大幅缩减模型参数量与体积，进一步提高在实际医疗场景中的响应效率与适用性。

（2）数据集构建与标注：高质量标注数据集是训练高性能深度学习模型的基石。然而，医学图像标注高度依赖专业医师的先验知识，过程耗时费力且成本昂贵，标注一致性和质量控制亦面临严峻挑战。为降低标注负担并提升数据可用性，未来可重点发展高效标注策略，例如融合人机交互的半自动标注工具（如基于配准的标签传播和交互式分割模型）、基于弱监督学习（如图像级标签或边界框提示）的方法，以及利用预训练模型进行迁移学习和跨域适应。此外，构建更大规模、多中心、多模态的医学图像数据集，涵盖不同设备、人群和疾病谱，对提升模型的泛化能力和临床适用性具有重要意义。

参考文献

[1]MINAEE S, BOYKOV Y, PORIKLI F, et al. Image segmentation using deep learning: A survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(7): 3523-3542.

[2]GOLAN R, JACOB C, DENZINGER J. Lung nodule detection in CT images using deep convolutional neural networks[C]//2016 international joint conference on neural networks (IJCNN). IEEE, 2016: 243-250.

[3]CHRIST P F, ETTLINGER F, GRÜN F, et al. Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks[J]. arXiv preprint arXiv:1702.05970, 2017.

[4]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[5]Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arxiv preprint arxiv:2010.11929, 2020.

[6]Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[7]Huang X, Deng Z, Li D, et al. Missformer: An effective medical image segmentation transformer[J]. arxiv preprint arxiv:2109.07162, 2021.

[8]Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation[J]. arxiv preprint arxiv:2102.04306, 2021.

[9]Cao H, Wang Y, Chen J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 205-218.

作者简介：康运成（1999 年 11 月—），性别：男，民族：汉，籍贯：湖南省娄底市，职务/职称，学历：硕士研究生，单位：湖南工业大学生物与医学工程学院，研究方向：医学图像分割。