激光与光电子学进展 先进成像

从 U-Net 到 Transformer：混合模型在医学图像分割中的应用进展

尹艺晓，马金刚*，张文凯，姜良
山东中医药大学医学信息工程学院，山东 济南 250355

摘要 医学图像分割能够准确快速地提取图像中感兴趣结构，在医学影像诊断、疾病分析、手术规划等领域具有重要应用价值。传统的医学图像分割方法通常依赖于边缘检测、模板匹配技术、统计形状模型、活动轮廓和传统机器学习技术。然而，由于图像存在模糊、噪声、对比度低等问题，传统方法的准确性和鲁棒性受到限制。深度学习方法通过学习数据的不同抽象层次来逐渐提取特征，相较于传统方法具有高精度、自适应性强和可扩展性强等优势。为更好地进行医学图像分割辅助诊断研究，本文综述了卷积神经网络、Transformer 以及 U-Net 和 Transformer 混合结构在医学图像分割中的应用情况，并对它们进行了综合对比分析。通过可视化结果和图像评估指标，证实了这些模型在医学图像分割中的可行性。最后总结目前研究中存在的问题，并对未来的研究方向进行展望。

关键词 深度学习；医学图像分割；U-Net；Transformer；混合模型
中图分类号 O436 **文献标志码** A **DOI**：10.3788/LOP240875

From U-Net to Transformer: Progress in the Application of Hybrid Models in Medical Image Segmentation

Yin Yixiao, Ma Jingang*, Zhang Wenkai, Jiang Liang
School of Medical Information Engineering, Shandong University of Traditional Chinese Medicine, Jinan 250355, *Shandong China*

Abstract Medical image segmentation can accurately and quickly extract structures of interest in images and has major application value in medical imaging diagnosis, disease analysis, surgical planning, and other fields. Traditional medical image segmentation methods typically rely on edge detection, template matching techniques, statistical shape models, active contours, and traditional machine learning techniques. However, due to problems such as blur, noise, and low contrast in images, the accuracy and robustness of traditional methods are limited. Deep learning methods gradually extract features by learning different levels of abstraction of data. Compared with traditional methods, they have the advantages of high accuracy, strong adaptability, and strong scalability. To better conduct research on auxiliary diagnosis of medical image segmentation, this article reviews the application of convolutional neural networks, Transformer, and U-Net and Transformer hybrid structures in medical image segmentation, and conducts a comprehensive comparative analysis of these models. The feasibility of these models in medical image segmentation is confirmed through visualization results and image evaluation metrics. Finally, we summarize the existing problems in current research and present future research directions.

Key words deep learning; medical image segmentation; U-Net; Transformer; hybrid model

1 引 言

医学图像分割任务是计算机视觉和医学影像处理领域的一个关键任务，根据区域间的相似或不同把图像分割成若干区域，以便更精确地定位和分析感兴趣的解剖结构或病变区域。这一任务对于医生进行精准诊断、制定有效的治疗计划以及进行疾病监测方面至关重要。医学图像是指通过医学影像技术获取的用于

收稿日期：2024-03-12；**修回日期**：2024-05-24；**录用日期**：2024-06-03；**网络首发日期**：2024-06-11
基金项目：国家自然科学基金(81973981,82074579)、山东省研究生优质教育教学资源项目(SDYAL2022041)
通信作者：*ma_jingang@126.com

诊断、治疗和研究的图像，其类型包括计算机断层扫描（CT）、核磁共振成像（MRI）、超声波成像（US）、X射线（X-ray）、病理图像、正电子发射断层扫描图像以及特定应用的图像如裂隙图像、视网膜图像和皮肤镜图像等。医学图像分割任务的核心是实现对不同类型图像中不同结构的准确提取和分离。这些结构包括人体器官、肿瘤、血管系统等。例如：在肿瘤分割任务[1]中，关注点集中在肿瘤的精准定位和边界提取；在器官分割任务[2]中，则需要有效分离各种器官，以支持医学图像的全面分析。此外，医学图像也存在噪声、伪影、类不平衡等问题，这使得医学图像分割成为一项极具挑战性的任务[3]。

传统的图像分割技术有基于区域的分割方法和基于边界的分割方法，前者依赖于图像的空间局部特征，如灰度、纹理及其他像素统计特性的均匀性等，后者主要是利用梯度信息确定目标的边界。传统分割方法在计算效率和简单性方面具有优势，但在处理复杂图像时高度依赖医生和放射科专家的专业知识，不能满足医学图像分割的实际应用需求[4]。随着算法与硬件计算力的不断进步，深度学习已经在医学图像分割等多个领域中发挥巨大作用[5-6]。深度学习的本质是通过学习数据的不同抽象层次来逐渐提取特征[7]。相较于传统分割方法，深度学习方法具有高精度、自适应性强和可扩展性强等优势。早期基于深度学习的网络在图像识别和分类任务中表现良好，但在图像分割领域中缺乏性能卓越、有针对性的网络模型[8]。直到2014年，用于图像分割的全卷积神经网络（FCN）[9]以及基于卷积神经网络（CNN）[10]的DeepLab[11]和U-Net[12]等网络才相继出现。

FCN采用了跨层跳跃连接，将深层的语义信息和浅层的纹理信息相结合，在图像分割中实现了较为精准的分割。但该网络的上采样策略和较大的感受野导致对小物体的分割精度较低，其上采样过程通常难以充分恢复图像细节，尤其是边缘区域的精细信息。此外，FCN在深层特征提取过程中可能丢失重要的空间信息，进一步影响了分割的准确性[13]。DeepLab是在FCN的基础上提出的一种图像语义分割模型，其主要特点是引入空洞卷积和空间金字塔池化（ASPP）机制，以增强模型对全局上下文信息和不同尺度信息的感知能力。然而，空洞卷积的引入也提升了计算成本，并且模型涉及多个关键超参数，如空洞卷积的膨胀率，导致训练和推理的时间较长[14]。Ronneberger等[12]在2015年提出了U-Net网络，其独特的编码器-解码器搭配跳跃连接的结构迅速成为研究的热点。U-Net的优势在于能够精确捕捉图像的局部细节信息，并通过跳跃连接有效减少信息在传输过程中的损失，实现对复杂医学图像中细微结构的精准分割。基于这些核心优势，部分研究者对U-Net进行重要扩展和改进，包括对骨干网络、跳跃连接、瓶颈区域等的优化以及多结构设计策略的引入。代表性的改进网络如UNet++[15]、UNet3+[16]、3D U-NET[17]等进一步增强了模型处理医学图像中复杂性结构的能力。此外，如张欢等[18]对U-Net的改进及其在医学图像分割中的应用进行了综述，旨在总结U-Net架构在医学图像处理领域的优化方法和应用成果，以指导未来的研究方向。这些研究成果凸显了U-Net在医学图像分割领域的持续影响力和潜力。近年来，Transformer架构在医学图像分割等领域逐渐得到应用，其独特的自注意力机制在捕获远程依赖关系方面可弥补U-Net对特征提取的不足[19-20]。因此，部分研究者们提出了基于U-Net和Transformer的多种混合模型架构[21-22]，这些混合模型结合了U-Net在捕获局部特征方面的强大能力和Transformer在理解全局依赖关系方面的优势，显著提高了模型的表示能力和处理复杂数据的效率，已成为当前的研究热点。目前，国内外学者已在混合模型研究上取得众多成果。如Zhang等[23]对基于CNN和Transformer的混合模型CTransNet的分割性能进行评估，实验结果表明该模型在多个医学图像数据集上表现出良好的泛化能力。张玮智等[24]对从U-Net到Transformer的深度模型在医学图像分割中的应用进行综述，总结U-Net及其改进模型的创新设计和不足之处。

本文对常见的医学图像类型及评价指标进行介绍，并深入分析了现有模型在捕获长距离依赖关系方面所遇到的挑战。随后，本文论述了U-Net和Transformer算法及其相关的改进在医学图像分割中的应用，特别是将重点放在了U-Net与Transformer混合模型的研究进展上。通过将混合模型按单尺度串行、多尺度串行、交替串行、整体并行和层级并行的混合方式进行分类，探讨不同混合结构在处理复杂和异质性的医学图像分割任务中的优势。进一步地，对所提模型进行性能对比和分析。最后讨论目前医学图像分割中面临的挑战，并对未来的研究方向进行展望。

2　医学图像分割关键技术

医学图像分割的目的是在医学图像上进行像素级分类，进而准确地分割目标对象。在临床诊断中，将根据患者的病情选择合适的医学影像检查方法，提高整个治疗过程的效率。

2.1　常见的医学成像技术

本部分综合大量文献详细介绍了X-射线成像、计算机断层扫描、核磁共振成像、超声成像、数字病理成像这五种常见医学成像技术的工作原理及应用场景，不同类型的医学图像如图1所示。各项医学成像技术的具体特性，请参见表1。

1）X-射线成像

X-射线成像的基本原理是利用X射线穿透体内组

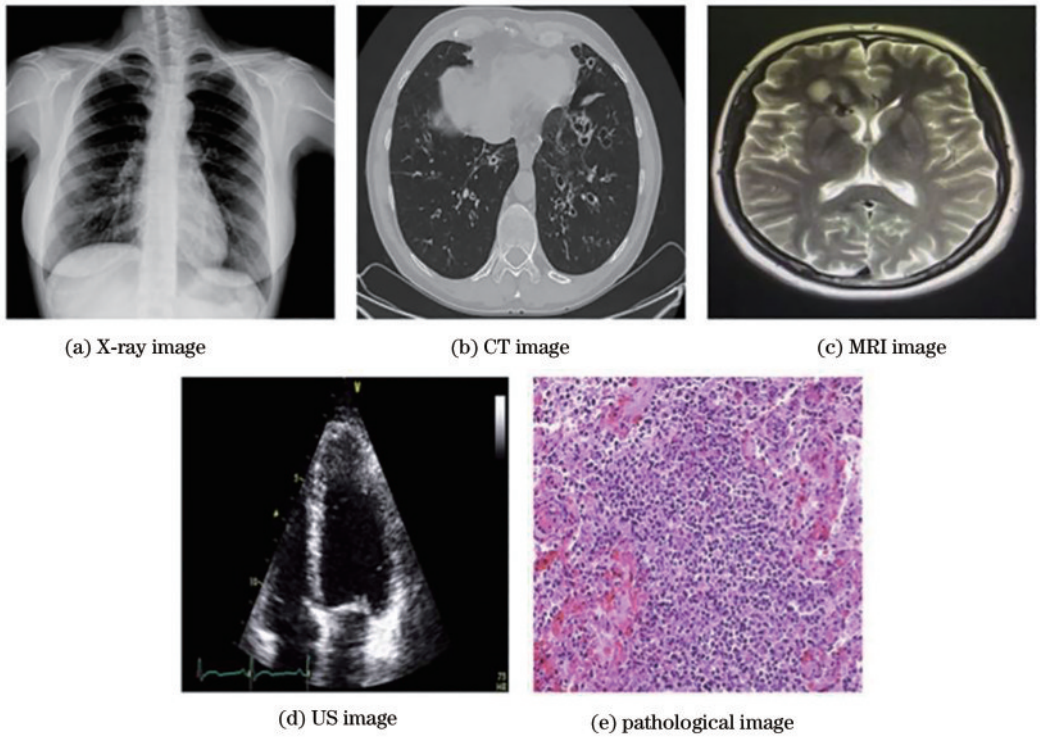

图 1　不同类型的医学图像

Fig. 1　Different types of medical images

表 1　五种成像方式对比表

Table 1　Comparison of five imaging methods

Imaging type	Imaging mechanism	Imaging objective	Advantage	Disadvantage	Common datasets
X-ray	Penetration of electromagnetic radiation into body tissues	Bones, lungs, heart	Fast, easy to operate, low cost, good contrast of negatives	Radiation, limited ability to image soft tissue, only 2D images available	COVID-19[33]
CT	Rotational X-ray fabrication of *in vivo* cross-sectional images	Any part of the body	High-resolution images with no overlap between tissue structure and lesion images	Radiating, prone to artifacts, average resolution for soft tissue	LiTS[34]
MRI	Strong magnetic fields and radio waves	Soft tissues, brain, joints	Non-invasive, no radiation risk, high resolution to soft tissue, ability to cut layers directly in any direction	Time-consuming, costly, tight inspection space and limited by metal implants	BraTS[35]
US	High-frequency sound waves	Soft tissues, blood flow, organs	Non-invasive, non-radiation-risked, real-time imaging	Poor imaging of gases and bony structures, high dependence on manual labor	EchoNet-Dynamic[36]
Digital pathology	High-resolution scans under an optical microscope	Tissues, cells	High-resolution images, digital storage	Requires advance sampling, which is costly and time-consuming	SEED[37]

织，根据不同组织如骨骼、肌肉等对X射线的吸收率差异来生成图像[25]。其中：高密度组织对X射线的吸收较多，在图像中显示为较亮的颜色；低密度组织吸收较少，显示为灰色或黑色，如图1(a)所示。该技术因其高效、成本相对低廉以及设备普遍性而被广泛应用于医学领域，尤其擅长检测骨折、诊断肺部疾病等。然而，X射线成像具有一定的辐射风险，对软组织的成像能力较弱，且主要提供二维图像，无法提供完整的三维视图。

2）计算机断层扫描图像

计算机断层扫描技术通过围绕患者旋转的X射线源和探测器从多个角度捕捉穿过身体的X射线数据，然后利用计算机重建成人体内部结构的详细横截面图像。CT成像的优点在于能够快速获得高分辨率的图像，适用于诊断各种类型的内部伤害和疾病，如脑出血、肿瘤等。但CT成像的成本较高，图像易受伪影的影响，且成像过程涉及电离辐射，频繁接受CT扫描可

能会增加癌症风险[26-27]。同时CT成像在区分软组织结构方面存在局限，如评估喉部的CT扫描难以区分软组织结构和确定软骨侵犯程度[28]。

3）核磁共振成像

核磁共振成像是一种利用磁共振现象从人体内部获取电磁信号并重构成体内结构图像的技术，能够生成横截面、矢状面、冠状面以及多种倾斜面的详细图像。MRI能较好地检出人体组织成分中水含量的情况，因此适用于软组织成像，对脊柱、颅脑、关节等方面的显示极佳。MRI检测过程无需使用造影剂且没有电离辐射，对人体不产生负面影响；同时避免了X射线与骨皮质的相互作用，从而消除了在CT检查中可能产生的伪影，因此其检测肿瘤边界的准确度更高[29]。但MRI也存在对患者运动敏感、成像时间长、成本较高等缺点。

4）超声成像

超声成像技术是一种使用声波来检查和捕捉体内器官和组织图像的医学成像技术。它基于超声波在不同介质中的传播速度和反射特性的差异来定性分析病变的物理属性，精确测定人体器官的位置、尺寸和形态，并提供详细的组织解剖图像。超声成像属于非侵入性、无痛、无辐射风险的成像手段，广泛应用于消化系统、心血管系统等的医疗领域，尤其适用于乳腺癌等疾病的检测和腋窝淋巴结的术前评估，对乳腺癌患者的分期诊断至关重要[30]。然而，超声图像的分割结果往往受到诊断者主观判断的影响，可能导致诊断结果与实际情况存在偏差[31]。

5）数字病理成像

数字病理成像技术通过高分辨率扫描仪将传统病理玻片转换为数字图像。这些切片通常来源于人体病变组织或细胞，经过染色处理以增强组织结构的可视性。病理图像具有高分辨率，能够展现组织、细胞乃至亚细胞结构的详细信息，使医生能够在细胞和分子层面上理解疾病的本质，是疾病诊断的“金标准”。例如在腺癌、肝癌等各类癌症的诊断中，数字病理成像能够提供肿瘤细胞的详细图像，帮助医生评估肿瘤的分级、分期以及预后。但病理图像数据同时也存在标注成本高、标注困难等现实因素，并极易引发标注样本缺乏、标注质量不佳等问题[32]。

2.2　医学图像分割的关键问题

医学图像分割技术的发展对计算机辅助诊断、临床应用和医疗健康等领域的研究具有极其重要的作用。然而在处理医学图像中广泛分布的病变区域和复杂解剖结构时，模型面临着捕获长距离依赖关系的困难。长距离依赖指的是图像中相距较远的区域之间存在的关联或依赖关系，这些关系对于理解图像的整体结构和上下文非常关键。例如，在早产新生儿脑室的分割预测[38]中，网络需要准确区分图像中的掩码和背景像素，由于图像背景分布不均，学习背景像素间的长距离依赖关系可以帮助网络避免将其归类为掩码，减少误报。然而，卷积核固定的大小和有限的感受野使得每次卷积操作仅能覆盖图像中的局部区域。如果要捕获长距离依赖关系，则需较深的网络或较大的滤波器，从而导致模型参数增加。虽然部分研究在U-Net网络中使用了图像金字塔[39]、空洞卷积[40]和注意力机制[41]等方法对长距离依赖关系进行建模，但图像金字塔增加了模型的计算量；空洞卷积存在的网格效应会降低模型的分割性能；使用注意力机制的模型获取全局感受野大多通过全局池化操作，难以提供像素级别的注意力。此外，医学图像还面临边缘不清晰、对比度低、目标尺寸不一等问题。因此，开发能够有效处理长距离依赖关系的医学图像分割模型，成为了推动该领域发展的关键挑战之一。

近年来，Transformer因其强大的长距离依赖捕捉能力被引入医学图像分割网络中，特别是与U-Net进行结合，体现了其处理复杂图像分割任务时的优异性能。例如在多模态医学图像分割任务中，这种结合模型展示了处理多源数据并捕捉更全面信息的能力[42]。但该类模型通常计算密集，限制了其在资源受限环境中的应用。如何创新地融合U-Net和Transformer模型且平衡效率与性能，推动医学诊断和治疗领域的创新与发展仍是未来研究的关键。

2.3　医学图像分割常用评价指标

评价指标用于衡量医学图像分割模型的性能，常用的有Dice、交并比（IoU）、准确度（ACC）、灵敏度（SE）。Dice用于衡量预测结果与真实标签的重叠程度，取值范围0～1，值越大表示重叠度越高；IoU用于比较预测结果与真实标签的相似程度，取值范围0～1，值越大表示相似程度越高；准确度用于衡量预测结果的整体准确性，是正确预测的样本数与总样本数之比；灵敏度用于衡量模型正确识别出正样本的能力，灵敏度越高说明模型检测病变时的漏检率越低。以上评价指标的数学表达式如表2所示。

表2　分割性能评价指标及数学描述

Table 2　Segmentation performance evaluation indexes and mathematical description

Evaluation index	Mathematical description
Dice	$\frac{2\lvert A\cap B\rvert}{\lvert A\rvert+\lvert B\rvert}$
IoU	$\frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$
ACC	$\frac{R_{TP}+R_{TN}}{R_{TP}+R_{TN}+R_{FP}+R_{FN}}$
SE	$\frac{R_{TP}}{R_{TP}+R_{FN}}$

Notes: A is the predicted result, B is the real label, $\lvert A\cap B\rvert$ is the size of the intersection between predicted result and true label, $\lvert A\cup B\rvert$ is the size of the union between predicted result and true label, R_{TP} is true positive, R_{TN} is true negative, R_{FP} is false positive, R_{FN} is false negative.

3　深度学习在图像分割中的应用

本节将从U-Net、Transformer以及U-Net+Transformer混合模型这三个方面进行介绍，对这些深度学习方法的改进以及在医学图像分割中的应用情况进行梳理，并对各个模型进行对比分析。

3.1　基于U-Net的分割研究

3.1.1　U-Net的结构

在医学图像分割任务中，Ronneberger等[12]提出的U-Net网络架构是最成功的方法之一。该架构包含编码器、对称的解码器以及跳跃连接。编码器使用CNN架构作为收缩路径来提取图像特征、降低分辨率，收缩路径共有4个子块，每个子块由两次连续的3×3卷积、ReLU激活函数和用于下采样的最大池化层组成。解码器由含有上采样操作的卷积块构成扩展路径来修复图像细节信息、定位分割对象边界，并逐步恢复特征图的空间分辨率。

编码器和解码器之间的跳跃连接是U-Net网络的重要组件。跳跃连接操作在每个编码器层结束时，将该层的特征图与对应解码器层的特征图进行连接，允许网络直接访问低级别和高级别的特征信息，有助于解决信息丢失问题，使网络更好地理解图像的上下文和细节信息。U-Net的网络架构如图2所示。

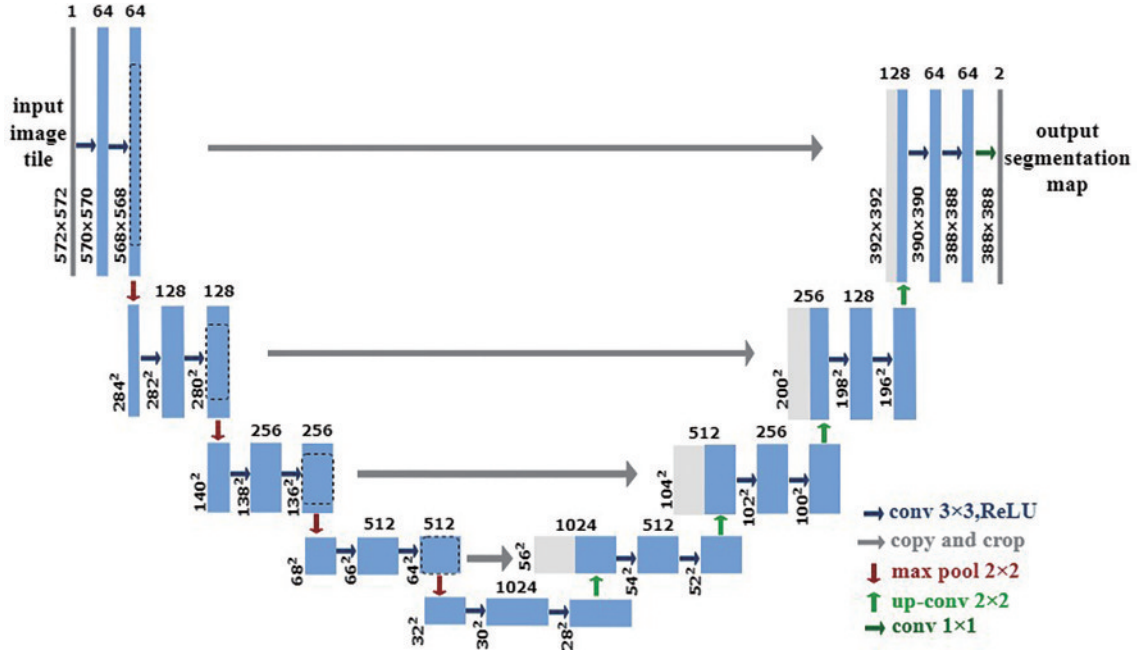

图2　U-Net结构图[12]

Fig. 2　U-Net structure diagram[12]

3.1.2　U-Net在医学图像分割中的改进及应用

为了充分提取细节和全局特征，加强特征信息的融合，或打破U-Net在处理噪声、伪影、类不平衡、梯度消失等问题的局限性，研究者们基于U-Net的基本架构进行了诸多尝试和改进，主要包括结构和非结构上的改进，以适应更广泛的应用需求和挑战。

3.1.2.1　结构改进

U-Net结构上的改进主要集中于对骨干网络、跳跃连接以及瓶颈区域等的优化。这些改进通常在U-Net的基础上增加残差块、密集连接块、Inception块、注意力机制等网络构件或者将子模块中的卷积替换为可变形卷积、扩张卷积、循环卷积等，旨在扩展和增强原有结构中的关键环节。例如Zhou等[15]在2018年提出了UNet++网络，该网络采用密集跳跃连接策略替代原始U-Net中跳跃连接，解决了浅层和深层特征之间的语义鸿沟问题。紧接着，Huang等[43]提出了UNet 3+，该模型采用全尺度跳跃连接和深度监督机制，整合不同尺度特征，提高对图像语义的理解和分割精度。AR-UNet[44]为解决原卷积和下采样操作可能造成的梯度消失和结构信息丢失等问题，将瓶颈部分原先的卷积层模块替换为空洞卷积模块，以提高物体间相关性，使得血管分割更加精确。Liu等[45]则在瓶颈部分加入空洞金字塔卷积模块来提取图像的全局结构特征并降低计算量。Oktay等[46]提出的Attention U-Net网络采用Attention Gate[47]注意力机制对下采样的输出结果进行过滤，消除跳跃连接中的噪声和无关信息，提升了提取关键特征的准确性。Alom等[48]在U-Net的基础上利用循环卷积层(RCLs)和带有残差单元的RCLs代替原有的正向卷积层，在参数量不变的情况下提升分割性能。另一方面，基于U-Net和3D影像设计得到3D-UNet[49]，该网络采用3D卷积操作代替二维卷积，可从稀疏注释的体积图像中学习特征；引入新的目标函数，以解决类别不平衡问题并保持紧密的空间关系。一些研究还通过改变网络结构来适应特定的应

用需求，以实现更优的图像分割性能。例如Fu等[50]在U形卷积网络的基础上增加多尺度输入层和侧边输出层，构建了用于视盘和视杯分割的M-Net模型架构，实现多层次感受野的融合并支持深层监督。

3.1.2.2　非结构改进

U-Net及其改进网络能够出色处理医学图像分割任务还得益于其非结构性的改进。这些改进通常涉及数据处理技术、训练策略、损失函数设计等方面，旨在提升模型性能、加速训练过程。例如杨鑫等[51]为了避免神经网络可能产生的过拟合现象，对图像采用了随机剪切、翻转、灰度扰动和形状扰动等增强技术以实现增加数据量。王士奇[52]针对脑肿瘤MRI图像存在的样本不平衡问题设计了一种混合损失函数，该函数将Dice损失函数与焦点损失函数有效地结合起来，并通过实验确定比例系数，有效地解决了脑肿瘤MRI图像样本不平衡问题。Isensee等[53]在提出的nnU-Net网络中采用了实例归一化(IN)方法，该方法的标准化过程独立于样本之间的通道数和批次大小，仅对每个独立样本内的像素进行标准化处理，从而帮助网络更好地适应不同样本之间的特征差异，同时有效避免了批量大小对网络性能的潜在影响。同样黄泳嘉等[54]在使用改进型U-Net网络进行肝部分割时采用了组归一化(GN)来代替常用的批量归一化(BN)，从而降低批量尺寸过小对网络性能的负面影响。

综上所述，U-Net模型及其改进版本的出现标志着医学图像分割领域取得重要进步，U-Net的优点包括：1)通过对结构的细微调整或对非结构的改进，U-Net能够高效处理各种复杂性的图像数据。这种能力不仅体现在二维图像分割任务上，其变体也成功扩展到三维图像分割和多模态图像处理等更复杂的应用场景中，显示出了极高的适应性和灵活性。2)U-Net因其深层结构和卷积操作的设计能够有效捕捉图像的局部细节信息，跳跃连接则确保了编解码器间的信息流畅传递，保留了高分辨率的细节。这使得其在边缘检测和细小结构的识别上表现出色。3)U-Net通过数据增强和网络结构的设计，能够在有限的训练样本下学习到有效的特征表示，从而在数据稀缺的图像分割领域表现优秀。

但U-Net也存在以下缺点：1)医学图像分割需要精确识别和划分图像中每个像素所属的结构或组织，这就要求模型对局部细节和整体结构之间的复杂关系有更深入的理解。然而，卷积核有限的感受野难以捕捉远程关系和全局信息，限制了模型性能的进一步提升。2)U-Net及其改进版本因其深层网络结构和跳跃连接通常需要较大的计算资源，尤其是在处理大尺寸图像或进行三维图像分割时。这种高资源需求限制了U-Net的应用。3)对于分辨率较高的图像，U-Net可能需要采用分块处理或降采样等方法，从而可能导致细节信息的丢失或增加处理的复杂度。

3.2　基于Transformer的分割研究

Vaswani等[55]于2017年提出Transformer模型，该模型采用自注意力机制作为核心组件，并引入层归一化、前馈神经网络(FFN)和残差结构等模块，在自然语言处理的各类任务中表现突出。随着深度学习技术的不断发展，Transformer的思想被应用在计算机视觉和医学图像分割等领域，Transformer克服了CNN在处理长距离依赖关系时的限制，取得了较优的分割效果。

3.2.1　Transformer的结构

Transformer模型的整体结构可拆分为输入部分、输出部分、编码器和解码器等4个部分，其中编码器和解码器两部分都由多层结构组成。Transformer结构如图3所示。Transformer的主要模块如下。

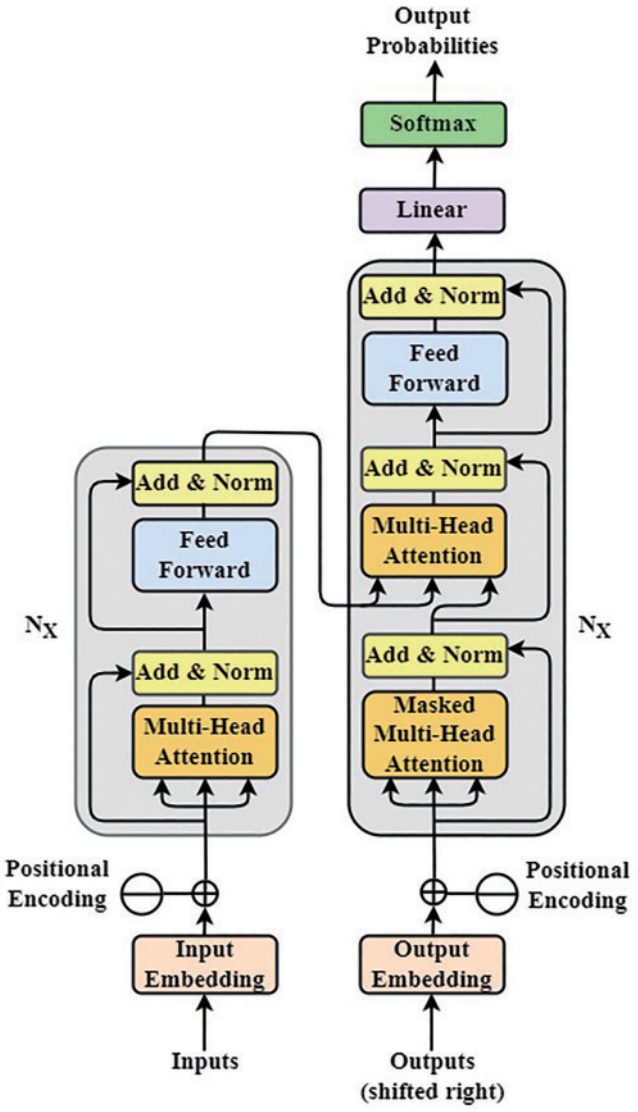

图3　Transformer结构图[55]

Fig. 3　Transformer structure diagram[55]

1) 自注意力机制

自注意力机制是Transformer处理序列数据的关键组成部分。它使模型能够在处理输入序列时对不同

位置的信息分配不同的注意力权重，从而将每个元素与其他元素相关联。这一过程如下。

①计算查询矩阵（$\boldsymbol{Q}$）、键矩阵（$\boldsymbol{K}$）和值矩阵（$\boldsymbol{V}$）：

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}_{\boldsymbol{Q}},\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}_{\boldsymbol{K}},\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}_{\boldsymbol{V}} \tag{1}$$

式中：$\boldsymbol{X}$表示输入向量；$\boldsymbol{W}_{\boldsymbol{Q}}$、$\boldsymbol{W}_{\boldsymbol{K}}$、$\boldsymbol{W}_{\boldsymbol{V}}$分别表示学习到的权重矩阵。

②计算注意力权重，公式如下：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

式中：d_k表示键矩阵的维度；$\sqrt{d_k}$表示缩放因子。

2）多头注意力机制

为了提高模型对不同子空间特征的捕捉能力，自注意力机制被扩展成多个注意力头的形式。每个头学习不同的查询、键、值的映射，最后通过拼接或加权求和的方式结合多个头的输出。具体计算公式如下：

$$H_{\text{head}\,i}=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}\boldsymbol{W}_i^{K},\boldsymbol{V}\boldsymbol{W}_i^{V}),$$
$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(H_{\text{head}\,1},\cdots,H_{\text{head}\,h})\boldsymbol{W}_{\text{out}} \tag{3}$$

式中：$i=1,2,\cdots,H$；$\boldsymbol{W}_{\text{out}}$表示输出层的权重矩阵；$\boldsymbol{W}_i^{Q}$，$\boldsymbol{W}_i^{K}$，$\boldsymbol{W}_i^{V}$表示第$i$个头的权重矩阵；Concat()表示矩阵拼接操作；Attention()表示注意力输出。

3）位置编码

由于自注意力机制不考虑元素的顺序关系，位置编码被引入以提供序列中元素的位置信息。位置编码的形式包括相对位置编码和绝对位置编码。相对位置编码可捕捉序列中元素间的相对距离关系；绝对位置编码通过数学函数为序列中的每个位置生成固定的位置编码。位置编码使得Transformer能够在建模序列时考虑元素的顺序关系，从而更好地捕捉全局上下文信息。

4）前馈神经网络

前馈神经网络通过全连接层和激活函数对每个位置的表示进行非线性映射和变换，有助于模型学习更复杂的特征表示。具体的公式如下：

$$N_{\text{FFN}}=\max(0,\boldsymbol{X}\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{4}$$

式中：$\boldsymbol{W}_1$和$\boldsymbol{W}_2$表示权重矩阵；$\boldsymbol{b}_1$和$\boldsymbol{b}_2$偏置向量。

5）层标准化和残差连接：

Transformer在每个子层后采用层标准化进行规范化，并通过残差连接将输入直接添加到输出，有助于解决训练过程中的梯度消失问题。

3.2.2　Transformer在医学图像分割中的应用

Dosovitskiy等[56]首次将Transformer拓展至医学图像分类任务，引入ViT（vision Transformer）将图像分割成一组图像块，并将每个图像块作为输入序列传递到Transformer编码器，以获得图像的全局表示。

基于ViT在医学图像分类任务中的成功，部分研究人员开始尝试在图像分割任务中使用Transformer。Liu等[57]针对图像分割和目标检测任务设计得到Swin Transformer架构，该架构将Transformer block中的标准多头自注意力模块（MSA）替换为Window-MSA和Shifted Window-MSA，且保持其他部分不变。该窗口机制可处理不同分辨率的特征图，在分割任务中取得较优性能。Cao等[58]提出了一个基于Swin Transformer块的U形编解码器结构，命名为Swin-Unet。该模型使用Swin Transformer模块代替二维卷积块作为特征表示和远程语义信息交互学习的基本单元。实验结果表明，该模型在多器官和心脏分割数据集上具有良好的分割精度和鲁棒泛化能力，但也存在缺乏对局部特征信息提取的问题。在此基础上，Wang等[59]提出C2Former网络，该网络重新设计了交叉卷积自注意机制算法，对长距离和短距离依赖关系进行建模，提高了对语义特征的理解能力。基于Swin Transformer的网络需在大规模数据集上进行预训练，针对此问题，Huang等[60]提出一种从零开始训练的MISSFormer网络，该网络采用增强Transformer模块克服卷积直接嵌入前馈神经网络所带来的特征识别限制，并进行有区别的特征表示。DS-TransUNet[61]进一步扩展了Swin-Unet，采用双尺度编码器子网来提取不同语义尺度的粗粒度和细粒度特征；同时设计Transformer Interactive Fusion模块，建立不同尺度特征之间的全局依赖关系。

由上述分析可知，Transformer模型具有以下优点：1）Transformer模型的自注意力机制可以捕获图像中任意两点间的依赖关系，从而在处理医学图像时考虑到全局信息。2）Transformer模型具有较强的可扩展性，能够处理不同大小的图像和不同分辨率的输入，适用于多种类型的医学图像数据。但Transformer模型也存在以下缺点：1）Transformer模型通常需要大量的标注数据进行训练，以充分利用其复杂的模型结构。然而，医学图像分割领域中高质量的标注数据稀缺，限制了该模型的应用。2）Transformer模型在捕捉图像的局部细节特征方面存在局限，对于需要精细局部特征解析的任务，如精确的边缘检测或小对象分割等，其性能不如专门设计的CNN模型显著。

综上所述，将Transformer与U-Net的优势相结合显得尤为重要，探索U-Net和Transformer相结合的模型开辟了医学图像分割领域研究的新方向。

3.3　基于U-Net和Transformer混合模型的分割研究

近年来，部分研究者将U-Net和Transformer进行结合，充分利用CNN在捕捉局部特征上的优越性以及Transformer对全局依赖性的处理能力，实现对局部和全局信息的整合，提高对医学图像的分割性能。下面将U-Net和Transformer的结合方式分为单尺度串行、多尺度串行、交替串行、整体并行和层级并行等方面进行介绍，并对这些方法在医学图像分割中的应用情况进行梳理。

3.3.1　串行结合

3.3.1.1　单尺度串行结合

部分研究在U形结构中设计先CNN后Transformer的顺序结构。CNN和Transformer之间采用单尺度串行结合方式，即Transformer接收来自CNN处理后输出的单一尺度图像数据。该结构先通过CNN获得图像的局部细节和纹理信息，然后在Transformer中实现对图像的全局关系建模。图4为单尺度串行结合概念性结构图。

TransUNet模型[62]是首个将U-Net和Transformer进行串联结合的U形网络，其模型结构如图5所示。该模型在编码器阶段使用CNN获取详细的高分辨率空间信息；Transformer模块串联在CNN之后，以CNN的特征映射作为输入提取全局上下文信息；最后将局部特征与自关注特征相结合，实现精确定位。在此基础上，Liu等[63]提出了TransUNet＋模型，该模型通过分数矩阵的列向量来重新设计跳跃连接，增强跳跃特征，从而提高分割性能。部分研究者将Transformer模型串联在双流融合模型之后，如Lu等[64]在U形结构的编码器前端采用两个基于ResNet的独立分支网络，提取眼底图像的临床知识和局部信息，编码器后端利用Transformer模块对双分支特征进行融合并进行全局建模。

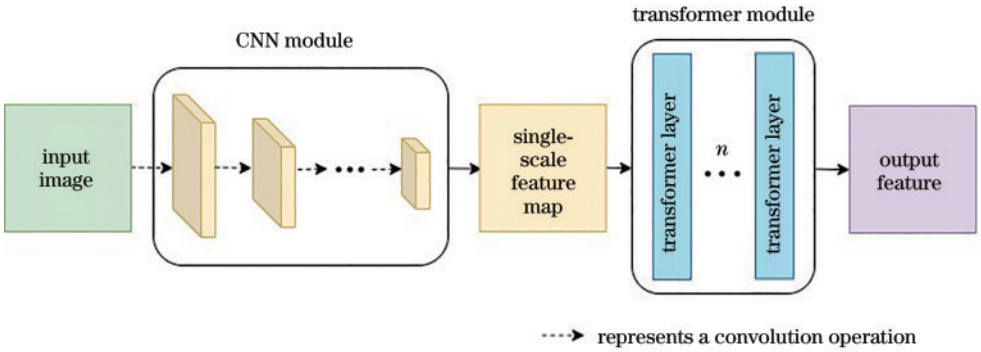

图4　单尺度串行结合概念性结构图

Fig. 4　Conceptual structure diagram of single-scale serial combination

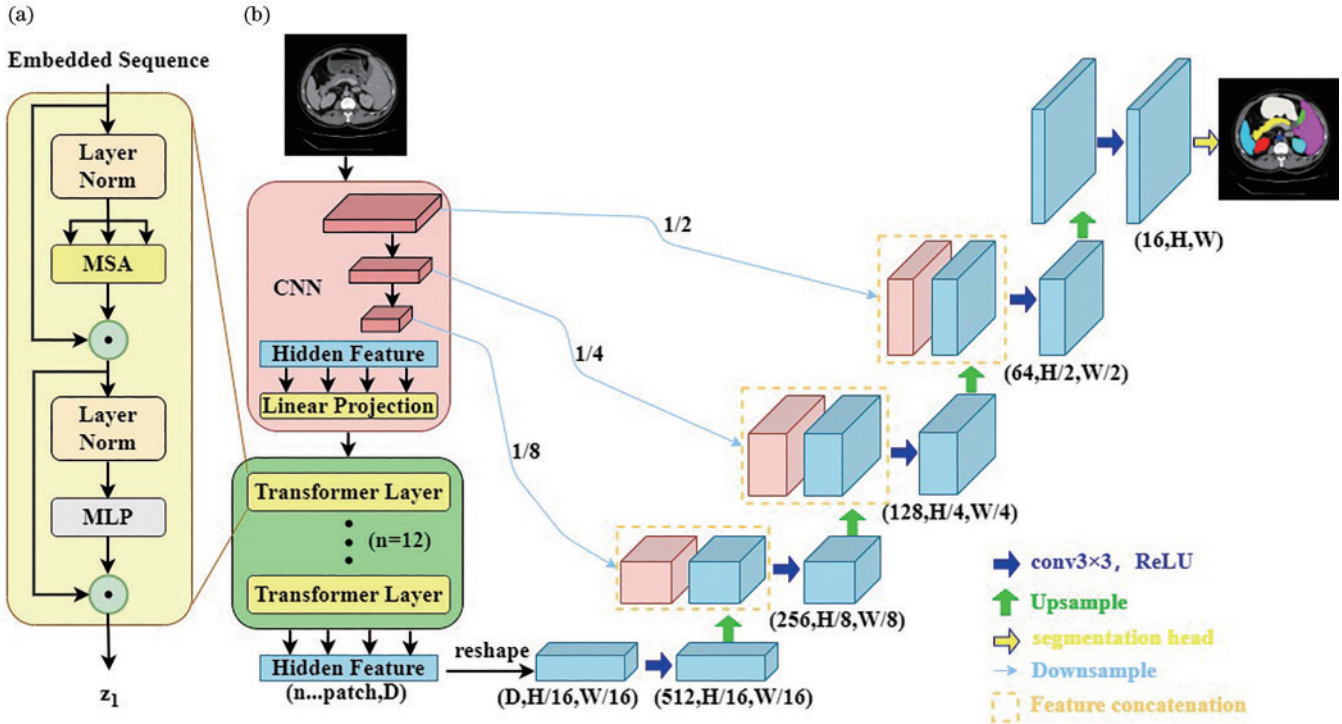

图5　TransUNet结构图[62]

Fig. 5　TransUNet structure diagram[62]

还有一些研究人员在U形结构的编解码器之间插入Transformer模块，以降低模型参数量。李佳松等[65]将Transformer层放置在U形结构的底部，以捕获全局上下文信息，并将其作为CNN提取局部特征的补充。Gai等[66]在编码器中采用多个堆叠残差块提取特征信息，在U形结构底部利用ViT模块对提取的特征

进行过滤，降低参数量，解决不相关背景信息对目标分割的干扰。此外，Li 等[67]研究发现增加更多的 Transformer 层并不能提高分割精度，从而提出了在 CNN 后仅串联一个 Transformer 层的 MultiIB-TransUNet 模型。该模型将多个 IB 模块集成到一个 Transformer 层中以压缩无关特征。实验证明，该模型在降低参数量的同时精度也优于 TransUNet 模型。

为了充分利用底层特征增强全局特征，减少编码和解码阶段之间的语义差距，UNETR 模型[68]将 Transformer 层作为特征提取的编码器，将卷积层作为解码器，以充分利用 Transformer 的远程依赖和 CNN 的归纳偏差。孙开鑫等[69]将 Swin Transformer 模块作为编码器捕获全局上下文信息和处理区域分割任务，CNN 作为解码器逐步恢复底层特征并定位局部特征，实现更准确的语义分割。

综上所述，单尺度串行结合利用了 CNN 和 Transformer 的优势，使得模型能够有效捕捉图像的局部特征和全局信息。这种混合模型的结构简单且易于理解，被广泛用于医学图像分割等任务中。

3.3.1.2　多尺度串行结合

单尺度串行缺乏对输入数据中不同尺度信息的学习，使得模型对多尺度特征的感知和表达能力较差。基于此，部分研究者基于 CNN 和 Transformer 设计多尺度串行结合模型。在该类模型中，CNN 采用不同大小的卷积核或多层级的卷积操作从输入图像中提取多尺度特征，然后输入到 Transformer 中进行融合并实现全局关系建模，从而提升模型获取不同尺度信息的能力。图 6 为多尺度串行结合概念性结构图。

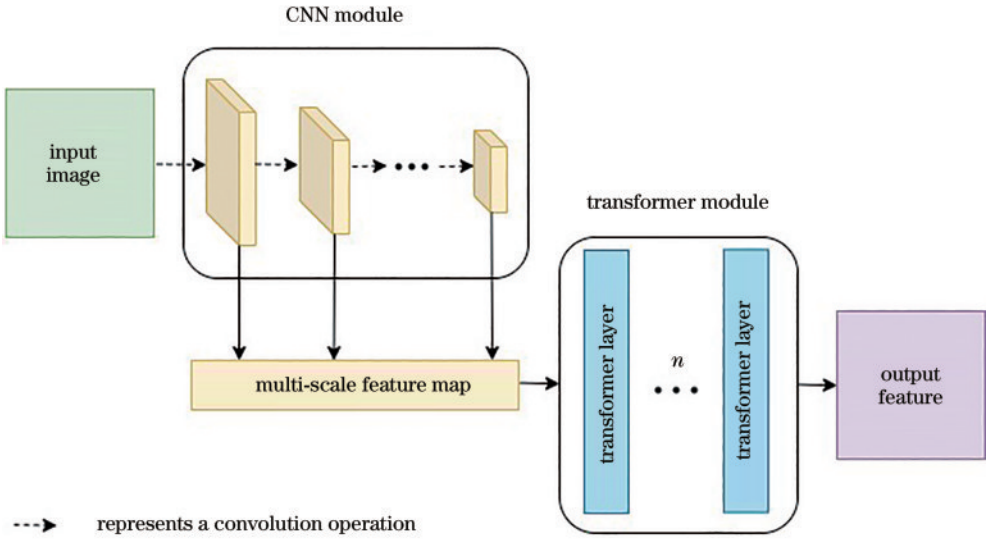

图 6　多尺度串行结合概念性结构图

Fig. 6　Conceptual structure diagram of multi-scale serial combination

Xie 等[70]提出了一种混合 U 形框架 CoTr，以实现三维医学图像分割。该模型在编码器中采用 CNN 结构提取多尺度特征映射，串联高效可变形 Transformer 模块以序列到序列的方式处理嵌入位置编码的扁平多尺度特征映射，可保留高分辨率信息并降低计算和空间的复杂性。CoTr 结构图如图 7 所示。Ji 等[71]提出了 MCTrans 网络，该网络将 CNN 提取多尺度特征标记反馈到 Transformer-Self-Attention 模块，构建多尺度上下文，并引入可学习的代理嵌入对语义关系和特征增强进行建模。该模型解决了跨尺度依赖、不同类别之间的语义对应问题。Wang 等[72]提出一种基于 Transformer 的特征融合网络（TFNet），该网络在编码器阶段将 CNN 和 Transformer 进行多尺度串行，以实现特征的多尺度融合和全局建模。然而部分超声序列缺乏轴向信息，尽管使用 Transformer 结构细化特征，但由于缺乏逐像素的空间依赖关系和较高的计算成本，无法高效地提取上下文信息。因此，Chi 等[73]将多尺度交叉注意转换器模块应用到二维 UNet 的跳跃连接中，以保留多个尺度的详细结构信息；同时利用 3D Transformer UNet 进行帧间特征提取；最后将帧内和帧间特征结合，实现准确的甲状腺分割。

上述方法的分割性能均有所提高，但仅利用了部分多尺度特征，且仅在低分辨率水平上提取全局特征，不能充分发挥自注意机制的作用。为保持 CNN 特征的多个不同分辨率表示，Yan 等[74]设计了一种基于 Transformer 的高分辨率网络（TransHRNet），该网络编码器利用 CNN 提取局部信息，将不同分辨率的特征映射并行馈送到每个 Transformer 流中，利用 Transformer 在每个分支中捕获不同的上下文信息并在分支间重复交换信息，实现多个尺度特征自适应地学习远程依赖关系。Hatamizadeh 等[75]将 Swin Transformer 作为编码器学习输入体积的序列表示，并提取其不同层次的编码表示，通过跳过连接与解码器合并预测最终的分割，增加了模型学习远程依赖关系

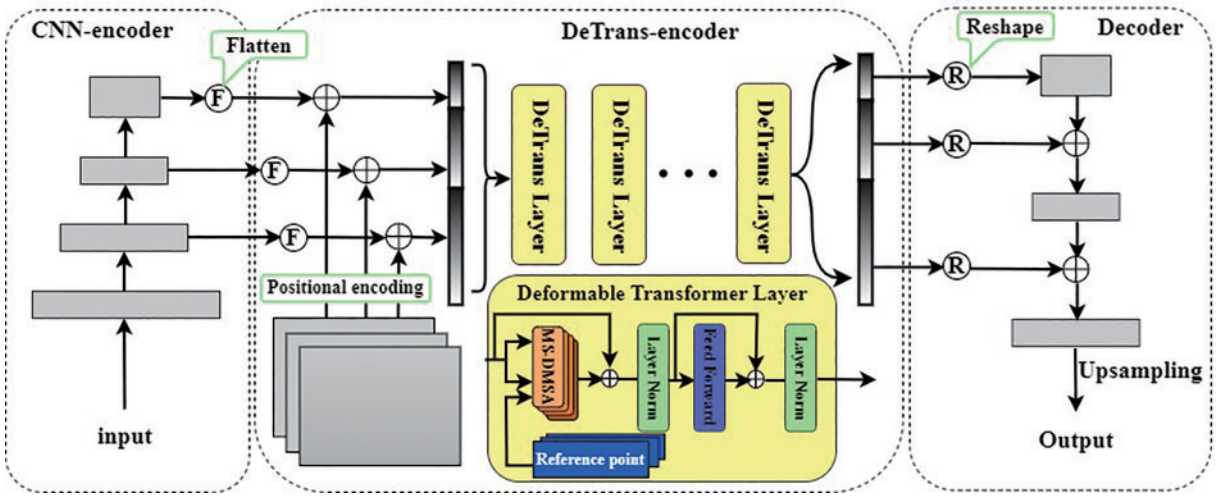

图 7　CoTr结构图[70]

Fig. 7　CoTr structure diagram[70]

和在多个尺度捕获全局上下文信息的能力。

总之，多尺度串行结合方法实现了对输入数据中不同尺度信息的学习，弥补了单尺度串行结合方法对多尺度信息学习的不足，提高了模型对多尺度信息的感知和表达能力。

3.3.1.3　交替串行结合

单次串行结构的网络模型在医学图像分割任务中存在局限性，部分研究者在U形结构的编解码器之间交替使用CNN和Transformer，在更大范围内全面捕获图像的语义信息。该设计策略是在每个CNN模块提取图像特征后，将特征图传递到Transformer模块进行全局关系建模。接着，整合后的特征图再传递到下一个CNN模块，重复上述过程。图8为交替串行结合概念性结构图。

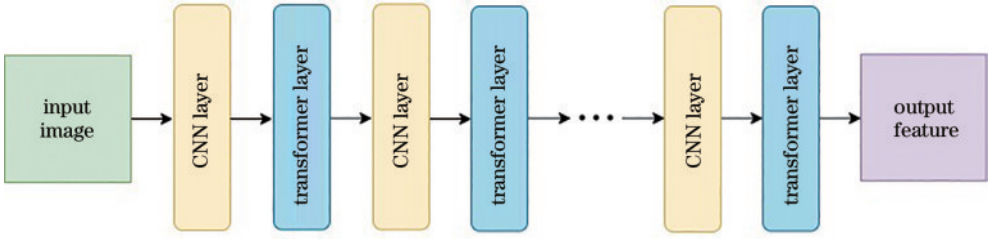

图 8　交替串行结合概念性结构图

Fig. 8　Conceptual structure diagram of alternating serial combination

He等[76]设计了一个名为HCTNet的分割网络，该网络在编码器中将CNN模块和Transformer编码器块（TEBlock）交替串行，共同提取不同尺度乳腺超声图像的全局特征；在解码器中使用残差连接聚合不同语义尺度的上下文特征映射；在跳跃连接中引入空间交叉注意模块，减小编解码器之间的语义差异，提高超声图像乳腺病变的分割精度。Wang等[77]在U形结构的编码器阶段引入混合注意块，该模块是由最大池化残差卷积块和2层高效Transformer块构成，实现在每个层次上提取特征；解码阶段则通过交叉融合块融合不同层次的高级语义特征，并通过多个残差卷积块恢复到原始分辨率特征。Zhang等[78]则将交替方式应用到了U形网络解码器阶段，设计了一种名为Star-shaped Window Transformer的模块，构成SWTRU网络。该网络的编码器保留了U-Net结构并重新设计了全尺寸跳跃连接；解码器则是由S2Win Transformer和基于CNN滤波集成机制交替组成，并在其最后一层进行特征集成，解决Transformer在高分辨率下性能不佳的问题。

部分研究将交替方式同时应用到了编解码器，从而获得更强的层次特征提取能力和更全面的信息融合效果。Gao等[79]提出了UTNet，其网络结构如图9所示。该网络将编解码器中每个分辨率的底部卷积替换为Transformer模块，在不同层次上实现扩展操作，从而更好地捕获远程依赖关系。Zhou等[80]提出了名为nnFormer的交错编码器架构，该架构在编解码器阶段采用卷积和自注意操作交替结合，通过卷积层对空间信息进行编码，利用Transformer层对全局上下文进行编码。Jiang等[81]提出了一种三维医学图像分割方法SwinBTS，该方法利用3D Swin Transformer模块和卷积模块交替构成编解码器，并在两个模块之间添加相邻特征连接增强（NFCE）模块，以增强特征信息，减少

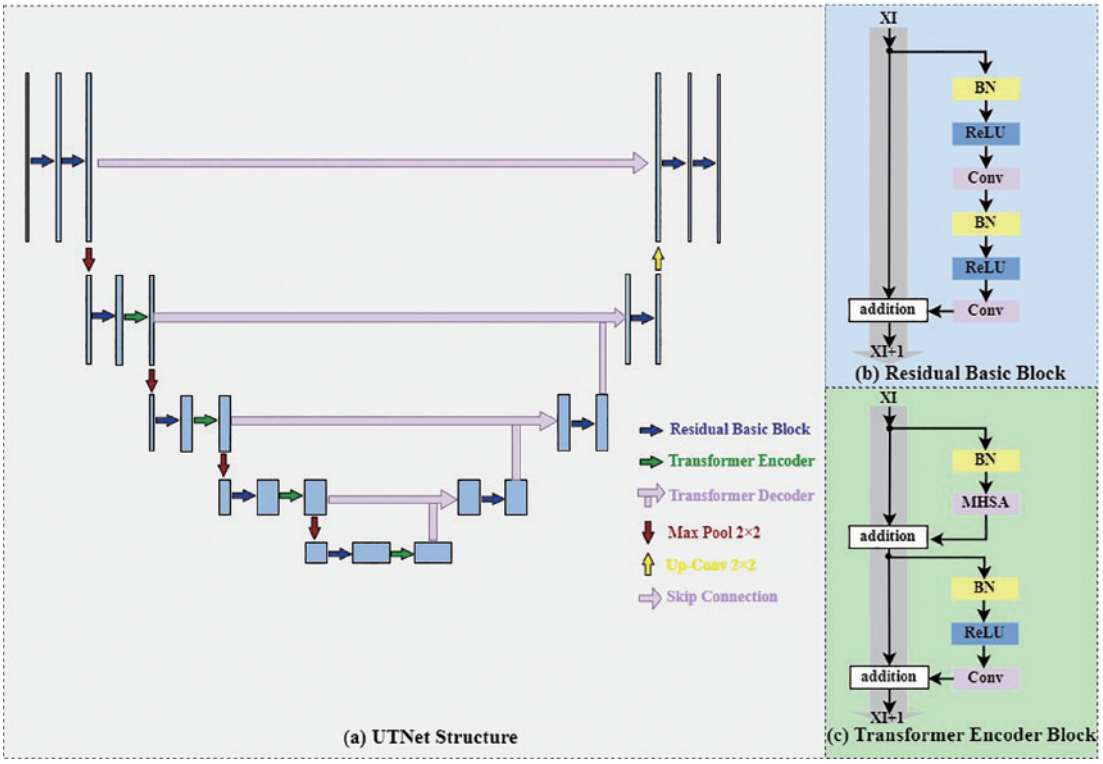

图 9　UTNet结构图[79]

Fig. 9　UTNet structure diagram[79]

信息丢失。实验结果表明，SwinBTS在脑肿瘤分割方面优于最先进的3D算法。

综合而言，交替串行结合方式通过在编码器和解码器之间交替使用CNN和Transformer，从而在每个层级更全面地捕获图像的语义信息。这种层次化的处理方式不仅提高了医学图像分割的准确性，还增强了模型对复杂医学场景的适应能力。

3.3.2　并行结合

3.3.2.1　整体并行结合

串行结合的优势是能够实现CNN和Transformer层对图像的顺序处理，从而使得特征提取更加有针对性。然而，部分串行结合网络中不同模块间可能存在信息传递的限制，造成性能瓶颈。鉴于此，部分研究者提出了并行结构，利用CNN和Transformer分别处理输入图像，产生两组特征图。通过融合策略将两组特征图结合，形成综合特征表示。该类模型在信息传递方面更为高效。图10为整体并行结合概念性结构图。

Zhang等[82]提出了一种名为Transfuse的模型。该模型利用CNN和Transformer两个并行分支提取不同分辨率的特征，同时捕获全局依赖关系和低级空间细

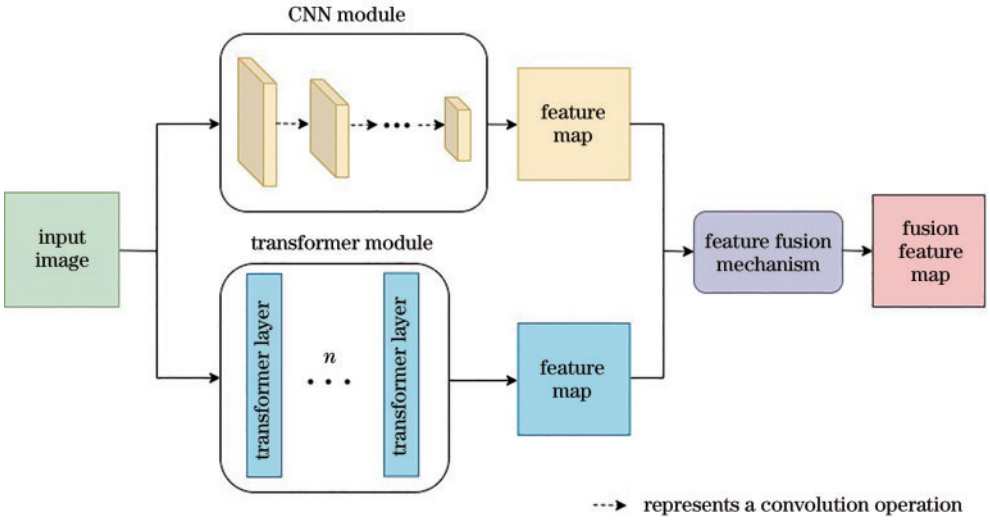

图 10　整体并行结合概念性结构图

Fig. 10　Conceptual structure diagram of overall parallel combination

节，并设计特征融合模块实现两种特征的融合，在二维和三维医学图像集上均取得不错的结果。Wang等[83]设计了TransFusionNet架构，该架构引入基于Transformer的全局特征提取编码器和基于多层SEBottleNet堆叠的局部残差网络编码器，共同提取CT图像的语义和空间特征；同时还引入了融合Transformer和CNN的特征提取模块和边缘提取模块，实现肝脏肿瘤和血管的精确分割。为避免分割精度损失，Zhang等[84]提出了一种双并行U形网络DuPNet，如图11所示。该模型由Transformer和CNN组成双支路编码器，通过拼接和融合双编码器提取的特征，获取直肠癌分割的全局和局部信息。同时在跳跃连接中设计特征自适应块，捕获不同尺度特征，该模型的平均IoU达到89.34%。Yuan等[85]设计了一种基于CNN和Transformer的互补网络。该网络在编码器阶段设计Swin Transformer和ResNet双编码路径产生互补特征，并利用跨域融合块进行特征融合；解码器采用Swin Transformer提高远程依赖关系的表示能力。

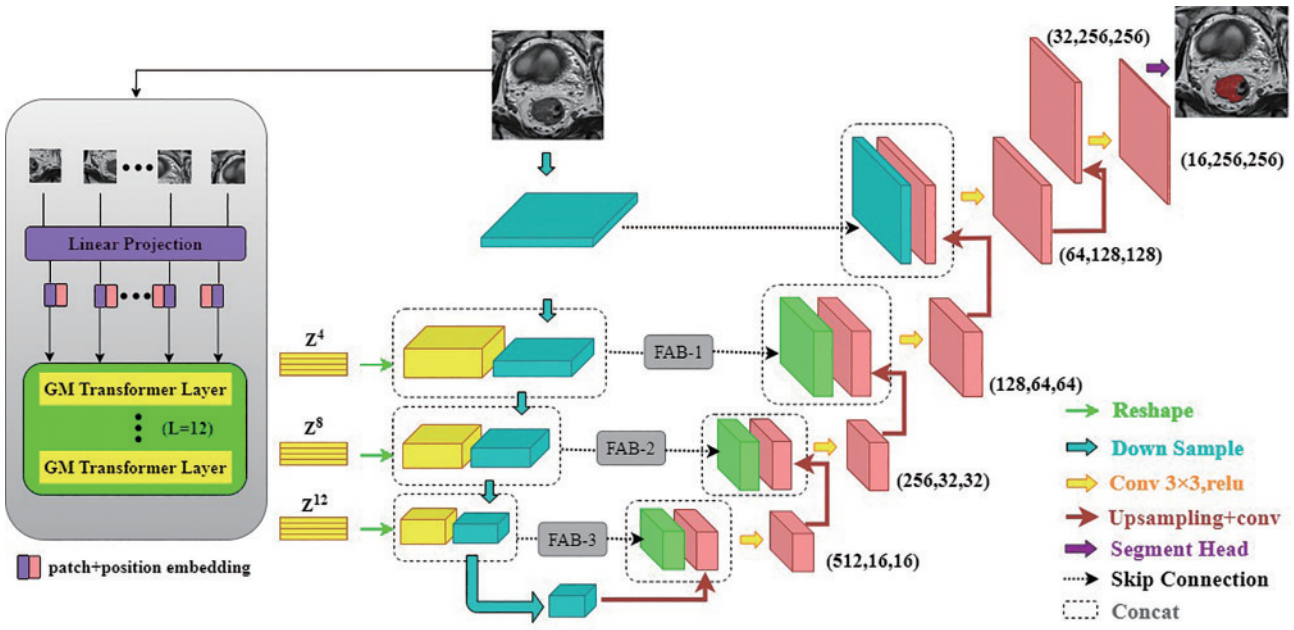

图11　DuPNet结构图[84]

Fig. 11　DuPNet structure diagram[84]

以上研究在编码器阶段利用CNN和Transformer并行结构提取不同特征并进行融合，实现了对全局和局部信息的充分捕获。另一些研究则提出了带有双解码器的U形结构，在解码层中学习更为详尽的图像特征信息和语义细节。Zhang等[86]设计的HSNet网络采用了PVT编码器生成底层特征，并通过CSA模块抑制底层特征的噪声信息；在解码器阶段则利用Transformer和CNN构成的双分支结构，从而有效地表示特征并在混合语义上恢复细节信息。Huang等[87]提出了带有双解码器的混合模型TDD-UNet来分割COVID-19感染区域。该网络解决了卷积局部化的问题，并引入背景分割分支改善传统单分支分割网络的偏差，从而在解码层中细化病灶的边界，得到更准确的分割结果。

综上所述，整体并行结合方式在U形结构中引入独立的CNN和Transformer分支，以专注于各自擅长的任务，并通过特征融合或其他机制协同工作，充分结合两者的优势。该方式能够更好地平衡局部和全局信息，使得模型更具适应性和灵活性。

3.3.2.2　层级并行结合

部分研究则将双分支结构应用到编解码器的每个层级中。具体来说，每个层级的CNN和Transformer分支并行处理图像后产生两组特征图，并通过融合策略形成综合特征表示，然后传递给下一个层级的双分支结构。该结构能够更好地在每个层级提取图像的全局和局部特征，更全面地理解不同层级的图像内容。图12为层级并行结合概念性结构图。

李擎等[88]提出一种双分支网络模型UConvTrans，其结构如图13所示。该模型在编解码器的每个层级利用CNN和Transformer双分支提取局部特征和全局上下文信息，保留细节信息的同时抑制了图像中噪声和背景区域的干扰；其次设计了特征融合模块将二者提取的特征交互融合并传递给下一个层级。Cai等[89]将模型中的特征提取子模块设计成CNN和Transformer的并行结构，其中基于CNN的Conv Block3D模块负责学习图像中的短距离依赖信息，基于Swin Transformer的Swin Block3D模块负责学习图像中的远距离依赖信息，随后进行特征融合并将输出

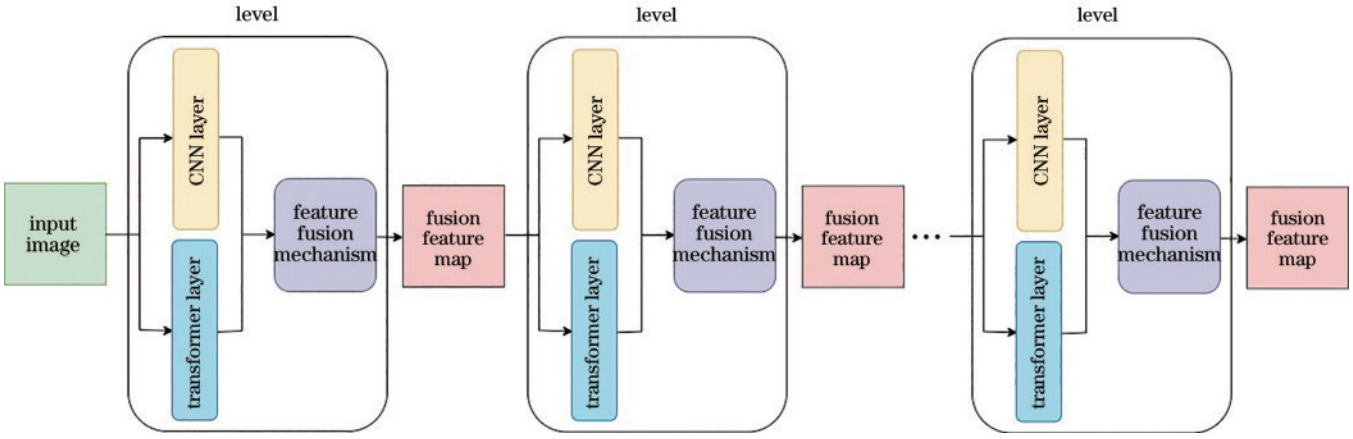

图 12　层级并行结合概念性结构图

Fig. 12　Conceptual structure diagram of hierarchical parallel combination

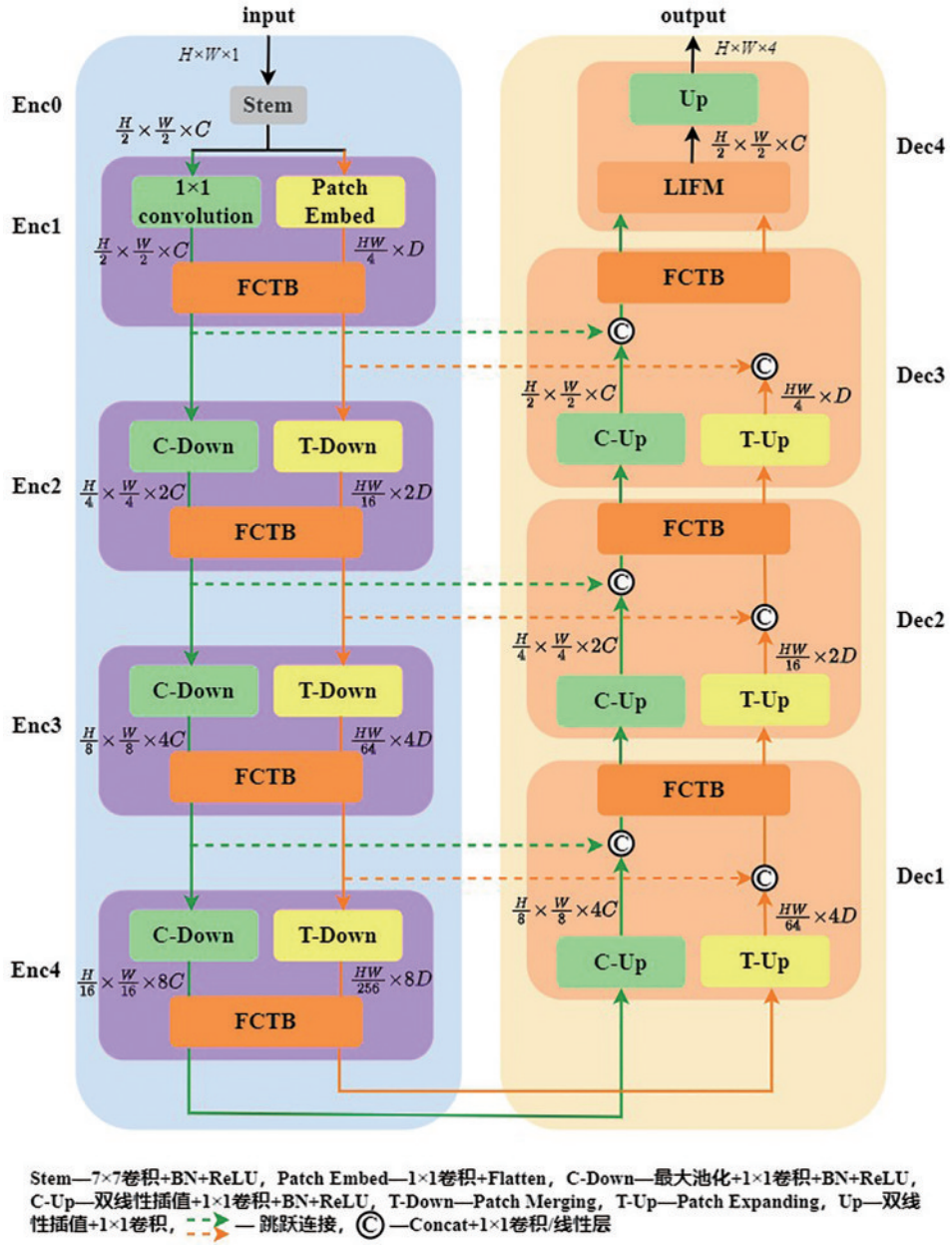

图 13　UconvTrans结构图[88]

Fig. 13　UconvTrans structure diagram[88]

的临时特征图像传递到下一个双分支模块。同样，Chong 等[90] 设计的 P-Transformer 模块使用 Transformer 层和卷积层并行结合进行特征提取，使网络能够模拟局部和全局信息，避免重复信息干扰。Zhang 等[91]针对腮腺肿瘤复杂多变的轮廓特征设计了 Transformer 与 CNN 双分支模块，应用到编码器各个层级，在每个层级更好地提取图像的全局和局部特征。

Liang 等[92]提出了并行网络 TransConver，该网络采用基于 CNN 和 Transformer 的 TCInception 并行模块堆叠而成，以提取局部和全局信息。通过 CAFGL 机制对局部和全局信息进行交叉关注融合，并利用跨注意融合跳跃连接机制消除编解码器特征之间的语义差异，实现更好的特征融合。Gai 等[93]提出了一种名为 RMTF-Net 的网络，该网络在编码器中利用 Mix Transformer 对图像进行处理得到多尺度特征图，并分层馈送到区域卷积神经网络（RCNN）中。同时，RCNN 将上一层级的特征映射输出与 MiT 的特征映射拼接在每个残差块上，并输入到下一个残差块；在特征解码器中，采用全局特征集成模块，利用全局关注特征丰富上下文。

通过在编解码器的每个层级中应用双分支结构，提升了对复杂场景和微小结构的分割能力，同时使网络更加灵活，能够适应不同尺度和层级的特征，有效地克服了传统单一结构模型在处理医学图像时的局限性，取得了更为精准的分割效果。

U-Net 和 Transformer 的混合模型已被广泛应用于医学图像分割任务中，取得了较好的成果。但也存在计算复杂度高、信息传递损失等问题。未来研究中用于医学图像分割的混合模型仍待进一步改进。表 3 从策略思想、优缺点和模型复杂度等角度对 5 种 U-Net＋Transformer 混合模型进行总结。

表 3　不同结合策略的总结

Table 3　Summary of different combination strategies

Combination strategy	Typology	Main idea	Advantage	Disadvantage	Model complexity
Serial	Single-scale serial	CNN and Transformer serially combined at a single scale	Easy to implement and understand	A loss of information between the CNN and the Transformer	Low
	Multi-scale serial	CNN serially combine with Transformers at different scales	Availability of multi-scale information	Increases complexity of managing different scales and detailed information may be lost when lowering resolution	Middle
	Alternating serial	Alternate between CNN and Transformer throughout the network	More balanced utilization of CNN and Transformer capabilities to improve information flow at each stage	Need to manage the interaction layer, which increases complexity and can cause information bottlenecks or redundancy if not carefully designed	High
Parallelism	Global parallelism	CNN and Transformer process the image in parallel and finally fuse the features	Taking full advantage of both networks, spatial and sequence information can be efficiently captured	Need to design effective feature fusion mechanisms with increased computation	High
	Hierarchical parallelism	CNN and Transformer are processed in parallel at multiple layers of the model, with feature fusion at each layer	Enhanced model generalization using two networks at each level to fully learn features	Need to design an effective feature fusion mechanism, complex structure and difficult to tune	High

4　算法分析与比较

深度学习已成为借助医学图像进行分割研究的热点。本文对基于 U-Net、Transformer、U-Net＋Transformer 的改进模型进行归纳总结，表 4 展示了部分改进方法的思想及优缺点。

基于 U-Net、Transformer 的医学图像分割方法在不同任务和图像类型下表现良好。表 5 对上述网络进行总结，从分割任务、图像类型、Dice、IoU 和 ACC 5 个方面进行对比分析。

表 6 对单尺度串行、多尺度串行、交替串行、整体并行和层级并行等 5 种类型的 U-Net＋Transformer 混合模型进行总结，从分割任务、图像类型、Dice、IoU 和 ACC 等方面进行对比分析。

表 4　部分改进模型的思想和优缺点对比

Table 4　Idea and comparison of advantages and disadvantages of partially improved models

Method		Model	Ideological improvement	Advantage	Disadvantage
U-Net		UNet++[15]	A dense jump connection strategy is used to introduce a deep supervision mechanism to motivate the network to better learn and utilize hierarchical features	Deeper and multi-scale feature extraction and fusion are realized to improve segmentation performance	Large number of model parameters and high computational complexity
		Attention U-Net[46]	AttentionGate attention mechanism is used to eliminate noise and irrelevant information of jump connections and extract key features	It effectively captures important local and global information in the image and improves the accuracy and performance of image segmentation	Model segmentation performance is improved to a lesser extent
Transformer		Swin-Unet[58]	Use Swin Transformer module instead of U-Net's 2D convolutional block	Multi-scale feature fusion, parameters and computational efficiency are optimized for applicability	Lack of extraction of local feature information
		C2Former[59]	Designing cross-convolutional self-attention mechanism algorithms to improve semantic feature understanding	Integration of multi-scale feature information and filtering of interfering information	Loss of feature information
U-Net+Transformer	Single-scale serial	TransUNet[62]	Local features are extracted using a CNN and the Transformer module is strung after the CNN to extract global contextual information	Perceiving global information for direct processing of sequential data, adaptation to multi-scale tasks	High computational complexity
		MultiIB-TransUNet[67]	A single Transformer layer is used to extract global features to reduce the number of parameters, and multiple IB blocks are utilized to compress noise and improve robustness	Reduced amount of model parameters	Compressing the relevant features, the segmentation accuracy is relatively reduced
	Multi-scale serial	CoTr[70]	Transformer in the encoder receives the feature maps at multiple scales in the CNN and introduces a deformable self-attention mechanism that focuses on some of the key sampling points	Computational and spatial complexity is reduced; multi-scale processing of 3D feature maps is realized	Insufficient generalizability
		HTUNet[73]	2D U-Net design MSCAT module is used to jump connections and extract intra-frame features, and the 3D Transformer UNet extracts inter-frame features	Intra- and inter-frame feature fusion for improved segmentation performance	Failure to delineate the nodal region from the gland is computationally expensive
	Alternating serial	HCTNet[76]	TEBlocks module is designed to learn global context information and combine it with CNN blocks to extract features	Combining CNN inductive bias for spatial association modeling and Transformer capabilities for remote dependency modeling	Boundary detail segmentation is not effective

表 4（续）

Method		Model	Ideological improvement	Advantage	Disadvantage
U-Net＋Transformer	Alternating serial	UTNet[79]	Replace the last convolution in each resolution of U-Net with the Transformer module	Combines convolutional and self-attentive mechanisms that are easy to understand and use	Poor segmentation performance in large-scale training tasks
	Global parallelism	Transfuse[82]	CNN and Transformer extract features in parallel, and BiFusion module is designed to fuse features from both branches	Capturing global information while maintaining sensitivity to low-level context, with strong fusion representation	Transformer layer is less efficient and insufficiently generalized
		HSNet[86]	Encoder and decoder are connected through an interactive attention mechanism, and the decoder has a two-branch structure	Discriminative remote dependencies can be generated, detail features can be recovered, and generalization ability is high	High model complexity
	Hierarchical parallelism	UconvTrans[88]	Each level uses a two-branch structure and a feature fusion module is designed to pass the fused features to the next level	Number of model parameters and the amount of computation are less, which better balances the segmentation accuracy and efficiency	Not applicable to 3D medical image segmentation where slice information is richer
		RMTF-Net[93]	Encoder combines Mix Transfomer and RCNN structures, and the feature decoder is designed with a GFI module to re-fuse the feature information extracted by the encoder	Network boundary coding is more capable and a local global balanced feature is available at each level of the encoder	Not extended to 3D image segmentation

表 5　基于 U-Net、Transformer 的分割方法对比

Table 5　Comparison of segmentation methods based on U-Net and Transformer

Method	Model	Split task	Image type	Dice /%	IoU /%	ACC /%
U-Net	UNet＋＋[15]	Nucleus/colon polyp/liver/pulmonary nodule segmentation	Microscopy/RGB video/CT/CT	—	92.52/32.10/82.90/77.21	—
	UNet 3＋[43]	Liver/spleen segmentation	CT/CT	96.75/96.20	—	—
	Attention U-Net[46]	Pancreatic segmentation	CT	81.48	—	—
	R2U-Net[48]	Retinal vascular/skin lesion/pulmonary nodule segmentation	Fundus image/dermoscopy/CT	—	—/94.21/99.18	97.12/94.24/99.18
Transformer	Swin-Unet[58]	Abdominal multi-organ/heart segmentation	CT/MRI	79.13/90.00	—	—
	C2Former[59]	Abdominal multi-organ/heart / skin cancer segmentation	CT/MRI/dermoscopy	83.22/91.42/86.78	—	—
	MISSFormer[60]	Abdominal multi-organ/heart segmentation	CT/MRI	81.96/90.86	—	—
	DS-TransUNet[61]	Polyp/dermatological/glandular/nuclear segmentation	Endoscopy/dermatoscopy/Pathology/microscopy	93.5/—/87.19/—	88.9/85.23/78.45/86.12	—

表 6　基于 U-Net＋Transformer 的分割方法对比

Table 6　Comparison of segmentation methods based on U-Net＋Transformer

Combination strategy	Typology	Model	Split task	Image type	Dice /%	IoU /%	ACC /%
Serial connection	Single-scale serial	TransUNet[62]	Multi-organ/cardiac segmentation	CT/MRI	77.48/89.71	—	—
		TransUNet＋[63]	Multi-organ/glandular/heart segmentation	CT/filmstrip/MRI	81.57/90.42/90.47	—/82.69/—	—
		PKRT-Net [64]	Videocups and videodiscs segmentation	Fundus image	91.20/97.66	—	—
		TU-Net[65]	Vascular segmentation	US	—	92.00/85.00/67.00	—
		GL-Segnet[66]	Rectal adenocarcinoma cell/skin lesion/glioma/thoracic organ segmentation	Pathology/dermoscopy/CT/X-ray	93.10/91.50/93.10/96.90	87.30/85.80/87.70/94.20	—
		MultiIB-TransUNet[67]	Mammary gland segmentation	US/CT	—/81.83	67.75/—	—
		UNETR[68]	Abdominal multi-organ/brain tumor/spleen segmentation	CT/MRI/CT	89.10/71.10/96.40	—	—
		UMSTC[69]	Microscope image	Microscopy	82.50	76.50	—
	Multi-scale serial	CoTr[70]	Cranial vault multi-organ segmentation	CT	85.00	—	—
		Multi-compound Transformer[71]	Cell/colon/skin lesion segmentation	Microscopy/endoscopy/dermoscopy	68.40/92.30/90.35	—	—
		TFNet[72]	Mammary gland segmentation	US	87.90	78.40	—
		HTUNet[73]	Thyroid segmentation	US	98.59	97.26	—
		TransHRNet[74]	Abdominal multi-organ/brain tumor/spleen segmentation	CT/MRI/CT	86.70/72.50/97.40	—	—
		Swin UNETR[75]	Brain tumor segmentation	MRI	91.30	—	—
	Alternating serial	HCTNet[76]	Mammary gland segmentation	US	97.23	94.63	97.41
		Feature integration network[77]	Abdominal multi-organ/prostate segmentation	CT/MRI	92.45/81.63	—	—
		SWTRU[78]	Liver and tumor/skin lesion/glioma segmentation	CT/Dermoscopy/MRI	97.20/90.40/89.70	94.90/—/—	—
		UTNet[79]	Heart segmentation	MRI	88.30	—	—
		nnFormer[80]	Brain tumor/abdominal multi-organ/ heart segmentation	MRI/CT/MRI	86.40/86.57/92.06	—	—
		SwinBTS[81]	Brain tumor segmentation	MRI	81.15	81.10	—
Parallel connection	Global parallelism	Transfuse[82]	Polyp/skin lesion/hip/prostate segmentation	Endoscopy/Dermoscopy/X-ray/MRI	94.20/87.20/—/—	89.70/—/—/—	—/94.40/—/—

表 6（续）

Combination strategy	Typology	Model	Split task	Image type	Dice /%	IoU /%	ACC /%
Parallel connection	Global parallelism	TransFusionNet[83]	Liver tumor/vascular segmentation	CT/CT	96.10/90.10	92.70/85.40	—
		DuPNet[84]	Rectal cancer segmentation	MRI	98.22	89.34	—
		CTC -Net[85]	Multi-organ/heart segmentation	CT/MRI	78.41/90.77	—	—
		HSNet[86]	Polyp segmentation	Endoscopy	92.60	87.70	—
		TDD-UNet[87]	COVID-19 pneumonia segmentation	CT、X-ray	78.94	—	96.34
	Hierarchical parallelism	UconvTrans[88]	Heart segmentation	MRI	89.60	—	—
		Swin Unet3D[89]	Brain tumor segmentation	MRI	90.50	—	—
		P-TransUNet[90]	Polyp/nucleus/glandular segmentation	Endoscopy/Microscopy/Pathology	93.52/93.63/95.93	88.93/88.75/91.42	—
		PCT[91]	Parotid tumor segmentation	US	91.51	84.34	—
		TransConver[92]	Brain tumor segmentation	MRI	86.32	—	—
		RMTF-Net[93]	Brain tumor segmentation	MRI	93.50	88.20	—

由表 5 和表 6 可知：U-Net 变体网络在细胞核、结肠息肉和视网膜血管等多个任务中表现出色。基于 Transformer 的模型在腹部多器官、心脏等任务中取得较优分割性能。混合模型中，在串行连接方面，单尺度串行结合网络 TransUNet 在多器官和心脏分割任务中表现出良好的性能，TU-Net 在血管分割任务中取得了较高的 IoU 分数。多尺度串行结合方法如 TFNet 和 HTUNet 在乳腺和甲状腺分割任务中显示出有效的多尺度信息利用。此外，交替串行连接的模型，如 HCTNet 在乳腺分割任务上的 Dice 分数高达 97.23%。在并行连接方面，Transfuse 通过整体并行方式适应多任务场景，TransFusionNet 在 CT 图像上实现了肝脏肿瘤和血管的高效分割。层级并行的模型 SWTRU 在肝脏和肿瘤分割任务中分别达到了 97.20% 的 Dice 分数和 94.90% 的 IoU 分数。其他模型如 TU-Net、DuPNet 和 RMTF-Net 等在不同分割任务中展现出各自的优势，为医学图像分割提供了多样而灵活的结合方式。

综上所述，这些模型在不同的医学图像分割任务中展现了各自的优势和适用性，但每种方法仍然存在一些挑战和局限性，如标注需求、模型复杂度和适用性等，需根据具体应用场景进行适当选择。

5　总结与展望

5.1　总结

本文系统阐述了 U-Net、Transformer 及 U-Net+Transformer 混合模型在医学图像分割领域的应用，并对 5 种不同的混合模型策略进行深入分析。其中：串行结合方式采用分阶段逐步深入的方法，精确挖掘提炼图像特征；并行结合方式则通过同步处理实现对图像局部细节和全局上下文的全面理解。混合模型结合了 CNN 的局部特征提取能力和 Transformer 的全局建模能力，克服了传统卷积神经网络在长距离依赖、全局信息捕捉及大规模数据处理方面的局限性。同时，通过整合局部和全局特征，混合模型提高了对复杂结构的识别能力，并在处理高分辨率图像时表现出色，显著提升了分割精度。这些融合策略的成功应用证明了多模型融合在提升医学图像分割任务精度和鲁棒性方面的巨大潜力，为未来医学图像分割技术的发展指明了方向。总之，U-Net+Transformer 混合模型在医学图像分割任务中已取得巨大进步，但从长远发展来看，依然存在以下亟待解决的问题：

1）医学图像数据集稀缺。混合模型需要大量数据集作为训练支撑，才能充分发挥 Transformer 模块捕捉长距离依赖的优势。但医学图像的标注费时费力，故高质量、精确标注的医学图像数据相对稀缺。此外，图像中病变区域或器官组织的面积相对较小，例如皮肤病变、甲状腺结节等，导致目标区域的像素数量远少于其他区域，产生类别不平衡问题。这种情况下，微小的边界定位误差也可能严重影响混合模型的性能。

2）信息丢失与融合困难。混合模型中存在模块的串行和并行设计。在串行设计中，CNN 提取的局部特征传递到 Transformer 时需要进行序列化和块处理，使得部分重要的细节信息，如心脏的边缘轮廓、肝脏的颗粒状纹理等无法被充分捕捉和利用，从而导致信息丢失，影响分割精度；在并行设计中，CNN 的卷积特征和 Transformer 的注意力特征具有不同的表示方式和尺度，从而导致两种特征的有效融合具有挑战性，需要设计复杂的融合机制。

3）训练与调优困难。混合模型结合了CNN和Transformer两种结构，训练过程复杂且难以调试，容易出现梯度消失或爆炸的问题，导致训练过程中的不稳定性。此外，CNN和Transformer对于不同超参数的敏感性存在差异，模型运行过程中需对两种模块的参数进行联合调优，优化过程复杂，需要大量实验来找到最佳超参数组合。

4）模型计算复杂度高。混合模型需同时维护U-Net的卷积层和Transformer的自注意力机制，导致计算复杂度高。例如，TransFusionNet[83]设计了基于CNN和Transformer的特征提取模块，虽然增强了细节特征的提取能力，但模型推理速度受限。同时，Transformer模块处理图像任务需要将图像序列化，形成较长的序列。尽管ViT提出了图像块序列，但图像切割后计算量仍然较大。

5）模型泛化能力不足。混合模型大多专注于单一器官或肿瘤的分割任务，训练数据集与新数据集之间存在特征分布的差异，导致混合模型在其他数据集或任务上的泛化能力不足。例如，UConvTrans[88]在二维图像分割中表现良好，但不适用于三维医学影像分割。

5.2　展望

近年来，U-Net＋Transformer混合模型在医学图像分割领域的应用愈发广泛，结合目前混合模型的发展现状和所面临的挑战，未来研究建议从以下几个方面继续深入：

1）解决医学图像数据集稀缺问题。需要建立高质量标注数据集，同时也需要将混合模型与有限标注数据的高效学习方法结合起来，如小样本学习方法。该方法涵盖了元学习、迁移学习、自监督学习，以及对抗性训练等策略，通过增强模型的适应性、充分利用未标注数据以及创造新训练样本来实现小样本数据的精准分割。

2）解决信息丢失和融合困难问题。可以在混合模型的不同模块或层级之间引入动态路由机制，该机制能够根据输入数据的特征动态调整信息流的路径，最大限度地减少信息丢失。特征融合机制可以结合动态加权融合、注意力机制等方法，实现对特征融合过程的灵活控制。目前图神经网络在处理非结构化数据和图像数据中取得显著的进展，应探索将图神经网络应用于特征关系学习和信息传递过程中，以解决信息丢失和特征融合问题。

3）解决训练和调优困难问题。可以对混合模型采用分步训练调优策略，对CNN和Transformer模块分别训练，使其在各自的任务上收敛到较好的状态，然后将两个模块进行联合训练，通过微调来调整两者之间的交互，以提高整体性能。此外，也可以利用自动超参数优化方法，如贝叶斯优化、网格搜索或进化算法等，高效找到最佳的超参数组合。

4）解决计算复杂度高的问题。可以采用网络剪枝技术减少混合模型中不重要的通道、层级和分支或减少其参数数量，如卷积层或Transformer的注意力权重等。同时，基于轻量化卷积神经网络如MobileNet、EfficientNet和优化的Transformer变体如Efficient Transformer设计轻量化混合模型，保持性能的同时减小模型规模和计算负担。此外，还可以采用量化、知识蒸馏等先进技术来有效解决模型计算复杂度高的问题。

5）解决模型泛化能力不足问题。可以将医学先验知识与混合模型结合起来，在CNN模块的输入层或中间层加入先验知识编码。例如，将解剖结构模板与原始图像堆叠在一起作为模型的输入，或者设计独立的通道来处理医学先验知识，然后将先验知识通道和混合模型通道的特征进行拼接或融合，以提升模型的泛化能力。此外，SAM等大模型广泛适用于医学图像分析领域，未来应探索将大模型与混合模型结合起来，以提升混合模型在像素级分类中的准确性和泛化能力。

总结而言，本文综述了U-Net、Transformer以及U-Net＋Transformer混合模型在医学图像分割中的应用，并深入探讨了不同模型策略的性能。同时，本文也指出了混合模型在数据集稀缺、信息传递、模型泛化等方面的挑战。为此，本文提出了一系列针对性的研究方向，包括引入小样本学习方法、动态路由机制、医学先验知识，以及利用大模型等，旨在推动医学图像分割技术的进步。随着深度学习技术的不断发展，可以预见，未来医学图像分割领域将迎来更加全面和深入的新阶段。

参　考　文　献

[1] 张建新，刘冬伟，张睦卿，等. 多尺度非局部自注意力MRI脑肿瘤分割网络[J]. 计算机系统应用，2024，33(2): 143-150.
Zhang J X, Liu D W, Zhang M Q, et al. Multi-scale non-local self-attention MRI brain tumor segmentation network[J]. Computer Systems & Applications, 2024, 33(2): 143-150.

[2] 步洪禧，何利文. 基于MAU-Net的CT多器官分割[J]. 计算机系统应用，2024，33(3): 103-110.
Bu H X, He L W. Multi-organ segmentation of CT based on MAU-Net[J]. Computer System Application, 2024, 33(3): 103-110.

[3] Malhotra P, Gupta S, Koundal D, et al. Deep neural networks for medical image segmentation[J]. Journal of Healthcare Engineering, 2022, 2022: 9580991.

[4] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[5] 周金治，胡震，郭莉莉，等. 基于GAN-DAUnet的肝脏CT图像肿瘤分割算法[J]. 中国医学物理学杂志，2023，40(8): 971-976.
Zhou J Z, Hu Z, Guo L L, et al. Liver CT image tumor

segmentation algorithm based on GAN-DAUnet[J]. Chinese Journal of Medical Physics, 2023, 40(8): 971-976.

[6] 徐蓬泉, 梁宇翔, 李英. 融合多尺度语义和剩余瓶颈注意力的医学图像分割[J]. 计算机工程, 2023, 49(10): 162-170.
Xu P Q, Liang Y X, Li Y. Medical image segmentation fusing multi-scale semantic and residual bottleneck attention[J]. Computer Engineering, 2023, 49(10): 162-170.

[7] 邬硕, 汪海涛, 姜瑛, 等. 基于混合神经网络的脑部MRI图像语义分割算法[J]. 重庆邮电大学学报(自然科学版), 2022, 34(3): 423-432.
Wu S, Wang H T, Jiang Y, et al. Semantic segmentation algorithm of brain MRI image based on hybrid neural network[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2022, 34(3): 423-432.

[8] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229-1251.
Zhou F Y, Jin L P, Dong J. Review of convolutional neural network[J]. Chinese Journal of Computers, 2017, 40(6): 1229-1251.

[9] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[10] Fukushima K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological Cybernetics, 1980, 36(4): 193-202.

[11] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[EB/OL]. (2014-12-22) [2024-02-05]. http://arxiv.org/abs/1412.7062v4.

[12] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[M]//Navab N, Hornegger J, Wells W M, et al. Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science. Cham: Springer, 2015, 9351: 234-241.

[13] 刘云鹏, 蔡文立, 洪国斌, 等. 应用图像块和全卷积神经网络的肩关节MRI自动分割[J]. 中国图象图形学报, 2018, 23(10): 1558-1570.
Liu Y P, Cai W L, Hong G B, et al. Automatic segmentation of shoulder joint in MRI by using patch-wise and full-image fully convolutional networks[J]. Journal of Image and Graphics, 2018, 23(10): 1558-1570.

[14] 杨志秀, 韩建宁, 于本知, 等. 一种改进的DeepLab V3+的医学图像分割方法[J]. 国外电子测量技术, 2021, 40(9): 18-23.
Yang Z X, Han J N, Yu B Z, et al. Improved medical image segmentation method based on DeepLab V3 +[J]. Foreign Electronic Measurement Technology, 2021, 40(9): 18-23.

[15] Zhou Z W, Siddiquee M M R, Tajbakhsh N, et al. UNet++: a nested U-Net architecture for medical image segmentation[EB/OL]. (2018-07-18) [2024-05-06]. https://arxiv.org/abs/1807.10165.

[16] 钟经纬. PRA-UNet3+: 全尺度跳跃连接CT肝脏图像分割模型[J]. 软件导刊, 2023, 22(2): 15-20.
Zhong J W. PRA-UNet3+: full-scale connected CT liver image segmentation model[J]. Software Guide, 2023, 22(2): 15-20.

[17] 李林静, 侯军浩, 吴建峰, 等. 改进3D U-NET在CT影像分割中的应用研究[J]. 现代信息科技, 2021, 5(21): 105-107, 111.
Li L J, Hou J H, Wu J F, et al. Research on application of improved 3D U-NET in CT image segmentation[J]. Modern Information Technology, 2021, 5(21): 105-107, 111.

[18] 张欢, 仇大伟, 冯毅博, 等. U-Net模型改进及其在医学图像分割上的研究综述[J]. 激光与光电子学进展, 2022, 59(2): 0200005.
Zhang H, Qiu D W, Feng Y B, et al. Improved U-Net models and its applications in medical image segmentation: a review[J]. Laser & Optoelectronics Progress, 2022, 59(2): 0200005.

[19] 杨鹤, 柏正尧. CoT-TransUNet: 轻量化的上下文Transformer医学图像分割网络[J]. 计算机工程与应用, 2023, 59(3): 218-225.
Yang H, Bai Z Y. CoT-TransUNet: lightweight context Transformer medical image segmentation network[J]. Computer Engineering and Applications, 2023, 59(3): 218-225.

[20] 袁媛, 陈明惠, 柯舒婷, 等. 基于集成卷积神经网络和Vit的眼底图像分类研究[J]. 中国激光, 2022, 49(20): 2007205.
Yuan Y, Chen M H, Ke S T, et al. Fundus image classification research based on ensemble convolutional neural network and vision transformer[J]. Chinese Journal of Lasers, 2022, 49(20): 2007205.

[21] 胡晓阳, 李哲. 基于卷积神经网络和Transformer的肝脏CT图像分割方法[J]. 中国医学物理学杂志, 2023, 40(4): 423-428.
Hu X Y, Li Z. Liver CT image segmentation method based on CNN and Transformer[J]. Chinese Journal of Medical Physics, 2023, 40(4): 423-428.

[22] Cheng J H, Liu J, Kuang H L, et al. A fully automated multimodal MRI-based multi-task learning for glioma segmentation and IDH genotyping[J]. IEEE Transactions on Medical Imaging, 2022, 41(6): 1520-1532.

[23] Zhang Z X, Jiang S H, Pan X H. CTransNet: convolutional neural network combined with transformer for medical image segmentation[J]. Computing and Informatics, 2023, 42(2): 392-410.

[24] 张玮智, 于谦, 苏金善, 等. 从U-Net到Transformer: 深度模型在医学图像分割中的应用综述[J/OL]. 计算机应用: 1-23[2024-03-01]. http://kns.cnki.net/kcms/detail/51.1307.tp.20231026.1648.002.html.
Zhang W Z, Yu Q, Su J S, et al. From U-Net to Transformer: a review of the application of deep models in medical image segmentation[J/OL]. Computer Applications: 1-23[2024-03-01]. http://kns.cnki.net/kcms/

detail/51.1307.tp.20231026.1648.002.html.

[25] Ou X Y, Chen X, Xu X N, et al. Recent development in X-ray imaging technology: future and challenges[J]. Research, 2021, 2021: 9892152.

[26] 王鸿飞, 马士青, 闵雷, 等. 基于图像分割和全变分的肺CT图像增强[J]. 中国激光, 2022, 49(20): 2007210.
Wang H F, Ma S Q, Min L, et al. Lung CT image enhancement based on image segmentation and total variational[J]. Chinese Journal of Lasers, 2022, 49(20): 2007210.

[27] 穆根, 张振辉, 石玉娇. 生物医学影像中的光声成像技术[J]. 中国激光, 2022, 49(20): 2007208.
Mu G, Zhang Z H, Shi Y J. Photoacoustic imaging technology in biomedical imaging[J]. Chinese Journal of Lasers, 2022, 49(20): 2007208.

[28] Khodrog O A A A. CT及MRI影像组学在喉癌诊断中的应用[D]. 长春: 吉林大学, 2023.
Khodrog O A A A. The application of CT and MRI radiomics in laryngeal cancer diagnosis[D]. Changchun: Jilin University, 2023.

[29] Weimar E A M, Huang S H, Lu L, et al. Radiologic-pathologic correlation of tumor thickness and its prognostic importance in squamous cell carcinoma of the oral cavity: implications for the eighth edition tumor, node, metastasis classification[J]. American Journal of Neuroradiology, 2018, 39(10): 1896-1902.

[30] 韩悦, 张永寿, 郭依廷, 等. 乳腺癌腋窝淋巴结超声图像分割算法研究[J]. 南京师大学报(自然科学版), 2021, 44(4): 122-126, 134.
Han Y, Zhang Y S, Guo Y T, et al. Research on ultrasound image segmentation algorithm for axillary lymph node with breast cancer[J]. Journal of Nanjing Normal University (Natural Science Edition), 2021, 44(4): 122-126, 134.

[31] 崔珂, 田启川, 廉露. 基于U-Net变体的医学图像分割算法综述[J]. 计算机工程与应用, 2024, 60(11): 32-49.
Cui K, Tian Q C, Lian L. Review of medical image segmentation algorithm based on U-Net variants[J]. Computer Engineering and Applications, 2024, 60(11): 32-49.

[32] 李涵生. 面向临床诊断的病理图像检测算法研究[D]. 西安: 西北大学, 2022.
Li H S. Research on pathological image detection algorithm for clinical diagnosis[D]. Xi'an: Northwest University, 2022.

[33] 姜杨, 刘成, 丁其川, 等. 基于双注意力机制的COVID-19病灶CT图像分割方法[J]. 东北大学学报(自然科学版), 2023, 44(9): 1259-1268.
Jiang Y, Liu C, Ding Q C, et al. Segmentation of COVID-19 CT images based on dual attention mechanism[J]. Journal of Northeastern University (Natural Science), 2023, 44(9): 1259-1268.

[34] Bilic P, Christ P, Li H B, et al. The liver tumor segmentation benchmark (LiTS) [J]. Medical Image Analysis, 2023, 84: 102680.

[35] 陈柏年, 韩雨童, 何涛, 等. 基于级联动态注意力U-Net的脑肿瘤分割方法[J]. 计算机科学, 2023, 50(S2): 1031-1037.
Chen Bonian, Han Yutong, He Tao, et al. Brain tumor segmentation method based on cascaded dynamic attention U-Net [J]. Computer Science, 2023, 50 (S2): 1031-1037

[36] Dai W H, Li X M, Ding X P, et al. Cyclical self-supervision for semi-supervised ejection fraction prediction from echocardiogram videos[J]. IEEE Transactions on Medical Imaging, 2023, 42(5): 1446-1461.

[37] 张婷, 秦涵书, 赵若璇. 基于多尺度注意力融合网络的胃癌病理图像分割方法[J]. 电子技术应用, 2023, 49(9): 46-52.
Zhang T, Qin H S, Zhao R X. Gastric cancer pathological image segmentation method based on multi-scale at-tention fusion network[J]. Application of Electronic Technique, 2023, 49(9): 46-52.

[38] Valanarasu J M J, Oza P, Hacihaliloglu I, et al. Medical transformer: gated axial-attention for medical image segmentation[M]//de Bruijne M, Cattin P C, Cotin S, et al. Medical image computing and computer assisted intervention-MICCAI 2021. Lecture notes in computer science. Cham: Springer, 2021, 12901: 36-46.

[39] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.

[40] 郭宁, 柏正尧. 注意力机制下密集空洞卷积的肺部图像分割[J]. 中国图象图形学报, 2021, 26(09): 2146-2155.
Guo Ning, Bai Zhengyao. Lung image segmentation using dense dilated convolution under attention mechanism [J]. Chinese Journal of Image and Graphics, 2021, 26 (09): 2146-2155

[41] Huang Z L, Wang X G, Huang L C, et al. CCNet: criss-cross attention for semantic segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2019: 603-612.

[42] 周涛, 侯森宝, 陆惠玲, 等. C^2 Transformer U-Net: 面向跨模态和上下文语义的医学图像分割模型[J]. 电子与信息学报, 2023, 45(5): 1807-1816.
Zhou T, Hou S B, Lu H L, et al. C^2 Transformer U-Net: a medical image segmentation model for cross-modality and contextual semantics[J]. Journal of Electronics & Information Technology, 2023, 45(5): 1807-1816.

[43] Huang H M, Lin L F, Tong R F, et al. UNet 3+: a full-scale connected UNet for medical image segmentation [C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 4-8, 2020, Barcelona, Spain. New York: IEEE Press, 2020: 1055-1059.

[44] 胡扬涛, 裴洋, 林川, 等. 空洞残差U型网络用于视网膜血管分割[J]. 计算机工程与应用, 2021, 57(7): 185-191.
Hu Y T, Pei Y, Lin C, et al. Atrous residual U-Net for retinal vessel segmentation[J]. Computer Engineering and

Applications, 2021, 57(7): 185-191.

[45] Liu C B, Xie H T, Zhang S C, et al. Misshapen pelvis landmark detection with local-global feature learning for diagnosing developmental dysplasia of the hip[J]. IEEE Transactions on Medical Imaging, 2020, 39(12): 3944-3954.

[46] Oktay O, Schlemper J, Le Folgoc L, et al. Attention U-Net: learning where to look for the pancreas[EB/OL]. (20108-04-11)[2024-05-06]. http://arxiv.org/abs/1804.03999v3.

[47] Zhang J X, Jiang Z K, Dong J, et al. Attention gate ResU-net for automatic MRI brain tumor segmentation [J]. IEEE Access, 2020, 8: 58533-58545.

[48] Alom M Z, Hasan M, Yakopcic C, et al. Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation[EB/OL]. (2018-02-20)[2021-04-20]. https://arxiv.org/abs/1802.06955.

[49] Qamar S, Jin H, Zheng R, et al. A variant form of 3D-UNet for infant brain segmentation[J]. Future Generation Computer Systems, 2020, 108: 613-623.

[50] Fu H Z, Cheng J, Xu Y W, et al. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation[J]. IEEE Transactions on Medical Imaging, 2018, 37(7): 1597-1605.

[51] 杨鑫, 李学妍, 张晓婷, 等. 基于自适应Unet网络的鼻咽癌放疗危及器官自动分割方法[J]. 南方医科大学学报, 2020, 40(11): 1579-1586.
Yang X, Li X Y, Zhang X T, et al. Segmentation of organs at risk in nasopharyngeal cancer for radiotherapy using a self-adaptive Unet network[J]. Journal of Southern Medical University, 2020, 40(11): 1579-1586.

[52] 王士奇. 基于改进U-Net的脑肿瘤MRI图像分割研究[D]. 长春: 长春工业大学, 2023.
Wang S Q. Research on MRI image segmentation of brain tumor based on improved U-Net[D]. Changchun: Changchun University of Technology, 2023.

[53] Isensee F, Jaeger P F, Kohl S A A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation[J]. Nature Methods, 2021, 18(2): 203-211.

[54] 黄泳嘉, 史再峰, 王仲琦, 等. 基于混合损失函数的改进型U-Net肝部医学影像分割方法[J]. 激光与光电子学进展, 2020, 57(22): 221003.
Huang Y J, Shi Z F, Wang Z Q, et al. Improved U-Net based on mixed loss function for liver medical image segmentation[J]. Laser & Optoelectronics Progress, 2020, 57(22): 221003.

[55] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[EB/OL]. (2017-06-12) [2024-02-05]. https://arxiv.org/abs/1706.03762.

[56] Dosovitskiy A, Beyer L, Kolesnikov A , et al. An image is worth 16 x 16 words: transformers for image recognition at scale[EB/OL]. (2020-10-22) [2024-05-06]. https://arxiv.org/abs/2010.11929.

[57] Liu Z, Lin Y T, Cao Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2021: 9992-10002.

[58] Cao H, Wang Y Y, Chen J, et al. Swin-Unet: Unet-like pure transformer for medical image segmentation[M]// Karlinsky L, Michaeli T, Nishino K. Computer vision-ECCV 2022 workshops. Lecture notes in computer science. Cham: Springer, 2023, 13803: 205-218.

[59] Wang J, Zhao H Y, Liang W, et al. Cross-convolutional transformer for automated multi-organs segmentation in a variety of medical images[J]. Physics in Medicine & Biology, 2023, 68(3): 035008.

[60] Huang X H, Deng Z F, Li D D, et al. MISSFormer: an effective medical image segmentation transformer[EB/OL]. (2021-09-15) [2024-05-06]. https://arxiv. org/abs/2109.07162.

[61] 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.

[62] Chen J N, Lu Y Y, Yu Q H, et al. TransUNet: transformers make strong encoders for medical image segmentation[EB/OL]. (2021-02-08) [2024-05-06]. https://arxiv.org/abs/2102.04306.

[63] Liu Y H, Wang H, Chen Z G, et al. TransUNet+: redesigning the skip connection to enhance features in medical image segmentation[J]. Knowledge-Based Systems, 2022, 256: 109859.

[64] Lu S, Zhao H, Liu H R, et al. PKRT-Net: prior knowledge-based relation transformer network for optic cup and disc segmentation[J]. Neurocomputing, 2023, 538: 126183.

[65] 李佳松, 曹洪帅, 舒丽霞, 等. 基于Transformer的血管内超声图像分割方法[J]. 北京生物医学工程, 2023, 42(1): 16-20, 51.
Li J S, Cao H S, Shu L X, et al. Transformer-based intravascular ultrasound image segmentation method[J]. Beijing Biomedical Engineering, 2023, 42(1): 16-20, 51.

[66] Gai D, Zhang J Q, Xiao Y S, et al. GL-Segnet: global-local representation learning net for medical image segmentation[J]. Frontiers in Neuroscience, 2023, 17: 1153356.

[67] Li G J, Jin D H, Yu Q, et al. MultiIB-TransUNet: Transformer with multiple information bottleneck blocks for CT and ultrasound image segmentation[J]. Medical Physics, 2024, 51(2): 1178-1189.

[68] Hatamizadeh A, Tang Y C, Nath V, et al. UNETR: transformers for 3D medical image segmentation[C]// 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 3-8, 2022, Waikoloa, HI, USA. New York: IEEE Press, 2022: 1748-1758.

[69] 孙开鑫, 刘斌, 苏曙光. 一种融合CNN和Swin Transformer的医学显微图像分割模型[J]. 计算机科学, 2023, 50(S2): 1023-1030.
Sun K X, Liu B, Su S G. A medical microscopic image segmentation model combining CNN and Swin Transformer[J]. Computer Science, 2023, 50(S2): 1023-1030.

[70] Xie Y T, Zhang J P, Shen C H, et al. CoTr: efficiently bridging CNN and transformer for 3D medical image segmentation[M]//de Bruijne M, Cattin P C, Cotin S, et al. Medical image computing and computer-assisted intervention- MICCAI 2021. Lecture notes in computer science. Cham: Springer, 2021, 12903: 171-180.

[71] Ji Y F, Zhang R M, Wang H J, et al. Multi-compound transformer for accurate biomedical image segmentation [EB/OL]. (2021-06-28)[2024-05-06]. https://arxiv.org/abs/2106.14385.

[72] Wang T, Lai Z H, Kong H. TFNet: transformer fusion network for ultrasound image segmentation[M]// Wallraven C, Liu Q, Nagahara H. Pattern recognition. Lecture notes in computer science. Cham: Springer, 2022, 13188: 314-325.

[73] Chi J N, Li Z L, Sun Z Y, et al. Hybrid transformer UNet for thyroid segmentation from ultrasound scans[J]. Computers in Biology and Medicine, 2023, 153: 106453.

[74] Yan Q S, Liu S Q, Xu S H, et al. 3D Medical image segmentation using parallel transformers[J]. Pattern Recognition, 2023, 138: 109432.

[75] Hatamizadeh A, Nath V, Tang Y C, et al. Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images[M]//Crimi A, Bakas S. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Lecture notes in computer science. Cham: Springer, 2022, 12962: 272-284.

[76] He Q Q, Yang Q J, Xie M H. HCTNet: a hybrid CNN-transformer network for breast ultrasound image segmentation[J]. Computers in Biology and Medicine, 2023, 155: 106629.

[77] Wang F, Wang B. Boundary-guided feature integration network with hierarchical transformer for medical image segmentation[J]. Multimedia Tools and Applications, 2024, 83(3): 8955-8969.

[78] Zhang J Y, Liu Y, Wu Q H, et al. SWTRU: star-shaped window transformer reinforced U-Net for medical image segmentation[J]. Computers in Biology and Medicine, 2022, 150: 105954.

[79] Gao Y H, Zhou M, Metaxas D N. UTNet: a hybrid transformer architecture for medical image segmentation [EB/OL]. (2021-07-02)[2024-05-06]. https://arxiv.org/abs/2107.00781.

[80] Zhou H Y, Guo J S, Zhang Y H, et al. nnFormer: volumetric medical image segmentation via a 3D transformer[J]. IEEE Transactions on Image Processing, 2023, 32: 4036-4045.

[81] Jiang Y, Zhang Y, Lin X, et al. SwinBTS: a method for 3D multimodal brain tumor segmentation using swin transformer[J]. Brain Sciences, 2022, 12(6): 797. [PubMed]

[82] Zhang Y, Liu H, Hu Q. Transfuse: fusing transformers and CNNs for medical image segmentation[M]//de Bruijne M, Cattin P C, Cotin S, et al. Medical image computing and computer-assisted intervention-MICCAI 2021. Lecture notes in computer science. Cham: Springer, 2021, 12901: 14-24.

[83] Wang X, Zhang X D, Wang G, et al. TransFusionNet: semantic and spatial features fusion framework for liver tumor and vessel segmentation under JetsonTX2[J]. IEEE Journal of Biomedical and Health Informatics, 2023, 27(3): 1173-1184.

[84] Zhang H T, Yang X T, Li D A, et al. Dual parallel net: a novel deep learning model for rectal tumor segmentation via CNN and transformer with Gaussian Mixture prior[J]. Journal of Biomedical Informatics, 2023, 139: 104304.

[85] Yuan F N, Zhang Z X, Fang Z J. An effective CNN and Transformer complementary network for medical image segmentation[J]. Pattern Recognition, 2023, 136: 109228.

[86] Zhang W C, Fu C, Zheng Y, et al. HSNet: a hybrid semantic network for polyp segmentation[J]. Computers in Biology and Medicine, 2022, 150: 106173.

[87] Huang X P, Chen J X, Chen M Z, et al. TDD-UNet: Transformer with double decoder UNet for COVID-19 lesions segmentation[J]. Computers in Biology and Medicine, 2022, 151: 106306.

[88] 李擎, 皇甫玉彬, 李江昀, 等. UConvTrans: 全局和局部信息交互的双分支心脏图像分割[J]. 上海交通大学学报, 2023, 57(5): 570-581.
Li Q, Huangfu Y B, Li J Y, et al. UConvTrans: a dual-flow cardiac image segmentation network by global and local information integration[J]. Journal of Shanghai Jiao Tong University, 2023, 57(5): 570-581.

[89] Cai Y M, Long Y Q, Han Z G, et al. Swin Unet3D: a three-dimensional medical image segmentation network combining vision transformer and convolution[J]. BMC Medical Informatics and Decision Making, 2023, 23(1): 33.

[90] Chong Y W, Xie N D, Liu X, et al. P-TransUNet: an improved parallel network for medical image segmentation[J]. BMC Bioinformatics, 2023, 24(1): 285.

[91] Zhang G, Zheng C H, He J F, et al. PCT: pyramid convolutional transformer for parotid gland tumor segmentation in ultrasound images[J]. Biomedical Signal Processing and Control, 2023, 81: 104498.

[92] Liang J J, Yang C H, Zeng M J, et al. TransConver: transformer and convolution parallel network for developing automatic brain tumor segmentation in MRI images[J]. Quantitative Imaging in Medicine and Surgery, 2022, 12(4): 2397-2415.

[93] Gai D, Zhang J Q, Xiao Y S, et al. RMTF-Net: residual mix transformer fusion net for 2D brain tumor segmentation [J]. Brain Sciences, 2022, 12(9): 1145.