

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2025)06-2161-26

论文引用格式: Shi J, Wang T T, Zhu Z Q, Zhao M F, Wang B X and An H. 2025. Deep learning-based medical image segmentation methods. Journal of Image and Graphics, 30(6): 2161-2186 (石军, 王天同, 朱子琦, 赵敏帆, 王炳勋, 安虹. 2025. 基于深度学习的医学图像分割方法综述. 中国图象图形学报, 30(6): 2161-2186) [DOI: 10. 11834/jig. 240467]

基于深度学习的医学图像分割方法综述

石军¹, 王天同¹, 朱子琦², 赵敏帆¹, 王炳勋¹, 安虹^{1*}

1. 中国科学技术大学计算机科学与技术学院, 合肥 230026; 2. 中国科学技术大学人工智能与数据科学学院, 合肥 230026

摘要: 医学图像分割是临床医学图像分析的重要组成部分, 目标是准确识别和分割医学图像中的人体解剖结构或病灶等感兴趣区域, 为临床疾病的诊断、治疗规划以及术后评估等应用场景提供客观、量化的决策依据。随着可用标注数据规模的不断增长, 基于深度学习的医学图像分割方法迅速发展, 展现出远超传统图像分割方法的精度和鲁棒性, 已成为该领域的主流技术。为了进一步提高分割精度, 大量研究集中在对分割模型的结构改进上, 产生了一系列结构迥异的分割方法。总的来说, 现有的基于深度学习的医学图像分割方法从模型结构上可以分为3类: 基于卷积神经网络(convolutional neural network, CNN)、基于视觉Transformer以及基于视觉Mamba。其中, 以U-Net为代表的基于CNN的方法最早广泛应用于各类医学图像分割任务。这类方法一般以卷积操作为核心, 能够有效提取图像的局部特征。相比之下, 基于视觉Transformer的方法更擅长捕捉全局信息和长距离依赖关系, 能够更好地处理复杂的上下文信息。基于视觉Mamba的方法作为一种新兴架构, 具有全局感受野和线性计算复杂度的特点, 表现出巨大的应用潜力。为了深入了解基于深度学习的医学图像分割方法的发展脉络、优势与不足, 本文对现有方法进行系统梳理和综述。首先简要回顾上述3类主流分割方法的结构演进历程, 分析不同方法的结构特点、优势与局限性, 然后从算法结构、学习方法和任务范式等多方面深入探讨医学图像分割领域面临的主要挑战及机遇, 最后对基于深度学习的医学图像分割方法的未来发展方向和应用前景进行深入分析和讨论。

关键词: 深度学习(DL); 医学图像分割; 卷积神经网络(CNN); 视觉Transformer(ViT); 视觉Mamba

Deep learning-based medical image segmentation methods

Shi Jun¹, Wang Tiantong¹, Zhu Ziqi², Zhao Minfan¹, Wang Bingxun¹, An Hong^{1*}

1. School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China;

2. School of Artificial Intelligence and Data Science, University of Science and Technology of China, Hefei 230026, China

Abstract: Medical image segmentation is a crucial component of clinical medical image analysis, aiming to accurately identify and delineate anatomical structures or regions of interest, such as lesions, within medical images. This process provides objective and quantitative support for decision-making in disease diagnosis, treatment planning, and postoperative evaluation. In recent years, the rapid expansion of available annotated datasets has driven the swift development of deep learning-based medical image segmentation methods. These methods have demonstrated superior accuracy and robustness compared to traditional segmentation techniques, establishing themselves as the mainstream technology in the field. Extensive research has been dedicated to improving the structural designs of segmentation models and further enhancing segmentation accuracy, leading to the development of various segmentation approaches. Current deep learning-based medical

收稿日期: 2024-08-07; 修回日期: 2024-11-01; 预印本日期: 2024-11-08

* 通信作者: 安虹 han@ustc.edu.cn

基金项目: 中国科学院战略性先导科技专项资助(XDB0500102)

Supported by: The Strategic Priority Research Program of Chinese Academy of Sciences(XDB0500102)

image segmentation methods can be classified into three main structural categories: convolutional neural networks (CNNs), Vision Transformers, and Vision Mamba. As a representative neural network architecture, CNNs effectively capture spatial features in images due to their unique local receptive fields and weight-sharing mechanisms. These characteristics make CNNs particularly suitable for image analysis and processing tasks. Since 2015, CNN-based methods, such as U-Net, have dominated the field of medical image segmentation, consistently achieving state-of-the-art performance across various downstream segmentation tasks. Many studies have focused on modifying and innovating the U-Net structure to further improve segmentation accuracy, resulting in several derived segmentation methods. Despite these advancements, CNN-based methods are still limited by the inherent constraints of convolutional operators, particularly the local nature of their receptive fields. These limitations restrict the capability of the model to capture global contextual dependencies, which are crucial for handling complex medical images and performing fine-grained segmentation. While techniques such as attention mechanisms and specialized convolutions have been introduced to address these challenges and help the model focus on global information, their effectiveness remains limited. Since 2020, researchers have begun introducing Transformer architectures, originally developed for natural language processing (NLP), into computer vision tasks, including medical image segmentation. Vision Transformers use self-attention mechanisms to effectively model global dependencies, drastically improving the quality of semantic feature extraction and facilitating the segmentation of complex medical images. Transformer-based methods for medical image segmentation mainly include hybrid approaches that combine Transformer with CNN, as well as pure Transformer methods, each displaying unique advantages and disadvantages. Hybrid approaches capitalize on the strengths of CNN in local feature extraction while leveraging the capabilities of Transformer in modeling global context. This combination enhances segmentation accuracy while maintaining computational efficiency. However, these hybrid methods remain dependent on CNN structures, which may limit their performance in complex scenarios. In contrast, pure Transformer methods excel at capturing long-range dependencies and multiscale features, leading to remarkable improvements in segmentation accuracy and generalization. However, pure Transformer architectures typically require substantial computational resources and high-quality training data, posing challenges, especially in the medical field, where large-scale annotated datasets are often difficult to obtain. Despite the notable advantages of Transformer architectures in capturing long-range dependencies and global contextual information, their computational complexity grows quadratically with the length of the input sequence, which limits their applicability in resource-constrained environments. Researchers are developing new methods that can model global dependencies with linear time complexity to overcome this challenge. For instance, Mamba introduces a novel selective state-space model that uses a selection mechanism, hardware-aware algorithms, and a streamlined architecture to reduce computational complexity while maintaining efficient performance in long-sequence modeling. Since 2024, numerous studies have started applying the Mamba structure to medical image segmentation tasks, achieving promising results and positioning it as a potential alternative to traditional Transformer structures. The hybrid method, combining Mamba with CNN, enhances segmentation accuracy and robustness by leveraging the feature extraction capabilities of CNN alongside the capability of Mamba to handle long-range dependencies. However, this approach may introduce additional computational complexity. In contrast, pure Mamba methods excel in tasks that require global contextual information but still face limitations in capturing spatial features within images and may demand higher computational resources during training. Overall, this paper systematically reviews and analyzes the development trajectory, advantages, and limitations of deep learning-based medical image segmentation methods from a structural perspective for the first time. First, all surveyed methods are categorized into three structural classes. Then, a brief overview of the structural evolution of different segmentation methods is provided, and their structural characteristics, strengths, and weaknesses are analyzed. Subsequently, the major challenges and opportunities encountered in the field of medical image segmentation are analyzed from multiple perspectives, including algorithm structure, learning methods, and task paradigms. Finally, an in-depth analysis and discussion of future development directions and application prospects are conducted.

Key words: deep learning (DL); medical image segmentation; convolutional neural network (CNN); vision Transformer (ViT); vision Mamba

0 引言

医学图像分割(medical image segmentation)是临床医学图像分析的关键任务之一,目标是从医学图像中识别和分割出人体解剖结构、疾病病灶等感兴趣区域(region of interest, ROI)。实现准确、快速的医学图像自动分割对精准医疗的发展具有重要意义,可以为疾病诊断、治疗规划及术后评估等应用场景提供客观、量化的决策依据。传统的医学图像分割方法主要基于规则或统计特征(郭艳蓉等, 2013),如水平集、边缘检测、区域生长和聚类等分割方法(钱芸和张英杰, 2008)。然而,这些方法大多依赖于手工设计的特征和规则,难以应对不同患者间的解剖结构差异和成像条件变化,导致在处理复杂的医学图像或分割目标时常表现出局限性。

近年来,随着设备计算能力和可用标注数据规模的持续增长,深度学习(deep learning, DL)技术快速发展,并在医学图像分析领域,特别是医学图像分割任务中得到了广泛应用。与传统方法不同,基于深度学习的分割方法利用深度神经网络(deep neural network, DNN)从标注数据中自动提取和学习关键图像特征,能够实现对图像上下文信息的高效建模,从而获得更为准确的分割结果。目前,基于深度学习的医学图像分割方法凭借卓越的分割精度和鲁棒性,已逐渐取代传统方法,成为该领域的主流技术。为了进一步提高对医学图像的分割精度,大量研究工作集中在对分割模型的结构改进上,由此产生了一系列结构迥异的分割方法。总的来说,现有的基于深度学习的医学图像分割方法经历了3次里程碑式的结构演进。

自2015年以来,以U-Net(Ronneberger等, 2015)为代表的基于卷积神经网络(convolutional neural network, CNN)的方法开始广泛应用于医学图像分割任务,并取得令人满意的结果(Du等, 2020; Zhou等, 2021)。这类方法大多采用编码器—解码器(encoder-decoder)架构,主要利用卷积操作来提取图像特征,并通过下采样和上采样过程实现对分割图像的压缩和重建(Siddique等, 2021)。其核心优势在于能够高效提取图像的多尺度特征,并通过跳跃连接(skip connection)实现低级与高级语义特征的融合,有效保留了丰富的上下文信息和图像细节,从

而显著提高了分割精度。U-Net及其变体结构的成功应用证明了深度学习方法在医学图像分割任务中的巨大潜力,而U-Net则被视为医学图像分割领域中传统方法与深度学习方法的分水岭。此后,研究者为了进一步提高CNN分割模型的特征提取能力,在模型结构上做出了许多有意义的探索,如设计各种各样的注意力机制和更复杂的跳转连接方式。然而,由于卷积操作固有的局部感受野,基于CNN的分割方法无法有效利用图像的全局上下文信息,导致对复杂目标的分割精度受限。

为了弥补这个不足,自2020年起,许多研究开始探索视觉Transformer结构在医学图像分割任务中的应用。与CNN相比,视觉Transformer能够捕捉图像的全局上下文信息。实际上,Transformer结构最初起源于自然语言处理(natural language processing, NLP)领域,主要用于提取文本序列特征。后来的研究发现其所采用的自注意力机制(self-attention mechanism)同样适用于图像特征提取,进而衍生出vision Transformer(ViT)(Dosovitskiy等, 2020)、Swin Transformer(Liu等, 2021)等视觉Transformer结构。通过自注意力机制,基于视觉Transformer的分割方法能够有效处理复杂的全局上下文信息,在各种医学图像分割任务中展现出优越性能(Xiao等, 2023)。然而,基于视觉Transformer的方法往往需要更多的训练数据,并具有更高的计算复杂度,在一定程度上限制了其广泛应用。为了克服这一问题,许多研究试图在CNN和视觉Transformer之间找到平衡点,开发出了一系列的混合分割模型并取得了良好结果。

2023年底,Gu和Dao(2023)提出基于选择状态空间(selective state space, SSM)的Mamba结构。得益于特定设计的选择机制和硬件感知算法,Mamba结构不仅展现出与Transformer结构相媲美的序列建模能力,而且仅具有线性计算复杂度。在该工作的启发下,许多研究工作开始将Mamba结构应用于计算机视觉领域,其中也包括了医学图像分割任务。目前,基于视觉Mamba的医学图像分割方法主要采用两类结构:一种是纯Mamba结构,其主要以VSS(visual state space)(Liu等, 2024b)模块作为分割模型的基础构建单元,与视觉Transformer结构中的编码器类似;另一种是“Mamba+”混合结构,其中大多数采用“Mamba + CNN”组合,旨在以较低的计算代

价兼顾局部和全局特征提取。总的来说,尽管现有的基于视觉 Mamba 的医学图像分割方法在一些下游任务中已展现出颇具潜力的结果,但该结构的能力边界及其在更广泛任务中的应用效果仍有待进一步的探索。

为了全面了解基于深度学习的医学图像分割方法的发展脉络、优势与不足,本文首次从结构视角对现有方法进行系统梳理和综述。首先,将调研的所有方法按照结构分为3类,如图1所示。1)基于 CNN (CNN-based) 的分割方法;2)基于视觉 Transformer (Transformer-based) 的分割方法;3)基于视觉 Mamba (Mamba-based) 的分割方法。值得注意的是,针对包含两种或以上的混合分割结构,本文会按照视觉 Mamba、视觉 Transformer 和 CNN 的优先级对方法进行归类。例如,采用“Mamba + CNN”方法会被划分为 Mamba-based 类别。然后,简要回顾上述3类主流分割方法的结构演进历程,分析不同方法的结构特点、优势与局限性。接着,从算法结构、学习方法和任务范式等多方面深入探讨当前医学图像分割领域面临的主要挑战及机遇。最后,对未来发展方向和应用前景进行深入分析和讨论。

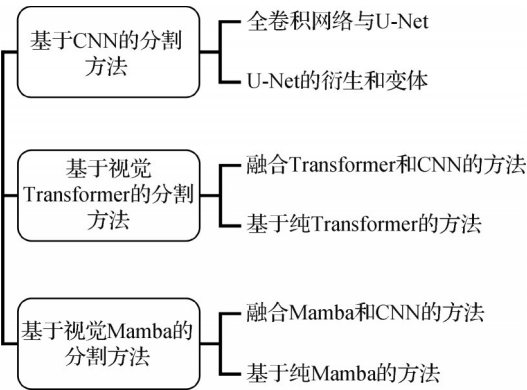


图1 分割方法分类导览
Fig. 1 Segmentation methods classification guide

1 基于 CNN 的分割方法

作为最典型的神经网络架构,卷积神经网络凭借其独特的局部感知和权重共享机制,能够有效捕捉图像中的空间特征,特别适用于图像分析和处理任务。自2015年以来,以U-Net为代表的基于卷积

神经网络的方法一直在医学图像分割领域中占据主导地位,长期在各类下游分割任务上保持着先进的性能(周涛等,2021)。为了进一步提高分割精度,大量的研究工作对U-Net结构进行了改进和创新,衍生出一系列的分割方法。本节首先回顾卷积神经网络的基本原理和发展脉络;然后从U-Net改进角度将现有的基于CNN的医学图像分割方法分为6类;最后对每类方法依次进行综述。

1.1 卷积神经网络的发展及其基本原理

卷积神经网络的研究可追溯至20世纪80年代,其设计灵感来源于生物视觉系统的结构与功能。然而,受限于当时的计算资源,该领域的研究进展相对缓慢。直到LeCun等人(1998)提出LeNet-5模型,这是首个在手写数字识别任务上取得显著效果的CNN模型。LeNet成功验证了CNN在视觉识别领域的有效性,是CNN研究的一个重要里程碑。

进入21世纪,随着设备计算能力的快速增长和大规模标注数据集的涌现,CNN的研究进入高速发展阶段。Krizhevsky等人(2012)提出的AlexNet在ImageNet大规模视觉识别挑战赛(ImageNet large-scale visual recognition challenge, ILSVRC)上取得突破性成绩,在计算机视觉领域掀起新一轮围绕CNN和深度学习的研究浪潮。此后,VGG (Visual Geometry Group) (Simonyan和Zisserman, 2015)、GoogLeNet (Szegedy等, 2015)、ResNet (residual neural network) (He等, 2016)等一系列更深、更复杂的网络结构相继提出,不断拓宽图像识别领域的性能边界。

卷积神经网络通常由输入层、隐藏层和输出层组成。其中隐藏层是CNN的核心,主要包括卷积层(convolutional layer)、池化层(pooling layer)和全连接层(fully connected layer)。卷积层是CNN实现特征提取的核心模块,它通过卷积操作处理输入数据。普通卷积操作如图2所示。

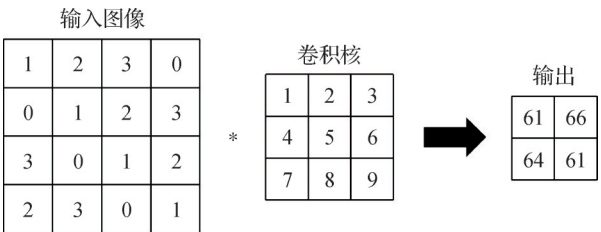


图2 普通卷积操作示意图
Fig. 2 Schematic diagram of ordinary convolution operation

在卷积操作中,卷积核会沿输入数据滑动,依次计算卷积核与其覆盖的局部数据之间的点积,从而产生新的特征。而池化层通常位于卷积层之间,负责对特征进行下采样以降低分辨率,同时保持特征的不变性。常见的池化操作有最大池化和平均池化。全连接层通常位于网络的最后几层,负责将之前网络层提取的特征与最后的输出进行映射。在全连接层中,每个节点都与前一层的所有节点相连,其本质是通过权重矩阵对输入特征进行线性变换,并结合非线性激活函数来学习输入特征与输出之间的复杂关系。

1.2 全卷积网络与 U-Net

在图像识别任务中取得的成功极大促进了 CNN 在其他图像分析任务,尤其是图像分割任务中的广泛应用。从算法的角度来看,图像分割的本质是对输入图像进行像素级别的分类,即将图像中的每个像素分配到预定义的语义类别中。立足于此,Long 等人(2015)针对图像分割任务首次提出全卷积网络(fully convolutional network, FCN),突破了当时“卷积网络+全连接层”的主流范式,创新性地使用卷积层代替全连接层作为像素分类器。这种设计不仅显著提高了模型的训练和推理速度,而且使模型能够接受任意大小的图像输入。为了进一步提高分割精度,FCN 采用跳转连接的方法,通过将不同尺度的中间特征图进行融合,使模型能同时兼顾深层的语义信息和浅层的图像细节。尽管 FCN 是将 CNN 应用于图像分割的开创性工作,但仍存在一些问题。例如,FCN 在逐层下采样过程中会损失一些空间分辨率和细节信息,而简单的上采样操作无法恢复这些细节,这可能会导致分割结果中出现模糊或不准确的边界。

同年,Ronneberger 等人(2015)针对医学图像分割任务提出 U-Net 模型,也采用了全卷积网络的设计思想。值得一提的是,U-Net 至今仍然是医学图像分割领域中的重要基线模型,其结构如图 3 所示。作为最具代表性的全卷积网络,FCN 与 U-Net 不约而同地使用“连续卷积+池化”的方式对图像进行特征编码。这种设计不仅可以减少计算量,同时也能扩大模型的感受野(respect-field)。更大的感受野意味着模型可以利用更丰富的上下文信息,有利于改善特征质量。此外,卷积和池化操作的结合可以使模型具备一定程度的平移和尺度不变性,能够提高

模型的鲁棒性。与 FCN 相比,U-Net 所做的主要改进在于:采用对称的 U 型编码器—解码器结构,并通过跳转连接的方式将编码器的浅层细节特征与解码器的深层语义特征进行融合。这种结构不仅提高了特征信息的利用率,同时增强了网络对图像细节的恢复能力。因此,U-Net 在多个医学图像分割任务上均展现出优于 FCN 的性能。特别是对于边界不清晰和结构复杂的医学图像,U-Net 能够提供更精准的分割结果。

总的来说,FCN 和 U-Net 在医学图像分割领域具有不可忽视的贡献。FCN 提出的全卷积架构为深度学习在图像分割中的应用奠定了基础,而 U-Net 采用的编码器—解码器结构和细节特征保留机制则显著提升了模型分割精度。目前,U-Net 已经广泛应用于各类医学图像分割任务,如肺脏分割(Liu 等, 2022a)、放疗靶区勾画(Samarasinghe 等, 2021)、心脏超声图像分析(Almajalid 等, 2018)、眼底血管分割(Wang 等, 2018)等。从计算机断层扫描(computerized tomography, CT)、核磁共振成像(magnetic resonance imaging, MRI)到 X 射线、显微镜检查等,U-Net 几乎适用于所有的医学图像模态(Siddique 等, 2021)。U-Net 的成功应用,不仅证明了深度学习技术,尤其是 CNN 在临床应用中的巨大潜力,还推动了医学图像分割领域的快速发展。

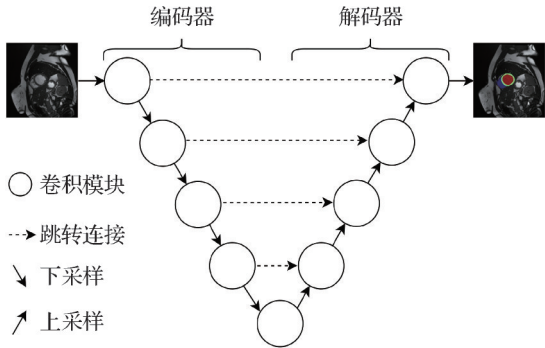


图 3 U-Net 结构

Fig. 3 Structure of U-Net model

1.3 U-Net 的衍生与变体

1.3.1 U-Net 的三维扩展

U-Net 的有效性在医学图像分割领域引起了广泛关注,研究者开始围绕其结构展开进一步的探索与优化。因此,在 U-Net 之后陆续出现了许多类似的 U 型分割结构。其中,为了将 U-Net 更好地应用

于 3D 医学图像(如 CT、MR (magnetic resonance) 图像)分割,Çiçek 等人(2016)提出 3D U-Net。该模型是 U-Net 的简易三维扩展,将二维的卷积、池化、上采样等核心操作均替换为三维形式。与 U-Net 相比,3D U-Net 的优势在于能够直接以完整的 3D 图像数据作为输入而无需 2D 切片,从而可以利用多视角的全局上下文信息进行分割。此外,其也避免了切片处理可能带来的空间信息丢失和上下文不一致的问题。紧接着,V-Net(Milletari 等,2016)在 3D U-Net 的基础上进行改进,使用大步长的卷积层代替编码器中的所有池化操作。同时,受到 ResNet(He 等,2016)的启发,V-Net 在编码器和解码器中都采用残差连接操作,即将上一层的特征图与当前层的输出进行求和以得到新的特征图。这样不仅可以加速模型的收敛,也能保留足够的原始特征信息。

1.3.2 基于模块连接增强的 U-Net

大量的研究工作(He 等,2016;Huang 等,2017)已经证明,增强卷积模块之间的连接不仅可以提高特征提取的质量,还能一定程度上缓解网络加深时可能出现的梯度消失问题,有利于提高模型的收敛效果。受到这些工作的启发,Res-UNet(Xiao 等,2018)和 H-DenseUNet(Li 等,2018)分别从不同角度对 U-Net 的卷积模块进行连接增强的改进。其中,Res-UNet 给 U-Net 的每个卷积模块都增加残差连接操作,整体结构如图 4 所示。

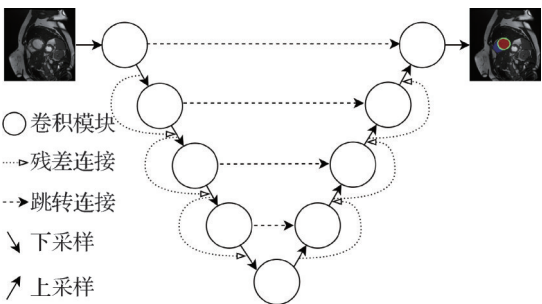


图 4 Res-UNet 结构
Fig. 4 Structure of Res-UNet

在具体实现时,通常将 ResNet 直接作为 U-Net 的主干网络。与之类似,H-DenseUNet 采用 DenseNet(Huang 等,2017)提出的 Dense 模块作为网络的基本组成单元,该模块中每个卷积层的输出都会与之前所有层的输出连接在一起。这两种连接增强策略不仅可以提高特征的重用性和梯度的流动

性,还允许使用更深的网络作为 U-Net 的编码器和解码器,这对提高 U-Net 的分割性能有一定作用。此外,2018 年还出现一系列更复杂的连接增强模型,如 R2U-Net(Alom 等,2018)尝试将残差连接、循环卷积神经网络(Liang 和 Hu,2015)与 U-Net 进行融合,以进一步提高其分割精度。总的来说,尽管这些模型在一些医学图像分割任务中取得了一定的性能提升,但也引入了更高的计算复杂度。

1.3.3 基于特征融合优化的 U-Net

U-Net 提出的对称跳转连接是一种关键的特征融合机制,负责将编码器的浅层特征传递给解码器。这种设计的核心目的是在高级语义特征提取过程中充分融合浅层的空间信息和纹理特征,从而在解码阶段更精确地重建边缘等细节。具体地,U-Net 中的跳跃连接采用的是一种最简单的实现方式,即将编码器和解码器中对应相同分辨率的特征图直接按通道维度拼接起来。尽管该方法在一定程度上提高了融合特征的质量和模型的分割精度,但仍存在着优化的可能。这是因为编码器和解码器的特征在语义上存在显著差异,或称为“语义鸿沟”。这一问题的根源在于编码器和解码器在特征提取上的侧重点有所不同。其中,浅层编码器主要捕捉图像的局部细节,如边缘和纹理特征,而深层解码器则倾向于提取更高级的语义特征,如物体的类别和形状。这种语义层次上的不匹配导致了特征的简单合并无法实现有效融合,从而影响模型在医学图像分割任务中的表现。为了解决这一问题,研究者开始尝试改进原有的简单跳跃连接。

Zhou 等人(2018)提出 UNet++,以进一步提升 U-Net 的分割性能。UNet++ 主要对跳转连接过程进行改良,试图优化特征融合过程以缓和编码器和解码器间存在的语义鸿沟问题。该方法背后的基本假设是:当解码器和编码器的特征映射在语义上更加相似时,网络能够更有效地完成学习任务。为此,UNet++ 对编码器的特征进行渐进式的丰富处理,提高语义层次的同时也使模型能够更有效地捕捉前景对象的细节特征,结构如图 5 所示。

UNet++ 在原始的跳跃连接中引入一系列嵌套、密集的短程连接,使解码器各阶段的高级语义特征均能充分利用来自编码器的多尺度低级特征。从结构角度,UNet++ 可以视做多个不同尺寸 U-Net 的嵌套。此外,UNet++ 在训练过程中引入深度监督机制

(deep supervision mechanism), 即从跳转连接中提取模型的多个中间输出共同用于损失计算。这种设计不仅能够实现对每个嵌套 U-Net 的独立监督学习, 同时便于模型的剪枝操作, 可以在使用时根据具体任务需求自由选择模型结构。在多个开源数据集上的结果充分证明了 UNet++ 的有效性。

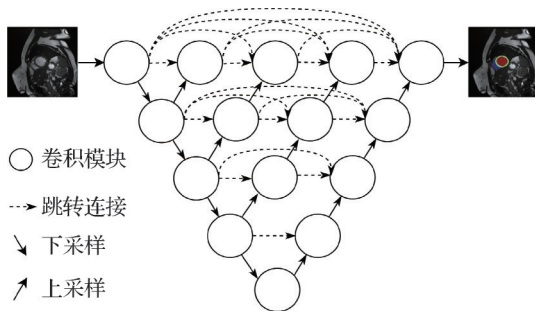


图5 UNet++结构

Fig. 5 Structure of UNet++

Huang 等人 (2020b) 在 UNet++ 的基础上提出 UNet 3+, 在保留深度监督机制的同时, 对跳转连接进一步优化。UNet 3+ 的创新点在于采用全尺寸跳转连接取代 UNet++ 中使用的短程连接, 并且移除了所有的中间计算层。全尺寸跳转连接不仅能够直接将编码器各层的低级语义特征与解码器的高级语义特征进行融合, 而且避免了中间层引入的额外计算。此外, 与 UNet++ 的监督策略不同, UNet 3+ 采用的深度监督机制覆盖解码器阶段的所有输出。这些改进不仅提升了分割精度, 还显著减少了模型的参数量, 从而提高了模型的训练和推理速度, 其模型结构如图 6 所示。

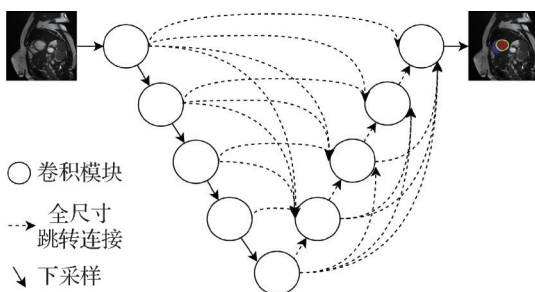


图6 UNet 3+结构

Fig. 6 Structure of UNet 3+

总的来说, 作为 U-Net 特征融合优化的代表性工作, UNet++ 和 UNet 3+ 在跳转连接机制上的创新, 使其在许多医学图像分割任务上取得了优异表现,

同时也能满足不同应用场景下对精度和计算效率的需求。除此之外, 还有 MFP-UNet (multi-feature pyramid UNet) (Moradi 等, 2019)、MultiResUNet (Ibtehaz 和 Sohel Rahman, 2020) 以及 DFA-Net (deep feature aggregation network) (杨兵 等, 2021) 等一系列改良结构, 它们在原始跳转连接的基础上, 利用多尺寸的卷积使具有不同语义的跨尺度特征能够更好地融合, 通过提高融合特征质量以进一步改善网络分割结果。

1.3.4 基于注意力机制的 U-Net

为了在编码器—解码器的架构下进一步优化模型的特征表达能力, 一些工作开始尝试引入注意力机制 (attention mechanism)。这种机制通过为图像的不同位置赋予不同的权重, 使模型能够集中处理图像中的关键区域, 有助于提高分割的准确性和效率。例如, Attention U-Net (Oktay 等, 2018) 提出一种新的注意门 (attention gate, AG) 模型, 可以自动学习聚焦于不同形状和大小的目标结构。通过加入 AG 模块, 模型能够隐式学习抑制输入图像中的不相关区域, 同时突出显示对特定任务有用的显著特征。具体来说, 该模型将多个注意力门模块置于跳转连接阶段, 此模块接收本层的编码器阶段的特征图和低一层的解码器阶段的特征图作为输入。注意力门的主要功能是通过对上述两个输入进行一系列操作, 计算出一个注意力系数, 最后输出这个注意力系数与本层的编码器阶段的特征图的乘积。该注意力门模块使得解码过程中拼接的特征图更关注特征图中的某些重点区域, 通过减少图像中无关部分的特征响应并强调与任务密切相关的关键特征, 有效提升了模型对细节的敏感度和整体的预测精度。这种方法优化了模型的注意力分配, 确保了重点区域得到更多的关注, 从而增强了模型的性能。此外, 由于使用了 Soft-Attention, 与 Hard-Attention 相比, 它的可微特性使得模型可以更好地通过学习来获得注意力模块中的权重参数。后续的 CASU-Net (connection sensitive attention) (Li 等, 2019) 对 Attention U-Net 的 AG 结构进行优化, 调整了其参数更新的方式, 使模型性能得到进一步提升, 证明了优化注意力机制的有效性。

随着注意力机制的持续发展, 出现了各种不同类型的注意力机制, 每一种都有其独特的优势和局限性。研究人员开始将各种各样的注意力机制融入

到U-Net等经典模型中,以提升分割效果。具体来说,与U-Net结合的主流注意力机制有以下几种:

1)空间注意力(spatial attention)。这种机制专注于图像中的空间位置信息,根据像素之间的相对位置来调整注意力权重。在医学图像分割任务中,空间注意力机制得到了广泛应用,因为其可以帮助模型更好地关注病变区域,从而提高分割精度。例如,SA-UNet(spatial attention UNet)(Guo等,2021)在U-Net的基础上引入了空间注意力机制,通过计算特征图中每个位置的重要性权重,使得模型能够更加关注重要区域,从而增强特征表示能力。具体来说,模型通过卷积操作生成一个注意力图,将该注意力图与原始特征图逐元素相乘,以突出重要区域并抑制不相关背景,提高模型的表示能力。不同于Attention U-Net,SA-UNet使用注意力机制计算最低分辨率层(即瓶颈部分)的特征图的注意力,从而达到减少计算量的目的。类似的工作(Fu等,2021; Mei等,2021;Fu等,2023)还将注意力机制应用到肿瘤或GTV(gross target volume)的分割中,同样取得了不错的效果。

2)通道注意力(channel attention)。该注意力机制源自Hu等人(2018)面向图像分类任务提出的SENet(squeeze-and-excitation networks)。其核心思想是采用一种新的“特征重标定”策略显式地建模特征通道之间的相互依赖关系。具体来说,就是通过学习的方式来自动获取每个特征通道的重要程度,然后依照这个重要程度去放大有用的特征并抑制与当前任务无关的特征,这有助于增强卷积神经网络中不同通道的特征表示。而在医学图像分割模型中,不同通道可能对应不同的组织结构或病变信息。通过学习通道之间的权重,通道注意力机制可以自适应地选择最相关的通道,它同时也可以帮助网络自动选择最具信息量的通道,减少冗余信息的影响,提高分割性能。Channel-attention U-Net(Huang等,2020a)和AnatomyNet(Zhu等,2019)在U-Net的基础上引入通道注意力模块分别用于实现高精度的食管癌和头颈部解剖结构的分割。此外,SUNet++(Zhang等,2023)则使用通道注意力来增强对3D医学图像中小目标的分割精度。

3)全局注意力(global attention)。该注意力机制的核心思想是通过提取和利用图像的全局上下文信息,构建特征的全局依赖关系,使得模型不仅能关

注某个局部区域,还能理解整体图像中的相对位置和全局特征。它能提供更全面的语义信息,有助于理解复杂的结构和全局上下文,使模型的特征表示能力得到提升,进而提高图像分类、目标检测和分割等任务的性能。这种机制特别适合处理包含复杂结构(如形态差异较大)的医学图像。GAU-Net(global attention)(Gan等,2021)以及DGFAU-Net(global feature attention upsample U-Net)(Peng等,2021)都是通过在U-Net中嵌入全局注意力模块来提高模型的全局特征表示能力,从而优化分割效果。

4)局部注意力(local attention)。与全局注意力机制相反,局部注意力机制则主要关注图像的局部区域,强调图像中各个像素或小区域之间的细粒度关系。这种方法对于捕捉局部特征和细节非常有用,特别是在处理复杂的、具有细微边界的目标时。局部注意力能够帮助模型更精确地捕捉轮廓,关注局部的纹理、颜色或形状变化。Li等人(2022)提出双重局部注意力模块以捕捉细粒度特征,提高了U-Net在视网膜血管分割任务中的表现。

5)混合注意力(hybrid attention)。上述注意力机制的侧重点各有不同,而混合注意力机制的目标则是通过结合不同注意力机制,以进一步提高模型的表征能力。其中,较为典型的混合注意力机制有CBAM(convolutional block attention module)(Woo等,2018)以及BAM(bottleneck attention module)(Park等,2018),CBAM将空间和通道注意力机制的计算结果进行串联,而BAM则是将这两种注意力机制的结果直接进行相加。类似的工作还有Focusnet(Kaul等,2019)、ScleraSegNet(Wang等,2019a)和RDA-Unet(residual double attention Unet)(徐宏伟等,2020),均是通过改进不同注意力机制的混合方式来提高医学图像分割的精度。张学峰等人(2023)则将全局注意力和局部注意力结合,提出分组注意力使得模型既能在每个小组内捕捉到局部的信息,又能通过跨组的注意力机制学习全局的信息。

总的来说,注意力机制是一种简单、有效的模型表征能力优化方法,能够让模型在处理复杂数据时自动聚焦于最相关的特征或区域,减少了冗余信息的干扰。这不仅有助于改善模型的表现性能,还使得其决策过程具有一定的可解释性,便于进行优化和改进。因此,结合不同注意力机制的U-Net在医学图像分割任务中得到了广泛应用,并取得了不俗

的表现。

1.3.5 基于特殊卷积的U-Net

卷积算子是CNN类分割模型基础组成部件,承担着图像特征提取的关键任务。在原始的U-Net和3D U-Net架构中,均采用标准的二维(2D)或三维(3D)卷积算子来处理输入图像,通过多层卷积操作逐步提取图像中的有用特征。然而,标准卷积算子在实际应用中存在若干显著的局限性。一方面,标准卷积算子具有固定的感受野,这意味着每个卷积核在图像上的滑动窗口大小和形状是预先设定且不变的。这种固定的感受野限制了模型在捕捉不同尺度和复杂结构特征时的灵活性。特别是在处理具有多尺度特征的医学图像时,固定感受野难以适应各种尺寸和形状的目标,导致模型在特征提取过程中可能无法充分捕捉到细微或大范围的关键信息。另一方面,随着特征维度(如通道数)的增加,标准卷积算子的计算量和参数量呈指数级增长。这不仅导致训练过程中的计算资源消耗显著增加,还使得模型在推理阶段的效率受到严重影响。

为了克服上述挑战,研究者提出多种改进的卷积操作,如空洞卷积、深度可分离卷积和可变形卷积等。这些特殊卷积算子通过优化感受野、减少参数量和计算复杂度,显著提升了模型的效率和灵活性,使其更适合实际应用中的复杂需求。

1) 空洞卷积(dilated or atrous convolution)。为了在不增加计算复杂度的情况下扩大感受野,从而提升信息利用率,Yu和Koltun(2015)提出空洞卷积(又称扩张卷积)。通过引入“扩张率(dilation rate)”参数,空洞卷积在卷积核中增加元素间的间距,有效地扩大了感受野。例如,使用扩张率为2的 3×3 卷积核,其感受野扩展至 5×5 ,而参数量和计算量保持不变。除了DeepLabv1(Chen等,2018a)等经典图像分割模型,空洞卷积在医学图像分割领域也得到了广泛应用。诸如ACNN(adaptive convolutional neural network)(Zhou等,2020)结合级联空洞卷积模块、残差学习和实例归一化,并提出优化扩张率的设置方法;CE-Net(context encoder network)(Gu等,2019)在上下文提取模块中引入了密集空洞卷积模块,结合多种不同扩张率空洞卷积,以增强对不同尺寸目标的特征提取能力。

2) 深度可分离卷积(depth-wise convolution)。Howard等人(2017)在MobileNet中提出一种高效的

卷积操作——深度可分离卷积,旨在减少卷积神经网络的计算量和参数量。该方法将标准卷积分解为深度卷积和逐点卷积两步:深度卷积(depth-wise convolution)独立处理每个输入通道以提取空间特征,每个卷积核仅作用于单一通道;逐点卷积(1×1 卷积)则通过线性组合不同通道的输出。与标准卷积相比,深度可分离卷积显著降低了计算复杂度和参数数量,提高了网络的计算效率。基于深度可分离卷积,Wang和Liu(2021)在DeepLabv3+(Chen等,2018b)的基础上提出一种用于胃癌分割的模型;X-Net(Qi等,2019)在脑卒中病灶分割领域表现突出,同样采用深度可分离卷积设计。此外,LV-Net(Lei等,2020)将深度可分离卷积应用于三维图像分割网络V-Net,实现了模型的轻量化目标。

3) 可变形卷积(deformable convolution)。Dai等人(2017)提出可变形卷积(deformable convolution),通过在标准卷积的每个采样点上添加矢量偏移量(offset),使得原本规则的矩形卷积核能够变形为不规则形状,如图7所示。这些偏移量是通过学习获得的,使得卷积核能根据不同输入自适应调整形状。相比于固定采样位置的标准卷积,可变形卷积在处理复杂形状的物体时表现出更优的效果,特别适用于需要处理不规则形状器官和病灶的医学图像分割任务。例如,DUNet(Jin等,2019)将可变形卷积集成到U-Net框架中,用于捕捉不同形状和尺寸的视网膜血管。

综上所述,通过引入空洞卷积、深度可分离卷积和可变形卷积等特殊卷积操作,U-Net模型的表征能力得到了显著提升。这些改进不仅优化了模型的感受野和计算效率,还增强了其在处理复杂和多样化医学图像分割任务中的适应性和精确性。不过,选择和设计合适的卷积算子仍需根据具体任务和应用场景进行权衡,以实现性能与效率的最佳平衡。

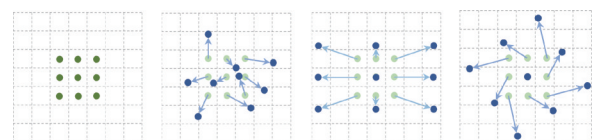


图7 标准卷积与可变形卷积

Fig. 7 Standard vs. deformable convolution

1.3.6 基于非结构性改良的U-Net

除了对U-Net进行结构上的优化,非结构性方

面(如数据增强、损失函数设计等)的改进研究也取得了大量成果。其中,nnU-Net(no new U-Net)(Isensee等,2021)是U-Net在非结构性改进方面最具代表性的成果。Isensee等人(2021)指出,在医学影像领域,不同应用场景下的分割任务需要不同的配置。过多的人为调整网络结构可能导致模型对特定数据集的过拟合,因此,非结构性改进在分割任务中往往具有更显著的影响。

nnU-Net采用一种独特的方法,与以往侧重于模型结构改进的方法不同,它更加关注模型训练和推理全过程中的流水线配置,包括数据预处理、训练过程中超参数的设定以及推理时的策略选择等。首先,nnU-Net会提取数据集的“指纹信息”(fingerprint),即数据集的基本特征,如图像维度、分辨率以及像素强度范围等。通过分析这些指纹信息,nnU-Net能够在无需人工干预的情况下自动选择最适合该数据集的预处理方法,主要包括重采样和非零值裁剪等操作。

此外,nnU-Net提供了3种不同的模型结构作为可选方案,分别为2D U-Net、3D U-Net以及级联的U-Net。在读取数据集的指纹信息后,框架会自动为2D U-Net和3D U-Net进行相应的配置,包括训练所需的所有超参数和数据增强方法。总体而言,nnU-Net是一种自适应的深度学习方法,能够根据不同的数据集和具体任务自动选择最优的网络架构和超参数设置,从而显著提升分割性能并减少人为调参的需求。从结果上来看,nnU-Net通过系统化和自动化的配置流程,在非结构性改进方面为U-Net系列模型带来了显著的性能提升,展示了其在医学图像分割任务中的广泛适用性和优越性。

2 基于视觉Transformer的分割方法

尽管基于CNN的医学图像分割方法已取得显著成果,但卷积算子固有的局部感受野限制了其捕捉全局上下文依赖关系的能力,这在处理复杂的医学图像和细粒度分割目标时尤为明显。虽然引入注意力机制和特殊卷积等技术在一定程度上缓解了这一问题,提升了模型对全局信息的关注,但效果仍然有限。自2020年以来,研究者开始将源自NLP领域的Transformer结构(Vaswani等,2017)引入计算机视觉任务(Dosovitskiy等,2020),其中也包括医学图像

分割。视觉Transformer通过其自注意力机制,能够有效建模全局依赖关系,显著提升语义特征的提取质量,对复杂医学图像的分割具有重要的促进作用。

本节首先介绍Transformer结构的基本原理及其在视觉任务中的优势;随后将现有的基于Transformer的医学图像分割方法划分为两类:一类是与CNN结合的混合结构;另一类是纯Transformer结构;最后对每类方法进行详细综述,探讨其在医学图像分割中的应用与效果。

2.1 视觉Transformer的基本原理

Transformer是一种深度学习模型,主要依赖于多头自注意力(multi-head self-attention)机制来处理序列数据(如文本)。最初,Transformer被设计用于NLP领域,旨在克服传统序列模型(如循环神经网络(recurrent neural network, RNN))在捕捉长距离依赖关系和实现并行计算方面的局限性。凭借其架构优势,Transformer不仅在NLP领域取得了显著成功,还在其他应用领域中展现出巨大的潜力。由此,Transformer逐渐被视为CNN的有力竞争者,并被尝试应用于以往由CNN主导的计算机视觉任务中。

与CNN相比,Transformer具有若干显著优势。首先,Transformer的自注意力机制无需对输入数据的长度和宽度设限,能够更灵活地适应不同长度和规模的输入序列。这使得Transformer在处理多样化数据时表现出更高的适应性。其次,自注意力机制允许在每个时间步进行并行计算,从而显著加快了模型的训练速度。最后,Transformer具备更强的全局依赖建模能力,能够更有效地捕捉序列中的上下文信息,提升了模型对复杂模式和长距离关系的理解能力。ViT(Dosovitskiy等,2020)的出现,标志着Transformer在计算机视觉领域取得了突破性进展。ViT通过将图像划分为固定大小的图块(patchs)并将其视为序列输入,成功应用于图像分类任务,展示了Transformer在视觉任务中的强大潜力。ViT不仅挑战了CNN在计算机视觉领域的主导地位,还为后续基于Transformer的视觉模型奠定了基础,推动了计算机视觉技术的发展。

如图8所示,ViT首先将输入图像划分为若干可自定义大小的图块,随后将展平后的图块通过线性投影层转换为图块嵌入(patch embedding)。由于ViT最初设计用于图像分类任务,因此在图块嵌入序列的首部添加了一个特殊的“类别”标识符,该标

识符对应的输出即为最终的图像类别预测。接着,所有的图块嵌入与相应的位置编码相加,输入到Transformer编码器模块中。最后,通过一个多层感知机(multi-layer perceptron, MLP)分类头生成最终的分类结果。实验证明,当预训练数据集规模足够大时,ViT能够克服对局部性和平移不变性等归纳偏置的依赖,在多种计算机视觉下游任务中表现优于传统CNN模型。目前,大多数基于Transformer的视觉模型采用ViT作为基础模块进行改进,可以认为在视觉领域中,使用Transformer即等同于使用ViT。

在ViT的基础上,Swin Transformer(Liu等,2021)创新地引入滑动窗口机制,使模型能够学习跨窗口的信息。此外,通过下采样层,Swin Transformer能够处理高分辨率图像,从而在降低计算量的同时,兼顾全局和局部信息的提取。相比于ViT,Swin Transformer显著减少了计算复杂度,其复杂度与输入图像尺寸呈线性关系。随着模型深度的增加,Swin Transformer逐步合并图像块,构建层次化的

Transformer结构,使其成为适用于图像分类、目标检测及语义分割等多种任务的通用视觉骨干网络。这一创新为纯Transformer在视觉领域的应用树立了成功范例,并推动了相关研究的深入发展。

随着Transformer在视觉领域的应用日益成熟,研究者开始探索在模型中融合Transformer与CNN的方法,以在保留CNN对医学图像处理所需的重要归纳偏置的同时,增强模型的全局建模能力。因此,出现了各种各样的混合视觉Transformer模块(hybrid vision transformer, HVT),其核心思想是将Transformer中的多头自注意力机制与CNN中的卷积操作相结合,从而更有效地建模图像中的局部和全局关系。相关研究(Chen等,2021; Hatamizadeh等,2022)表明,纯ViT模型由于仅依赖多头自注意力机制,可能忽视一些细节,影响分割的精确性。而HVT模型通过结合ViT和CNN的优点,不仅能够更好地理解数据中的长距离依赖信息,还能有效捕捉局部上下文,因而在多种任务,尤其是医学图像处理领域表现出色。

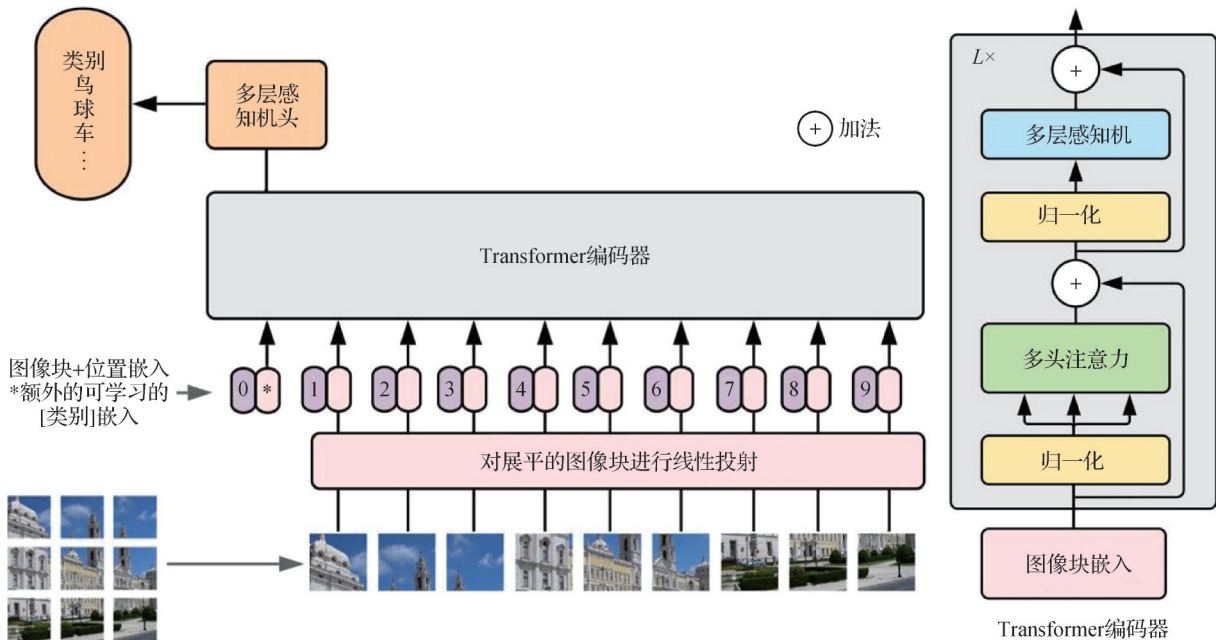


图8 ViT结构

Fig. 8 Structure of ViT

当前,基于Transformer的医学图像分割方法主要分为两类:一类融合了Transformer和CNN,既可以将Transformer作为视觉序列编码模块,与CNN交叉或并行使用,又可将Transformer用做注意力机制或特征融合模块以增强现有的CNN架构;另一类则

采用纯粹的视觉Transformer结构来实现端到端的分割。以下对这些方法进行分类和详细介绍。

2.2 融合Transformer和CNN的方法

如前所述,Transformer结构的优势在于能够建模全局上下文的依赖关系,而CNN则侧重于提取局

部感受野内的语义特征。为了进一步提升模型的特征提取能力,从而提高分割性能,一个直观的思路是将二者结合起来。然而,采用何种融合方式才能充分发挥 Transformer 和 CNN 的各自优势,是一个关键问题。自 2020 年以来,众多研究围绕这一问题展开,结合编码器—解码器架构,提出多种融合 Transformer 与 CNN 的混合结构,并在下游医学图像分割任务中取得了令人满意的成果。

从融合方式的角度来看,现有方法主要可以分为两类:第 1 类是将 Transformer 模块作为注意力机制或特征融合模块直接嵌入现有的 CNN 分割模型,以提升特征的质量。这类方法通常在 CNN 的特征提取过程中引入 Transformer 模块,通过自注意力机制捕捉全局上下文信息,进而增强特征表示。第 2 类方法则是采用 Transformer 的序列编码方式对图像进行编码,然后在特征提取阶段交叉或并行使用 Transformer 和卷积算子,充分利用两者的优势。这类方法通过并行或交替的特征提取路径,既保留了 CNN 在局部特征提取方面的优势,又发挥了 Transformer 在全局建模上的能力。

2.2.1 将 Transformer 作为注意力机制

TransUNet (Chen 等, 2021) 是将 Transformer 与 CNN 融合应用于医学图像分割的早期尝试。基于传统的 U-Net, TransUNet 首先对 CNN 编码器得到的特征图进行序列编码,然后利用一个包含多个 Transformer 层的子模块以增强特征,最后由 CNN 解码器得到预测输出。具体结构如图 9 所示,该模型在编码阶段同时使用了 CNN 和 Transformer,其中的 Transformer 模块可以视做特殊的“全局注意力机制”,用于对 CNN 提取的特征进行优化。这种结构在一定程度上兼备了全局上下文关系的建模能力和局部细节处理能力,有利于提升分割精度。

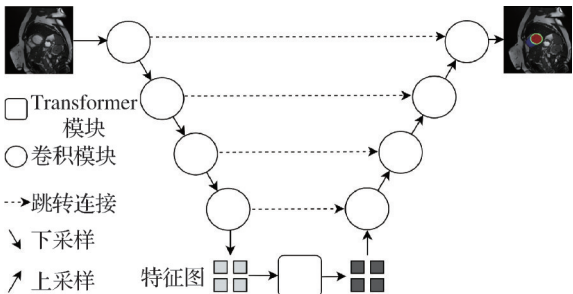


图9 TransUNet 结构

Fig. 9 Structure of TransUNet

在 TransUNet 的基础上, TransNUNet (Yang 和 Tian, 2022) 在保持编码器结构不变的前提下引入卷积注意力模块 CBAM 来优化解码器的上采样过程; TransClaw U-Net (Yao 等, 2021) 则在原始编码器的基础上对其跳转连接部分做了改进, 增强了特征信息的传递与融合。总体而言, 这些改进方法均采用了 CNN 来提取图像的浅层特征, 然后利用 Transformer 对 CNN 处理后的特征进行全局建模。尽管这种简单的融合方式在部分任务中实现了一定的性能提升, 但其效果有限, 仍需要探索更高效的融合策略, 以进一步提升分割性能。

2.2.2 将 Transformer 作为特征融合模块

除了优化隐藏层的特征之外, 一些研究尝试利用 Transformer 对 U-Net 的跳转连接进行优化, 主要通过融合多尺度的编码器特征, 以缓解连接时的“语义鸿沟”问题。例如, UCTransNet (Wang, 2022a) 提出通道 Transformer (channel transformer, CTrans) 模块, 并将其作为 U-Net 的跳转连接模块, 结构如图 10 所示。

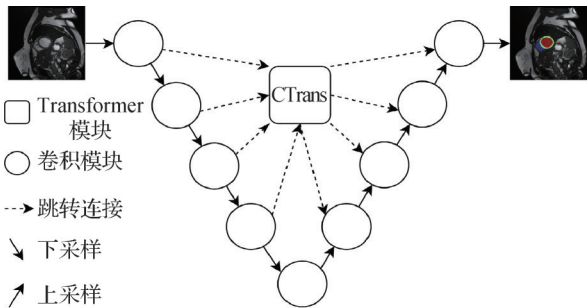


图10 UCTransNet 结构

Fig. 10 Structure of UCTransNet

该模块主要由通道交叉融合 Transformer 模块 (cross channel Transformer, CCT) 和通道交叉注意力模块 (cross channel attention, CCA) 组成。其中, CCT 采用交叉注意力机制, 从通道角度对多尺度的编码特征进行融合, 确保对全局上下文信息的充分利用; 而 CCA 则用于解码器阶段的特征优化, 通过动态调整特征权重以进一步增强 CCT 的融合特征。与朴素的跳转连接相比, 由 CTrans 构成的跳跃连接模块理论上能够更有效地融合和利用编码器阶段提取的多尺度特征, 从而有利于缓解连接时的语义鸿沟问题。CoTr (Xie 等, 2021) 也采用了类似的思路, 但不同之处在于其使用了可变形 Transformer (deformable Transformer) 作为跳转过程中的特征融合模块。与

标准Transformer相比,可变形Transformer仅对选定的关键采样点进行自注意力操作,能够显著减少计算负担。总的来说,尽管通过Transformer优化U-Net的跳跃连接在一定程度上提升了特征融合质量和分割精度,但也带来了新的问题。一方面,引入Transformer模块会显著增加计算复杂度;另一方面,这些方法在不同尺度特征的融合过程中,如何有效平衡全局上下文信息与局部细节信息仍然是一个挑战。

2.2.3 将Transformer作为混合编码模块

为了充分发挥Transformer和CNN的结构优势,大量的研究致力于在编码模块中融合二者的特性,由此形成了一系列的混合编码模块。从组合方式上,这些混合编码模块主要采用串行组合和并行组合两种策略。

串行组合是指将CNN和Transformer模块按顺序连接,通常是先通过CNN提取初步特征,再利用Transformer进行特征依赖关系的全局建模。这种方法能够逐步深化特征表示,提高特征语义层次。例如,UTNet(Gao等,2021)提出一种高效的自注意力机制,并在编码器—解码器架构中交替使用Transformer和卷积模块进行特征提取,结构如图11所示。类似的工作还有MT-UNet(multi-task UNet)(Wang等,2022b)、ConvFormer(Lin等,2023)等,均采用串行组合的方式,实现了特征表示的逐层深化和语义增强。

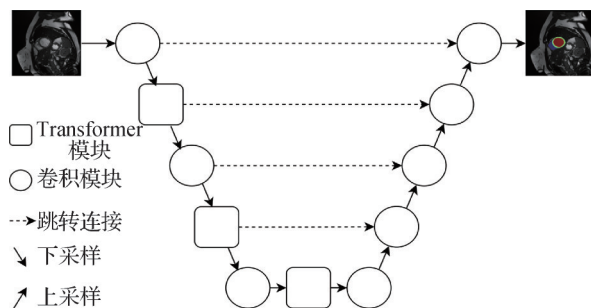


图11 UTNet结构

Fig. 11 Structure of UTNet

并行组合则是将CNN和Transformer模块同时作用于同一输入特征,通过并行处理得到不同类型的特征表示,并在层内进行特征融合。其中,CNN负责提取局部细节特征,而Transformer则捕捉全局上下文依赖信息,二者特征的融合能够实现信息的互补。例如,UNet-2022(Guo等,2022)提出一种非同构的混合编码模块,主要关注于平衡CNN和

Transformer在特征提取过程中的权重分配差异。该研究指出,Transformer的自注意力机制主要在同一通道内为不同空间位置计算动态权重,不同位置具有不同的权重,而同一位置的所有通道权重则保持一致。相比之下,CNN的卷积操作为不同通道分配不同的权重,但在同一通道的所有空间位置共享相同的权重。为了结合二者的优势,UNet-2022采用创新的权重分配机制,以平衡空间动态性和通道动态性,并通过逐深度卷积有效降低了模型的参数量。CSU-Net(context spatial U-Net)(Chen等,2022a)、FcTC-Unet(fine-grained combination of Transformer and CNN)(Qiao等,2022)等均采用并行组合的方式,不过组合粒度和阶段各有不同。

综上所述,串行组合和并行组合各有优势,前者在逐步深化特征表示方面表现出色,后者则在同时利用局部与全局特征方面具有优势。此外,针对Transformer模块带来的计算负担,研究者通过创新的模块设计和计算优化策略,进一步提升了混合编码模块的效率和实用性。这些结合Transformer与CNN的混合编码方法为提升如U-Net等模型在医学图像分割任务中的性能提供了有力支持。

2.2.4 将Transformer作为序列编码器

上述研究大多仍采用CNN模块对输入图像进行初始编码,Transformer则主要用于后续特征的处理与优化。然而,ViT(Dosovitskiy等,2020)已经证明,将图像进行序列化编码同样能够获得有效的特征表示。因此,越来越多的研究开始直接将Transformer作为医学分割网络的序列编码器,以充分发挥其在全局特征建模和长距离依赖捕捉方面的优势,从而提升图像分割的精度和性能。从网络架构的角度来看,当前的研究工作主要分为单分支和多分支两类。单分支方法指的是网络仅采用单一的序列编码分支,而多分支方法则包含多个编码分支,例如结合CNN分支与Transformer分支。多分支方法旨在通过多样化的特征提取路径,进一步增强模型的表现能力和性能。

作为典型的单分支方法,UNETR(Hatamizadeh等,2022)在3D U-Net的基础上引入Transformer构成的序列编码器,如图12所示。UNETR首先将3D图像转化为patch序列,并通过线性投射层将这些patch投影到嵌入空间中。接着,在添加位置编码后,这些嵌入特征被输入到由多个Transformer层组

成的编码器中。在进行跳转连接时,UNETR 选择编码器中不同 Transformer 层的输出,并通过 3D 转置卷积使这些特征图的尺寸与对应上采样过程中的特征图保持一致,随后进行拼接。受 Swin Transformer 的启发, Hatamizadeh 等人 (2022) 进一步提出 Swin UNETR, 采用 Swin Transformer 代替标准 ViT 作为编码器模块,以优化模型的表征能力并降低计算复杂度。类似地, ITUnet (integration of Transformers and Unet) (Kan 等, 2022)、UNesT (Yu 等, 2023) 等工作均是在采用单分支序列编码器的基础上,对网络结构进行针对性改进,以实现更好的分割效果。

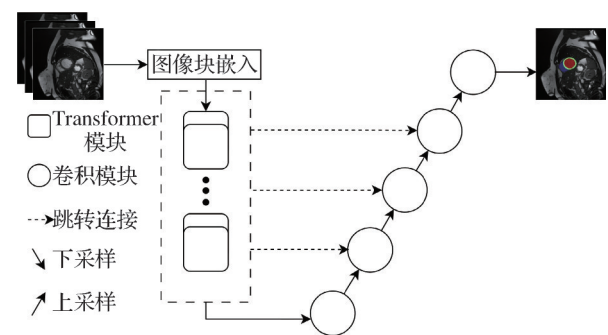


图 12 UNETR 结构
Fig. 12 Structure of UNETR

为了进一步提升模型的表征能力,部分研究开始采用多分支编码架构。例如, MedT (medical Transformer) (Valanarasu 等, 2021) 提出局部—全局 (local-global, LoGo) 训练方法和门控轴向注意力机制。具体地, MedT 将单分支编码器结构分为全局分支和局部分支两个部分。全局分支将整个输入图像通过一个层次较浅的 U 型编码器—解码器结构,生成全局分割结果;而局部分支则将图像划分为若干 patch 后,输入到一个层次较深的 U 型编码器—解码器结构中,再将各 patch 对应的输出结果重新组合,生成局部分割结果。最后,对两个分支的输出进行逐元素相加,并通过卷积层生成最终的预测结果。LoGo 方法通过多分支设计增强了模型在特征提取阶段对局部和全局特征的综合提取能力,而门控轴向注意力机制则在高效计算的基础上进一步提高了低分辨率图像的分割精度。与之类似, TransFuse (Zhang 等, 2021)、DHUNet (dual-branch hierarchical) (Wang 等, 2023) 和 H-DenseFormer (Shi 等, 2023) 等分割模型均采用了多分支特征提取器,并且分别在不同的编码分支中采用 CNN 和 Transformer 结构,以

充分发挥这两种结构的表征优势。其中, TransFuse 提出一种新颖的特征融合模块,用于整合来自 CNN 和 Transformer 分支的特征; DHUNet 对整体结构进行改进,分别采用 Swin Transformer 和 ConvNeXt (Liu 等, 2022b) 作为特征提取模块; H-DenseFormer 设计了一种基于密集连接的轻量 Transformer 模块,有效地降低了计算复杂度。

尽管这些方法已经在多个任务上取得显著效果,但仍存在一些挑战。首先,与 CNN 相比,基于视觉 Transformer 的序列编码器通常需要更多的训练数据来学习归纳偏置,而大规模的标注数据意味着高昂的成本。其次,多分支结构不可避免地引入了额外的计算开销,增加了模型的复杂性和推理时间。最后,不同编码器分支之间可能存在语义鸿沟问题,即不同分支提取的特征在语义空间中存在差异,如何有效融合这些异质特征将是一个新的挑战。

2.3 基于纯 Transformer 的方法

除了将 Transformer 作为子模块集成到 CNN 结构中以提升模型性能外,部分研究还创新性地采用纯视觉 Transformer 结构来完成端到端的医学图分割任务。例如, MISSFormer (medical image segmentation transformer) (Huang 等, 2021) 提出一种分层的编码器—解码器网络结构,主要包含两个关键设计:增强的 Transformer 块 (enhanced Transformer block) 和增强的 Transformer 上下文桥 (enhanced Transformer context bridge)。前者针对性地改进了前馈网络,强化了模型对长距离依赖关系的捕捉能力,并通过补充局部上下文信息,使隐藏特征更具有区分度;而后者突破了传统的仅建模全局信息的方法,通过结合增强的 Transformer 块能够更有效地提取由分层 Transformer 编码器生成的多尺度特征的长距离依赖关系和局部上下文信息。这些设计使 MISSFormer 在医学图像分割任务中表现出色,特别是在捕获有关复杂分割目标的上下文信息方面。此外, Swin-Unet (Cao 等, 2022) 仍采用类 U-Net 的编码器—解码器结构,将所有的卷积模块都替换为由 Swin Transformer 构成的子模块,并通过块合并 (patch merging) 和块扩充 (patch expanding) 的方式进行下采样和上采样。值得注意的是, Swin-Unet 在编码器和解码器阶段均采用了相同的纯 Swin Transformer 结构。类似的纯 Transformer 分割网络还包括 Segmenter (Strudel 等, 2021)、TransDeepLab (Azad 等, 2022) 等。这

些研究表明,基于纯Transformer的U形编码器—解码器网络在医学图像分割任务上,能够媲美甚至优于纯CNN方法以及Transformer与CNN的混合方法,这为医学图像分割领域带来了新的可能性。

总的来说,基于Transformer的医学图像分割方法,包括Transformer与CNN的混合方法以及纯Transformer方法,各自展现出了独特的优势和劣势。混合方法通过结合CNN在局部特征提取方面的优势与Transformer在全局上下文建模中的能力,能够在保持计算效率的同时提升分割精度。然而,这类方法仍然依赖于CNN的结构,可能限制了其在复杂场景中的表现。相比之下,纯Transformer方法能够更有效地捕捉长距离依赖关系和多尺度特征,可以显著提高分割的准确性和泛化性。但纯Transformer架构通常需要大量的计算资源和高质量的训练数据,而在医学领域中获取大规模标注数据具有较高的难度和成本。

3 基于视觉Mamba的分割方法

虽然Transformer结构在捕捉长距离依赖关系和全局上下文信息方面展现出显著优势,但其计算复杂度随着输入序列长度的增加呈二次增长,限制了在资源受限环境(如移动设备和嵌入式系统)中的应用。为克服这一挑战,研究者致力于开发能够以线性时间复杂度建模全局依赖关系的新方法。其中,Mamba(Gu和Dao,2023)提出一种新的选择性状态空间模型,通过引入选择机制、硬件感知算法以及更简单的架构,能够在保持高效长序列建模性能的同时,显著降低模型计算复杂度。因此,自2024年以来,许多研究工作开始将Mamba结构引入医学图像分割任务,并取得了不错的结果,Mamba结构有望取代Transformer结构。本节首先介绍了Mamba结构的基本原理及其优势;随后将现有的基于Mamba的医学图像分割方法划分为两类:一类是与其他结构结合的混合结构,另一类是纯Mamba结构;最后对每类方法进行概括描述,分析其应用效果。

3.1 Mamba结构的发展及其基本原理

Mamba是一种基于选择性结构化状态空间模型(selective structured state space model, S6)的架构,旨在通过创新性的参数化方法和硬件感知的并行化算法,在实现高效序列建模能力的同时,保持线性计算

复杂度。Mamba架构由状态空间模型(state space model, SSM)发展而来,期间经历过两次重要迭代:结构化状态空间序列模型(structured state space for Sequence modeling, S4)以及Mamba。

作为Mamba架构的前身,状态空间模型(state space model, SSM)(Gu等,2021)最早被提出用于时间序列分析和动态系统建模。SSM的核心思想是通过引入隐含的系统状态来描述观测数据的动态特性和随机波动。其基本结构由状态方程和观测方程组成。状态方程描述系统内部状态的演变,通常以递推关系或差分方程表示,反映系统的内部动力学。观测方程则将系统状态与实际观测数据关联起来,将不可直接观测的系统状态映射到可观测的数据上。通过这两种方程的结合,SSM能够有效捕捉和预测时间序列数据中的潜在模式和变化。本质上,SSM基于当前观测数据及先前的隐状态来预测系统的未来状态,其工作原理与循环神经网络(RNN)相似,均通过时间序列数据更新隐含状态。然而,与RNN不同,SSM在状态估计和预测时采用贝叶斯推理和卡尔曼滤波等方法,具有更为严格的数学框架和理论基础。图13展示了SSM的基本结构,其中, A 、 B 、 C 、 D 为权重矩阵,分别用于描述状态转移、输入对状态的影响、状态到观测的映射以及输入对观测的影响。在实际应用中,这些权重矩阵是可学习的,使模型能够准确反映系统的动态行为。

虽然SSM为序列建模提供了新方法,但在实际应用中存在诸多限制,尤其在语言建模和蛋白质结构预测等任务中,数据通常映射到离散空间,使基于连续状态空间的SSM不再适用。为此,S4引入了离散化的SSM。具体而言,S4通过零阶保持技术将离散输入数据转换为SSM可处理的连续数据,并通过固定时间步采样将SSM生成的连续输出序列转化为离散形式,实现了离散输入输出的映射。此外,SSM在训练效率上存在显著不足,因其计算依赖于状态的逐步更新。针对这一问题,S4设计了一种新的重参数化方法,使得在推理过程中可采用与RNN类似的循环结构表示SSM。离散化后的SSM可视为针对时间步的嵌套计算,通过重参数化构造等效的卷积核,利用卷积算子统一计算每个时间步的输出,从而实现并行训练。S4在训练阶段采用CNN结构,推理阶段则使用RNN结构,兼具两者在速度上的优势。为了进一步提升对长序列数据的处理能力,S4引入

基于 HiPPO 的处理方法,通过优化的多项式基函数有效压缩和表示长序列数据,显著增强了长时间依赖关系的建模能力。SSM 到 S4 的演进不仅在模型结构上实现了创新,还在处理长序列数据方面取得了突破,使 S4 在复杂序列建模任务中表现出更强的适应性和表现力。

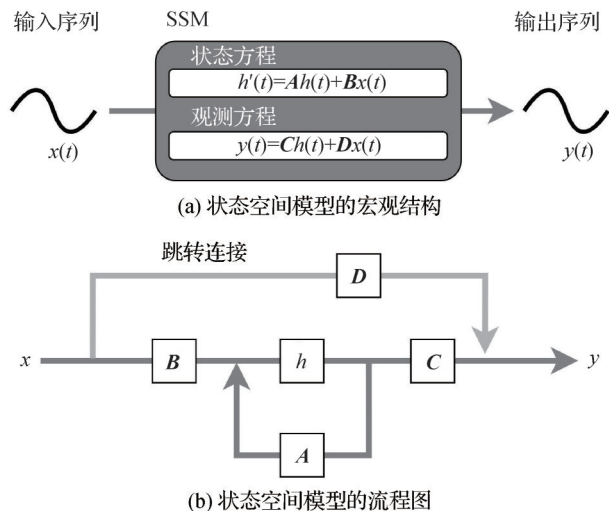


图 13 SSM 结构
Fig. 13 Structure of state space model
((a) macro-structure of SSMs; (b) flowchart of SSMs)

针对 SSM 和 S4 中 A 、 B 、 C 三个权重矩阵在模型推理阶段保持不变,导致模型在不同输入下无法进行针对性推理的限制,Mamba 提出一种创新性的解决方案。具体而言,Mamba 通过引入选择性机制,使得模型能够在不改变可学习参数的情况下,根据不同的输入动态生成不同的 A 、 B 、 C 权重矩阵,从而实现对多样化输入的适配。这一机制通过选择性地调整权重矩阵的值,使模型能够更灵活地应对不同类型和特征的输入数据,显著提升了模型的建模能力和泛化性能。然而,选择性机制的引入使得 Mamba 的迭代过程变得动态,无法再依赖于 CNN 的固定卷积核进行高效计算。为解决此问题,Mamba 采用了并行扫描算法,实现了计算过程的高度并行化。此外,Mamba 借鉴了 Flash Attention (Dao 等, 2022) 中的访存感知技术,优化了模型的内存访问模式,使其更好地适配 GPU (graphic process unit) 等高效并行计算硬件结构。这些优化措施不仅提升了模型的计算效率,还大幅减少了计算资源的消耗,使 Mamba 在实际应用中能够实现快速且高效的推理过程。

为进一步增强模型的表达能力和计算效率,

Mamba 结合了 SSM 架构与门控多层感知机 (gated multi-layer perceptron, GMLP), 设计了新的计算模块。这一模块通过引入门控机制,有效地控制了信息流动,提升了模型在处理复杂序列数据时的灵活性和准确性。同时,Mamba 利用卷积算子增强了局部上下文的建模能力,使得模型能够更好地捕捉序列中的局部依赖关系,从而在多种序列建模任务上显著提升了预测精度。此外,Mamba 在设计过程中注重模型的可扩展性和模块化,使其能够方便地集成到不同的应用场景中。通过优化权重矩阵的生成和计算过程,Mamba 不仅保持了高效的计算性能,还确保了模型在面对大规模数据和复杂任务时的稳定性和可靠性。

综上所述,Mamba 通过引入选择性机制、并行扫描算法以及结合门控多层感知机和卷积算子等创新性设计,不仅提升了模型在多样化输入下的适应能力和计算效率,还在多种序列建模任务中实现了更高的预测精度和鲁棒性。这些改进使得 Mamba 成为当前先进序列建模方法中的重要成员,并开始广泛应用于如医学图像分割等计算机视觉任务。与 Transformer 类似,目前基于 Mamba 的医学图像分割方法也可以划分为两类:一类是融合 Mamba 的混合结构;另一类是纯 Mamba 结构。

3.2 融合 Mamba 和 CNN 的方法

Mamba 的独特设计使其在计算效率和长序列建模方面优于 Transformer 结构,因而迅速应用于各类任务中。自 Mamba 架构问世以来,研究人员不断基于其结构提出多种新方法,专注于提升医学图像分割的精度和效率。目前,大部分针对医学图像分割任务的 Mamba 模型致力于将 Mamba 与现有的 CNN 架构相融合,以兼顾强大的特征提取能力和高效的计算性能。例如,U-Mamba (Ma 等, 2024b) 将 Mamba 作为核心模块集成到 U 型编码器—解码器架构中,如图 14 所示。U-Mamba 在编码器阶段采用 Mamba 与 CNN 结合的新型结构,取代传统 U-Net 中的标准卷积模块,以增强特征提取质量。在解码器阶段,U-Mamba 结合了包含标准卷积的残差模块和转置卷积,确保了高分辨率特征的有效重建和细节信息的精准恢复。

另一个代表性的工作是 SegMamba (Xing 等, 2024), 该模型同样采用了 U 型结构,但在编码器阶段引入了 TSMamba (tri-oriented spatial Mamba) 模块。

传统的基于Transformer的3D模型在处理输入的3D数据时,通常需要通过下采样操作来降低计算复杂度,然而这一过程可能导致关键信息的丢失。为此,SegMamba通过TSMamba模块结合门控空间卷积,实现了在高维输入下对全局特征的高效建模。这一设计不仅有效地保留了重要的空间信息,还能够显著提升模型在复杂三维医学影像中的分割性能。解码器和跳转连接部分则延续了U-Mamba的设计理念,采用包含标准卷积的残差模块和直接跳转连接。此外,研究者们还探索了将Mamba应用于模型其他方面的可能性。例如,LightM-UNet(Liao等,2024)首次将Mamba引入轻量化模型设计,旨在减少模型参数量和计算资源需求,同时保持较高的分割精度。而其他的基于Mamba的混合方法,如ProMamba、H-vmunet、Swin-UMamba等(Xie等,2024;Wu等,2024;Liu等,2024a;Wang等,2024a;Hao等,2024)也成功应用于各种医学图像分割任务。这些方法通过结合Mamba的长序列处理能力显著提升了模型在不同医学图像数据集上的表现。

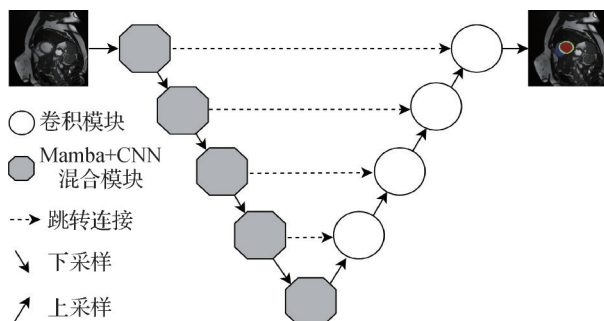


图14 U-Mamba结构

Fig. 14 Structure of U-Mamba

3.3 基于纯Mamba的方法

除了融合Mamba与其他结构之外,另一种思路是利用纯Mamba模块来构建分割模型。Liu等人(2024b)提出的VMamba在Mamba的基础上设计了一种新型的通用视觉模型基础模块——视觉状态空间(visual state space, VSS)模块,使Mamba成为继Transformer之后,第2种可供选择的通用视觉模块。由于Mamba最初是为处理具有顺序的语言序列而设计的,其在处理非顺序性的图像数据时存在一定的局限性。为此,VSS利用二维选择性扫描(2d-selective-scan, SS2D)模块替代了Transformer中的多头自注意力机制,使其在处理图像数据时能够更有效地捕捉空间特征。随后,Ruan等人(2024)提出

VM-UNet,完全采用VSS模块来构建编码器和解码器,使模型能够高效地利用长距离的上下文信息,VM-UNet模型结构如图15所示。类似地,Wang等人(2024b)和Sanjid等人(2024)分别提出Mamba-UNet和Mamba-HUNet,在部分分割任务上均取得了令人满意的结果。

这些研究证明,仅使用Mamba模块构建的分割模型,不仅能够充分发挥Mamba在处理长序列数据方面的优势,还能有效捕捉图像中的空间特征信息。通过与U-Net等经典分割架构的结合,基于VSS模块的模型在保持高精度分割性能的同时,优化了计算效率和模型结构的灵活性。此外,这些方法还展示了Mamba在不同医学图像分割场景中的广泛适用性,进一步证明了其作为通用视觉模块的潜力。

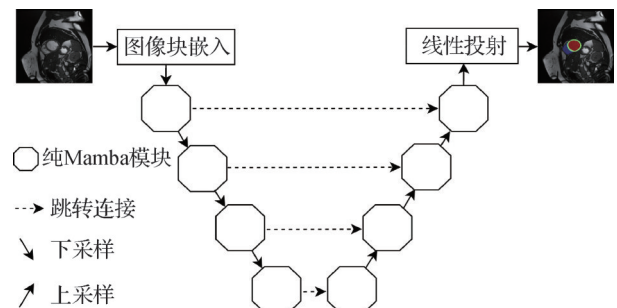


图15 VM-UNet结构

Fig. 15 Structure of VM-UNet

总的来说,基于Mamba的医学图像分割方法尚处于早期研究阶段,但其在提升分割精度和效率方面展现出显著潜力。将Mamba与CNN结合的混合方法,通过融合CNN的特征提取能力与Mamba处理长距离依赖关系的优势,能够更有效提升分割精度和鲁棒性。然而,这类方法在集成过程中可能增加计算复杂度。此外,纯Mamba方法更适合需要全局上下文信息的分割任务,在处理图像的空间特征时仍存在不足,且训练过程可能需要更多的计算资源。此外,纯Mamba模型在适应二维图像数据方面相较于成熟的CNN架构效果尚未完全显现。

4 不同医学图像语义分割算法对比

如上所述,医学图像语义分割算法包括基于CNN的分割方法、基于视觉Transformer的分割方法和基于视觉Mamba的分割方法。表1对比了不同分割算法。

表 1 不同医学图像的语义分割算法对比
Table 1 Comparison of different medical image segmentation algorithms

分类	方法	年份	发表	模态	亮点	
基于 CNN 的方法	全卷积网络 与 U-Net	FCN	2015	CVPR	2D	全卷积化分割,跳转连接
		U-Net	2015	MICCAI	2D	对称的 U 型编码器—解码器结构,采用跳转连接
	U-Net 的衍生 和变体	3D U-Net	2016	MICCAI	3D	采用三维的卷积、池化
		V-Net	2016	3DV	3D	引入残差连接,采用卷积层替换池化层
		H-DenseUNet	2018	TMI	2D	引入密集连接
		Res-UNet	2018	ITME	2D	引入残差连接
		R2U-Net	2018	arXiv	2D	引入循环残差机制
		UNet++	2018	DLMIA	2D	采用密集的跳转连接、嵌套多层网络结构和深度监督
		UNet3+	2020	ICASSP	2D	采用全尺寸跳转连接
		Attention U-Net	2018	MIDL	2D	在跳转连接中加入注意力门控机制
		SA-UNet	2021	ICPR	2D	在瓶颈部分加入空间注意力
		Channel-attention U-Net	2020	IEEE Access	2D	引入通道注意力
		Focusnet	2019	ISBI	2D	采用混合注意力机制
		nnU-Net	2021	Nature Methods	2D/3D	自动的网络结构和流程配置
		CE-Net	2019	TMI	2D	引入密集空洞卷积模块
		ACNN	2020	ICRA	2D	引入级联的空洞卷积模块
		LV-Net	2020	ICASSP	3D	采用深度可分离卷积
		X-Net	2019	MICCAI	2D	采用深度可分离卷积
		DUNet	2019	KBS	2D	采用可变形卷积
		基于视觉 Transformer 的方法	融合 Transformer 和 CNN 的方法	TransUNet	2021	arXiv
TransNUNet	2022			ICPECA	2D	在 TransUNet 的解码器中添加 CBAM 模块
TransClaw U-Net	2021			arXiv	2D	改进了 TransUNet 的跳转连接
CoTr	2021			MICCAI	3D	使用 Transformer 处理 CNN 特征提取中每一层的输出
MedT	2021			MICCAI	2D	使用全局—局部训练策略,根据模块深度区分不同分支
TransFuse	2021			MICCAI	2D	采用多分支结构,并行组合 CNN 和 Transformer
DHUNet	2023			BSPC	2D	采用双分支结构融合 Swin Transformer 和 ConvNeXt
UNet-2022	2022			MICAD	2D	并行组合 Transformer 和 CNN
UNETR	2022			WACV	3D	采用 Transformer 作为编码器
Swin UNETR	2021			MICCAI	3D	将 UNETR 的编码器替换为 Swin Transformer
H-DenseFormer	2023			MICCAI	2D/3D	采用多分支结构,提出基于密集连接的轻量 Trans- former
UCTransNet	2022			AAAI	2D	采用 Transformer 作为跳转连接和特征融合模块
U-Transformer	2021			MICCAI	2D	提出由多头自注意力和交叉注意力构成的 Trans- former
UTNet	2021			MICCAI	2D	结合 CNN 和 Transformer,提出一种新型注意力机制
基于纯 Transformer 的方法	Swin-Unet		2022	ECCV	2D	采用纯 Swin Transformer 构成分割网络
	MISSFormer		2021	arXiv	2D	提出增强的 Transformer 模块
	Segmenter		2021	ICCV	2D	采用纯 Transformer 解码器

续表 1 不同医学图像的语义分割算法对比
Continued Table 1 Comparison of different medical image segmentation algorithms

分类	方法	年份	发表	模态	亮点
基于 Mamba 的方法	U-Mamba	2024	arXiv	2D/3D	融合 CNN 与 Mamba
	Swin-UMamba	2024	MICCAI	2D	采用 ImageNet 的预训练权重
	ProMamba	2024	arXiv	2D	融合 Mamba 和 Prompt 技术
	H-vmunet	2024	arXiv	2D	提出 H-SS2D 模块
	LightM-UNet	2024	arXiv	3D	采用 Mamba 轻量化 UNet
	SegMamba	2024	MICCAI	3D	提出三维空间导向的 Mamba
	VM-UNet	2024	arXiv	2D	在 U 型结构中引入纯 Mamba
	Mamba-UNet	2024	arXiv	2D	在 U 型结构中引入纯 Mamba
	Mamba-HUNet	2024	arXiv	3D	融合 Mamba-UNet 和分层上采样网络

5 医学图像分割的挑战和机遇

在医学图像分割技术迅速发展的当下,研究者依然面临许多挑战,同时也迎来了众多亟待探索的机遇。尽管本文重点聚焦于基于深度学习的医学图像分割方法在算法结构方面的进展,但仅通过结构上的改进,很难在分割精度、效率和泛化性方面取得显著提升。本文认为,当前的挑战和机遇主要集中在算法结构、学习方法与数据以及任务范式 3 个方面。

5.1 算法结构

算法结构的设计会直接影响最终的分割精度和计算效率,而当前的分割方法在处理复杂图像和实现轻量化与实时推理方面仍面临诸多挑战。

5.1.1 面向复杂图像和目标

尽管现有方法在处理边界清晰的图像时表现优异,但在分割复杂图像和目标方面仍存在显著不足,尤其是在面对多模态、低对比度、高噪音图像以及形态多变的分割目标时存在诸多挑战。例如,采用磁共振成像(MRI)观察脑肿瘤时通常会涉及多参数扫描序列,如何从多模态 MRI 数据中精准分割肿瘤区域一直是一个挑战。此外,计算机断层扫描(CT)图像中常见的噪声和伪影也会显著增加分割任务的复杂性,导致分割算法在识别和界定目标区域时的准确性下降。类似地,超声图像由于其固有的高噪音和边界模糊特性,对分割算法的鲁棒性提出严峻考验。因此,面对这些复杂图像和分割目标,提升分割

算法的性能需要在模型架构上进行针对性的创新,例如采用多尺度特征提取技术和多模态融合技术,以更有效地捕捉和区分细微的图像特征。

5.1.2 轻量化与实时推理

在临床实践中,实时推理是对医学图像分割算法的核心要求。然而,目前的工作大多侧重于提高模型的精度,而忽略了计算效率,导致现有方法通常具有庞大的参数规模和缓慢的推理速度,这使其在计算资源受限的环境中难以应用。因此,设计高效且轻量的模型架构或探索模型压缩技术具有重要意义。例如,对于 CNN 而言,一种常见的轻量化策略是改造卷积操作,例如采用深度可分离卷积等高效卷积代替标准卷积模块,从而实现减少参数数量的目的。对于基于 Transformer 的分割网络,也有着类似于轴向压缩注意力(Wan 等,2023)等新型模块出现,可以显著减少标准 Transformer 模块计算量。至于 Mamba 类模型,由于 Mamba 本身就具有训练和推理速度上的优势,可以直接看做是一种轻量化的手段。此外,还可以采用一些通用的模型压缩技术(如网络剪枝和模型量化等)对现有方法进行优化。总的来说,以轻量化与实时推理为目标的医学图像分割算法设计仍处于早期阶段,有待进一步探索。

5.2 学习方法

基于强监督学习的医学图像分割方法的性能严重依赖于可用标注数据的规模和质量,而获取高质量、大规模的医学图像标注数据会产生巨大的时间和经济成本。因此,探索弱监督、域适应等低代价的分割方法具有重要意义。

5.2.1 弱监督学习

弱监督学习(weakly supervised learning)方法为医学图像分割提供了一种高效的解决方案,能够在利用不完全、不精确或存在噪声的标注数据的情况下,依然实现令人满意的分割性能。弱监督学习的核心理念在于通过多种策略弥补标注数据的不足,具体包括:利用部分标注数据进行训练(不完全监督)、采用粗粒度标注数据指导细粒度分割任务(不精确监督)及设计鲁棒算法以应对带有噪声的标注数据(不正确监督)。具体而言,不完全监督方法通过结合少量精确标注的数据与大量未标注的数据,利用精确标注样本引导模型的学习过程,同时通过模型自身的预测补充未标注数据,从而提升整体分割性能。在不精确监督方面,常见的方法包括利用图像级标签来引导像素级的分割任务,或通过区域级标注来指导边界的精细分割,这些方法在缺乏精确标注的情况下,仍然能有效地实现细粒度的分割结果。针对不正确监督,研究者们设计了噪声处理技术和鲁棒优化策略,使模型能够在存在标注错误的情况下,依然保持较高的学习效果和分割准确性。

5.2.2 域适应

域适应(domain adaptation)旨在通过在源域(例如通用数据集)上训练模型,将其有效迁移至目标域(如特定医学图像数据集)。该方法的核心目标在于缓解源域与目标域之间的数据分布差异,从而确保模型在目标域上依然能够保持优异的性能,即便目标域的数据分布与源域存在显著不同。域适应技术在医学图像分割中尤为重要,因为不同的成像设备、成像条件以及患者个体差异等因素常常导致图像特征的显著变化,进而使得模型在迁移至新数据集时性能下降。域适应通过解决源域和目标域之间的数据分布不一致问题,使得模型能够在新的目标域上维持良好的分割性能。这不仅提升了模型的适应性和泛化能力,还显著减少了对大规模标注数据的依赖。在实际应用中,常见的域适应技术包括对抗性训练(adversarial training)、生成对抗网络(generative adversarial network, GAN)以及特征对齐(feature alignment)等。这些方法通过不同的机制有效地缩小源域与目标域之间的分布差距,从而增强模型在目标域上的表现。

5.3 任务范式

当前的大多数医学图像分割方法都是为某一特

定的医学图像分割任务而设计的。然而,面向多样化的医学影像数据和复杂的实际临床场景,发展更具通用性的分割方法变得愈发重要。

5.3.1 交互式分割

区别于端到端的全自动分割,交互式分割方法旨在引入实时的人机交互操作来指导分割过程。用户可以通过点击、笔画或其他形式的输入,提供关于分割目标的指示信息,从而获得分割结果。在交互式分割中,模型会根据用户输入实时调整分割结果,使其逐步接近用户期望。例如,在进行器官分割时,用户可以点击目标器官内部的一些关键点,模型将这些点作为指导,通过迭代优化实现对器官的准确分割。交互式分割的另一个优势在于能够快速识别和纠正分割算法的错误,通过用户的反馈,实时修正分割结果。这种方法特别适用于需要高精度和高可信度的复杂医学图像分割任务,如肿瘤区域和重要器官的精确分割。交互式分割不仅能将用户的知识和经验融入分割过程,使分割结果更可控,而且可以应用于更广泛的任务场景,而不局限于特定数据集或分割目标。

5.3.2 通用分割

随着医学图像分割任务的多样化,研究者们逐渐意识到单一任务模型难以应对各种不同的分割需求。通用分割方法,旨在探索能够适应多种分割任务和多种图像模态的通用模型,实现一种“一模多用”的分割解决方案。通过在多任务和多模态医学图像数据上进行联合训练,使模型能够适应不同的分割任务,克服任务特定方法的局限性。该类方法的核心在于多任务学习和多模态融合:多任务学习使模型能够同时处理异质分割任务,通过共享底层特征提升整体性能;多模态融合则整合CT、MRI和超声等不同成像模态的数据,从而提高分割的准确性和稳定性。例如,UniverSeg(Butoi等,2023)作为一种通用分割模型,仅凭借若干提示图像一标注对即可对查询图像实现高精度分割,展示了通用分割范式的巨大潜力。可以说,通用分割模型的成功有望为医学图像分割领域带来革命性变革,显著提升算法的适用性和实用价值。

综上所述,尽管医学图像分割领域存在着诸多挑战,但也伴随着巨大的发展机遇,有望进一步推动该领域的持续发展,为临床疾病诊疗提供更为高效和可靠的智能辅助工具。

6 结 语

在临床疾病诊疗中,医学图像分割发挥着至关重要的作用。近年来,基于深度学习的方法已成为该领域的主流技术。然而,结合最新深度学习技术进行系统性综述的文献仍然较为匮乏,尤其是针对最新模型架构的深入探讨更为罕见。为填补这一空白,本文从模型结构的角度,深入分析了当前医学图像分割领域中3种最具代表性的架构:卷积神经网络(CNN)、视觉Transformer以及视觉Mamba,并系统探讨了这些结构的特性及其应用前景。

具体而言,本文首先将调研的相关方法按照结构划分为上述3类;随后重点介绍每类中最具代表性的模型及其核心原理,旨在帮助研究者全面理解这些模型的优势与局限;最后,本文从算法结构、学习方法和任务范式等多个维度,深入探讨了当前医学图像分割领域所面临的主要挑战与机遇,为未来新方法的探索指明潜在方向。本文不仅有助于学者深入掌握这些模型的基本原理和创新点,还为相关研究提供了有价值的参考,有助于推动医学图像分割领域的新方法创新与发展。随着算法和硬件的不断进步,以及新兴技术的融合与发展,有望实现更加高效、准确和可靠的医学图像分割,从而为精准医疗的发展提供坚实的技术支持。

参考文献(References)

- Almajalid R, Shan J, Du Y D and Zhang M. 2018. Development of a deep-learning-based method for breast ultrasound image segmentation//Proceedings of the 17th IEEE International Conference on Machine Learning and Applications (ICMLA). Orlando, USA: IEEE: 1103-1108 [DOI: 10.1109/ICMLA.2018.00179]
- Alom M Z, Hasan M, Yakopcic C, Taha T M and Asari V K. 2018. Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/1802.06955.pdf>
- Azad R, Heidari M, Shariatnia M, Aghdam E K, Karimijafarbigloo S, Adeli E and Merhof D. 2022. TransDeepLab: convolution-free transformer-based DeepLab v3+ for medical image segmentation//5th International Workshop on Predictive Intelligence in Medicine. Singapore, Singapore: Springer: 91-102 [DOI: 10.1007/978-3-031-16919-9_9]
- Butoi V I, Ortiz J J G, Ma T Y, Sabuncu M R, Guttag J and Dalca A V. 2023. UniverSeg: universal medical image segmentation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 21438-21451 [DOI: 10.1109/ICCV51070.2023.01960]
- Cao H, Wang Y Y, Chen J, Jiang D S, Zhang X P, Tian Q and Wang M N. 2022. Swin-Unet: Unet-like pure transformer for medical image segmentation//Proceedings of 2022 European Conference on Computer Vision. Tel Aviv, Israel: Springer: 205-218 [DOI: 10.1007/978-3-031-25066-8_9]
- Chen J J, Lu Y Y, Yu Q H, Luo X D, Adeli E, Wang Y, Lu L, Yuille A L and Zhou Y Y. 2021. TransUNet: transformers make strong encoders for medical image segmentation [EB/OL]. [2024-07-20]. <https://arxiv.org/pdf/2102.04306.pdf>
- Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2018a. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848 [DOI: 10.1109/TPAMI.2017.2699184]
- Chen L C, Zhu Y K, Papandreou G, Schroff F and Adam H. 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 801-818 [DOI: 10.1007/978-3-030-01234-2_49]
- Chen Y, Yin M, Li Y and Cai Q. 2022a. CSU-Net: a CNN-transformer parallel network for multimodal brain tumour segmentation. Electronics, 11(14): #2226 [DOI: 10.3390/electronics11142226]
- Chen Y P, Dai X Y, Chen D D, Liu M C, Dong X Y, Yuan L and Liu Z C. 2022b. Mobile-former: bridging mobileNet and Transformer//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5260-5269 [DOI: 10.1109/CVPR52688.2022.00520]
- Çiçek Ö, Abdulkadir A, Lienkamp S S, Brox T and Ronneberger O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation// Proceedings of the 19th International Conference on Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016. Athens, Greece: Springer: 424-432 [DOI: 10.1007/978-3-319-46723-8_49]
- Dai J F, Qi H Z, Xiong Y W, Li Y, Zhang G D, Hu H and Wei Y C. 2017. Deformable convolutional networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 764-773 [DOI: 10.1109/ICCV.2017.89]
- Dao T, Fu D Y, Ermon S, Rudra A and Ré C. 2022. FlashAttention: fast and memory-efficient exact attention with IO-awareness. Advances in Neural Information Processing Systems. 35: 16344-16359 [DOI: 10.48550/arXiv.2205.14135]
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Housley N. 2020. An image is worth 16 × 16 words: transformers for image recognition at scale [EB/OL].

- [2024-08-07]. <https://arxiv.org/pdf/2010.11929.pdf>
- Du G T, Cao X, Liang J M, Chen X L and Zhan Y H. 2020. Medical image segmentation based on U-Net: a review. *Journal of Imaging Science and Technology*, 64(2): #jst0710 [DOI: 10.2352/J.ImagingSci.Technol.2020.64.2.020508]
- Fu X H, Bi L, Kumar A, Fulham M and Kim J. 2021. Multimodal spatial attention module for targeting multimodal PET-CT lung tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 25(9): 3507-3516 [DOI: 10.1109/JBHI.2021.3059453]
- Fu Z J, Li J J and Hua Z. 2023. MSA-Net: multiscale spatial attention network for medical image segmentation. *Alexandria Engineering Journal*, 70: 453-473 [DOI: 10.1016/j.aej.2023.02.039]
- Gan X L, Wang L D, Chen Q, Ge Y J and Duan S K. 2021. GAU-Net: U-Net based on global attention mechanism for brain tumor segmentation. *Journal of Physics: Conference Series*, 1861(1): #012041 [DOI: 10.1088/1742-6596/1861/1/012041]
- Gao Y H, Zhou M and Metaxas D N. 2021. UTRNet: a hybrid transformer architecture for medical image segmentation//*Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*. Strasbourg, France: Springer: 61-71 [DOI: 10.1007/978-3-030-87199-4_6]
- Gu A and Dao T. 2023. Mamba: linear-time sequence modeling with selective state spaces [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/2312.0075.pdf>
- Gu A, Goel K and Ré C. 2021. Efficiently modeling long sequences with structured state spaces [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/2111.00396.pdf>
- Gu Z W, Cheng J, Fu H Z, Zhou K, Hao H Y, Zhao Y T, Zhang T Y, Gao S H and Liu J. 2019. CE-Net: context encoder network for 2D medical image segmentation. *IEEE Transactions on Medical Imaging*, 38(10): 2281-2292 [DOI: 10.1109/TMI.2019.2903562]
- Guo C L, Szemenyei M and Yi Y G. 2021. SA-UNet: spatial attention U-Net for retinal vessel segmentation//*Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*. Milan, Italy: IEEE: 1236-1242 [DOI: 10.1109/ICPR48806.2021.9413346]
- Guo J S, Zhou H Y, Wang L S and Yu Y Z. 2022. UNet-2022: exploring dynamics in non-isomorphic architecture//*Proceedings of 2022 International Conference on Medical Imaging and Computer-Aided Diagnosis*. Singapore, Singapore: Springer: 465-476 [DOI: 10.1007/978-981-16-6775-6_38]
- Guo Y R, Jiang J G, Hao S J, Zhan S and Li H. 2013. Medical image segmentation based on statistical similarity feature. *Journal of Image and Graphics*, 18(2): 225-234 (郭艳蓉, 蒋建国, 郝世杰, 詹曙, 李鸿. 2013. 统计相似度特征的医学图像分割. *中国图象图形学报*, 18(2): 225-234) [DOI: 10.11834/jig.20130215]
- Hao J, He L and Hung K F. 2024. T-Mamba: frequency-enhanced gated long-range dependency for tooth 3D CBCT segmentation [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/2404.01065.pdf>
- Hatamizadeh A, Nath V, Tang Y C, Yang D, Roth H R and Xu D G. 2021. Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images//*7th International MICCAI Brainlesion Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Cham, Switzerland: Springer: 272-284 [DOI: 10.1007/978-3-031-08999-2_22]
- Hatamizadeh A, Tang Y C, Nath V, Yang D, Myronenko A, Landman B, Roth H R and Xu D G. 2022. UNETR: transformers for 3D medical image segmentation//*Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision*. Waikoloa, USA: 574-584 [DOI: 10.1109/WACV51458.2022.00181]
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, USA: IEEE [DOI: 10.1109/cvpr.2016.90]
- Howard A G, Zhu M L, Chen B, Kalenichenko D, Wang W J, Weyand T, Andreetto M and Adam H. 2017. MobileNets: efficient convolutional neural networks for mobile vision applications [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/1704.04861.pdf>
- Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]
- Huang G, Liu Z, Van Der Maaten L and Weinberger K Q. 2017. Densely connected convolutional networks//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, USA: 2261-2269 [DOI: 10.1109/CVPR.2017.243]
- Huang G H, Zhu J W, Li J J, Wang Z W, Cheng L L, Liu L Z, Li H J and Zhou J. 2020a. Channel-attention U-Net: channel attention mechanism for semantic segmentation of esophagus and esophageal cancer. *IEEE Access*, 8: 122798-122810 [DOI: 10.1109/ACCESS.2020.3007719]
- Huang H M, Lin L F, Tong R F, Hu H J, Zhang Q W, Iwamoto Y, Han X H, Chen Y W and Wu J. 2020b. UNet 3+: a full-scale connected UNet for medical image segmentation//*Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE: 1055-1059 [DOI: 10.1109/ICASSP40776.2020.9053405]
- Huang X H, Deng Z F, Li D D and Yuan X G. 2021. MISSFormer: an effective medical image segmentation transformer [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/2109.07162.pdf>
- Ibtehaz N and Sohel Rahman M. 2020. MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks*, 121: 74-87 [DOI: 10.1016/j.neunet.2019.08.025]
- Isensee F, Jaeger P F, Kohl S A A, Petersen J and Maier-Hein K H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2): 203-211 [DOI: 10.1038/s41592-020-01008-z]

- Jin Q G, Meng Z P, Pham T D, Chen Q, Wei L Y and Su R. 2019. DUNet: a deformable network for retinal vessel segmentation. *Knowledge-Based Systems*, 178: 149-162 [DOI: 10.1016/j.knosys.2019.04.025]
- Kan H Y, Shi J, Zhao M F, Wang Z H, Han W T, An H, Wang Z Y and Wang S. 2022. ITUNet: integration of transformers and Unet for organs-at-risk segmentation//*Proceedings of the 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Glasgow, Scotland, United Kingdom: IEEE: 2123-2127 [DOI: 10.1109/EMBC48229.2022.9871945]
- Kaul C, Manandhar S and Pears N. 2019. Focusnet: an attention-based fully convolutional network for medical image segmentation//*Proceedings of the 16th IEEE International Symposium on Biomedical Imaging*. Venice, Italy: IEEE: 455-458 [DOI: 10.1109/ISBI.2019.8759477]
- Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84-90 [DOI: 10.1145/3065386]
- LeCun Y, Bottou L, Bengio Y and Haffner P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278-2324 [DOI: 10.1109/5.726791]
- Lei T, Zhou W Z, Zhang Y X, Wang R S, Meng H Y and Nandi A K. 2020. Lightweight V-Net for liver segmentation//*Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE: 1379-1383 [DOI: 10.1109/ICASSP40776.2020.9053454]
- Li R R, Li M M, Li J C and Zhou Y T. 2019. Connection sensitive attention U-NET for accurate retinal vessel segmentation [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/1903.05558.pdf>
- Li X M, Chen H, Qi X J, Dou Q, Fu C W and Heng P A. 2018. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging*, 37(12): 2663-2674 [DOI: 10.1109/TMI.2018.2845918]
- Li Y, Zhang Y, Liu J Y, Wang K, Zhang K, Zhang G S, Liao X F and Yang G. 2022. Global transformer and dual local attention network via deep-shallow hierarchical feature fusion for retinal vessel segmentation. *IEEE Transactions on Cybernetics*, 53(9): 5826-5839. [DOI: 10.1109/TCYB.2022.3194099]
- Liang M and Hu X L. 2015. Recurrent convolutional neural network for object recognition//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Boston, USA: IEEE: 3367-3375 [DOI: 10.1109/cvpr.2015.7298958]
- Liao W B, Zhu Y H, Wang X Y, Pan C W, Wang Y S and Ma L T. 2024. LightM-UNet: Mamba assists in lightweight UNet for medical image segmentation [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/2403.05246.pdf>
- Liu X, Yan Z Q, Deng X B, Zheng C S and Yu L. 2023. ConvFormer: plug-and-play CNN-style transformers for improving medical image segmentation//*Proceedings of the 26th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vancouver, Canada: Springer: 642-651 [DOI: 10.1007/978-3-031-43901-8_61]
- Liu J R, Yang H, Zhou H Y, Xi Y, Yu L Q, Li C, Liang Y, Shi G M, Yu Y Z, Zhang S T, Zheng H R and Wang S S. 2024a. Swin-UMamba: Mamba-based UNet with imageNet-based pretraining//*Proceedings of the 27th International Conference on Medical Image Computing and Computer Assisted Intervention*. Marrakesh, Morocco: Springer: 615-625 [DOI: 10.1007/978-3-031-72114-4_59]
- Liu W F, Luo J X, Yang Y, Wang W L, Deng J K and Yu L. 2022a. Automatic lung segmentation in chest X-ray images using improved U-Net. *Scientific Reports*, 12(1): #8649 [DOI: 10.1038/s41598-022-12743-y]
- Liu Y, Tian Y J, Zhao Y Z, Yu H T, Xie L X, Wang Y W, Ye Q X, Jiao J B and Liu Y F. 2024b. VMamba: visual state space model [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/2401.10166.pdf>
- Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin Transformer: hierarchical vision Transformer using shifted windows//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*. Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]
- Liu Z, Mao H Z, Wu C Y, Feichtenhofer C, Darrell T and Xie S N. 2022b. A convNet for the 2020s//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, USA: IEEE: 11976-11986 [DOI: 10.1109/CVPR52688.2022.01167]
- Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Boston, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]
- Ma J, Li F F and Wang B. 2024b. U-Mamba: enhancing long-range dependency for biomedical image segmentation [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/2401.04722.pdf>
- Mei H C, Lei W H, Gu R, Ye S, Sun Z W T, Zhang S C and Wang G T. 2021. Automatic segmentation of gross target volume of nasopharynx cancer using ensemble of multiscale deep neural networks with spatial attention. *Neurocomputing*, 438: 211-222 [DOI: 10.1016/j.neucom.2020.06.146]
- Milletari F, Navab N and Ahmadi S A. 2016. V-Net: fully convolutional neural networks for volumetric medical image segmentation//*Proceedings of 2016 International Conference on 3D Vision*. Stanford, USA: IEEE: 565-571
- Moradi S, Oghli M G, Alizadehasl A, Shiri I, Oveisi N, Oveisi M, Maleki M and Dhooge J. 2019. MFP-Unet: a novel deep learning based approach for left ventricle segmentation in echocardiography. *Physica Medica*, 67: 58-69 [DOI: 10.1016/j.ejmp.2019.10.001]
- Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla N Y, Kainz B, Glocker B and

- Rueckert D. 2018. Attention U-Net: learning where to look for the pancreas [EB/OL]. [2024-08-07].
<https://arxiv.org/pdf/1804.03999.pdf>
- Park J, Woo S, Lee J Y and Kweon I S. 2018. BAM: Bottleneck attention module//Proceedings of 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]
- Peng D L, Yu X, Peng W J and Lu J P. 2021. DGFAU-Net: global feature attention upsampling network for medical image segmentation. *Neural Computing and Applications*, 33: 12023-12037. [DOI: 10.1007/s00521-021-05908-9]
- Qi K H, Yang H, Li C, Liu Z Y, Wang M Y, Liu Q G and Wang S S. 2019. X-Net: brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies//Proceedings of the 22nd International Conference on Medical Image Computing and Computer Assisted Intervention—MICCAI 2019. Shenzhen, China: Springer: 247-255 [DOI: 10.1007/978-3-030-32248-9_28]
- Qian Y and Zhang Y J. 2008. Level set methods and its application on image segmentation. *Journal of Image and Graphics*, 13(1): 7-13 (钱芸, 张英杰. 2008. 水平集的图像分割方法综述. *中国图象图形学报*, 13(1): 7-13) [DOI: 10.11834/jig.20080102]
- Qiao L, Liu Q, Shi J, Zhao M F, Kan H Y, Wang Z H, An H, Xiao C G and Wang S. 2022. FcTC-UNet: fine-grained combination of transformer and CNN for thoracic organs segmentation//Proceedings of the 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Glasgow, United Kingdom: IEEE: 4749-4753 [DOI: 10.1109/EMBC48229.2022.9870880]
- Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]
- Ruan J C, Li J C and Xiang S C. 2024. Vm-Unet: vision Mamba UNet for medical image segmentation [EB/OL]. [2024-08-07].
<https://arxiv.org/pdf/2402.02491.pdf>
- Samarasinghe G, Jameson M, Vinod S, Field M, Dowling J, Sowmya A and Holloway L. 2021. Deep learning for segmentation in radiation therapy planning: a review. *Journal of Medical Imaging and Radiation Oncology*, 65(5): 578-595 [DOI: 10.1111/1754-9485.13286]
- Sanjid K S, Hossain M T, Junayed M S S and Uddin M M. 2024. Integrating Mamba sequence model and hierarchical upsampling network for accurate semantic segmentation of multiple sclerosis lesion [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/2403.17432.pdf>
- Shi J, Kan H Y, Ruan S L, Zhu Z Q, Zhao M F, Qiao L, Wang Z H, An H and Xue X D. 2023. H-DenseFormer: an efficient hybrid densely connected transformer for multimodal tumor segmentation//Proceedings of the 26th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Vancouver, Canada: Springer: 692-702 [DOI: 10.1007/978-3-031-43901-8_66]
- Siddique N, Paheding S, Elkin C P and Devabhaktuni V. 2021. U-net and its variants for medical image segmentation: a review of theory and applications. *IEEE Access*, 9: 82031-82057 [DOI: 10.1109/ACCESS.2021.3086020]
- Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: Computational and Biological Learning Society: 1-14 [DOI: 10.48550/arXiv.1409.1556]
- Strudel R, Garcia R, Laptev I and Schmid C. 2021. Segmenter: transformer for semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7262-7272 [DOI: 10.1109/ICCV48922.2021.00717]
- Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1-9 [DOI: 10.1109/CVPR.2015.7298594]
- Valanarasu J M J, Oza P, Hacıhaliloglu I and Patel V M. 2021. Medical transformer: gated axial-attention for medical image segmentation//Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention—MICCAI 2021. Strasbourg, France: Springer: 36-46 [DOI: 10.1007/978-3-030-87193-2_4]
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010
- Wan Q, Huang Z L, Lu J C, Yu G and Zhang L. 2023. SeaFormer: squeeze-enhanced axial Transformer for mobile semantic segmentation//Proceedings of 2023 International Conference on Learning Representations
- Wang C Y, He Y, Liu Y F, He Z F, He R and Sun Z N. 2019a. ScleraSegNet: an improved U-net model with attention for accurate sclera segmentation//Proceedings of 2019 International Conference on Biometrics (ICB). Crete, Greece: IEEE: 1-8 [DOI: 10.1109/ICB45273.2019.8987270]
- Wang H N, Cao P, Wang J Q and Zaiane O R. 2022a. UCTransNet: rethinking the skip connections in U-Net from a channel-wise perspective with transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3): 2441-2449 [DOI: 10.1609/aaai.v36i3.20144]
- Wang H Y, Xie S A, Lin L F, Iwamoto Y, Han X H, Chen Y W and Tong R F. 2022b. Mixed transformer U-Net for medical image segmentation//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore,

- Singapore: IEEE: 2390-2394 [DOI: 10.1109/ICASSP43922.2022.9746172]
- Wang J and Liu X P. 2021. Medical image recognition and segmentation of pathological slices of gastric cancer based on Deeplab v3+ neural network. *Computer Methods and Programs in Biomedicine*, 207: #106210 [DOI: 10.1016/j.cmpb.2021.106210]
- Wang J H, Chen J T, Chen D and Wu J. 2024a. Large window-based Mamba UNet for medical image segmentation: beyond convolution and self-attention [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/2403.07332v1.pdf>
- Wang L, Pan L R, Wang H T, Liu M T, Feng Z C, Rong P F, Chen Z and Peng S L. 2023. DHUnet: dual-branch hierarchical global - local fusion network for whole slide image segmentation. *Biomedical Signal Processing and Control*, 85: #104976 [DOI: 10.1016/j.bspc.2023.104976]
- Wang X C, Li W, Miao B Y, He J, Jiang Z W, Xu W, Ji Z Y, Hong G and Shen Z M. 2018. Retina blood vessel segmentation using a U-Net based Convolutional neural network//*Proceedings of 2018 Procedia Computer Science: International Conference on Data Science ICDS 2018*. 8-9
- Wang Z Y, Zheng J Q, Zhang Y C, Cui G and Li L. 2024b. Mamba-UNet: UNet-like pure visual Mamba for medical image segmentation [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/2402.05079.pdf>
- Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: convolutional block attention module//*Proceedings of 2018 European Conference on Computer Vision (ECCV)*. Munich, Germany: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]
- Wu R K, Liu Y H, Liang P C and Chang Q. 2024. H-vmunet: high-order vision Mamba UNet for medical image segmentation [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/2403.13642.pdf>
- Xiao H G, Li L, Liu Q Y, Zhu X H and Zhang Q H. 2023. Transformers in medical image segmentation: a review. *Biomedical Signal Processing and Control*, 84: #104791 [DOI: 10.1016/j.bspc.2023.104791]
- Xiao X, Lian S, Luo Z M and Li S Z. 2018. Weighted Res-UNet for high-quality retina vessel segmentation//*Proceedings of the 9th International Conference on Information Technology in Medicine and Education (ITME)*. Hangzhou, China: IEEE [DOI: 10.1109/itme.2018.00080]
- Xie J H, Liao R F, Zhang Z A, Yi S D, Zhu Y S and Luo G B. 2024. ProMamba: prompt-Mamba for polyp segmentation [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/2403.13660.pdf>
- Xie Y T, Zhang J P, Shen C H and Xia Y. 2021. CoTr: efficiently bridging CNN and transformer for 3D medical image segmentation//*Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*. Strasbourg, France: SpringerL 171-180 [DOI: 10.1007/978-3-030-87199-4_16]
- Xing Z H, Ye T, Yang Y J, Liu G and Zhu L. 2024. SegMamba: long-range sequential modeling Mamba for 3D medical image segmentation//*Proceedings of the 27th International Conference on Medical Image Computing and Computer Assisted Intervention*. Marrakesh, Morocco: Springer: 578-588 [DOI: 10.1007/978-3-031-72111-3_54]
- Xu H W, Yan P X, Wu M, Xu Z Y and Sun Y B. 2020. Automated segmentation of cystic kidney in CT images using residual double attention motivated U-Net model. *Application Research of Computers*, 37(7): 2237-2240 (徐宏伟, 闫培新, 吴敏, 徐振宇, 孙玉宝. 2020. 基于残差双注意力 U-Net 模型的 CT 图像囊肿肾脏自动分割. *计算机应用研究*, 37(7): 2237-2240) [DOI: 10.19734/j.issn.1001-3695.2019.03.0092]
- Yang B, Liu X F and Zhang J. 2021. Medical image segmentation based on deep feature aggregation network. *Computer Engineering*, 47(4): 187-196 (杨兵, 刘晓芳, 张纠. 2021. 基于深度特征聚合网络的医学图像分割. *计算机工程*, 47(4): 187-196) [DOI: 10.19678/j.issn.1000-3428.0057330]
- Yang X N and Tian X L. 2022. TransUNet: using attention mechanism for whole heart segmentation//*Proceedings of the 2nd IEEE International Conference on Power, Electronics and Computer Applications (ICPECA)*. Shenyang, China: IEEE: 553-556 [DOI: 10.1109/ICPECA53709.2022.9719101]
- Yao C, Hu M H, Zhai G T and Zhang X P. 2021. TransClaw U-Net: claw U-Net with transformers for medical image segmentation [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/2107.05188.pdf>
- Yu F and Koltun V. 2015. Multi-scale context aggregation by dilated convolutions [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/1511.07122.pdf>
- Yu X, Yang Q, Zhou Y C, Cai L Y, Gao R Q, Lee H H, Li T, Bao S X, Xu Z B, Lasko T A, Abramson R G, Zhang Z Z, Huo Y K, Landman B A and Tang Y C. 2023. UNesT: local spatial representation learning with hierarchical transformer for efficient medical segmentation. *Medical Image Analysis*, 90: #102939 [DOI: 10.1016/j.media.2023.102939]
- Zhang L, Zhang K J and Pan H W. 2023. SUNet++: a deep network with channel attention for small-scale object segmentation on 3D medical images. *Tsinghua Science and Technology*, 28(4): 628-638 [DOI: 10.26599/TST.2022.9010023]
- Zhang X F, Zhang S, Zhang D H and Liu R. 2023. Group attention-based medical image segmentation model. *Journal of Image and Graphics*, 28(10): 3231-3242 (张学峰, 张胜, 张冬晖, 刘瑞. 2023. 引入分组注意力的医学图像分割模型. *中国图象图形学报*, 28(10): 3231-3242) [DOI: 10.11834/jig.220748]
- Zhang Y D, Liu H Y and Hu Q. 2021. TransFuse: fusing transformers and CNNs for medical image segmentation//*Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Strasbourg, France: Springer: 14-24 [DOI: 10.1007/978-3-030-87193-2_2]
- Zhou H Y, Guo J S, Zhang Y H, Yu L Q, Wang L S and Yu Y Z. 2021.

nnFormer: interleaved transformer for volumetric segmentation [EB/OL]. [2024-08-07]. <https://arxiv.org/pdf/2109.03201.pdf>

- Zhou T, Dong Y L, Huo B Q, Liu S and Ma Z J. 2021. U-Net and its applications in medical image segmentation: a review. *Journal of Image and Graphics*, 26(9): 2058-2077 (周涛, 董雅丽, 霍兵强, 刘珊, 马宗军. 2021. U-Net网络医学图像分割应用综述. *中国图象图形学报*, 26(9): 2058-2077) [DOI: 10.11834/jig.200704]
- Zhou X Y, Zheng J Q, Li P C and Yang G Z. 2020. ACNN: a full resolution DCNN for medical image segmentation//*Proceedings of 2020 IEEE International Conference on Robotics and Automation (ICRA)*. Paris, France: IEEE: 8455-8461 [DOI: 10.1109/ICRA40945.2020.9197328]
- Zhou Z W, Rahman Siddiquee M M, Tajbakhsh N and Liang J M. 2018. UNet++: a nested U-Net architecture for medical image segmentation//*4th International Workshop on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Granada, Spain: Springer: 3-11 [DOI: 10.1007/978-3-030-00889-5_1]
- Zhu W T, Huang Y F, Zeng L, Chen X M, Liu Y, Qian Z, Du N, ,

Fan W and Xie X H. 2019. AnatomyNet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical Physics*, 46(2): 576-589 [DOI: 10.1002/mp.13300]

作者简介

石军,男,博士后,主要研究方向为医学影像认知计算、AI for Science。E-mail: shijun18@mail.ustc.edu.cn

安虹,通信作者,女,教授,主要研究方向为高性能计算和医学影像认知计算。E-mail: han@ustc.edu.cn

王天同,男,硕士研究生,主要研究方向为医学影像认知计算。E-mail: wangtiantong23@mail.ustc.edu.cn

朱子琦,男,博士研究生,主要研究方向为医学影像认知计算和大模型应用技术。E-mail: tally@mail.ustc.edu.cn

赵敏帆,男,博士研究生,主要研究方向为医学影像分析和多模态大模型。E-mail: zmf@mail.ustc.edu.cn

王炳勋,男,硕士研究生,主要研究方向为医学影像认知计算。E-mail: wbxun@mail.ustc.edu.cn