语义扩散对齐的多尺度感知医学图像分割方法

赖晓婷, 张静*

(华东理工大学信息科学与工程学院 上海 200237)
(jingzhang@ecust.edu.cn)

摘 要: 有效的多尺度特征表示对于准确地分割医学图像中大小不一的病变区域具有重要意义. 针对现有方法未能充分挖掘不同目标的多尺度信息导致难以实现精确分割的问题, 提出一种语义扩散对齐的多尺度感知医学图像分割方法. 首先从局部和全局 2 个角度探索多尺度上下文信息的感知能力, 构建多尺度编码器和解码器. 其中, 多尺度编码器利用局部多尺度自注意力机制和全局重调整来提取多尺度特征, 以捕获图像中不同目标的信息, 多尺度解码器则通过上采样恢复空间分辨率并保留细节信息, 得到更准确的分割结果; 为了进一步增强特征语义表示, 提出语义扩散对齐模块, 实现了低级特征与高级特征的语义对齐, 获得更具鉴别性的融合特征. 在 Synapse 和 ACDC 数据集上的实验结果表明, 所提方法的平均 Dice 相似系数分别达到 82.42%和 92.25%, 性能优于大多数主流分割方法. 该方法的相关代码链接为 https://github.com/MiniCoCo-be/MSPSN.

关键词: 医学图像分割; 多尺度感知; 语义扩散; 特征对齐; 注意力机制
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023-00604

Semantic Diffusion Alignment-Based Multi-Scale Perception for Medical Image Segmentation

Lai Xiaoting and Zhang Jing*

(*Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai* 200237)

Abstract: Effective multi-scale feature representation is crucial for accurately segmenting lesions of varying sizes in medical images. Addressing the challenge of existing methods failing to fully exploit multi-scale information for different targets, a novel multi-scale perception medical image segmentation method was proposed based on semantic diffusion alignment. It first explored the perceptual ability of multi-scale contextual information from both local and global perspectives to construct multi-scale encoder and decoder. The multi-scale encoder utilized a local multi-scale self-attention mechanism and global fine-tuning to extract features at multiple scales, capturing information from different targets in the image. The multi-scale decoder restored spatial resolution through upsampling while preserving detailed information to achieve more precise segmentation results. Furthermore, in order to further enhance the semantic representation of features, a semantic diffusion alignment module was proposed, which realizes the semantic alignment of low-level features and high-level features, thus obtaining more discriminative fusion features. Finally, experimental results on the Synapse dataset and ACDC dataset demonstrate that the proposed method achieves average Dice similarity coefficients of 82.42% and 92.25% respectively, outperforming most existing methods. Additionally, the relevant code can be found at https://github.com/MiniCoCo-be/MSPSN.

收稿日期: 2023-09-13; 修回日期: 2024-06-12. **基金项目**: 上海市自然科学基金(22ZR1418400). **赖晓婷**(1998—), 女, 硕士研究生, 主要研究方向为医学图像处理、计算机视觉; **张静**(1978—), 女, 博士, 副教授, 硕士生导师, CCF 会员, 论文通信作者, 主要研究方向为医学图像处理、视觉问答、图像描述、小样本学习.

Key words: medical image segmentation; multi-scale perception; semantic diffusion; feature alignment; attention mechanism

医学图像分割是医学图像分析领域中的一项重要任务, 旨在从医学图像中识别和分割出感兴趣区域, 如器官、组织结构和病变区域等. 这些准确的分割结果对于后续的临床诊断、治疗规划和疾病研究具有重要意义[1].

近年来, 随着深度学习的发展, 涌现出许多基于深度学习的医学图像分割方法. Ronneberger 等[2]开创性地提出对称性全卷积网络 U-Net, 其包括编码器、跳跃连接和解码器 3 部分. 得益于跳跃连接, 低级和高级图像特征能够有效地融合, 实现了强大的语义表示学习, 提升了图像分割的准确性[3]. 目前, U-Net 在医学图像分割任务得到广泛应用, 并出现了大量基于 U-Net 的变体[4]. 然而, 由于 U-Net 及其变体都采用基于卷积层的架构, 其在学习多尺度信息方面存在一定的局限性.

对于分割这种密集预测任务, 多尺度特征表示具有至关重要的作用, 其用于区分各种尺寸的对象或区域[5]. 在医学图像中这一点尤为重要, 因为误分割可能对后续的诊断结果产生不良影响. 为了缓解这个问题, 一些学者提出并行结构模块[6]来增强捕获多尺度特征的能力. Zhao 等[7]采用多尺度空洞卷积思想, Chen 等[8]使用空间金字塔池化, 以生成不同尺度的信息表示. 受到 Transformer[9]的启发, 还有一些研究人员将自注意力机制引入医学图像分割中, 以提高感知上下文信息的能力. Lin 等[10]和 Chen 等[11]在编码器的最后阶段利用其设计的注意力机制, 提高上下文建模能力; Zhang 等[12]将图像缩放到不同尺度输入到模型中, 从而获取不同尺度的特征. 为了进一步增强网络提取特征的语义丰富性, Li 等[13]提出门控全融合架构, 以完全连接的方式使用门选择性地融合来自多个级别的特征, 实现增强的语义特征表示; Wang 等[14]从通道角度提出多尺度通道交叉融合的 Transformer 子模块, 以及一个用于引导融合的通道级交叉注意力子模块来替代原始的跳过连接, 有效解决医学图像精准自动分割中的语义差距问题, 极大地提高了分割性能.

尽管上述方法在分割任务中展现了良好的结果, 但仍然存在一些挑战: 在多尺度信息建模能力方面, 虽然上述方法取得了一些进展, 但没有充分地考虑存在的大尺度对象和小尺度对象的信息学习, 不能有效地学习多尺度特征; 虽然现有方法在一定程度上增强了语义表示, 但由于医学图像前景和背景信息相似, 简单地过滤信息或信息交互, 可能导致低层和高层特征之间的语义信息无法有效地互相补充, 从而导致分割性能不尽如人意.

针对上述问题, 本文提出一种语义扩散对齐的多尺度感知医学图像分割方法, 有效地提高了分割效果. 为了同时关注到不同大小的尺度信息, 提出局部-全局多尺度感知模块(local-global multi-scale perception module, LGMSPM), 并基于此构建多尺度编码器和多尺度解码器, 以有效地捕获不同器官大小的细节信息和上下文信息; 从语义扩散角度出发, 提出语义扩散对齐模块(semantic diffusion alignment module, SDAM), 利用像素之间的空间关系和多尺度信息分析减少局部噪声的干扰, 并以高级特征语义为引导, 在不同层级上实现语义信息的层级扩散, 实现低级特征语义与高级特征语义的对齐融合; 使用深度监督(deep supervision, DS)方法[15-16]优化模型学习过程, 实现准确地分割医学图像.

1　相关工作

1.1　多尺度特征学习方法

近年来, 深度学习在医学图像分割中取得了显著的发展, 其中, 多尺度特征表示学习具有至关重要的作用. Cai 等[17]采用密集连接的思想[18], 以密集的方式增强上下文信息, 以获取多尺度特征表示. 还有一些嵌入到各级网络中的并行结构. Al-Masni 等[19]提出一种使用膨胀卷积和金字塔池化模块的上下文多尺度多层级网络, 在多个空间尺度上融合全局上下文特征. 为了进一步增强多尺度特征表示, Li 等[20]提出了一种多尺度输入网络, 使用多输入融合模块和多尺度空洞空间金字塔模块增强获取多尺度信息的能力, 进一步提高了对小血管区域的分割能力.

除了上述添加不同感受野的并行结构, 以及在输入尺度方面增强多尺度特征表示外, 还有一些学者利用自注意机制实现特征增强. Lin 等[21]利用基于分层 Swin Transformer[22]的双尺度编码器替代传统的编码器结构, 学习多尺度特征表示; 文献[23-25]将 Transformer 集成到 U 型架构的编码器和解码器中, 进一步增强特征表示学习能力.

1.2　多层级特征融合方法

尽管多尺度特征表示学习的增强在一定程度

上提高了分割性能，但由于下采样的使用导致细节信息丢失使分割性能受限．为了解决这个问题，Ronneberger 等[2]和 Long 等[26]使用跳跃连接，利用浅层特征图补充丢失的空间信息．为了进一步考虑不同层级特征图之间存在的语义差距，Zhou 等[27]直接用嵌套的密集跳跃连接取代普通的跳跃连接．

受到 ViT[28]的启发，Wang 等[14]在跳跃连接上提出有效的 Transformer 模块，以弥补编码器中下采样带来的细节损失；Li 等[29]通过引入 Transformer 的注意力解码器，在局部窗口中基于来自下采样路径的局部和详细特征对特征进行上采样，得到更好的特征表示．

1.3 基于扩散模型的医学图像分割方法

近年来，针对图像分割的研究广泛关注利用生成模型的方法．其中，去噪扩散概率模型(denoising diffusion probabilistic models，DDPM)[30]通过逐步添加噪声来迭代地改进数据样本，以学习数据的概率分布．近年来的研究表明，DDPM 学习得到的表示具有对于分割任务至关重要的高级语义信息[31-32]．

Brempong 等[33]提出解码器去噪预训练(decoder denoising pretraining，DDeP)方法，通过对加噪图像进行去噪的预训练得到解码器的初始参数，之后进行常规的监督训练，显著地提高了语义分割的准确性．Amit 等[34]则将含有随机噪声的二值分割图与输入图像相加输入到去噪网络中进行多次迭代，通过取多次迭代的平均值得到最终的分割结果．与 Amit 等[34]的方法不同，Wolleb 等[35]提出一种直接在真实分割标签上添加噪声的方法，将噪声标签与原始图像进行拼接再输入到去噪网络中，生成预测的分割掩码．Baranchuk 等[32]通过激活 DDPM 的中间表示，证明 DDPM 能够有效地捕获图像的语义信息，并将其用于分割任务，实现在少量训练样本下的准确分割．

上述扩散分割模型通过添加噪声，使模型在去除噪声的过程中学习到图像特征分布，最终生成分割结果．受到这种加噪去噪方法在特征优化学习方面的启发，本文提出的 SDAM 将低级语义特征视作噪声，并将其添加到高级特征上采样的过程中，以学习更精确的特征恢复，进一步提高模型对语义信息的学习能力．

尽管上述方法在医学图像分割方面取得了一定的进展，但在处理不同尺度目标时仍存在一些局限性．为了解决上述问题，本文提出一种语义扩散对齐的多尺度感知医学图像分割方法，首先提出 LGMSPM，利用残差卷积层学习局部特征表示，再结合局部多尺度注意力机制增强多尺度学习能力，最后引入全局 token 进行特征重校准，得到更好的多尺度特征表示；还提出了一个 SDAM，通过交叉注意力分析像素空间关系，实现低级特征和高级特征更好的融合，得到更具鉴别性的语义特征．

2 本文方法

2.1 整体框架

本文提出一种语义扩散对齐的多尺度感知医学图像分割方法，它是一种对称的 U 型编码器解码器网络结构，其整体框架如图 1 所示．其中，主干部分包括对称的基于 LGMSPM 的多尺度编码器和多尺度解码器，如图中 b 和 d 部分所示；基于 SDAM 的跳跃连接，如 c 部分所示；初级特征学习模块如 a 部分所示，由 ResNet 中的 2 个残差块构成；预测模块如 e 部分所示，由 1×1 卷积实现．

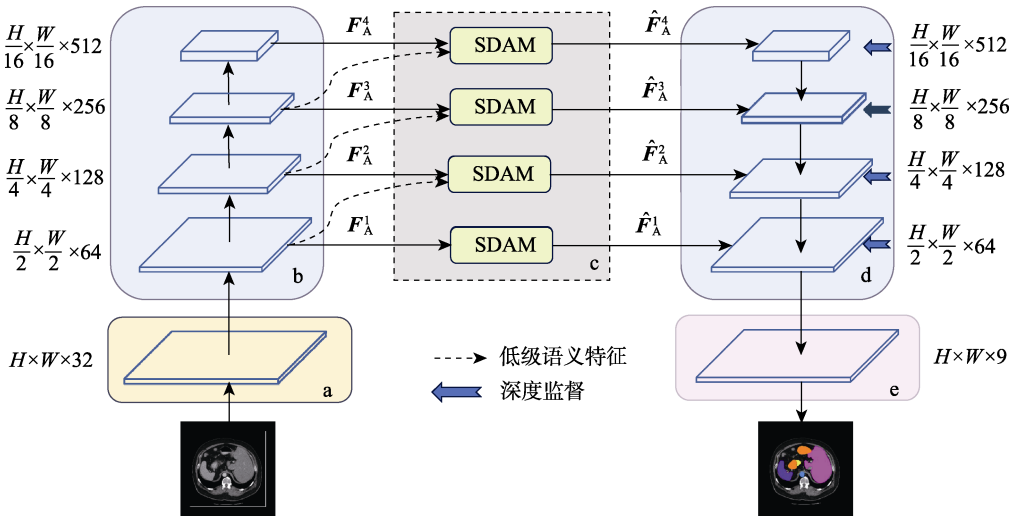


图 1 语义扩散对齐的多尺度感知医学图像分割方法

给定一幅输入图像，本文方法主要步骤如下.

Step1. 通过初级特征学习模块进行初步特征提取.

Step2. 利用多尺度编码器，通过局部多尺度感知模块(multi-scale perception module, MSPM)和全局重调整模块实现深层多尺度特征的互补学习.

Step3. 跳跃连接机制通过交叉注意力的自适应学习，实现低级特征与高级特征的语义对齐，获得更具鉴别性的融合特征，注意，跳跃连接第 1 层位于初步提取特征阶段，只接收来自初级特征学习模块的单一输入，而其他层则采用相邻特征作为输入.

Step4. 利用多尺度解码器对融合的特征进行上采样，有效地恢复空间分辨率和语义信息，得到与原图大小相同且具有丰富语义的特征图.

Step5. 通过像素级预测模块得到最终的分割结果. 在训练阶段，本文还采用 DS 机制[15-16]对解码器输出的每层特征添加额外的监督，以学习更具判别性的特征.

2.2　LGMSPM 学习

多尺度特征学习有利于提高密集预测任务的性能[5]. 受到 PSPNet[7]中空间金字塔池化的启发，本文将空间金字塔池化应用于自注意机制的键和值，以感知多尺度的信息. 结合局部特征学习的优点，本文提出的 LGMSPM 由局部特征学习模块、多尺度感知模块和全局重调整模块组成，其结构如图 2 所示. 基于此，本文构建多尺度编码器和多尺度解码器作为网络的主干结构，以获取适应多尺度的特征表示.

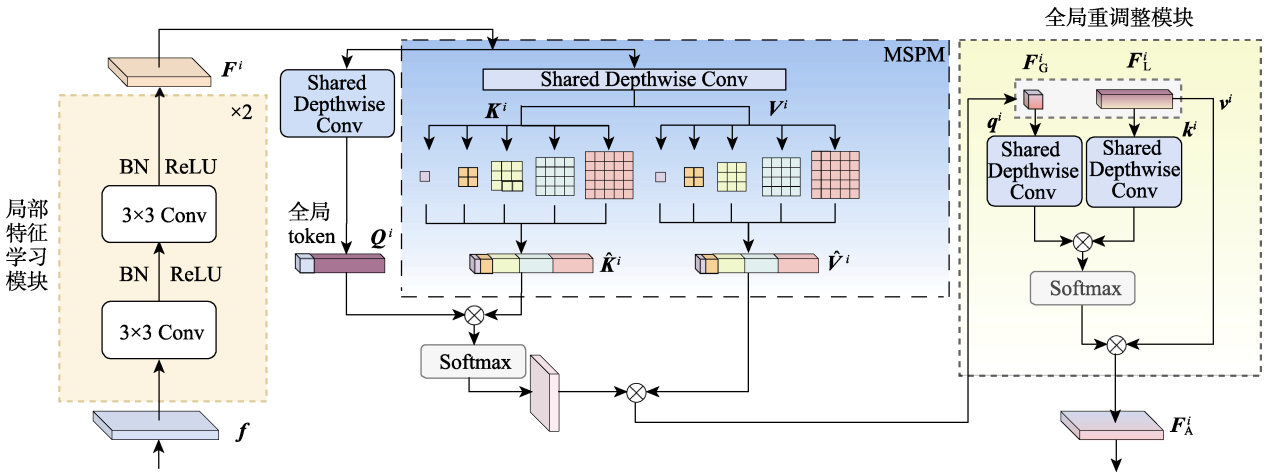


图 2　LGMSPM

给定 $\boldsymbol{f}$ 为输入特征，由图 1 中 a 部分的初级特征学习模块生成，作为后续流程的输入特征，再利用局部特征学习模块通过卷积操作捕捉图像的局部信息；多尺度感知模块采用局部多尺度注意力机制，动态地学习和融合不同尺度的特征；全局重调整模块关注全局的多尺度信息，并通过注意力机制对先前学习到的特征进行调整和优化，进一步增强多尺度特征的表示和区分能力.

给定一幅输入图像 $X \in \mathbb{R}^{H\times W\times C}$，其中，$H$，$W$ 和 C 分别表示图像的高度、宽度和通道数. 假设 $\left\{\boldsymbol{F}^i \in \mathbb{R}^{H_i\times W_i\times C_i}, i=1,2,\cdots,4\right\}$ 为第 i 层多尺度编码器中局部特征学习模块得到的特征图，$[H/2,W/2,64]$，$[H/4,W/4,128]$，$[H/8,W/8,256]$，$[H/16,W/16,512]$ 分别是对应的特征图尺寸. 多尺度解码器与多尺度编码器类似，之后利用多尺度感知模块从局部角度提取多尺度信息. 对于第 i 层输入特征 $\boldsymbol{F}^i$，先将该特征转换为映射 $\boldsymbol{Q}^i$，$\boldsymbol{K}^i$ 和 $\boldsymbol{V}^i$，计算公式为

$$\begin{cases}\boldsymbol{Q}^i=\boldsymbol{F}^i W_1\\ \boldsymbol{K}^i,\ \boldsymbol{V}^i=\boldsymbol{F}^i W_2\end{cases} \tag{1}$$

其中，W_1 和 W_2 分别通过使用 Depthwise Conv[36] 3×3 和 Shared Depthwise Conv[36] 3×3 实现，它们都是将卷积的 groups 参数设置为输入通道数，从而实现对特征图的每个通道进行独立的卷积操作. 与 Depthwise Conv[36] 操作不同，Shared Depthwise Conv[36]操作在不同的位置上共享相同的卷积核. 再通过不同比率的自适应下采样对键 $\boldsymbol{K}^i$ 和值 $\boldsymbol{V}^i$ 实现多尺度的信息分析. 计算公式为

$$\begin{cases}\boldsymbol{K}_j^i=\varphi_j\left(\boldsymbol{K}^i\right)\\ \boldsymbol{V}_j^i=\varphi_j\left(\boldsymbol{V}^i\right)\end{cases} \tag{2}$$

其中，$j\in\{1,2,3,4,5\}$；φ_j 表示不同比率的自适应下采样操作，比率设置参考实验细节部分. 最后分别将得到的 $\boldsymbol{K}_j^i$ 和 $\boldsymbol{V}_j^i$ 进行拼接，获得包含多尺度信息的键 $\hat{\boldsymbol{K}}^i$ 和值 $\hat{\boldsymbol{V}}^i$.

为了互补地利用多尺度全局信息，在查询映射中引入一个可学习的全局 token，用 $\boldsymbol{\theta}_{\text{glo}}$ 表示，将其与 $\boldsymbol{Q}^i$ 拼接得到最终的查询为

$$\hat{\boldsymbol{Q}}^i=\text{Concat}\left(\boldsymbol{\theta}_{\text{glo}},\boldsymbol{Q}^i\right).$$

局部多尺度自注意机制的输出为

$$\hat{\boldsymbol{F}}^i=\text{Softmax}\left(\frac{\hat{\boldsymbol{Q}}^i\hat{\boldsymbol{K}}^{i\text{T}}}{\sqrt{d}}\right)\hat{\boldsymbol{V}}^i.$$

其中，d 表示 $\hat{\boldsymbol{K}}^i$ 信道维数，$\sqrt{d}$ 表示近似的归一化. 之后利用全局重调整模块对上述得到的特征 $\hat{\boldsymbol{F}}^i$ 进行特征增强. 先将 $\hat{\boldsymbol{F}}^i$ 按通道划分出全局特征 $\boldsymbol{F}_{\text{G}}^i$ 和局部注意力特征 $\boldsymbol{F}_{\text{L}}^i$，然后分别将 $\boldsymbol{F}_{\text{G}}^i$ 和 $\boldsymbol{F}_{\text{L}}^i$ 映射为 $\boldsymbol{q}^i$ 和 $\boldsymbol{k}^i,\boldsymbol{v}^i$，公式为

$$\begin{cases}\boldsymbol{q}^i=\boldsymbol{F}_{\text{G}}^iW_{\text{g}}\\ \boldsymbol{k}^i,\boldsymbol{v}^i=\boldsymbol{F}_{\text{L}}^iW_{\text{l}}\end{cases}.$$

其中，W_{g} 和 W_{l} 分别通过 Depthwise conv[36] 3×3 实现，将 $\boldsymbol{q}^i$ 和 $\boldsymbol{k}^i$ 进行矩阵相乘得到相似性权重；最后与 $\boldsymbol{v}^i$ 相乘得到调整后的多尺度自适应特征为

$$\boldsymbol{F}_{\text{A}}^i=\text{Softmax}\left(\frac{\boldsymbol{q}^i\boldsymbol{k}^{i\text{T}}}{\sqrt{d}}\right)\boldsymbol{v}^i.$$

其中，d 表示 $\boldsymbol{k}^i$ 的通道维数.

2.3 SDAM

虽然 LGMSPM 能够有效地学习多尺度信息，但不同层级间的语义差距依然存在，使分割性能受到一定限制. 受到去噪扩散模型的启发，本文方法引入 SDAM，以实现低级语义特征向高级语义特征的扩散传播和特征对齐. SDAM 结构如图 3 所示，该模块接收多尺度编码器中不同层级的特征，包括低级语义特征 $\boldsymbol{F}_{\text{A}}^{i-1}$ 和高级语义特征 $\boldsymbol{F}_{\text{A}}^i$ 作为输入. 首先通过语义扩散模块实现不同层级特征的语义对齐，得到融合后的 2 个特征 $\boldsymbol{F}_{\text{A}}^{i-1\prime}$ 和 $\boldsymbol{F}_{\text{A}}^{i\prime}$；然后将这 2 个融合特征与原始的低级语义特征 $\boldsymbol{F}_{\text{A}}^{i-1}$ 和经过双线性插值上采样的 $\boldsymbol{F}_{\text{A}}^{i\prime\prime}$ 进行自适应融合 (adapating fusion, AF)，得到最终增强的特征 $\boldsymbol{F}_{\text{E}}^i$.

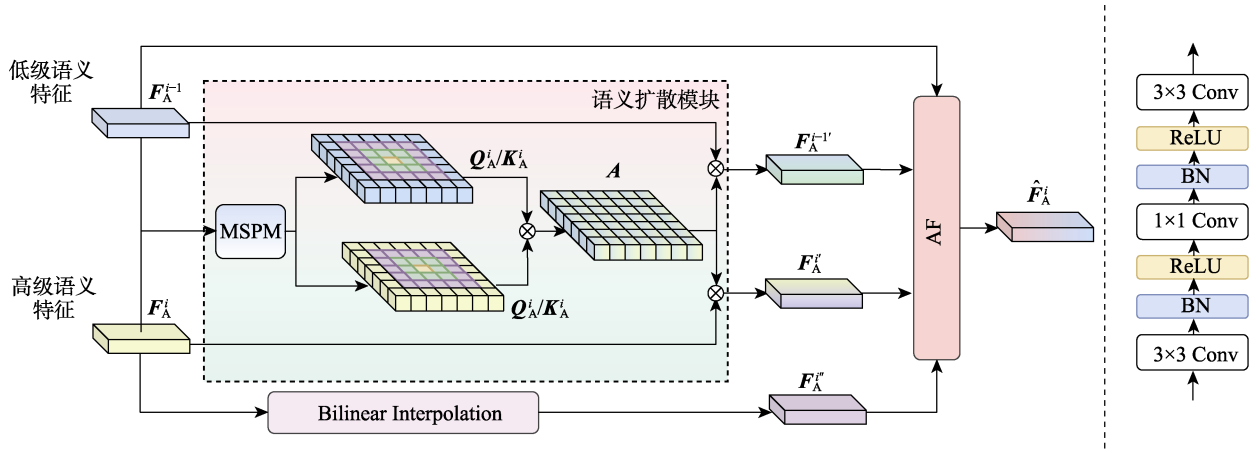


图 3 SDAM

给定相邻特征图 $\boldsymbol{F}_{\text{A}}^{i-1}$ 和 $\boldsymbol{F}_{\text{A}}^i$，先对 $\boldsymbol{F}_{\text{A}}^i$ 进行双线性插值上采样到与 $\boldsymbol{F}_{\text{A}}^{i-1}$ 一样的空间大小；然后通过 MSPM 分别感知其局部多尺度细节信息，得到对应 $\boldsymbol{Q}_{\text{A}}^{i-1},\boldsymbol{K}_{\text{A}}^{i-1}$ 和 $\boldsymbol{Q}_{\text{A}}^i,\boldsymbol{K}_{\text{A}}^i$，计算公式与式(1)(2)相同；再基于空间位置分析低级语义特征和高级语义特征的相似性，得到语义相似性矩阵

$$\boldsymbol{A}=\frac{\boldsymbol{Q}_{\text{A}}^{i-1}\left(\boldsymbol{K}_{\text{A}}^i\right)^{\text{T}}}{\sqrt{d}}.$$

其中，d 表示键 $\boldsymbol{K}_{\text{A}}^i$ 的通道维度. 由于矩阵 $\boldsymbol{A}$ 通过查询和键的点积来衡量语义相似度，为了简化计算，可以将 $\boldsymbol{A}^{\text{T}}$ 作为高级语义特征和低级语义特征间的相似性矩阵，得到相似度信息；然后根据语义相似性得到不同位置间的扩散关系，并与相应特征相乘，实现不同位置特征间的扩散，公式为

$$\boldsymbol{F}_{\text{A}}^{i-1\prime}=\text{Softmax}\left(\boldsymbol{A}\right)\times\boldsymbol{F}_{\text{A}}^{i-1}W_1,$$

$$\boldsymbol{F}_{\text{A}}^{i\prime}=\text{Softmax}\left(\boldsymbol{A}^{\text{T}}\right)\times\boldsymbol{F}_{\text{A}}^iW_2.$$

最后将扩散特征 $\boldsymbol{F}_{\text{A}}^{i-1\prime}$ 和 $\boldsymbol{F}_{\text{A}}^{i\prime}$ 与低级语义特征 $\boldsymbol{F}_{\text{A}}^{i-1}$ 和上采样特征 $\boldsymbol{F}_{\text{A}}^{i\prime\prime}$ 拼接在一起，作为 AF 网络的输入，得到最终的对齐特征为

$$\hat{\boldsymbol{F}}_{\text{A}}^i=\text{AF}\left(\text{Concat}\left(\boldsymbol{F}_{\text{A}}^{i-1},\boldsymbol{F}_{\text{A}}^{i-1},\boldsymbol{F}_{\text{A}}^{i\prime},\boldsymbol{F}_{\text{A}}^{i\prime\prime}\right)\right).$$

其中，AF 通过 2 个连续的 3×3 卷积，中间加一个

1×1 卷积实现. 最后将对齐特征 $\hat{\boldsymbol{F}}_{\mathrm{A}}^{i}$ 输入到多尺度解码器中恢复至原图分辨率大小. 多尺度解码器和多尺度编码器一样，由 LGMSPM 构成. 特征 $\hat{\boldsymbol{F}}_{\mathrm{A}}^{i}$ 先由局部特征学习模块提取局部信息，再进行多尺度信息分析，得到分割特征；最后经过 1×1 卷积得到像素级预测结果.

3　实验及结果分析

3.1　实验数据

本文在 2 个公开的医学图像数据集进行实验，包括腹部 CT 多器官 Synapse 数据集(简称 Synapse 数据集)[37]和心脏磁共振图像 ACDC 数据集(简称 ACDC 数据集)[38]. Synapse 数据集包含 30 个腹部 CT 扫描案例，共计 3 779 幅轴向临床 CT 图像；涉及 8 个腹部器官的分割，包括主动脉、胆囊、脾脏、左肾、右肾、肝脏、胰腺和胃；训练时，数据集被分为 18 个训练案例和 12 个测试案例. ACDC 数据集包含来自不同患者的 100 幅 MRI 扫描图像；每个病例都包括左心室，右心室和心肌的分割标签；数据集被分为 70 个训练样本、10 个验证样本和 20 个测试样本.

评估指标采用 95% Hausdorff 距离(Hausdorff distance at 95% quantile, HD95)和 Dice 相似系数(Dice similarity coefficient, DSC).

3.2　实验细节

模型架构设置主要是对 LGMSPM 中的池化设置，根据输入尺寸的不同而变化. 本文采用基于设定池化率的池化操作，编码器中，每个阶段池化设置分别为[12,16,20,24,56]，[6,8,10,12,18]，[3,4,5,6,14]和[1,2,3,4,7]；解码器与对应的编码器阶段设置类似.

针对实验超参数，本文使用 PyTorch 框架进行模型搭建，并在一块具有 12 GB 内存的单个 NVIDIA 2080 GPU 上进行训练. 输入图像的尺寸为 224×224 像素，总共进行 300 个训练轮次，每个批次数据设为 6；使用 Adam 优化器，设置学习率为 10^{-4}，动量参数为(0.900,0.999)，权重衰减为 10^{-4}. 为了展示网络训练情况，绘制了训练损失曲线，图 4 所示为在 Synapse 和 ACDC 数据集上进行 300 次迭代后的损失收敛情况. 可以看出，随着训练的进行，损失逐渐减少并最终收敛，表明模型在训练数据上获得了较好的拟合效果.

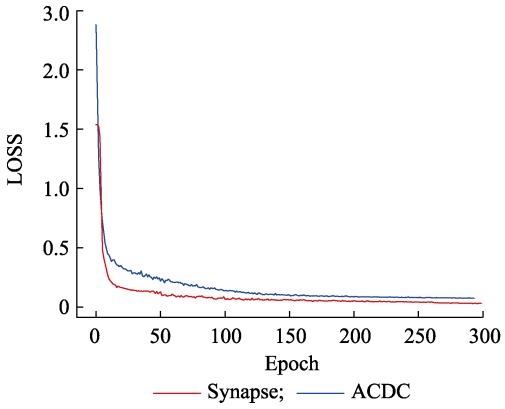


图 4　本文方法在不同数据集迭代训练损失曲线

3.3　消融实验

为了验证本文方法中所提出的 LGMSPM，SDAM, DS 策略和 AF 这 4 个要素对模型性能的影响，在 Synapse 数据集上进行消融实验，以 ResNet-50 残差块构建的 U-Net 模型作为 Baseline，并采用 DSC 和 HD95 作为评价指标. 消融实验结果如表 1 所示. 可以看出，LGMSPM 对模型性能的贡献最显著，与 Baseline 模型相比，在平均 DSC 提高 2.38 个百分点，平均 HD95 降低 7.65 个百分点，在一定程度上表明多尺度特征表示学习在分割任务中的重要性；添加了 SDAM 的模型性能也有所改善，在平均 DSC 上提高 0.60 个百分点，但平均 HD95 却增长了 1.87 个百分点，这是由于利用交叉注意力机制使模型更加关注器官语义特征，而模糊了边界；对于特征的融合，本文采用的 AF 方法的平均 DSC 稍微提高了性能，但在平均 HD95 上表现略差，这是由于多个特征融合使得边界信息

表 1　Synapse 数据集上各模块组合的消融实验结果

方法	模块				平均/%	
	LGMSP	SDAM	AF	DS	DSC ↑	HD95 ↓
Baseline					78.37	24.05
Baseline+LGMSPM	√				80.75	**16.40**
Baseline+LGMSPM+SDAM	√	√			81.35	18.27
Baseline+LGMSPM+SDAM+AF	√	√	√		81.69	19.65
Baseline+LGMSPM+SDAM+AF+DS	√	√	√	√	**82.42**	17.81

注. “↑”表示该指标越高越好，“↓”表示该指标越低越好；粗体表示最优值.

变得不明确; DS 策略的应用也带来了性能的提升, 平均 DSC 提高 0.73 个百分点, 平均 HD95 降低 1.84 个百分点. 因此, 消融实验结果进一步验证了本文方法中各模块的有效性.

3.4 对比实验

为了验证本文方法的有效性, 将其与一些主流分割方法进行实验对比, 包括 TransUNet[11], Att-UNet[39], Swin-Unet[40], UNETR[41], ST-Unet[42], HiFormer[43]等.

(1) 在 Synapse 数据集上进行多器官分割实验, 结果如表 2 所示. 可以看出, 与 U-Net[2]相比, 本文方法的平均 DSC 提高 5.57 个百分点, 平均 HD95 降低 21.89 个百分点, 尤其在分割大型或广泛分布的器官(如胃和胰腺)方面有显著改善; 与 CS-Unet[44]和 ConvWin-Unet[45]相比, 本文方法在平均 DSC 上分别提高 0.21 个百分点和 3.03 个百分点, 在平均 HD95 上, 分别降低 9.21 个百分点和 3.58 个百分点. 可以看出 本文方法在大多数器官的分割方面都有一定的改进, 表明多尺度信息感知有助于学习鉴别性区域并产生更好的分割结果.

表 2 不同方法在 Synapse 数据集上的实验结果比较 %

方法	DSC ↑									平均 HD95 ↓
	主动脉	胆囊	左肾	右肾	肝脏	胰腺	脾脏	胃	平均	
U-Net[2]	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58	76.85	39.70
TransUNet[11]	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62	77.48	31.69
UCTransNet[14]									79.12	26.75
MT-UNet[24]	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81	78.59	26.59
Att-UNet[39]	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75	77.77	36.02
Swin-Unet[40]	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60	79.13	21.55
UNETR[41]	**89.99**	60.56	85.66	**84.80**	94.46	59.25	87.81	73.99	79.56	22.97
ST-Unet[42]	85.68	69.05	**85.81**	73.04	95.13	60.23	89.15	72.78	78.86	20.37
HiFormer[43]	87.03	68.61	84.23	78.37	94.07	60.77	90.44	**82.03**	80.69	19.14
CS-Unet[44]	88.40	**72.59**	85.28	79.52	94.35	**70.12**	**91.06**	75.72	82.21	27.02
ConvWin-Unet[45]	84.37	69.89	84.85	82.08	93.69	58.84	87.32	74.04	79.39	21.39
本文	88.87	66.11	85.52	83.57	**95.34**	68.94	90.07	80.98	**82.42**	**17.81**

注. “↑”表示该指标越高越好, “↓”表示该指标越低越好; 粗体表示最优值.

(2) 为了更清晰地展示实验结果, 还绘制了箱线图, 图 5 所示为 11 种方法在 Synapse 数据集上的分割结果. 可以看出, 本文方法与 HiFormer[43]在各个器官的 DSC 分布较为集中, 优于大部分方法, 且本文方法上分位指标和下分位指标均高于 HiFormer[43], 表明其分割性能较好, 并且具有良好的稳定性和优越性.

(3) 在 ACDC 数据集上评估本文方法, 实验结果如表 3 所示. 可以看出, 本文方法优于所有对比方法; 与基于卷积网络的方法相比, 本文方法在平均 DSC 上比 U-Net[2]和 Att-UNet[39]分别提高 4.70 个百分点和 5.50 个百分点, 这种性能差异可以通过 LGMSPM 产生更丰富的多尺度特征来解释; 与近期的基于 Transformer 的方法, 如 MT-UNet[24]和 CS-Unet[44]相比, 本文方法在平均 DSC 上提高 1.82 个百分点和 0.88 个百分点, 再次证明了多尺度表示学习模块的有效性. 为了更直观地展示各个器官的分割情况, 还绘制了在 ACDC 数据集上的箱线图, 如图 6 所示. 可以看出, 基于卷积网络的模型, 如 U-Net[2]和 Att-UNet[39]在各个器官的 DSC 上分布较为分散, 相比之下, 本文方法和 CS-Unet[44]的 DSC 分布较为集中, 这也在一定程度上表明引入注

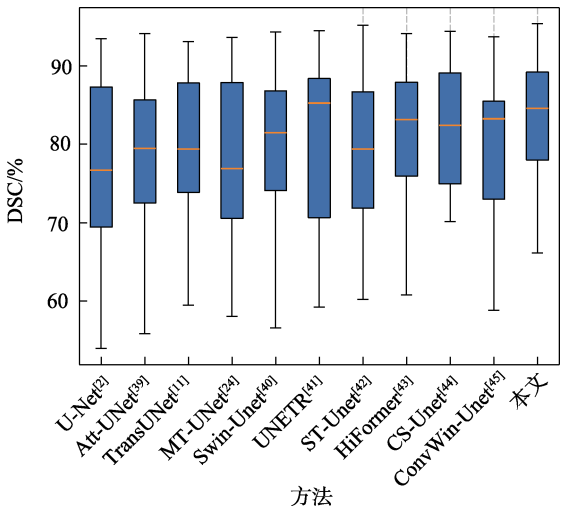


图 5 11 种方法在 Synapse 数据集上的分割结果

表 3　在 ACDC 数据集上的实验结果比较　%

方法	DSC ↑			
	右心室	心肌	左心室	平均
U-Net[2]	87.10	80.63	94.92	87.55
TransUNet[11]	88.86	84.53	95.73	89.71
UTNet[22]	88.20	83.50	93.10	88.30
MT-UNet[24]	86.64	89.04	95.62	90.43
Att-UNet[39]	87.58	79.20	93.47	86.75
Swin-Unet[40]	88.55	85.62	95.83	90.00
CS-Unet[44]	89.20	89.47	95.42	91.37
本文	**90.66**	**89.99**	**96.10**	**92.25**

注. “↑”表示该指标越高越好；粗体表示最优值.

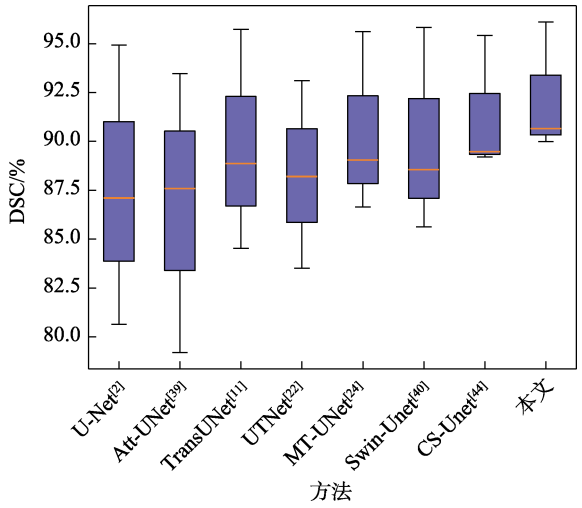


图 6　8 种方法在 ACDC 数据集上的分割结果

意力机制有助于提高分割性能；与其他方法相比，本文方法的下分位数指标较高，表明其能够实现更高的分割准确率，并且具有较好的鲁棒性.

3.5　方法复杂度分析

为了深入分析本文方法的性能，将其与一些主流方法的参数量和浮点运算数进行对比实验，结果如表 4 所示. 其中，浮点运算数是根据输入图像大小为 3×224×224 像素计算得到的运算次数. 可以看出，U-Net 具有最小的参数量和浮点运算数，分别为 19.62M 和 25.14G；基于 Transformer 的方法如 TransUNet[11]和 Swin-Unet[40]在参数量和浮点运算数上都较高，分别为 105.28M 和 56.7G, 83.3M 和 67.3G；相比之下，本文方法在参数量和浮点运算数上相对较低，分别为 56.47M 和 49.16G, 表明其能够实现较小的模型复杂度，同时达到较好的分割效果.

表 4　6 种方法复杂度比较

方法	参数量	浮点运算数
U-Net[2]	19.62M	25.14G
TransUNet[11]	105.28M	56.70G
UCTransNet[14]	66.43M	32.96G
MT-UNet[24]	79.08M	44.73G
Swin-Unet[40]	83.30M	67.30G
本文	56.47M	49.16G

3.6　可视化分析

为了进一步验证本文方法的有效性，在 2 个数据集上进行分割结果可视化比较. 图 7 所示为 7 种方法在 Synapse 数据集上的一些样例结果，通过在图中使用红色框突出显示，凸显了本文方法在某些区域的分割性能优于其他方法. 可以看出：第 1 行的示例中，大部分方法，如 U-Net[2]，UCTransNet[14]

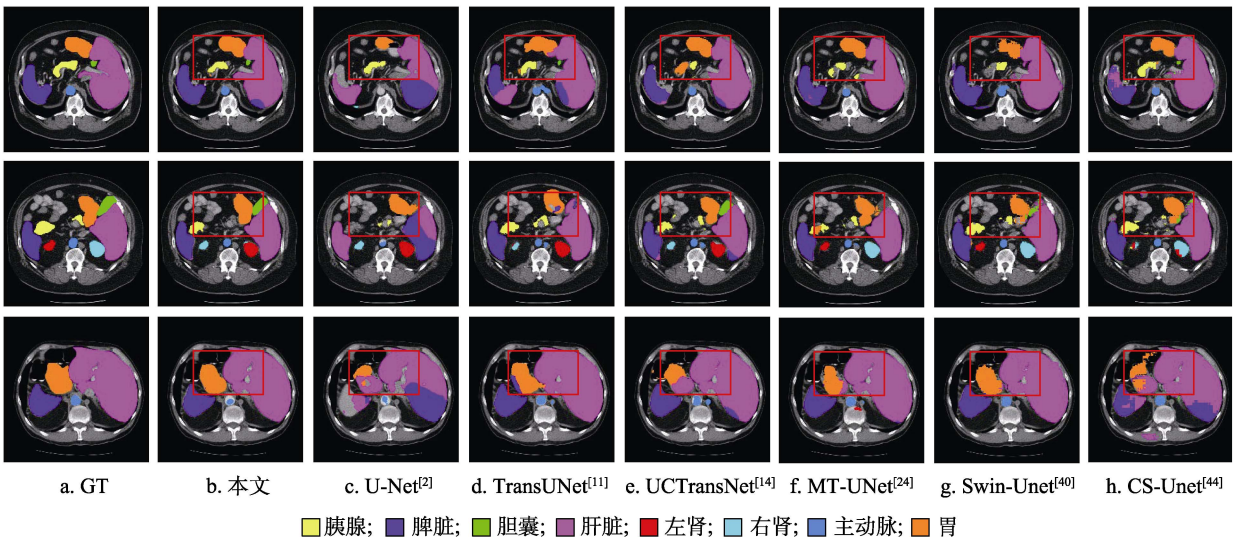


图 7　7 种方法在 Synapse 数据集上的可视化结果

和 MT-UNet[24]都忽略了对胆囊的分割，而 Swin-Unet[40]则将胆囊误分为肝脏；相比之下，本文方法能够准确地关注到胆囊并实现了良好的分割效果. 第 2 行示例中，与 U-Net[2], TransUNet[11], Swin-Unet[40]和 CS-Unet[44]等方法相比，本文方法能够更好地关注胃器官的整体形状信息，实现更准确的分割. 第 3 行示例中，U-Net[2], UCTransNet[14], MT-UNet[24]和 CS-Unet[44]等方法在肝脏和胃的分割中产生了混淆，而本文方法能够更好地区分它们. 这得益于本文方法中采用的 LGMSPM 能够更好地捕捉器官的形状，产生更准确的分割结果.

图 8 所示为 7 种方法在 ACDC 数据集上的分割结果可视化. 可以看出，第 1 行示例中，本文方法实现了较好的分割结果，而 MT-UNet[24]未能区分左心室、心肌和右心室的边界，导致分割结果不连续；第 2 行示例中，MT-UNet[24], Swin-Unet[40]和 CS-Unet[44]混淆了右心室和背景，导致分割不准确；本文方法能够更好地区分这些结构，实现更准确的分割.

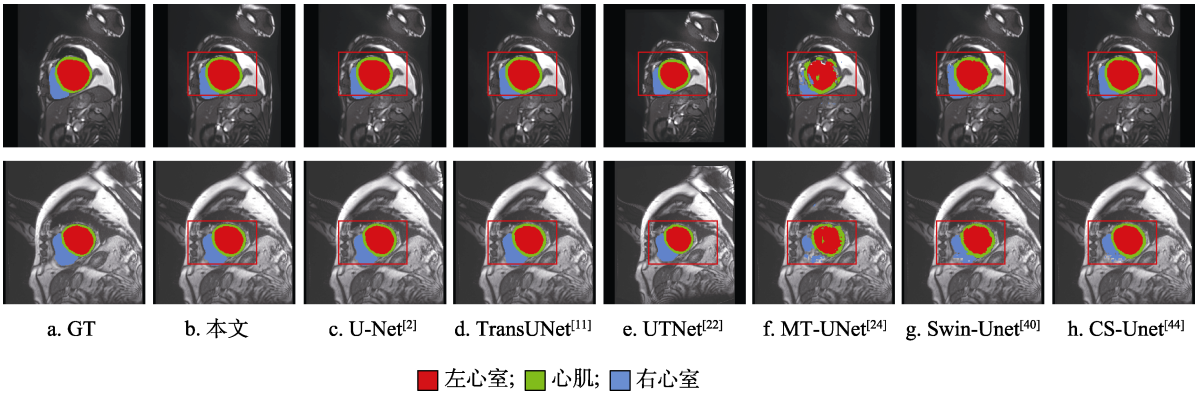


图 8 7 种方法在 ACDC 数据集上的可视化结果

通过这些可视化示例可以看出，本文方法在不同数据集上都展现出优越的分割性能，并且与其他方法相比，在特定情况下能够实现更准确的分割结果，进一步证明了其有效性和适用性.

4 结 语

本文提出一种语义扩散对齐的多尺度感知医学图像分割方法，旨在提高医学图像分割的准确性. 该方法在编码器和解码器的各个阶段引入 LGMSPM，不仅学习局部多尺度细节信息，还从全局角度进一步增强多尺度特征；并且引入了 SDAM，通过利用语义扩散思想，根据特征之间的空间关系分析相似性得到扩散关系，实现特征对齐. 在不同模态的数据集上进行大量实验的结果表明，本文方法展现出良好的性能和泛化能力：在 Synapse 数据集上，平均 DSC 达到 82.42%，平均 HD95 为 17.81%；在 ACDC 数据集上，平均 DSC 为 92.25%，优于大多数对比方法. 这些实验结果充分证明了本文方法的优越性. 但是，本文方法仍然存在一些不足之处：在计算复杂度方面相对较高，需要较长的训练时间和较高的计算资源；在边缘分割方面还不够准确. 未来将致力于研究轻量化的医学图像分割模型及边缘增强算法，探索更广泛的数据集和应用场景.

参考文献(References):

[1] Lee H, Chen Y P P. Image based computer aided diagnosis system for cancer detection[J]. Expert Systems with Applications, 2015, 42(12): 5356-5365

[2] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C] //Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2015: 234-241

[3] Wang R S, Lei T, Cui R X, *et al*. Medical image segmentation using deep learning: a survey[J]. IET Image Processing, 2022, 16(5): 1243-1267

[4] Yin Xiaohang, Wang Yongcai, Li Deying. Suvery of medical image segmentation technology based on U-Net structure improvement[J]. Journal of Software, 2021, 32(2): 519-550(in Chinese)
(殷晓航, 王永才, 李德英. 基于 U-Net 结构改进的医学影像分割技术综述[J]. 软件学报, 2021, 32(2): 519-550)

[5] Lee Y, Kim J, Willette J, *et al*. MPViT: multi-path vision transformer for dense prediction[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 7277-7286

[6] Zhang Rui, Li Jintao. A survey on algorithm research of scene

parsing based on deep learning[J]. Journal of Computer Research and Development, 2020, 57(4): 859-875(in Chinese)
(张蕊, 李锦涛. 基于深度学习的场景分割算法研究综述[J]. 计算机研究与发展, 2020, 57(4): 859-875)

[7] Zhao H S, Shi J P, Qi X J, *et al*. Pyramid scene parsing network[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 6230-6239

[8] Chen L C, Papandreou G, Kokkinos I, *et al*. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848

[9] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017

[10] Lin X, Yu L, Cheng K T, *et al*. The lighter the better: rethinking Transformers in medical image segmentation through adaptive pruning[J]. IEEE Transactions on Medical Imaging, 2023, 42(8): 2325-2337

[11] Chen J, Lu Y, Yu Q, *et al*. TransUNet: transformers make strong encoders for medical image segmentation[OL]. [2023-09-13]. https://arxiv.org/abs/2102.04306

[12] Zhang Z Z, Zhang W X. Pyramid medical transformer for medical image segmentation[J]. [2023-09-13]. https://arxiv.org/abs/2104.14702

[13] Li X T, Zhao H L, Han L, *et al*. Gated fully fusion for semantic segmentation[C] //Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 11418-11425

[14] Wang H N, Cao P, Wang J Q, *et al*. UCTransNet: rethinking the skip connections in U-Net from a channel-wise perspective with transformer[C] //Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022: 2441-2449

[15] Wang Y, Deng Z J, Hu X W, *et al*. Deep attentional features for prostate segmentation in ultrasound[C] //Proceedings of the 21st International Conference on Medical Image Computing and Computer Assisted Intervention. Heidelberg: Springer, 2018: 523-530

[16] Hu X W, Zhu L, Qin J, *et al*. Recurrently aggregating deep features for salient object detection[C] //Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 6943-6950

[17] Cai S J, Tian Y X, Lui H, *et al*. Dense-UNet: a novel multiphoton *in vivo* cellular image segmentation model based on a convolutional neural network[J]. Quantitative Imaging in Medicine and Surgery, 2020, 10(6): 1275-1285

[18] Huang G, Liu Z, Van Der Maaten L, *et al*. Densely connected convolutional networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2261-2269

[19] Al-Masni M A, Kim D H. CMM-Net: contextual multi-scale multi-level network for efficient biomedical image segmentation[J]. Scientific Reports, 2021, 11(1): Article No.10191

[20] Li X C, Song J Q, Jiao W Z, *et al*. MINet: multi-scale input network for fundus microvascular segmentation[J]. Computers in Biology and Medicine, 2023, 154: Article No.106608

[21] 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: Article No.4005615

[22] Liu Z, Lin Y T, Cao Y, *et al*. Swin transformer: hierarchical vision transformer using shifted windows[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 9992-10002

[23] Gao Y H, Zhou M, Metaxas D N. UTNet: a hybrid transformer architecture for medical image segmentation[C] //Proceedings of the 24th International Conference Medical Image Computing and Computer Assisted Intervention. Heidelberg: Springer, 2021: 61-71

[24] Wang H Y, Xie S A, Lin L F, *et al*. Mixed transformer U-Net for medical image segmentation[C] //Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Los Alamitos: IEEE Computer Society Press, 2022: 2390-2394

[25] Wang J L, Zhang H X, Yi Z. CCTrans: improving medical image segmentation with contoured convolutional transformer network[J]. Mathematics, 2023, 11(9): Article No.2082

[26] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3431-3440

[27] Zhou Z W, Rahman Siddiquee M M, Tajbakhsh N, *et al*. UNet++: a nested U-Net architecture for medical image segmentation[C] //Proceedings of the 4th International Workshop, and 8th International Workshop on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Heidelberg: Springer, 2018: 3-11

[28] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: transformers for image recognition at scale[OL]. [2023-09-13]. https://arxiv.org/abs/2010.11929

[29] Li Y J, Cai W T, Gao Y, *et al*. More than encoder: introducing transformer decoder to upsample[C] //Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. Los Alamitos: IEEE Computer Society Press, 2022: 1597-1602

[30] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2020: Article No.574

[31] Graikos A, Malkin N, Jojic N, *et al*. Diffusion models as plug-and-play priors[C] //Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2022: Article No.1070

[32] Baranchuk D, Rubachev I, Voynov A, *et al*. Label-efficient semantic segmentation with diffusion models[OL]. [2023-09-13]. https://arxiv.org/abs/2112.03126

[33] Brempong E A, Kornblith S, Chen T, *et al*. Denoising pretraining for semantic segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 4174-4185

[34] Amit T, Shaharbany T, Nachmani E, *et al*. SegDiff: image seg-

mentation with diffusion probabilistic models[OL]. [2023-09-13]. https://arxiv.org/abs/2112.00390

[35] Wolleb J, Sandkühler R, Bieder F, *et al*. Diffusion models for implicit image segmentation ensembles[C] //Proceedings of the 5th International Conference on Medical Imaging with Deep Learning. New York: PMLR, 2022: 1336-1348

[36] Howard A G, Zhu M L, Chen B, *et al*. MobileNets: efficient convolutional neural networks for mobile vision applications[OL]. [2023-09-13]. https://arxiv.org/abs/1704.04861

[37] Zhou Y Y, Li Z, Bai S, *et al*. Prior-aware neural network for partially-supervised multi-organ segmentation[OL]. [2023-09-13]. https://arxiv.org/abs/1904.06346

[38] Bernard O, Lalande A, Zotti C, *et al*. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?[J]. IEEE Transactions on Medical Imaging, 2018, 37(11): 2514-2525

[39] Oktay O, Schlemper J, Le Folgoc L, *et al*. Attention U-Net: learning where to look for the pancreas[OL]. [2023-09-13]. https://arxiv.org/abs/1804.03999

[40] Cao H, Wang Y Y, Chen J, *et al*. Swin-Unet: Unet-like pure transformer for medical image segmentation[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2022: 205-218

[41] Hatamizadeh A, Tang Y C, Nath V, *et al*. UNETR: transformers for 3D medical image segmentation[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2022: 1748-1758

[42] Zhang J, Qin Q G, Ye Q, *et al*. ST-Unet: swin transformer boosted U-Net with cross-layer feature enhancement for medical image segmentation[J]. Computers in Biology and Medicine, 2023, 153: Article No.106516

[43] Heidari M, Kazerouni A, Soltany M, *et al*. HiFormer: hierarchical multi-scale representations using transformers for medical image segmentation[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 6191-6201

[44] Liu Q Y, Kaul C, Wang J, *et al*. Optimizing vision transformers for medical image segmentation[C] //Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Los Alamitos: IEEE Computer Society Press, 2023: 1-5

[45] Feng X M, Wang T P, Yang X H, *et al*. ConvWin-UNet: UNet-like hierarchical vision Transformer combined with convolution for medical image segmentation[J]. Mathematical Biosciences and Engineering, 2023, 20(1): 128-144