引用格式：徐浩斌, 李振东, 武艺强, 等. 多种注意力机制的 AViT-UNet 高效医学影像分割方法 [J]. 电子科技大学学报, 2026, 55(3): 473-480.
XU H B, LI Z D, WU Y Q, et al. AViT-UNet: An efficient medical image segmentation method based on multiple attention mechanisms[J]. Journal of University of Electronic Science and Technology of China, 2026, 55(3): 473-480.

多种注意力机制的 AViT-UNet 高效医学影像分割方法

徐浩斌[1]，李振东[1,2*]，武艺强[1,2]，刘 昊[1,2]，李 帅[1,2]

（1. 宁夏大学信息工程学院，银川 750000；2. 宁夏东数西算人工智能与信息安全重点实验室，银川 750021）

摘要：针对现有的医学影像语义分割方法复杂度高、参数量大、精度低及无法在低配置及医院边缘部署设备等实际问题，提出了一种基于多种注意力机制的 vision transformer U-Net 型轻量级医学影像语义分割方法 AViT-UNet。首先设计了轻量化的卷积模块（LDB）并应用于编码–解码层的卷积模块，降低了模型的计算复杂度。其次，引入了自注意力机制模块 EMHA，在深层网络与瓶颈层进行应用，加强了分割精度。最后，针对跳跃连接与特征输入部分，网络使用通道注意力、空间注意力等机制，加强了残差连接与卷积深度，使分割结果更加精细。该方法有效弥补了 Transformer 的高计算量与卷积神经网络在捕获全局特征方面的不足，在轻量化网络的同时提高了语义分割的精度，使语义分割网络能够部署在配置有限的医疗设备和移动平台上。在 Synapse、GlaS 和 MoNuSeg 这 3 个公开医学影像语义分割基准数据集上进行多维度评测指标验证，结果证明了该方法具有一定的先进性和可行性。具体实现代码已上传至 https://github.com/shepherdxu/AViT-UNet。

关键词：语义分割；卷积神经网络；视觉 Transformer；注意力机制
中图分类号：TP391 **文献标志码：**A DOI: 10.12178/1001-0548.2025107

AViT-UNet: An efficient medical image segmentation method based on multiple attention mechanisms

XU Haobin[1], LI Zhendong[1,2*], WU Yiqiang[1,2], LIU Hao[1,2], and LI Shuai[1,2]

(1. School of Information Engineering, Ningxia University, Yinchuan 750000, China;
2. Ningxia Key Laboratory of Artificial Intelligence and Information Security for Channeling Computing Resources from the East to the West, Yinchuan 750021, China)

Abstract: Existing medical image semantic segmentation methods suffer from high computational complexity, large parameter counts, suboptimal accuracy, and inability to be deployed on low-resource and clinical edge devices. To address these issues, AViT-UNet, a lightweight vision transformer U-Net model incorporating multiple attention mechanisms is proposed to reduce model size and latency while maintaining competitive segmentation performance. Firstly, a lightweight convolutional module, lightweight dilated bottleneck (LDB), is designed in this model and applied to the convolutional module of the encoding-decoding layer, which significantly reduces the computational complexity of the model. Secondly, a self-attention mechanism module, efficient multi-head attention (EMHA), is invoked and applied in the deep network and bottleneck layer to enhance the segmentation accuracy. Finally, to enhance the fidelity of skip connections and feature fusion, the network integrates channel and spatial attention mechanisms to bolster residual pathways and deepen convolutional representations, yielding more precise segmentation outputs. This strategy effectively compensates the high computational demands of transformer-based models and the limited global receptive field of conventional convolutional neural networks. As a result, the proposed lightweight architecture achieves superior semantic segmentation accuracy while remaining suitable for deployment on resource-constrained medical devices and

收稿日期：2025−05−20
基金项目：宁夏回族自治区中央引导地方科技发展专项项目（2024FRD05050）；装备智能运用教育部重点实验室开放基金（AAIE-2023-0403）
作者简介：徐浩斌，主要从事人工智能、医学影像分析方面的研究。
*通信作者 E-mail: lizhendong@nxu.edu.cn

mobile platforms. The proposed method is validated on three publicly available medical image semantic segmentation benchmark datasets, Synapse, GlaS, and MoNuSeg, with multi-dimensional evaluation metrics. Experimental results fully prove that this method has a certain degree of advancement and feasibility. The specific implementation code of the method has been uploaded to https://github.com/shepherdxu/AViT-UNet.

Key words: semantic segmentation; convolutional neural networks; vision Transformer; attention mechanism

在现代医学领域中，医学影像是辅助医生诊断疾病和制定治疗方案的重要工具。随着深度学习技术的快速发展，基于卷积神经网络（convolutional neural networks, CNNs）的 U 型架构在语义分割领域取得了显著成功，并广泛应用于医学影像分割，在计算机辅助诊断和临床手术中发挥了关键作用[1-2]。文献 [3] 提出了关键的 U-Net 网络，该网络使用对称的编码器–解码器结构，对输入的特征图像进行降采样–上采样，并通过长跳跃连接以保留不同尺度的分辨率特征，实现高精度的语义分割，并以此衍生出诸多变种。文献 [4] 提出了 U-Net++，此方法在 U-Net 的基础上，引入更多的上采样节点和跳跃连接；文献 [5] 提出的 DeepLabV3+ 同样借鉴 U-Net 的思路，采用 Deeplabv3[6] 作为 encoder 模块；3D U-Net[7] 将网络应用于磁共振图像、CT 等 3D 图像，扩展了分割网络在医学领域的实践应用。这些模型通过优化 U-Net 实现了像素级别的精确分割，并在多种医学影像分割任务中表现出色。

然而，由于卷积操作的局部性[8]，其性能受限于有限的感受野和固有的归纳偏置[9]。因此，研究者开始关注自然语音处理（natural speech processing, NLP）领域的 Transformer 模型[10]。Transformer 凭借其长上下文捕捉能力，为计算机视觉领域提供了新的思路，由此诞生了 vision Transformer（ViT）[11]，弥补了 CNN 在长程依赖和全局信息建模上的不足。研究人员开始将 Transformer 引入 U-Net 网络，得益于 Transformer 的多头自注意力（multi-head self-attention）机制[10]，SwinUNet[12]、TransUNet[2] 等模型在医学影像分割中取得了良好效果，缓解了 CNN 的局限性。其中，文献 [2] 提出的 TransUNet 率先将 Transformer 结构用于医学图像分割工作的研究，该网络将重视全局信息的 Transformer 结构和底层图像特征的 CNN 一起进行混合编码，一定程度上提升 UNet 的分割效果。然而，Transformer 在视觉领域仍然存在局限性，其自注意力机制可能忽略低层特征（如边缘和纹理）[13]，而这些特征对精确分割物体边界至关重要[14]。因此 CNN-Transformer 模型被提出，如 HiFormer[13]，它将 CNN 的局部特征提取能力与 Transformer 的全局建模能力相结合，进一步提升了分割性能。至此，医学影像分割的技术已基本成熟。

值得注意的是，尽管 ViT 在视觉任务中表现出色，但其预训练需要大量的参数计算，基于 CNN 和 Transformer 的现有混合模型在计算复杂度和参数量上的高需求仍未得到充分优化[15-18]。TransUNet、DeepLabv3 和 Attn-UNet 等模型在 Synapse 医学影像语义分割数据集上的平均 DSC（dice similarity coefficient）约为 77%[19]，但计算复杂度（FLOPs）均超过 20 G，这种大量的资源需求限制了模型在医疗边缘设备（如嵌入式系统、移动设备或低端 GPU）上的部署，而这些边缘设备通常面临计算和内存的限制。因此，在保持语义分割精度的前提下，轻量化网络模型并降低模型的计算复杂度成为当下的研究热点。

受文献 [19] 启发，本文提出了一种高效的医院影像语义分割方法 AViT-UNet，该方法充分融合了 ViT 的长程依赖优势与 CNN 的局部捕捉能力，确保高精度语义分割；同时通过模型的轻量化设计降低了模型的计算复杂度，使模型能够在资源受限的边缘设备上部署。

1 方法描述

本文所提出的 AViT-UNet 方法整体框架如图 1 所示。模型采用了与 U-Net 相近的 U 型架构，由编码器、解码器、瓶颈层和跳跃连接 4 个部分构成。编码器与解码器部分主要由轻量化的扩张瓶颈模块（lightweight dilated bottleneck, LDB）和高效自注意力模块（efficient multi-head attention, EMHA）[20] 组成。模型前两层以卷积为主用于提取图像的局部特征，如图 2a 所示；后两层为全局与局部结合的注意力机制模块，同时捕获图像全局和局部信息，如图 2b 所示。输入图像首先经过输入特征提取模块（input feature module, IFM）进行初步处理，如图 3 所示。AViT-UNet 以每层 2 倍的降采样率降采样到瓶颈层，瓶颈层中的自注意力模块会接收降采样后的特征图，并将其传递给与编码器结构对称的解码器。解码器会接收来自 3 个不同维度的跳跃连

接特征图，逐层上采样后输出最终的分割结果。

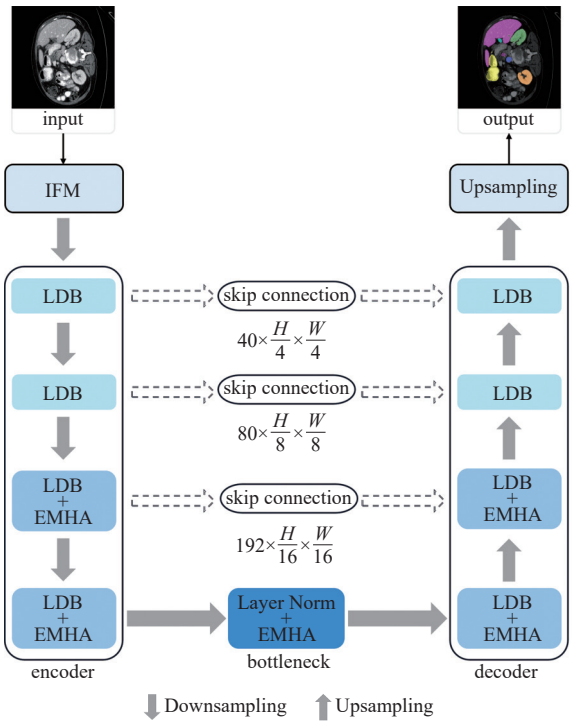


图 1　模型整体框架图

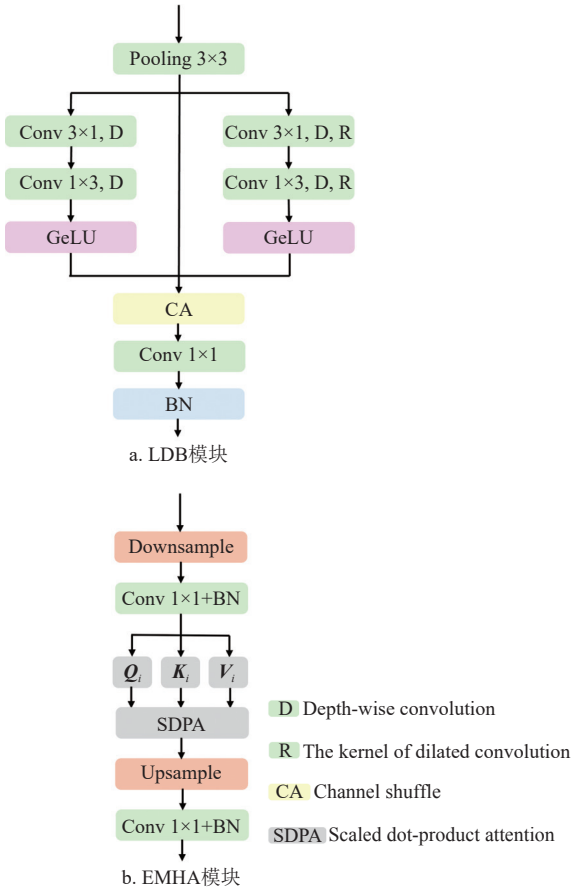


图 2　LDB 和 EMHA 模块设计示意图

1.1　输入特征模块

在 EViT-UNet[19] 以及 ResNet[21] 中，STEM 模块用于网络的起始部分，将输入的原始图像进行初步处理，其核心步骤经过两步 3×3 的卷积层，转换成为特征图并降低图像的分辨率，为后续的编码器提供了良好的初始特征[18]。然而，作为初始的输入模块，3×3 的卷积与两步归一化处理（batch normalization, BN）在通道数较大时会增加计算成本，同时连续两次下采样使图像空间分辨率快速缩小，导致在开始即丢失边缘细节信息，不利于高精度的语义分割。因此，在初始的输入特征模块中，本文将原始卷积替换为深度可分离卷积（depthwise separable convolution），如图 3 所示，减少了参数量，并引入了残差连接，在深度不变的情况下增强了特征传递。

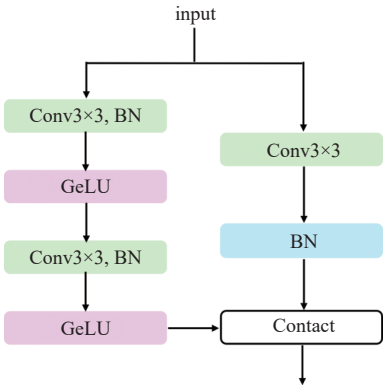


图 3　输入特征模块

1.2　上下采样模块

1.2.1　LDB 模块

受 EViT-UNet[19] 以及 ResNet[21] 的启发，本文设计了一种轻量化的扩张瓶颈模块 LDB，如图 2a 所示，LDB 的结构包括池化层、并行卷积路径、通道注意力机制。LDB 模块以一个 3×3 的平均池化操作作为起点，在编码过程中的池化操作可以有效扩大网络的感受野，同时降低图像的分辨率，有助于捕捉影像中器官的整体轮廓及病灶分布。

然而，降采样可能会导致精度降低，进而导致细节信息（如组织边缘或小结构）的丢失，而这些细节对于精确分割至关重要[21]。为此，在池化操作后，特征图像被送入两个不同的卷积分支，图 2a 中左侧部分负责局部特征的提取，图 2a 中右侧则采用膨胀卷积以及残差连接，减弱了特征提取中的信息丢失。两分支均添加了 3×1 与 1×3 的分解卷积，与 3×3 的常规卷积相比，分解卷积拥有

更加广阔的感受野和更少的参数量，有助于轻量化计算。

对于 CNN，大部分的特征信息都集中在通道中，不同的特征通道可能对应不同的信息类型（如边缘、纹理或语义区域等）[8-9]。为此，在模块中引入通道注意力机制（channel attention, CA）[22]，两个并行分支的输出特征图（通过拼接）被送入通道注意力机制 CA 中，CA 通过学习通道间的权重动态调整各通道的重要性。在通过一个恢复网络维度的 1×1 卷积以及 BN 后，LDB 模块将特征图输出。完整的数学表达如下：

$$X_{\text{left}} = \text{GeLU}(\text{Conv}_{1\times3,\,D}(\text{Conv}_{3\times1,\,D}(X_{\text{pool}}))) \tag{1}$$

$$X_{\text{right}} = \text{GeLU}(\text{Conv}_{1\times3,\,D,\,R}(\text{Conv}_{3\times1,\,D,\,R}(X_{\text{pool}}))) \tag{2}$$

$$X_{\text{CA}} = X \odot \sigma\left(\begin{matrix}\text{Unsqueeze}\\(\text{Conv}_{1\times1}(\text{Transpose}(\text{AVG}(X))))\end{matrix}\right) \tag{3}$$

$$\text{LDB}(X) = f_{\text{BN}}(f_{\text{Conv},\,1\times1}(f_{\text{CA}}(f_{\text{Concat}}(F_{\text{left}}, F_{\text{right}})))) \tag{4}$$

式中，left 和 right 分别对应深度可分离卷积的两个并行分支；⊙代表点积；AVG 代表全局平均池化。

1.2.2 EMHA 模块

本文为弥补 CNN 全局表示能力的局限性，在网络中设计了高效自注意力模块 EHMA 以实现长程依赖关系，如图 2b 所示。与 LDB 不同，EHMA 模块采用了以自注意力为主的全局特征捕捉模块，专注于捕捉图像的全局特征。

高分辨率下自注意力机制的训练与计算成本会急剧增加[11, 13, 15, 23]，EHMA 模块仅部署于网络深层的上下采样阶段，此设计不仅充分融合了全局与局部特征模块，同时降低了计算复杂度。EMHA 通过优化改进 ESRT[24] 自注意力模块，与原始 Transformer 占用的内存相比，ESRT 内存占用量减少了约 73.9%，且不降低原有效果[24]。同时，为了进一步轻量化，避免过大的内存占用和计算负载，EMHA 仅保留单个 ET（efficient transformer）作为胶囊网络置于网络中间。在 EHMA 中，查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$、值 $\boldsymbol{V}$ 矩阵被分割为 s 个片段，分别计算缩放点积注意力，生成 $O_1 \sim O_s$，再合并为整体输出 O。此方法借鉴了分组卷积思想，将大矩阵分解为小矩阵计算并整合，有效降低了资源消耗。经过以上优化，EHMA 模块在保持全局特征捕捉能力的同时，大幅减轻了计算负担[21]。缩放点积注意力的计算为：

$$O_i = \text{softmax}\left(\frac{Q_i K_i^{\text{T}}}{\sqrt{d}}\right) V_i \quad 1 \leqslant i \leqslant s \tag{5}$$

$$O = \text{Contact}(O_1, O_2, \cdots, O_s) \tag{6}$$

自注意力模块的核心计算可简化为：

$$\text{EMHA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}(\boldsymbol{Q}\boldsymbol{K}^{\text{T}} + ab) \cdot \boldsymbol{V} \tag{7}$$

式中，ab 是位置编码的可学习注意力偏差。

1.3 跳跃连接

在经典的 U-Net 网络中，跳跃连接通过融合编码器与解码器对应层级的特征图，弥补了在下采样和池化过程中的细节损失[3]。为了进一步提升 AViT-UNet 的分割性能，在跳跃连接阶段引入了卷积块注意力模块[25]。该模块是一种轻量级的注意力模块，融合了通道注意力和空间注意力。通道注意力机制通过对特征图的每个通道进行加权，突出对分割任务更重要的通道信息。该机制首先对输入特征图进行全局平均池化和全局最大池化，然后通过共享的多层感知机（multi layer perceptron, MLP）学习通道注意力权重，最后将权重应用到原始特征图的每个通道上。空间注意力机制则关注特征图的不同空间区域的重要性。空间注意力首先在通道维度上进行平均池化和最大池化，得到两个单通道的特征图，然后将它们拼接并通过一个卷积层学习空间注意力权重，最后将权重应用到原始特征图的每个空间位置上。两种注意力机制相结合，能够自适应地调整不同通道和空间位置的权重，从而更有效地融合来自编码器和解码器的特征。对应的注意力模块操作如下：

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \tag{8}$$

$$M_s(F') = \sigma(\text{Conv}([\text{AvgPool}(F'); \text{MaxPool}(F')])) \tag{9}$$

式中，F 和 F'分别表示特征图。

2 实验与结果

2.1 数据集

本文方法 AViT-UNet 在 Synapse、GlaS 和 MoNuSeg 这 3 个公开的医学影像分割基准测试数据集上进行了有效的验证。

Synapse 多器官分割数据集[26] 是一个用于深度学习测试的公开医学图像数据集，多用于语义分割的测试与评估。该数据集包含 30 例腹部 CT 扫描，并标注了 8 个重要结构，包括主动脉、胆囊、脾脏、左肾、右肾、肝脏、胰腺和胃，用于训练多器官分割模型。Synapse 数据集的评价指标为 DSC

（dice similarity coefficient），具体计算方式为：

$$\mathrm{DSC}=\frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{10}$$

式中，TP（true positive）为预测和标签均为当前类别的体素数；FP（false positive）为预测为当前类别但标签为其他类别的体素数；FN（false negative）为预测为其他类别但标签为当前类别的体素数。

GlaS（gland segmentation in colon histology images）[27] 数据集是一组用于结直肠腺体组织的分割任务数据集，目标是开发算法以自动分割组织学图像中的腺体结构。该数据集包括 165 张图像，这些图像来源于 16 个 H&E 染色的组织学切片，这些切片代表了 T3 或 T4 阶段的结直肠腺癌病例。每个切片来自不同的患者，并且切片在实验室中处理的时间也不相同。因此，这个数据集在染色分布和组织结构方面展现出高度的人与人之间的变异性。

MoNuSeg（multi-organ nuclei segmentation）[28] 数据集包含来自 44 个不同类型的 30 000 多个手动标注的细胞核，并对每个细胞核进行了像素级别的精确标注。该数据集是通过从 TCGA 档案中以 40 倍放大倍率下载并捕获的 H&E 染色组织图像创建的。

2.2　实验环境

本文的实验环境为 Python3.10 及 PyTorch2.0.1，输入图像尺寸大小为 3×224×224，批大小（batch_size）和训练周期（Epoch）分别为 32 和 70，迭代轮数为 300 轮，训练和测试显卡为单张 Nvidia V100 GPU。在 3 个基准数据集上优化器均采用 SGD 及其反向传播算法，在 GlaS 和 MoNuSeg 数据集的测试集上均采用了 5 折交叉方法进行实验验证。

2.3　实验结果对比

表 1 和表 2 为本文方法 AViT-UNet 在 3 个基准数据集上的分割结果对比。

表 1　AViT-Net 在 Synapse 数据集上的表现　　%

Methods	评价指标	重要结构							
	Average DSC	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
R50 U-Net[23]	74.68	87.74	63.66	80.60	78.19	93.74	56.90	85.87	74.16
U-Net[3]	76.85	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
R50 Att-U-Net[23]	75.57	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
Att-U-Net[23]	77.77	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
TransUNet[2]	77.48	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SwinUNet[12]	79.13	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
LeViT-UNet[29]	78.53	78.53	62.23	84.61	80.25	93.11	59.07	88.86	72.76
DeepLabv3[6]	77.63	88.04	66.51	82.76	74.21	91.23	58.32	87.43	73.53
HiFormer[13]	80.29	85.63	**73.29**	82.39	64.84	94.22	60.84	91.03	78.07
SelfReg + U-Net[3, 15]	80.34	88.74	71.78	85.32	80.71	93.80	62.22	84.78	75.39
SelfReg + SwinUNet[12, 15]	80.54	86.07	69.65	85.12	82.58	94.18	61.08	87.42	78.22
EViT-UNet[19]	80.87	87.13	66.53	**85.45**	**83.14**	94.92	62.92	89.66	77.18
本文方法	**82.01**	87.74	69.35	85.19	80.93	**95.17**	**65.01**	**91.67**	**81.04**

表 2　在 Glas 及 MoNuSeg 数据集上的表现　　%

Method	GlaS		MoNuSeg	
	DSC	IOU	DSC	IOU
U-Net[3]	85.45±1.25	74.78±1.67	76.45±2.62	62.86±3.00
UNet++[4]	87.56±1.17	79.13±1.70	77.01±2.10	63.04±2.54
AttUNet[23]	88.80±1.07	80.69±1.66	76.67±1.06	63.47±1.16
MRUNet[30]	88.73±1.17	80.89±1.67	78.22±2.47	64.83±2.87
TransUNet[2]	88.40±0.74	80.40±1.04	78.53±1.06	65.05±1.28
MedT[31]	85.92±2.93	75.47±3.46	77.46±2.38	63.37±3.11
SwinUNet[12]	89.58±0.57	82.06±0.73	77.69±0.94	63.77±1.15
UCTransNet[32]	90.18±0.71	82.96±1.06	79.08±0.67	65.50±0.91
SelfReg+SwinUNet[12, 15]	91.62±0.16	85.29±0.30	79.38±0.15	65.87±0.20
EViT-UNet[19]	92.44±0.23	86.50±0.38	79.27±0.24	65.87±0.21
本文方法	**92.82±0.20**	**86.72±0.35**	**79.36±0.30**	**65.88±0.25**

表 1 展示了本文方法在 Synapse 数据集上与其他先进方法的实验结果，本文方法的 Average DSC

达到了 82.01%，优于表中的其他对比方法，相较于 EViT-UNet 算法性能提高了 1.14%，并在 Liver（95.17%）、Pancreas（65.01%）、Spleen（91.67%）和 Stomach（81.04%）单器官分类中达到对比表的最高准确度。同时，在 Aorta、Gallbladder 等单结构分类任务中，AViT-UNet 的准确度均优于对比方法。

Synapse 数据集的分割效果如图 4a 组可视化分割对比结果所示，AViT-UNet 在 Stomach、Pancreas 等大面积单个器官上效果显著，尤其对器官的模糊边界做出了精确分割。同时对 Gallbladder 这类低对比度器官分类，AViT-UNet 捕获了比同类方法更多的特征信息，使图片结果更加清晰直观，缓解了边缘模糊、结果丢失等分割缺陷[14]。

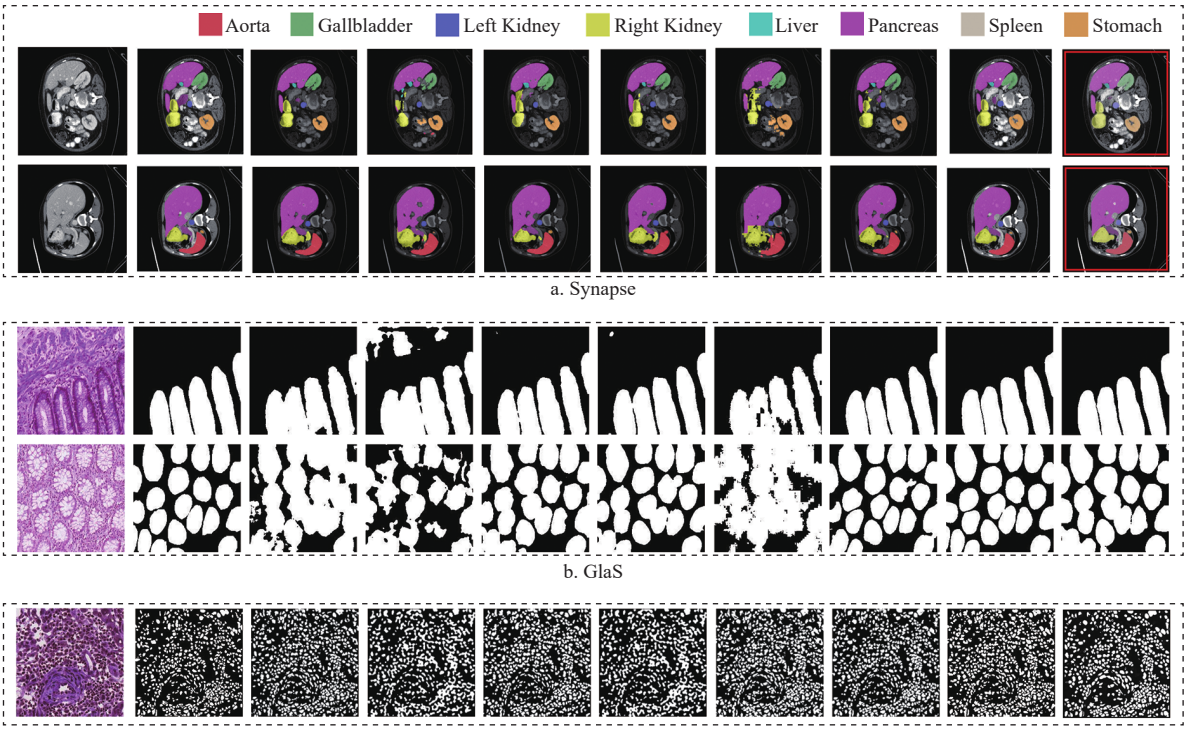


图 4 Synapse、GlaS 和 MuNuSeg 数据集的可视化分割结果对比

为了验证本文方法 AViT-UNet 的泛化能力和鲁棒性，在 Glas 和 MoNuSeg 数据集上均进行了实验并与其他先进算法进行了对比，对比结果及分割可视化结果分别如表 2 和图 4b、图 4c 分割可视化对比结果所示。从表 2 及图 4 的实验结果对比表明，本文方法在器官分割、腺体分割以及细胞分割领域均展现出了优异性能。

最重要的是，本文方法在提高网络性能的同时保持了极度轻量化。由图 5 所示，本文方法计算复杂度为 6.41 GMac，而分割精确度（Average DSC）达到了所有比较方法中的最优，实现了 57% per GMac 的计算性价比。实验表明，在 Snapdragon Gen2（45TOPS 算力）移动计算平台上理论帧率损失仅 0.4 帧/s，而每百万病例识别准确数增加约 10 000 例，达到了计算资源消耗和性能提升之间的最佳平衡，适合部署在边缘计算设备或移动端平台。

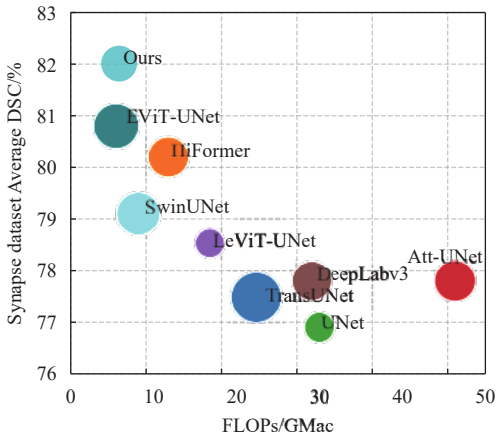


图 5 模型参数、计算复杂度与模型精度对比（圆圈大小代表不同方法模型的参数量大小）

2.4　消融实验

为了全面评估本文方法的各个模块与架构对网络性能产生的影响，进行了针对 Synapse 数据集的消融实验。作为参照标准，基线模型达到了 80.87% 的平均 Dice 相似系数（Average DSC）[17]，如表 3 所示。表 3 及图 6 表明，实施 4 种主要方法和模块的优化后，基线模型的准确度与精度均有了显著提高。在添加 LDB 模块后，胃部器官的分割精度有了显著的提升，但牺牲了部分需要全局信息的器官分割精度。为此，引入自注意力机制模块，将平均 DSC 提升至 81.11%，该模块的引入增强了模型对复杂结构（如胰腺和脾脏）的识别能力。同时，跳跃连接的改进进一步增强了类网络的特征融合能力，提升了胆囊与胰腺这类精细目标的分割结果。消融实验的结果表明，各个模块及方法的提出对分割模型均有着积极作用，使得 AViT-UNet 在 Synapse 数据集上的 Average DSC 达到了 82.01%。

表 3　基于 Synapse 多器官分割数据集的消融实验　　%

Methods	评价指标	重要结构							
	Average DSC	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
Baseline	80.87	87.13	66.53	85.45	83.14	94.92	62.92	89.66	77.1
+ LDB	80.90	87.98	66.69	82.22	79.41	94.81	64.29	90.38	81.41
+ Bottleneck	81.11	87.43	69.32	80.67	79.32	94.56	64.70	92.07	80.80
+ Skip Connection	81.60	88.34	69.91	83.45	82.64	94.88	66.01	89.15	78.45
+ EMHA	82.01	87.74	69.35	85.19	80.93	95.17	65.01	91.67	81.04

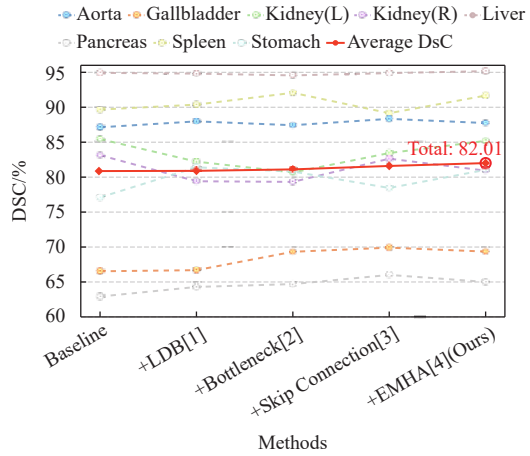


图 6　基于 Synapse 数据集的消融实验二维图

3　结束语

本文提出了一种基于 U-Net 架构的语义分割网络模型 AViT-UNet，用于资源受限和有高精度需求的医学影像语义分割场景。其中 LDB 模块与 EMHA 模块能够全面捕捉特征图像的上下文信息，跳跃连接与瓶颈层的改进，进一步提升了网络精度。实验对比结果表明，本文方法在多种医学影像语义分割基准数据集上取得了良好的效果，优于其他先进的语义分割方法，具有出色的性能和泛化能力。在保证精确度的同时，本研究对方法进行了轻量化设计，降低了计算复杂度，使其在医院的小型设备和移动设备上部署成为可能，实现了计算复杂度与网络性能之间的平衡。

参 考 文 献

[1] 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. [S.l.]: IEEE, 2022: 10.1109/wacv51458.2022.00181.

[2] CHEN J, LU Y, YU Q, et al. TransUNet: Transformers make strong encoders for medical image segmentation[EB/OL]. [2025-03-11]. https://arxiv.org/pdf/2102.04306.

[3] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing, 2015: 234-241.

[4] ZHOU Z W, RAHMAN SIDDIQUEE M M, TAJBAKHSH N, et al. UNet++: A nested U-Net architecture for medical image segmentation[C]//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Cham: Springer International Publishing, 2018: 3-11.

[5] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 833-851.

[6] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. [2025-03-11]. https://arxiv.org/pdf/1706.05587.

[7] ÇIÇEK Ö, ABDULKADIR A, LIENKAMP S S, et al. 3D U-Net: Learning dense volumetric segmentation from sparse annotation[C]//Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Cham: Springer International Publishing, 2016: 424-432.

[8] JIN Q, MENG Z, SUN C, et al. RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT

scans[J]. Front Bioeng Biotechnol, 2020, 8: 605132.
[9] AZAD R, FAYJIE A R, KAUFFMANN C, et al. On the texture bias for few-shot CNN segmentation[EB/OL]. [2025-03-11]. https://arxiv.org/pdf/2003.04052.
[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30. [S.l.]: Curran Associates, 2017: 5998-6008.
[11] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. [2025-03-11]. https://arxiv.org/pdf/2010.11929.
[12] CAO H, WANG Y Y, CHEN J, et al. Swin-unet: Unet-like pure transformer forMedical image segmentation[C]// Computer Vision – ECCV 2022 Workshops. Cham: Springer Nature Switzerland, 2023: 205-218.
[13] 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. Piscataway: IEEE, 2023: 6202-6212.
[14] 王雪, 李占山, 吕颖达. 基于多尺度感知和语义适配的医学图像分割算法[J]. 吉林大学学报（工学版）, 2022, 52(3): 640-647.
WANG X, LI Z S, LYU Y D. Medical image segmentation based on multi-scale context-aware and semantic adaptor[J]. Journal of Jilin University (Engineering and Technology Edition), 2022, 52(3): 640-647.
[15] ZHU W H, CHEN X W, QIU P J, et al. SelfReg-UNet: Self-regularized UNet for medical image segmentation [C]//Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Cham: Springer Nature Switzerland, 2024: 601-611.
[16] DALVI F, SAJJAD H, DURRANI N, et al. Analyzing redundancy in pretrained transformer models[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2020: 4908-4926.
[17] LI H, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient ConvNets[EB/OL]. [2025-03-11]. https://arxiv.org/pdf/1608.08710.
[18] LI L J. Self-regulated feature learning via teacher-free feature distillation[C]//Computer Vision – ECCV 2022. Cham: Springer Nature Switzerland, 2022: 347-363.
[19] LI X, ZHU W H, DONG X Z, et al. EViT-UNET: U-Net like efficient vision transformer for medical image segmentation on mobile and edge devices[C]//Proceedings of the IEEE 22nd International Symposium on Biomedical Imaging. [S.l.]: IEEE, 2025: 10981108.
[20] XU G A, LI J C, GAO G W, et al. Lightweight real-time semantic segmentation network with efficient transformer and CNN[J]. IEEE Transactions on Intelligent Transportation Systems, 24(12): 15897-15906.
[21] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 770-778.
[22] WANG Q L, WU B G, ZHU P F, et al. ECA-net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 11531-11539.
[23] OKTAY O, SCHLEMPER J, FOLGOC L L, et al. Attention U-Net: Learning where to look for the pancreas [EB/OL]. [2025-03-11]. https://arxiv.org/pdf/1804.03999.
[24] LU Z, LI J, LIU H, et al. Supplementary material of transformer for single image super-resolution[EB/OL]. [2025-03-11]. https://openaccess.thecvf.com/content/CVPR2022W/NTIRE/supplemental/Lu_Transformer_for_Single_CVPRW_2022_supplemental.pdf.
[25] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional Block attention module[C]//Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.
[26] LANDMAN B, XU Z, IGELSIAS J, et al. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge[EB/OL]. [2025-03-12]. https://www.synapse.org/Synapse:syn3193805.
[27] SIRINUKUNWATTANA K, PLUIM J P W, CHEN H, et al. Gland segmentation in colon histology images: The glas challenge contest[J]. Medical Image Analysis, 2017, 35: 489-502.
[28] KUMAR N, VERMA R, SHARMA S, et al. A dataset and a technique for generalized nuclear segmentation for computational pathology[J]. IEEE Trans Med Imaging, 2017, 36(7): 1550-1560.
[29] XU G P, ZHANG X, HE X W, et al. LeViT-UNet: Make faster encoders withTransformer forMedical image segmentation[C]//Pattern Recognition and Computer Vision. Singapore: Springer Nature Singapore, 2024: 42-53.
[30] DING H W, CUI X H, CHEN L Y, et al. MRU-NET: A U-shaped network for retinal vessel segmentation[J]. Applied Sciences, 2020, 10(19): 6823.
[31] QI Q F, LIN L Y, ZHANG R, et al. MEDT: Using multimodal encoding-decoding network as in transformer for multimodal sentiment analysis[J]. IEEE Access, 2022, 10: 28750-28759.
[32] 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[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 2441-2449.

责任编辑 税 红