智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于分组注意力并行编码的医学图像分割网络

孟祥福, 杨雨卓, 张霄雁, 李帅

引用本文:
孟祥福, 杨雨卓, 张霄雁, 等. 基于分组注意力并行编码的医学图像分割网络[J]. 智能系统学报, 2026, 21(5): 1194-1210.
MENG Xiangfu, YANG Yuzhuo, ZHANG Xiaoyan, et al. Medical image segmentation network based on group attention parallel encoding[J]. CAAI transactions on intelligent systems, 2026, 21(5): 1194-1210.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202511010

DOI: 10.11992/tis.202511010
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20260707.0932.002

基于分组注意力并行编码的医学图像分割网络

孟祥福[1], 杨雨卓[1], 张霄雁[1], 李帅[2]
(1. 辽宁工程技术大学 电子与信息工程学院, 辽宁 葫芦岛 125105; 2. 辽宁省健康产业集团阜新矿总医院, 辽宁 阜新 123000)

摘 要: 针对医学图像分割任务, 传统单分支卷积神经网络 (convolutional neural network, CNN) 架构因感受野限制难以有效整合局部细节与全局语义, 导致多尺度结构建模不足及模态泛化能力弱。为此提出了一种分组注意力并行融合编码的分割网络, 采用 Transformer 和 CNN 并行双分支编码器, 分别提取图像的全局语义信息与局部细节信息。编码器后设计分组注意力并行融合模块, 解决了局部与全局语义割裂及多尺度结构冲突问题; 解码器中提出的空间−通道双注意力门控模块, 能够有效解决医学图像分割中的特征干扰问题, 提升分割精度。在 Synapse、ACDC 和 AVT 等数据集上开展了大量实验, Dice 相似系数分别为 84.86%、91.66% 和 88.79%, 95% 豪斯多夫距离分别为 13.54、1.20 和 4.02 mm。实验结果表明, 与现有主流医学图像分割模型相比, 所提方法在分割任务中表现出更高的分割精度和鲁棒性, 尤其在处理复杂解剖结构的器官数据时优势显著。

关键词: 医学图像分割; 双分支编码器; 分组注意力; 特征提取; Transformer; 卷积神经网络; 并行融合; 空间−通道解码
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 1673−4785(2026)05−1194−17

中文引用格式: 孟祥福, 杨雨卓, 张霄雁, 等. 基于分组注意力并行编码的医学图像分割网络 [J]. 智能系统学报, 2026, 21(5): 1194–1210.
英文引用格式: MENG Xiangfu, YANG Yuzhuo, ZHANG Xiaoyan, et al. Medical image segmentation network based on group attention parallel encoding[J]. CAAI transactions on intelligent systems, 2026, 21(5): 1194–1210.

Medical image segmentation network based on group attention parallel encoding

MENG Xiangfu[1], YANG Yuzhuo[1], ZHANG Xiaoyan[1], LI Shuai[2]
(1. School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China; 2. Liaoning Health Industry Group Fuxin Mine General Hospital, Fuxin 123000, China)

Abstract: For the task of medical image segmentation, traditional single-branch convolutional neural network (CNN) architectures, limited by their receptive fields, struggle to effectively integrate local details with global semantics. This results in insufficient modeling of multi-scale structures and weak generalization across modalities. To address these limitations, we propose a novel segmentation network with group attention parallel fusion encoding. It employs a parallel dual-branch encoder combining Transformer and CNN to extract global semantic information and local details from images, respectively. The group attention parallel fusion module following the encoder resolves the disconnection between local and global semantics as well as conflicts in multi-scale structures. Additionally, the spatial-channel dual-attention gating module in the decoder effectively tackles feature interference in medical image segmentation, thereby enhancing segmentation accuracy. Extensive experiments were conducted on the Synapse, ACDC and AVT datasets, yielding dice similarity coefficient (DSC) of 84.86%, 91.66%, and 88.79%, and 95% hausdorff distance (HD95) of 13.54, 1.20, and 4.02 mm, respectively. The results demonstrate that compared with existing major medical image segmentation models, our approach achieves higher segmentation accuracy and robustness, especially when dealing with complex anatomical structures of organ data.

Keywords: medical image segmentation; dual-branch encoder; group attention; feature extraction; Transformer; CNN; parallel fusion; spatial-channel decoding

收稿日期: 2025−11−06. **网络出版日期**: 2026−07−07.
基金项目: 辽宁省自然基金联合计划项目 (20240311).
通信作者: 孟祥福. E-mail: marxi@126.com.

医学图像分割作为计算机辅助诊断的关键环节, 对器官定量分析、手术规划和病灶筛查具有

重要临床价值。尽管当前主流分割架构取得显著进步, 但医学图像分割仍存在以下 3 方面关键局限。其一, 局部与全局语义割裂, 传统基于卷积神经网络 (convolutional neural network, CNN) 的模型 (如 U-Net[1]、UNet++[2]、ResUNet[3]) 受限于卷积核的局部感受野, 如 3×3 卷积仅覆盖 5×5 像素邻域, 难以建模器官间的长距离解剖约束[4], 典型表现为无法编码主动脉弓与肾动脉间严格的纵向位移约束及血流拓扑次序。其二, 多尺度结构冲突, 传统单一路径特征聚合策略无法兼顾微小目标的差异化建模需求[5-6], 如胰腺导管或血管末梢与大型腔室如心脏心室或肝脏实质之间的差异。其三, 模态泛化能力薄弱, 不同影像模态 (如 CT (computed tomography)、MRI (magnetic resonance imaging)) 的物理成像特性差异显著, MRI 的软组织高对比度与 CT 的骨结构显影特性存在显著差异, 导致模型跨模态迁移时性能不稳定。

为同时捕获局部细节与全局语义, 近期研究工作提出了 Transformer-CNN 混合架构[7]。其中串行方案 (如 TransUNet[8]) 通过 CNN 提取低级特征后输入 Transformer 编码器, 但多阶段串联会引发梯度消失的风险; 而并行方案[9] 中 CNN 分支输出的高分辨率局部特征图与 Transformer 分支提取的低维全局语义仅通过简单的通道拼接 (Concat) 或逐元素相加 (Add) 聚合, 这种操作本质上是线性变换, 会导致空间错位和语义失准, 具体表现为 CNN 分支敏感的器官边缘细节被全局语义淹没, Transformer 建模的长程拓扑无法精确投射至像素空间。从大量模型的实验结果来看, 在 Synapse 多器官数据集上, 胰腺的 Dice 相似系数 (dice similarity coefficient, DSC) 为 62.85%, 远低于肝脏的 94.71%, 这一性能差异源于 CNN 与 Transformer 融合方式对微小器官的抑制。

针对上述问题, 本文提出了一种新的网络架构, 采用 Transformer-CNN 双分支编码器分别提取全局与局部特征, 并通过分组注意力并行融合模块整合 4 组异构通路 (点、局部、中程、全局注意力), 实现多尺度特征自适应融合; 在解码阶段, 空间–通道双注意力模块结合通道注意力抑制干扰与可变形空间注意力锐化边界, 通过模态感知参数自适应兼容 CT/MRI 的刚弹性差异, 最终在单一模型中实现高效的多尺度解剖结构分割与跨模态泛化。

1　相关工作

1.1　基于 U-Net 架构的分割模型

医学图像分割领域的重大突破始于全卷积网络 (fully convolutional network, FCN[10]) 奠定像素级预测基础, 随后的 U-Net 凭借其对称编解码架构和创新的跳跃连接机制, 开创了医学影像细粒度分割的新范式[11]。为应对复杂的临床需求, 研究者相继提出 3 大改进方向: 在特征优化方面, DenseUNet[12] 通过密集连接增强特征复用, UNet++采用嵌套跳跃连接结合深度监督显著提升边界分割精度; 针对计算效率, LightM-UNet[13] 创新性引入视觉 Mamba 层, 在保持线性计算复杂度的同时实现长程空间依赖建模; 而面向多模态场景的模型 MM-Unet[14] 通过跨模态注意力机制强化异质区域识别。最新进展如 DiffU-Net[15] 将扩散模型生成能力融入分割流程, 进一步提升了模型鲁棒性。然而基于 U-Net 的医学图像分割方法在建模图像长程依赖关系方面表现不足, 对分散病变或大范围解剖结构的捕捉能力有限; 针对微小结构 (如毛细血管或早期病灶) 的分割精度明显下降, 易受跳跃连接中浅层特征的干扰而影响性能。

1.2　基于 Transformer-CNN 混合架构的分割模型

近年来, Transformer-CNN 混合架构已成为医学图像分割研究热点[16-17], 通过结合 CNN 的局部特征提取和 Transformer[18] 的全局建模优势, 显著提升了复杂结构的分割精度[19]。代表性工作包括: TransUNet 率先将 CNN 特征与 Transformer 注意力结合, 在胰腺分割中超越 U-Net; AD2Former[20] 采用 CNN 与 Transformer 交替编码和双解码器设计, 在 Synapse 数据集上 Dice 系数达到 83.18%, 较 Swin-Unet[21] 提升显著; HiFormer[22] 通过 Swin Transformer 与 CNN 双级融合实现高效多尺度交互, 计算复杂度降低 30%; SMAFormer[23] 创新性提出多维注意力机制, 在肝肿瘤分割中优于 Swin UNETR[24]。然而该类方法在整合 CNN 的局部特征和 Transformer 的全局上下文时, 采用简单的串行特征拼接或级联策略, 导致 CNN 的低层细节特征与 Transformer 的高阶语义特征难以充分对齐, 细粒度特征在跨模块传递过程中仍存在显著损失。

1.3　基于并行融合机制的分割模型

单一串行架构难以兼顾高频纹理细节与低频形态结构的实时协同建模, 这一挑战催生了并行融合机制的发展[25-26]。PCTNet[27] 利用注意力机制动态融合局部纹理与全局语义特征; DUNETR[28] 引入降噪模块提升精度; RBSTUNet[29] 通过双分支注意力融合实现结直肠息肉分割 Dice 系数达到 93.81%; DFLNet[30] 采用空间交叉注意力降低乳腺癌误检 41%。多尺度整合方面, ParaTrans-CNN[9] 通过构建金字塔结构结合不同分辨率特

征；ST-Unet[31] 优化跨层增强模块实现 3D 器官分割的全局−局部平衡。然而此类方法对数据质量和规模要求高，Transformer 分支在小样本医学数据集上容易过拟合，而 CNN 分支在面对跨模态数据或弱标注数据时的泛化能力受限，导致在标注稀缺的场景下分割精度下降。

2　GAPE-HTC 医学图像分割网络

本文提出了一种基于 Transformer-CNN 的分组注意力并行编码的医学图像分割网络 (group attention parallel encoding hybrid Transformer-CNN, GAPE-HTC)，整体架构如图 1 所示。GAPE-HTC 网络遵循编码器−解码器架构，由双分支并行编码器、分组注意力并行融合模块 (parallel layer-wise multi-scale group attention, PLMG)、空间−通道双注意力门控解码器 (spatial-channel dual attention gated decoder, SC-DAG) 等 3 部分组成，分别对应图 1(a)~(c)。

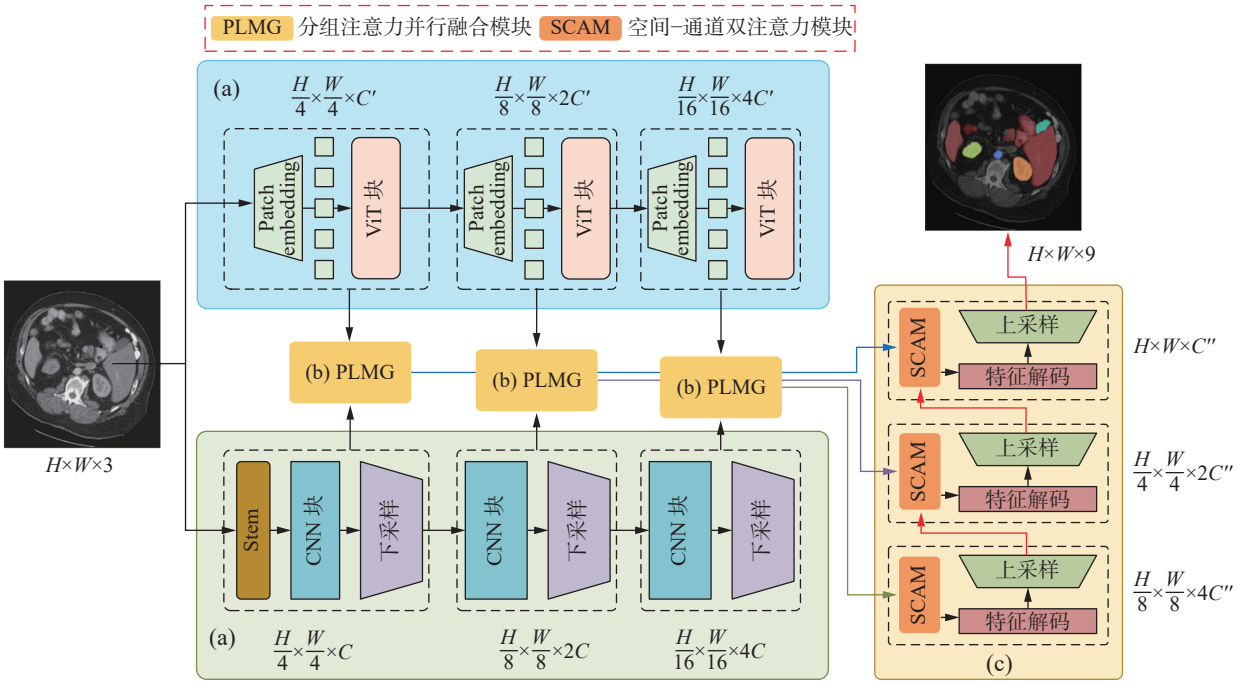


图 1　GAPE-HTC 整体架构

Fig. 1　Overall architecture of the GAPE-HTC

2.1　双分支并行编码器

与常见的 Transformer-CNN 串行架构不同，本文提出的双分支并行编码器采用并行结构，使用 Transformer 分支建模长程依赖关系以捕获全局上下文语义信息，利用 CNN 分支的局部感受野特性专注于提取边缘、纹理等底层局部特征，使 CNN 和 Transformer 分支可以相互补充，同时保留局部细节和全局上下文。编码器每个分支均采用 3 级结构，分别对医学图像特征进行多尺度特征提取，实现从局部细节到全局语义的层次化表示，充分挖掘了医学图像中不同尺度的解剖结构信息。

针对 CNN 分支，为有效降低模型的复杂度，选择 ResNet-34[32] 作为 CNN 分支的主干，并对其进行改造，保留了除池化层和全连接层的其他层组成新的主干网络，使用对应的预训练权重作为该分支的初始权重。如图 1(a) 所示，CNN 分支首先通过 Stem 层提取图像的初始特征，随后的每层分别通过 1/4、1/8 和 1/16 的降维因子实现可扩展特征提取，每层提取到的图像通道数增至上一层的 2 倍。CNN 分支可以表示为

$$\boldsymbol{F}_{\mathrm{s11}}^{\mathrm{CNN}}=\mathrm{Stem}(\boldsymbol{I})\in\mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times\frac{C}{2}}$$

$$\boldsymbol{F}_{\mathrm{s12}}^{\mathrm{CNN}}=\mathrm{S}_1\mathrm{C}(\boldsymbol{F}_{\mathrm{s11}}^{\mathrm{CNN}})\in\mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$$

$$\boldsymbol{F}_{\mathrm{s2}}^{\mathrm{CNN}}=\mathrm{S}_2\mathrm{C}(\boldsymbol{F}_{\mathrm{s1}}^{\mathrm{CNN}})\in\mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2C}$$

$$\boldsymbol{F}_{\mathrm{s3}}^{\mathrm{CNN}}=\mathrm{S}_3\mathrm{C}(\boldsymbol{F}_{\mathrm{s2}}^{\mathrm{CNN}})\in\mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 4C}$$

式中：$\boldsymbol{I}\in\mathbb{R}^{H\times W\times 3}$ 为 CNN 分支的输入图像，Stem(·) 表示提取到的图像初始特征，$\mathrm{S}_i\mathrm{C}(\cdot)$ 表示 CNN 分支第 i 层 ($i\in\{1,2,3\}$) 提取到的可扩展特征。CNN 分支共有 3 层，其中 $\boldsymbol{F}_{\mathrm{s11}}^{\mathrm{CNN}}$ 与 $\boldsymbol{F}_{\mathrm{s12}}^{\mathrm{CNN}}$ 共同组成第 1 层 $\boldsymbol{F}_{\mathrm{s1}}^{\mathrm{CNN}}$，而 $\boldsymbol{F}_{\mathrm{s2}}^{\mathrm{CNN}}$ 与 $\boldsymbol{F}_{\mathrm{s3}}^{\mathrm{CNN}}$ 分别为第 2、3 层。

针对 Transformer 分支，设计了 3 层特征提取模块，如图 1(a) 所示，每个特征提取模块包含 Patch Embedding 和 ViT(vision Transformer) 块[33] 两部分，每层特征提取模块的 Patch 大小、特征维度、MLP(multilayer perceptron) 维度和深度参数如表 1 所示。

表 1　特征提取模块参数

Table 1　Parameters for feature extraction module

层级	Patch大小	特征维度	MLP维度	深度
第1层	4×4	128	256	3
第2层	2×2	256	512	3
第3层	2×2	512	1 024	3

Transformer 分支输入图像大小为 $H\times W\times 3$，其中 $H=W=224$，通道数 $C=3$。每层中的 Patch Embedding 均采用无重叠划分方式进行下采样，将原始图像大小变为 $H/4\times W/4\times C'$、$H/8\times W/8\times 2C'$ 和 $H/16\times W/16\times 4C'$，其中 $C'=128$。后续将 Patch Embedding 部分提取到的特征输入到 ViT 块进一步处理，以增强特征的表达能力。Transformer 分支可以表示为

$$\boldsymbol{F}_{s1}^{ViT}=S_1V(PE_1(\boldsymbol{I}))\in\mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C'}$$

$$\boldsymbol{F}_{s2}^{ViT}=S_2V(PE_2(\boldsymbol{F}_{s1}^{ViT}))\in\mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2C'}$$

$$\boldsymbol{F}_{s3}^{ViT}=S_3V(PE_3(\boldsymbol{F}_{s2}^{ViT}))\in\mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 4C'}$$

式中：$\boldsymbol{I}\in\mathbb{R}^{H\times W\times 3}$ 为 Transformer 分支的输入图像，$PE_i(\cdot)$ 和 $S_iV(\cdot)$ 分别表示 Transformer 分支第 i 层 ($i\in\{1,2,3\}$) 的 Patch Embedding 操作和 ViT 块进一步提取到的特征。

这种多尺度特征提取结构的设计充分利用了 Transformer 的优势。通过逐层下采样，特征图尺寸逐渐减小，不仅显著降低了计算复杂度，而且模型在高层能够更容易地捕捉长距离依赖关系，进一步增强了模型的语义理解能力。

2.2　分组注意力并行融合模块

为了将 CNN 捕获到的边缘、纹理等细节信息与 Transformer 提取到的整体结构和语义信息进行有效结合，模型中设计了一个分组注意力并行融合模块 (PLMG)，如图 2 所示。分组注意力并行融合模块的核心思想是将特征提取和融合过程划分为 4 个独立的分组，每个分组以不同粒度对器官的特征进行处理，使其同时关注目标器官的边缘和纹理以及器官的整体形状和位置，从而提取和融合更加丰富的特征。

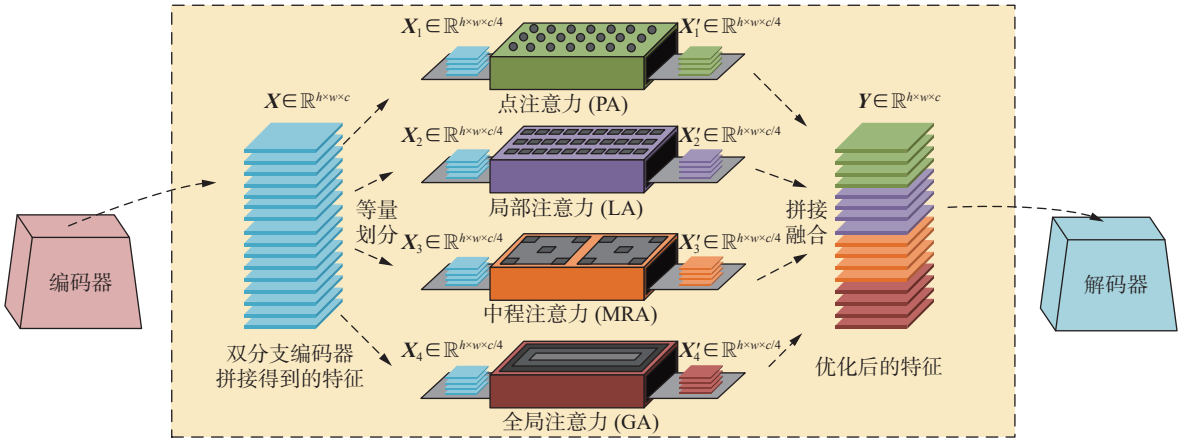


图 2　分组注意力并行融合模块 (PLMG)

Fig. 2　Group attention parallel fusion module (PLMG)

PLMG 模块将 CNN 分支提取到的局部特征 $S_iC(\cdot)$ 和 Transformer 分支提取到的全局特征 $S_iV(\cdot)$ ($i\in\{1,2,3\}$) 进行拼接，形成输入特征图 $\boldsymbol{X}\in\mathbb{R}^{h\times w\times c}$，随后将输入特征图等分成 4 个子特征图 $\boldsymbol{X}_j\in\mathbb{R}^{h\times w\times c/4}$，分别送入第 j 个独立的分组 ($j\in\{1,2,3,4\}$)，其中 h、w 和 c 分别表示输入特征图的高度、宽度和通道数。4 组通路分别使用点注意力 (point attention, PA)、局部注意力 (local attention, LA)、中程注意力 (mid-range attention, MRA) 和全局注意力 (global attention, GA) 从小粒度到大范围提取图像特征，能够将双分支编码器提取到的局部特征与全局语义特征深度融合，使 2 种特征在不同尺度上得到充分的互补和融合。表 2 给出了 3 层编码器中分组注意力并行融合模块的特征维度和模块深度信息。

表 2　分组注意力并行融合模块特征维度和模块深度信息

Table 2　Feature dimension and module depth information of the group attention parallel fusion module

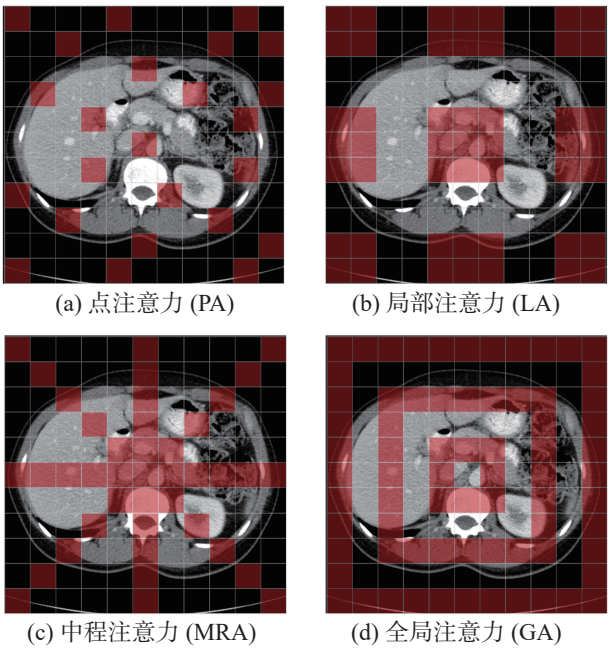
层级	特征维度	深度
第1层	256	4
第2层	512	2
第3层	1 024	1

PA 模块　如图 3(a) 所示，点注意力模块通过学习逐点的注意力权重，对输入特征进行加权，特别关注医学图像中高频细节信息，从而提升模

型对特征的细节表达能力。输入特征 $\boldsymbol{X}_1$ 首先经过特征转换与激活操作，使用一个步长为 1 卷积核大小为 1×1 的卷积层将输入特征维度从 $c/4$ 扩展到 c，随后使用 BN(batch normalization) 层[34]对特征进行归一化处理，再通过 GELU(Gaussian error linear unit) 激活函数[35]引入非线性，这一过程可以表示为

$$\boldsymbol{h} = \mathrm{Act}(\mathrm{Norm}(\mathrm{Conv}(\boldsymbol{X}_1)))$$

式中：$\boldsymbol{X}_1$ 为输入特征，Conv(·) 表示卷积操作，Norm(·) 表示归一化操作，Act(·) 表示激活函数。



(a) 点注意力 (PA)　(b) 局部注意力 (LA)

(c) 中程注意力 (MRA)　(d) 全局注意力 (GA)

图 3　4 种注意力模块可视化

Fig. 3　Visualization of four attention modules

随后，模块通过另一个步长为 1 卷积核大小为 1×1 的卷积层进行注意力权重生成操作，将特征维度从 c 压缩回 $c/4$，并使用 Sigmoid 激活函数生成注意力权重：

$$\boldsymbol{a} = \sigma(\mathrm{Conv}(\boldsymbol{h}))$$

式中：$\sigma(\cdot)$ 表示 Sigmoid 函数；$\boldsymbol{a}$ 为生成的注意力权重，其值域为 [0,1]。

最后，模块将输入特征 $\boldsymbol{X}_1$ 与注意力权重 $\boldsymbol{a}$ 相乘，得到加权后的特征，并与原始输入特征 $\boldsymbol{X}_1$ 残差连接，得到输出特征：

$$\boldsymbol{X}_1' = \boldsymbol{X}_1 \odot a + \boldsymbol{X}_1$$

式中：⊙表示逐元素相乘，$\boldsymbol{X}_1'$ 为该分组最终的输出特征。

LA 模块　如图 3(b) 所示，局部注意力模块通过局部卷积操作，聚焦于输入特征的局部区域，特别关注规则器官轮廓的特征增强，从而提升模型对这些关键区域的识别能力。输入特征 $\boldsymbol{X}_2$ 首先经过局部卷积操作，使用一个步长为 1，填充为 1，卷积核大小为 3×3 的卷积层进行局部处理，从而确保输出特征图的尺寸与输入特征图一致。后续使用 BN 层对特征进行归一化处理，最后通过 GELU 激活函数引入非线性，整个过程表示为

$$\boldsymbol{X}_2' = \mathrm{Act}(\mathrm{Norm}(\mathrm{Conv}(\boldsymbol{X}_2)))$$

式中：$\boldsymbol{X}_2$ 为输入特征，Conv(·) 表示卷积操作，Norm(·) 表示归一化操作，Act(·) 表示激活函数，$\boldsymbol{X}_2'$ 为该分组最终的输出特征。

MRA 模块　如图 3(c) 所示，中程注意力模块通过多尺度特征提取和注意力机制，捕捉器官间的相对位置信息，从而提升模型对器官间空间关系的理解和识别能力。在特征下采样部分，模块输入特征 $\boldsymbol{X}_3$ 首先会被一个步长和填充均为 1，池化核大小为 3×3 的最大池化层进行下采样操作，然后再应用 BlurPool[36]方法进行抗锯齿下采样，通过低通滤波来减少下采样过程中出现的混叠效应，保留器官的轮廓和结构信息，从而更好地关注中程尺度的特征。这一过程可以表示为

$$\boldsymbol{x}_{\mathrm{tem}} = \mathrm{BP}(\mathrm{MP}(\boldsymbol{X}_3))$$

式中：$\boldsymbol{X}_3$ 为输入特征，MP(·) 表示最大池化操作，BP(·) 表示抗锯齿下采样操作，$\boldsymbol{x}_{\mathrm{tem}}$ 为下采样后的特征。

在多尺度特征提取部分，模块通过水平和垂直方向的卷积操作，结合变换和反变换操作，能够提取多尺度的特征信息，进一步增强模型对器官间位置关系的建模能力。水平方向使用步长为 1、填充为 $(k/2,1)$、卷积核大小为 $k\times3$ 的分组卷积，垂直方向使用步长为 1、填充为 $(1,k/2)$、卷积核大小为 $3\times k$ 的分组卷积，其中 k 值设置为 11。通过变换操作对特征进行扩展，并通过反变换操作恢复特征的原始尺寸，相关操作可以表示为

$$\boldsymbol{x}_{\mathrm{h1}} = \mathrm{H}_{\mathrm{att1}}(\boldsymbol{x}_{\mathrm{tem}})$$

$$\boldsymbol{x}_{\mathrm{w1}} = \mathrm{V}_{\mathrm{att1}}(\boldsymbol{x}_{\mathrm{tem}})$$

$$\boldsymbol{x}_{\mathrm{h2}} = \mathrm{h}^{\mathrm{inv}}_{\mathrm{transform}}(\mathrm{H}_{\mathrm{att2}}(\mathrm{h}_{\mathrm{transform}}(\boldsymbol{x}_{\mathrm{tem}})))$$

$$\boldsymbol{x}_{\mathrm{w2}} = \mathrm{v}^{\mathrm{inv}}_{\mathrm{transform}}(\mathrm{V}_{\mathrm{att2}}(\mathrm{v}_{\mathrm{transform}}(\boldsymbol{x}_{\mathrm{tem}})))$$

式中：$\mathrm{H}_{\mathrm{att1}}(\cdot)$ 和 $\mathrm{V}_{\mathrm{att1}}(\cdot)$ 分别为水平分组卷积和垂直分组卷积，$\mathrm{H}_{\mathrm{att2}}(\cdot)$ 和 $\mathrm{V}_{\mathrm{att2}}(\cdot)$ 分别为水平扩展分组卷积和垂直扩展分组卷积，$\mathrm{h}_{\mathrm{transform}}(\cdot)$ 和 $\mathrm{v}_{\mathrm{transform}}(\cdot)$ 为扩展变换操作，$\mathrm{h}^{\mathrm{inv}}_{\mathrm{transform}}(\cdot)$ 和 $\mathrm{v}^{\mathrm{inv}}_{\mathrm{transform}}(\cdot)$ 为恢复变换操作。

随后，将上述提取的特征进行归一化处理，并生成权重特征图，使用 Sigmoid 函数将权重特征图归一化到 [0,1] 范围内，操作可以表示为

$$\boldsymbol{a} = \sigma(\mathrm{Norm}(\boldsymbol{x}_{\mathrm{h1}} + \boldsymbol{x}_{\mathrm{w1}} + \boldsymbol{x}_{\mathrm{h2}} + \boldsymbol{x}_{\mathrm{w2}}))$$

式中：Norm(·) 表示归一化处理，$\sigma(\cdot)$ 表示 Sigmoid

函数, $\boldsymbol{a}$为生成的权重特征图。

最后, MRA 模块将权重特征图$\boldsymbol{a}$通过最近邻插值上采样到与输入特征相同的尺寸, 将输入特征$\boldsymbol{X}_3$与上采样得到的特征图相乘得到最终输出特征, 可以表示为

$$\boldsymbol{X}_3' = \boldsymbol{X}_3 \odot \mathrm{F}_{\mathrm{inter}}(\boldsymbol{a}, (\boldsymbol{X}_3^h, \boldsymbol{X}_3^w))$$

式中: $\boldsymbol{X}_3^h$和$\boldsymbol{X}_3^w$分别表示输入特征$\boldsymbol{X}_3$的长和宽, $\mathrm{F}_{\mathrm{inter}}(\cdot)$表示插值上采样操作, $\boldsymbol{X}_3'$为该分组最终的输出特征。

GA 模块　如图 3(d) 所示, 全局注意力模块采用多头自注意力机制捕捉全局特征信息, 并建立全图的拓扑连接, 有效增强了特征的全局表达能力。输入特征$\boldsymbol{X}_4$通过2×2最大池化层进行下采样得到特征图$\boldsymbol{x}_{\mathrm{down}}$和索引$d$, 特征图$\boldsymbol{x}_{\mathrm{down}}$经过归一化处理后送入注意力机制进行全局特征建模得到加权特征图$\boldsymbol{x}_{\mathrm{weighted}}$, 最大反池化层使用索引$d$将加权特征图$\boldsymbol{x}_{\mathrm{weighted}}$上采样回原始分辨率, 得到增强后的特征图$\boldsymbol{X}_4''$。最后, 将输入特征$\boldsymbol{X}_4$与上采样得到的特征图$\boldsymbol{X}_4''$残差连接得到最终输出特征$\boldsymbol{X}_4'$, 相关操作可以表示为

$$(\boldsymbol{x}_{\mathrm{down}}, d) = \mathrm{MP}(\boldsymbol{X}_4)$$

$$\boldsymbol{X}_4' = \mathrm{MaxUnpool}(\mathrm{Att}(\mathrm{Norm}(\boldsymbol{x}_{\mathrm{down}})), d) + \boldsymbol{X}_4$$

式中: MP(·)表示最大池化操作, Norm(·)表示归一化操作, Att(·)表示注意力机制, MaxUnpool(·)表示最大反池化操作, $\boldsymbol{X}_4'$为该分组最终的输出特征。

最后, 将 4 组注意力的输出特征进行拼接融合, 生成融合模块最终的输出特征$\boldsymbol{Y} \in \mathbb{R}^{h \times w \times c}$。分组注意力并行融合模块根据编码器层级深浅的差异性设置不同的深度, 浅层特征包含更多的细节信息, 需要更复杂的处理来捕捉这些细节, 因此深度较深; 深层特征的分辨率较低, 但包含更高级别的语义信息, 较浅的网络结构能够高效地处理全局信息, 因此深度较浅, 具体配置信息见表 2。

2.3　空间−通道双注意力门控解码器

在多器官分割任务中, 由于器官之间的解剖结构黏连以及尺度多样性, 传统的解码器在特征提取和分割过程中容易受到干扰。特征图中不同器官的语义信息混淆, 难以同时处理不同尺度的特征, 导致分割精度下降。为了解决这些问题, 本文提出了一种空间−通道双注意力门控解码器 (spatial-channel dual attention gated decoder, SC-DAG), 通过层级化特征校准机制, 分别在通道维度抑制无关器官语义噪声, 在空间维度动态聚焦解剖边界。

SC-DAG 解码器核心组成部分是 SCAM 块 (spatial-channel dual attention module), SCAM 块在 CBAM(convolutional block attention module) 块[37] 的基础上进行了改进, 通过通道注意力块 (channel attention block, CAB) 和空间注意力块 (spatial attention block, SAB) 的协同作用, 显式地建模器官间的空间排斥和通道竞争关系, 从而有效抑制特征干扰, 动态聚焦解剖边界。SC-DAG 解码器共有 3 层, 每一层都包含 SCAM 块、特征解码块和上采样块 (如图 1(c) 所示)。SCAM 块的结构如图 4 所示。

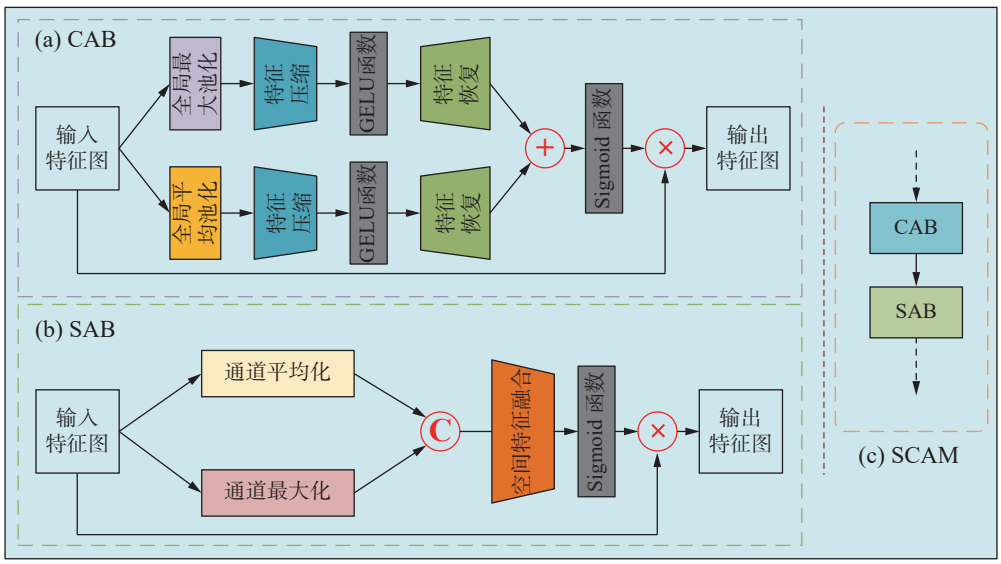


图 4　空间−通道双注意力模块 (SCAM)

Fig. 4　Spatial-channel dual attention module (SCAM)

CAB 块结构如图 4(a) 所示, 其核心思想是通过全局平均池化和全局最大池化分别捕获特征图的全局信息, 并通过特征压缩卷积和特征恢复卷积对通道特征进行加权, 最后通过 Sigmoid 激活

函数生成通道注意力权重。这些权重用于对输入特征图的每个通道进行加权得到特征图$\boldsymbol{F}_{\mathrm{CAB}}$，从而增强重要的通道特征，抑制不重要的通道特征。SCAM 块中的通道注意力块 (CAB) 采用了 GELU 激活函数，取代了 CBAM 块通道注意力机制中的 ReLU(rectified linear unit) 激活函数。这一改进显著增强了通道注意力的非线性表达能力，使模型能够更有效地捕捉复杂的特征关系，从而更好地适应医学图像分割任务的需求。

SAB 块的结构如图 4(b) 所示，其核心思想是先计算平均值特征图和最大值特征图，然后将这 2 个特征图拼接，最后使用空间特征融合卷积并经过 Sigmoid 激活函数处理生成空间权重特征图$\boldsymbol{F}_{\mathrm{SAB}}$。这些权重用于对输入特征图的每个空间位置进行加权，从而增强重要的空间区域，抑制不重要的空间区域。SCAM 块中的空间注意力块 (SAB) 在 CBAM 的空间注意力机制基础上进行了扩展。为了捕捉更丰富的空间信息，SAB 引入了多尺度特征融合机制，通过多个卷积层提取不同尺度的空间特征，并将这些特征加权融合。这种设计显著增强了模型对不同尺度特征的适应能力，特别是在处理多器官分割任务时，能够有效应对器官间的尺度差异和解剖结构黏连问题。此外，在将融合后的特征输入到第一个卷积层之前，使用 repeat 方法动态调整特征的通道数，以匹配卷积层的输入通道要求。这种设计确保了特征在通道维度上的适配性，同时避免了因通道数不匹配导致的错误。

对于经 SCAM 块处理后的特征图$\boldsymbol{F}_{\mathrm{SAB}}$，将其传入特征解码块和上采样块最终得到该层解码器的输出。其中，特征解码块由3×3卷积、BN 层和 ReLU 层组成，上采样块使用转置卷积进行上采样。3 层解码器分别采用 2、2、4 倍的升维因子，将特征图还原并得到分割掩码。

空间−通道双注意力门控解码器 SC-DAG 能够有效解决多器官分割中的特征干扰问题，提升分割精度。在 3 层解码器中的应用进一步增强了模型对不同尺度特征的处理能力。

3　实验结果与分析

3.1　实验数据集

Synapse 多器官数据集[8] 包含 30 例腹部 CT 扫描，共计 3 779 张大小为 512×512 像素的轴向增强对比切片。本工作遵循 TransUNet 的设置，18 例共 2 212 张 2D 轴向切片作为训练集，12 例 3D 格式的 CT 扫描作为测试集，对主动脉、胆囊、左肾、右肾、肝脏、胰腺、脾和胃等 8 个腹部器官进行分割。

ACDC 数据集[38] 包括 100 例心脏 MRI 扫描，涵盖左心室 (left ventricle, LV)、心肌层 (myocardium, MYO) 和右心室 (right ventricle, RV) 3 个亚器官结构。本工作遵循 TransUNet 的设置，采用随机划分的 70 例 MRI 扫描作为训练集 (共计 1 930 张轴向切片)，10 例作为验证集，20 例作为测试集。

AVT 数据集[39] 包含 56 例 CT 血管造影扫描数据，分别来源 3 个子集，分别为 KiTS Grand Challenge(K 子集，20 例病例)、Rider Lung CT(R 子集，18 例病例)、东阳医院 (D 子集，18 例病例)。本工作将 38 例 CT 血管造影数据 (15 044 张轴向切片) 作为训练集，18 例作为测试集。

3.2　实验环境与实验结果

所有实验均在单张显存为 80 GB 的 NVIDIA A800 上完成，采用 Python 编程语言使用 PyTorch 2.2.1 框架训练，在数据预处理阶段采用随机旋转和水平翻转进行数据增强。不同数据集的训练超参数如表 3 所示，所有数据集输入到模型的图像大小均为 224×224 像素。此外，实验为保证可复现性，均设置相同的随机种子，并开启确定性模式。

表 3　不同数据集的训练超参数

Table 3　Training hyperparameters for different datasets

数据集	优化器	初始学习率	批量大小	轮数
Synapse	AdamW	1×10^{-4}	6	400
ACDC	AdamW	1×10^{-4}	6	200
AVT	SGD	1×10^{-2}	4	200

为了同时考虑像素级分类准确性和区域分割的完整性，GAPE-HTC 网络采用了交叉熵损失函数和 Dice 损失函数的组合。在本实验任务中，将两种损失函数的权重均设置为 0.5，即

$$L = 0.5\cdot L_{\mathrm{CE}} + 0.5\cdot L_{\mathrm{Dice}}$$

式中：L_{CE}和L_{Dice}分别表示交叉熵损失值和 Dice 损失值。训练过程中，模型通过最小化上述组合损失函数来学习最优的分割参数，实现对目标器官进行准确分割。实验采用 Dice 相似系数和 95% 豪斯多夫距离作为分割任务的评价指标。

3.3　消融实验

在消融实验中，本工作使用了相同的输入图像尺寸 (224×224 像素)，并在同一数据集上进行训练和评估。为了确保结果的可靠性，所有实验均在相同环境下进行，训练轮数和优化器设置也保持一致。

3.3.1　参数消融

Transformer 模型的核心在于通过自注意力机制捕捉序列之间的关系, 而特征维度 (Dim) 是决定模型特征表示能力的关键参数。为了全面评估 Dim 对模型性能的影响, 本工作在 Synapse 数据集上设计了一系列实验, 分别将 Dim 设置为 16、32、64 和 128, 同时保持解码器中的 Transformer 层数配置 (L1) 和 PLMG 模块深度配置 (L2) 不变, 实验结果如表 4 所示。当 Dim 为 128 时, DSC 达到最高值, HD95 达到最低值, 模型在大多数器官的分割性能上表现最佳, 尤其是在肾脏、胰腺和胃的分割任务中。但胆囊的分割性能在 Dim 为 32 时表现最佳, 表明不同器官对 Dim 的敏感性不同。

表 4　Synapse 数据集上的参数消融实验结果

Table 4　The results of parameter ablation experiments on the Synapse dataset

Dim	L1	L2	平均值		DSC/%↑							
			DSC/%↑	HD95/mm↓	主动脉	胆囊	左肾	右肾	肝脏	胰腺	脾脏	胃
16	(3,3,3)	(4,2,1)	83.05	28.43	**89.43**	67.68	85.72	83.62	94.94	70.64	90.75	81.62
32	(3,3,3)	(4,2,1)	84.47	14.49	89.16	**74.26**	86.54	83.84	**95.54**	69.61	**92.27**	84.53
64	(3,3,3)	(4,2,1)	82.85	19.39	88.12	71.19	86.17	84.23	94.85	69.68	90.66	77.93
128	(2,3,3)	(4,2,1)	82.93	20.59	88.76	69.80	86.26	82.00	95.44	69.21	90.36	81.61
128	(3,3,4)	(4,2,1)	83.26	17.40	89.33	69.27	86.14	84.37	95.11	67.68	91.89	82.27
128	(3,4,3)	(4,2,1)	82.69	22.03	88.68	64.07	86.94	84.61	94.60	68.26	90.94	83.45
128	(3,3,3)	(4,1,1)	83.55	16.74	88.23	70.22	88.05	84.95	95.51	68.65	90.73	82.06
128	(3,3,3)	(2,2,1)	83.36	16.99	88.21	72.62	88.07	82.79	95.28	68.24	91.57	80.07
128	(3,3,3)	(4,2,2)	82.96	18.03	88.27	70.84	85.51	83.25	95.30	66.08	91.36	83.03
128	(3,3,3)	(4,2,1)	**84.86**	**13.54**	89.15	71.53	**88.18**	**86.18**	95.38	**72.28**	91.27	**84.87**

注: 加粗表示最优结果。

为了探讨 Transformer 层数对模型性能的影响, 本工作继续对 Transformer 层数进行了参数消融实验, 分别测试了 (2,3,3)、(3,3,3)、(3,3,4) 和 (3,4,3)4 种不同的层数配置。这些配置分别对应模型中 Transformer 分支 3 层解码器的 Transformer 层数。实验结果如表 4 所示, 表 4 中 L1 列表示解码器中的 Transformer 层数配置。当层数配置取 (3,3,3) 时, 模型在整体分割精度上表现最佳, DSC 达到 84.86%, HD95 减小到 13.54 mm, 显著优于其他配置。此外, 不同阶段的 Transformer 层数对模型性能的影响存在差异。例如, 增加浅层阶段的层数对分割精度的提升更为显著, 而增加中层和深层阶段的层数则对分割精度的提升不明显。

为了验证 PLMG 块不同深度配置对模型性能的影响, 本工作具体测试了 (4,1,1)、(2,2,1)、(4,2,2) 和 (4,2,1) 4 种配置, 这些配置分别对应表 2 中第 1、2、3 层的深度。实验结果如表 4 所示, 表中 L2 列表示分组注意力并行融合模块的深度配置, 当采用 (4,2,1) 的深度配置时, 模型的 DSC 和 HD95 指标达到最好。根据实验结果, 增加浅层和中层特征的深度可以分别显著提升模型对细节信息的捕捉能力, 从而提高分割精度; 而深层特征的深度对模型性能的影响较小。

3.3.2　结构消融

为了验证 GAPE-HTC 模型结构的有效性, 本工作依次在基础的 ResNet34 单分支 CNN 网络上集成 Transformer 分支、PLMG 模块和 SCAM 模块, 在 Synapse 数据集上的结构消融实验结果如表 5 所示。引入 Transformer 分支后, 各器官 DSC 指标显著提高, 且整体 DSC 指标增长 5.08 百分点, 但 HD95 指标下降 0.64 mm。PLMG 模块通过分组以不同粒度提取到更丰富的融合特征, 显著提升了主动脉、脾和胃等器官的分割精度, 尤其是对胃的分割实现了 5.92 百分点的提升。在解码器阶段, 本工作引入的 SCAM 模块在通道维度抑制无关器官语义噪声, 在空间维度动态聚焦解剖边界, 显著提高了多数器官的分割精度, 其中胆囊和胰腺的提升尤为突出, 分别达到 2.99 和 5.50 百分点。左肾和右肾的分割与注意力图如图 5 所示, 随模块的逐步增加, 可直观观察到网络的注意力更加专注, 更加近分割结果。

表 5　Synapse 数据集上的结构消融实验结果

Table 5　The results of structure ablation experiments on the Synapse dataset

模型	平均值		DSC/%↑							
	DSC/%↑	HD95/mm↓	主动脉	胆囊	左肾	右肾	肝脏	胰腺	脾脏	胃
ResNet34	76.50	21.24	87.07	62.40	86.44	72.86	93.93	50.30	85.94	73.05
+ Transformer	81.58	20.60	88.46	70.14	85.23	80.24	94.72	63.38	90.37	80.13
+ PLMG	83.60	18.82	**89.53**	68.54	87.55	83.91	94.95	66.78	**91.52**	**86.05**
+ CAB	83.90	17.13	88.60	**73.65**	88.17	85.23	95.37	69.89	90.29	80.02
+ SAB	**84.86**	**13.54**	89.15	71.53	**88.18**	**86.18**	**95.38**	**72.28**	91.27	84.87

注: 加粗表示最优结果, 下划线表示次优结果。

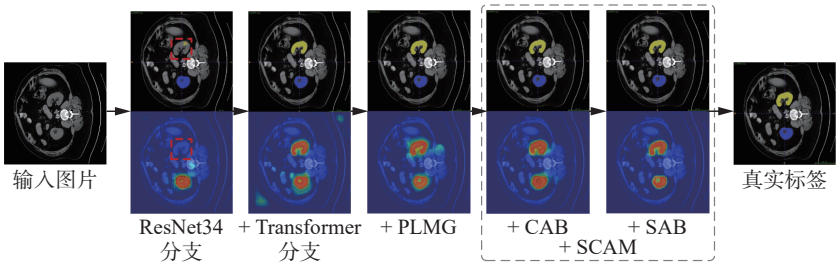


图 5　左肾和右肾的分割与注意力图

Fig. 5　Segmentation and attention maps of the left and right kidneys

3.3.3　PLMG 模块消融

为了深入探讨分组注意力并行融合模块 (PLMG) 中不同分组策略对模型性能的影响, 本文设计了一系列消融实验。实验数据集采用 Synapse 数据集, 消融实验策略为分别移除 PLMG 模块中的点注意力 (PA)、局部注意力 (LA)、中程注意力 (MRA) 和全局注意力 (GA) 模块, 以及逐步将 PLMG 模块添加完整, 从仅包含点注意力 (PA) 开始, 逐步增加其他分组模块。通过对比这些配置的性能, 分析各分组模块对模型分割精度和鲁棒性的影响。实验结果如表 6 所示, 通过对比不同分组模块组合的性能, 验证了各分组模块在模型中的重要性。实验结果表明, PLMG 模块中的每个分组模块都对模型性能有显著影响。

表 6　Synapse 数据集上的 PLMG 模块消融实验结果

Table 6　The results of ablation experiments with the PLMG module on the Synapse dataset

模块	平均值		DSC/%↑							
	DSC/%↑	HD95/mm↓	主动脉	胆囊	左肾	右肾	肝脏	胰腺	脾脏	胃
PA	80.91	39.25	87.84	62.58	81.90	81.33	94.93	68.51	90.16	80.06
PA+LA	81.77	20.50	87.97	64.55	86.96	84.50	95.11	66.07	89.66	79.38
PA+LA+MRA	82.51	21.94	88.33	69.54	85.14	80.20	94.82	66.06	**91.63**	84.38
PLMG	**84.86**	**13.54**	89.15	**71.53**	**88.18**	**86.18**	95.38	**72.28**	91.27	**84.87**
去除PA	83.09	18.97	87.85	68.68	87.47	83.18	95.57	68.09	90.51	83.35
去除LA	83.38	23.34	**89.36**	67.87	88.03	81.99	94.68	70.08	90.66	84.34
去除MRA	81.76	19.42	88.21	70.94	87.16	83.87	**95.58**	66.43	87.27	74.64
去除GA	82.88	22.08	87.65	69.51	87.86	83.82	94.84	68.30	90.92	80.10

注: 加粗表示最优结果, 下划线表示次优结果。

点注意力模块 (PA) 主要关注图像中的关键点信息, 移除后模型在胰腺的分割中表现较差, DSC 下降了 4.19 百分点。这说明点注意力模块在处理器官的关键点特征时具有独特作用, 能够提升模型对细节的感知能力。胰腺的结构较为复杂, 包含许多关键点信息, 点注意力模块能够捕

捉这些关键点, 从而在分割过程中提供更精确的定位和边界信息。

移除局部注意力模块 (LA) 后, 模型在胆囊和右肾等器官的分割中, DSC 分别下降了 3.66 和 4.19 百分点。这说明局部注意力模块能够有效捕捉器官的局部细节。胆囊和右肾的结构较为复杂, 其轮廓和内部结构需要高精度的特征提取来实现准确分割, 局部注意力模块通过增强这些关键区域的特征, 显著提升了模型对这些器官的识别能力。

移除中程注意力模块 (MRA) 后, 模型在胃的分割中, DSC 下降了 10.23 百分点。这是因为胃的分割需要模型不仅关注器官本身的细节, 还需要理解其与其他周围器官的空间位置关系。中程注意力模块通过多尺度特征提取和注意力机制, 能够捕捉器官间的相对位置信息, 从而提升模型对器官间空间关系的理解和识别能力。

移除全局注意力模块 (GA) 后, 模型在胰腺和胃的分割中表现下降, DSC 分别下降了 3.98 和 4.77 百分点。原因是胰腺和胃的分割不仅需要关注局部细节, 还需要理解这些器官在整个图像中的位置, 全局注意力模块能够捕捉到器官在整个图像中的位置关系和全局上下文信息, 从而提升模型对器官整体结构的理解和分割精度。

通过逐步加入点注意力、局部注意力、中程注意力和全局注意力模块, 模型的分割性能逐步提升, 最终在完整 PLMG 模块配置下达到最佳性能, DSC 为 84.86%, HD95 为 13.54 mm。这表明分组注意力机制能够有效整合局部细节和全局语义信息, 显著提升模型对不同尺度器官的分割精度和鲁棒性。

3.4　对比实验

为评估 GAPE-HTC 在多种医学图像分割任务中的性能, 分别在 Synapse、ACDC 和 AVT 数据集上, 与近十年来 CNN 网络、Transformer 网络以及混合 Transformer-CNN 的网络进行对比分析。

3.4.1　腹部多器官分割

Synapse 数据集上的评估结果如表 7 所示, GAPE-HTC 在 DSC 和 HD95 指标上均取得了最佳性能, 分别为 84.86% 和 13.54 mm。与 CNN 网络、Transformer 网络以及 Transformer-CNN 混合网络的最优模型相比, DSC 分别提升了 7.09、1.97 和 1.00 百分点。相比于次优模型, 本工作的模型在右肾和胰腺等具体器官上的 DSC 分数有明显的改善, 分别提高了 2.34 和 2.49 百分点。

表 7　Synapse 数据集上模型对比实验结果

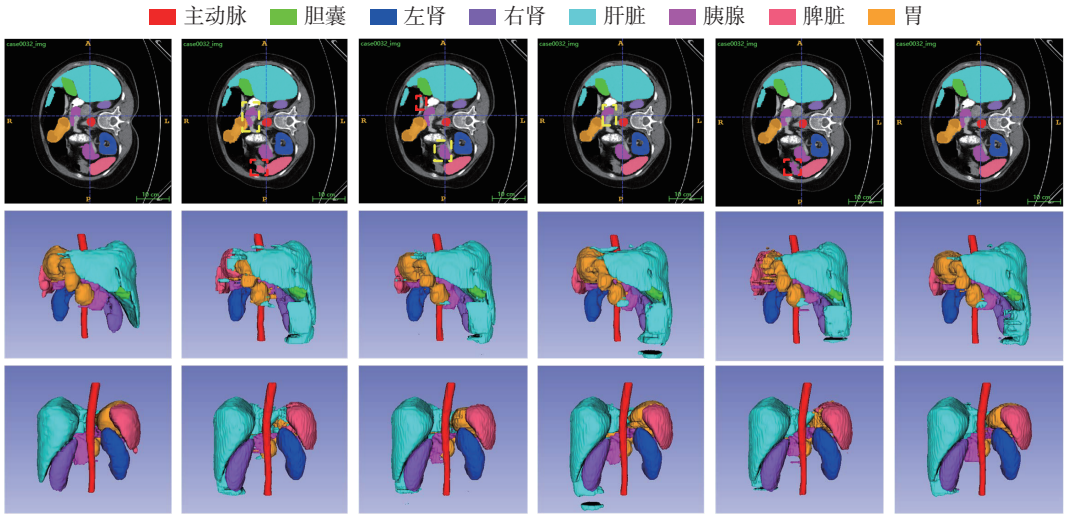
Table 7　Comparison experiment results of models on the Synapse dataset

网络结构	模型	平均值		DSC/%↑							
		DSC/%↑	HD95/mm↓	主动脉	胆囊	左肾	右肾	肝脏	胰腺	脾脏	胃
CNN	U-Net[1]	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
	AttU-Net[40]	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
	R50 U-Net[8]	74.68	36.87	87.47	66.36	80.60	78.19	93.74	56.90	85.87	74.16
Transformer	Swin-Unet[21]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
	TransDeepLab[41]	80.16	21.25	86.04	69.16	84.08	79.88	93.53	61.19	89.00	78.40
	MISSFormer[42]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
	DAE-Former[43]	82.43	17.46	88.96	<u>72.30</u>	86.08	80.88	94.98	65.12	91.94	79.19
	SSTrans-Net[44]	82.89	15.55	86.72	**72.87**	86.99	83.33	94.61	64.98	<u>92.23</u>	81.37
Mamba	VM-UNet[45]	81.08	19.21	86.40	69.41	86.16	82.76	94.17	58.80	89.51	81.40
	CCViM[46]	82.65	17.83	87.63	68.45	86.23	83.22	94.67	67.12	92.05	81.82
混合结构	TransUNet[8]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
	HiFormer[22]	80.39	<u>14.70</u>	86.21	65.69	85.23	79.77	94.61	59.52	90.99	81.08
	TransCeption[47]	82.24	20.89	87.60	71.82	86.23	80.29	95.01	65.27	91.68	80.02
	ParaTransCNN[9]	<u>83.86</u>	15.86	88.12	68.97	<u>87.99</u>	<u>83.84</u>	95.01	<u>69.79</u>	**92.71**	<u>84.43</u>
	SMFNet[48]	81.11	18.10	86.84	62.49	84.65	81.78	94.86	64.46	90.74	83.92
	AD2Former[20]	83.18	20.89	88.94	71.76	83.54	80.00	<u>95.16</u>	69.71	92.14	84.23
	GAPE-HTC	**84.86**	**13.54**	<u>89.15</u>	71.53	**88.18**	**86.18**	**95.38**	**72.28**	91.27	**84.87**

注: 加粗表示最优结果, 下划线表示次优结果。

为了进一步验证 GAPE-HTC 模型的优越性，本文对比了最新的基于高级 Transformer 的 SSTrans-Net[44] 模型和基于 Mamba 的 VM-UNet[45] 和 CCViM[46] 模型。实验结果表明，GAPE-HTC 在 DSC 和 HD95 指标上均优于这些最新的对比模型。具体而言，与 SSTrans-Net 相比，GAPE-HTC 的 DSC 提升了 1.97 百分点，HD95 降低了 2.01 mm。此外，VM-UNet 和 CCViM 作为基于 Mamba 的先进模型，在整体分割精度和鲁棒性上仍不如 GAPE-HTC。GAPE-HTC 通过并行架构的优势，能够更好地捕捉多尺度特征，并在特定器官的细节捕捉上表现出色。在胰腺分割任务中，GAPE-HTC 的 DSC 比 VM-UNet 和 CCViM 分别高出 13.48 和 5.16 百分点，这主要得益于 PLMG 模块能够精细地处理胰腺的复杂解剖结构，并通过分组注意力机制，有效捕捉胰腺的细微边界和内部结构，从而提升分割精度。

如图 6 所示，传统的 U-Net 方法、Swin-Unet 方法和 TransCeption 方法在分割脾、肝脏和胰腺时出现误分割的情况 (如图 6 中第 1 列红框所示)，U-Net 方法、Swin-Unet 方法和 TransUNet 方法对胰腺进行分割时还出现了边缘粗糙和部分漏检的现象 (如图 6 中第 1 列黄框所示)，而 GAPE-HTC 模型能够准确识别和分割以上的问题器官，从三维的结果更能看出，本文模型分割的器官结构更加完整，边缘更加平滑，表明其有更高的精确度。



(a) 真实标签　(b) U-Net　(c) Swin-Unet　(d) TransUNet　(e) TransCeption　(f) GAPE-HTC

图 6　Synapse 数据集上不同模型可视化分割结果

3.4.2　心脏器官分割

ACDC 数据集上的评估结果如表 8 所示，GAPE-HTC 模型的 DSC 为 91.66%，相比于其他模型展现出卓越的性能，但 HD95 与最优模型相比仍有一定差距。模型在左心室分割中效果大幅提升，相比于次优模型 DSC 提升了 1.05 百分点，相比于其他 Transformer-CNN 混合模型 DSC 最高提升了 1.97 百分点。

表 8　ACDC 数据集上模型对比实验结果

Table 8　Comparison experiment results of models on the ACDC dataset

网络结构	模型	平均值		DSC/%↑		
		DSC/%↑	HD95/mm↓	右心室	心肌	左心室
CNN	U-Net[1]	90.68	1.51	91.65	84.82	95.57
	AttU-Net[40]	90.90	1.52	91.33	85.79	95.58
Transformer	Swin-Unet[21]	90.25	1.32	90.67	84.92	95.16
	TransDeepLab[41]	90.41	1.12	91.45	84.54	95.23
	MISSFormer[42]	86.78	2.32	86.04	80.44	93.84
	DAE-Former[43]	89.78	1.91	89.91	84.38	95.04
	SSTrans-Net[44]	90.31	—	89.02	87.51	94.40

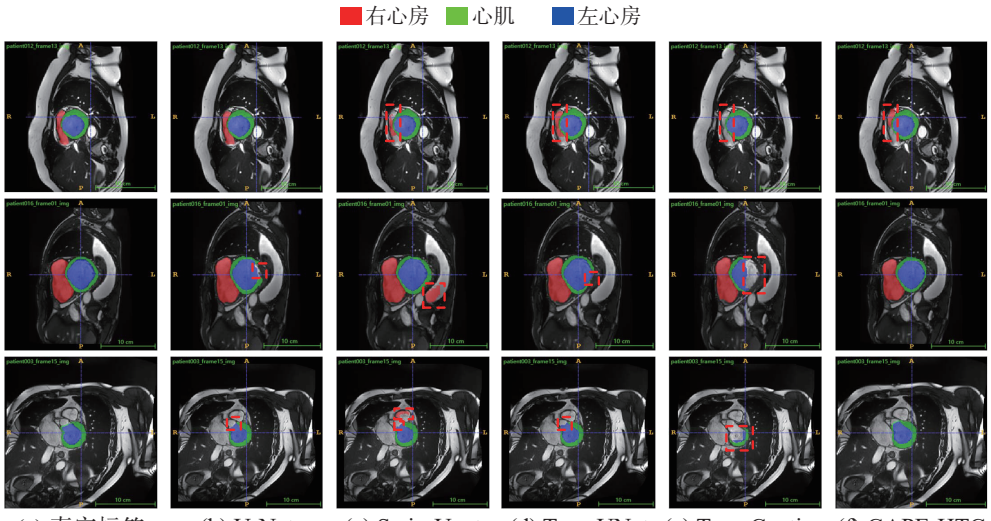
续表 8

网络结构	模型	平均值		DSC/%↑		
		DSC/%↑	HD95/mm↓	右心室	心肌	左心室
Mamba	PHMamba[49]	91.04	1.24	—	—	—
	SegMamba[50]	90.74	1.19	—	—	—
混合结构	TransUNet[8]	90.79	1.08	91.56	85.24	95.18
	HiFormer[22]	90.12	2.15	91.06	84.54	94.77
	TransCeption[47]	88.47	1.78	87.88	82.87	94.66
	ParaTransCNN[9]	91.31	1.16	**92.76**	85.84	95.34
	SMFNet[48]	91.52	**1.06**	89.98	**89.01**	95.57
	MCT-Net[51]	90.91	—	90.13	87.36	95.23
	GAPE-HTC	**91.66**	1.20	90.44	87.90	**96.63**

注: 加粗表示最优结果, 下划线表示次优结果, 横线表示实验缺失值。

在 ACDC 数据集的对比实验中, 还引入了最新的基于高级 Transformer 的 SSTrans-Net 模型以及基于 Mamba 的 PHMamba[49] 和 SegMamba[50] 模型。实验结果显示, GAPE-HTC 在 DSC 指标上优于所有对比模型。与 SSTrans-Net 相比, GAPE-HTC 在左心室的分割精度上有显著提升, DSC 提升了 2.23 百分点, 而 PHMamba 和 SegMamba 虽然在 HD95 指标上表现接近, 但在 DSC 指标上的表现仍不如 GAPE-HTC。这是因为 GAPE-HTC 的并行架构能够更好地处理心脏器官的复杂拓扑结构, 通过空间−通道双注意力模块 (SCAM) 有效抑制了特征干扰, 优化了分割边缘, 从而避免了分割不全的问题。

ACDC 数据集上的部分可视化结果如图 7 所示, 本文随机选择了 3 例测试集样本进行可视化对比, 从可视化结果可以观察到, GAPE-HTC 在心肌层和左心室分割中效果明显, 达到了最佳的分割性能, 其他方法在分割左心室时出现了分割不全的现象, 而本文提出的 GAPE-HTC 模型能够捕捉左心室和心肌层的拓扑结构, 优化心肌分割边缘, 从而解决分割不全的问题。



(a) 真实标签　(b) U-Net　(c) Swin-Unet　(d) TransUNet　(e) TransCeption　(f) GAPE-HTC

图 7　ACDC 数据集上不同模型可视化分割结果

Fig. 7　Visualize segmentation results of different models on the ACDC dataset

3.4.3　血管造影分割

AVT 数据集上的评估结果如表 9 所示, GAPE-HTC 模型以 88.79% 的 DSC 和 4.02 mm 的 HD95 指标在所有的模型中分割性能位列第一。本文模型在 DSC 指标上较 Swin-Unet 提升了 9.17 百分点, 相较于 TransUNet 在 HD95 指标上下降了 7.41 mm。在子数据集 R 和 K 上达到子数据集 DSC 的最大值, 在 D 数据集上 DSC 和 HD95 指标仅比最优方法分别低 0.18 和 0.07 百分点。

表 9　AVT 数据集上模型对比实验结果

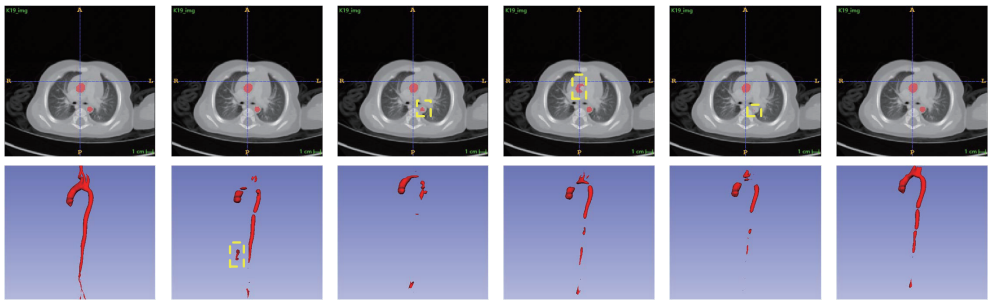
Table 9　Comparison experiment results of models on the AVT dataset

模型	平均值		D数据集		R数据集		K数据集	
	DSC/%↑	HD95/mm↓	DSC/%↑	HD95/mm↓	DSC(%)↑	HD95/mm↓	DSC(%)↑	HD95/mm↓
U-Net[1]	82.89	7.27	93.13	2.65	76.32	10.34	79.22	8.83
AttU-Net[40]	80.25	9.45	92.69	3.58	70.15	15.18	77.91	9.58
Swin-Unet[21]	79.62	13.76	92.02	2.19	69.41	26.48	77.45	12.62
TransDeepLab[41]	79.72	12.95	92.97	1.90	71.06	24.88	75.15	12.06
MISSFormer[42]	79.09	12.25	92.28	2.67	69.20	21.62	75.81	12.45
DAE-Former[43]	85.02	8.96	93.96	1.21	78.85	14.32	82.25	11.37
SSTrans-Net[44]	83.88	10.23	92.81	1.75	78.11	17.82	80.74	11.11
FE-MAANet[52]	86.25	6.05	94.07	1.07	84.09	7.42	82.25	9.68
TransUNet[8]	82.99	11.43	93.59	1.21	74.73	23.47	80.67	9.61
HiFormer[22]	82.53	7.35	93.35	1.41	75.56	11.19	78.69	9.45
TransCeption[47]	82.87	10.33	93.70	1.21	75.54	17.92	79.37	11.97
ParaTransCNN[9]	87.82	4.70	93.91	1.07	83.01	7.75	86.55	5.27
Atten-Nonlocal Unet[53]	86.94	4.78	**94.32**	**1.00**	80.93	9.03	85.58	4.33
GAPE-HTC	**88.79**	**4.02**	94.14	1.07	**84.98**	**7.07**	**87.24**	**3.90**

注: 加粗表示最优结果, 下划线表示次优结果。

实验同样对比了最新的基于高级 Transformer 的 SSTrans-Net 和 FE-MAANet[52] 模型。实验结果表明, GAPE-HTC 在 DSC 和 HD95 指标上均优于这些最新对比模型。具体而言, 与 SSTrans-Net 相比, GAPE-HTC 的 DSC 提升了 4.91 百分点, HD95 降低了 6.21 mm, 尤其在 R 数据集上 DSC 提升了 6.87 百分点, HD95 降低了 10.75 mm, 而 FE-MAANet 虽然在某些子数据集上 HD95 与 GAPE-HTC 表现接近, 但在整体分割精度上仍不如 GAPE-HTC。这表明 GAPE-HTC 在血管造影分割任务中也具有较高精度, 其优势在于其并行架构能够更好处理血管的复杂结构, 同时 GAPE-HTC 通过分组注意力并行融合模块 (PLMG) 能够更准确识别和分割血管结构, 避免了断层现象和异常分割。

在图 8 的 AVT 数据集可视化结果中, 有的方法在主动脉整体上出现了断层的现象, 例如 Swin-Unet 和 TransCeption 方法在分割中出现了巨大间隙, 有的方法不能精确识别并分割主动脉结构, 例如 U-Net 方法分割时出现了异常分割现象。而本文方法在整体上能够产生连续的主动脉结构, 其分割边缘清晰度明显优于其他方法。



(a) 真实标签　(b) U-Net　(c) Swin-Unet　(d) TransUNet　(e) TransCeption　(f) GAPE-HTC

图 8　AVT 数据集上不同模型可视化分割结果

Fig. 8　Visualize segmentation results of different models on the AVT dataset

GAPE-HTC 模型在多个数据集上表现出色, 主要归功于模型的双分支结构和注意力机制的设计。双分支结构有效解决了传统单分支架构中局部与全局语义割裂的问题, 使得模型在处理复杂

解剖结构时能够同时兼顾细节和整体, 从而显著提升了分割精度。PLMG 模块通过将特征提取和融合过程划分成 4 个独立分组, 能够以不同粒度对器官的特征进行处理。这种多尺度的特征提取与融合机制, 使得 CNN 分支提取的局部特征与 Transformer 分支提取的全局语义特征在不同尺度上得到充分的互补和融合。SCAM 模块通过结合通道注意力和空间注意力, 显式建模器官间的空间排斥与通道竞争关系, 从而有效抑制了特征干扰, 进一步提升了模型的鲁棒性和泛化能力。

3.5　模型性能分析

为了全面评估所提出的 GAPE-HTC 模型在精度和效率方面的表现, 本文在 Synapse 数据集上与其他主流医学图像分割模型进行了性能对比分析。对比结果如表 10 所示, 包括模型的参数量、计算复杂度 FLOPs(floating point operations)、实时性 FPS(frames per second)、分割精度 (DSC 和 HD95) 等评价指标。

表 10　Synapse 数据集上不同模型的性能对比

Table 10　Comparison of model performance on the Synapse dataset

模型	参数量/10^6↓	FLOPs/10^9↓	FPS/(帧/s)↑	DSC/%↑	HD95/mm↓
U-Net[1]	31.04	41.93	77.06	76.85	39.70
AttU-Net[40]	34.88	51.04	<u>78.81</u>	77.77	36.02
Swin-Unet[21]	<u>21.75</u>	**5.93**	41.14	79.13	21.55
TransDeepLab[41]	**21.14**	15.68	**127.23**	80.16	21.25
MISSFormer[42]	35.45	<u>7.28</u>	52.80	81.96	18.20
DAE-Former[43]	29.69	26.16	41.70	82.43	17.46
TransUNet[8]	105.28	24.73	61.38	77.48	31.69
HiFormer[22]	34.14	17.81	48.90	80.39	14.70
ParaTransCNN[9]	245.47	568.10	31.80	<u>83.86</u>	15.86
MixFormer[54]	123.35	108.32	30.50	82.64	**12.67**
GAPE-HTC	87.07	116.68	43.62	**84.86**	<u>13.54</u>

注: 加粗表示最优结果, 下划线表示次优结果。

GAPE-HTC 模型在 DSC 和 HD95 上均优于大多数对比模型, 分别达到了 84.86% 和 13.54 mm。这表明 GAPE-HTC 在处理复杂解剖结构时具有较高的分割精度, 能够满足高精度医学图像分割的需求。在计算复杂度方面, GAPE-HTC 模型的 FLOPs 为 116.68×10^9。这一数值虽然高于一些轻量级模型 (如 Swin-Unet, FLOPs 为 5.93×10^9), 但低于一些计算量较大的模型 (如 ParaTransCNN, FLOPs 为 568.1×10^9)。这表明 GAPE-HTC 在计算复杂度上进行了合理的权衡, 既保证了高精度又避免了过高的计算负担。在参数量方面, GAPE-HTC 模型参数量为 87.07×10^6, 显著低于一些高精度但参数量较大的模型 (如 MixFormer[54], 参数量为 123.35×10^6)。与这些模型相比, GAPE-HTC 在保持高精度的同时, 显著减少了参数量, 降低了模型的复杂度和存储需求。这不仅有助于模型的快速训练和部署, 还减少了计算资源的消耗, 提高了模型的可扩展性和适用性。在实时性方面, GAPE-HTC 模型的 FPS 为 43.62 帧/s, 虽然不及一些轻量级模型 (如 TransDeepLab, FPS 为 127.23 帧/s), 但仍然能够满足大多数临床场景的实时性需求, 特别是在对精度要求较高的场景中, GAPE-HTC 的实时性表现是可接受的。

总体而言, GAPE-HTC 模型在精度和效率之间取得了较好的平衡。与高精度但计算量较大的模型相比, 通过优化网络结构和注意力机制, 显著降低了计算复杂度, 同时保持了较高的分割精度。与轻量级但精度较低的模型相比, 通过引入分组注意力机制, GAPE-HTC 显著提升了分割精度, 尽管计算复杂度有所增加, 但仍然在可接受范围内。这种平衡使得 GAPE-HTC 模型适用于对精度要求较高但对实时性要求不严苛的临床场景。

4　结束语

本文提出了一种 GAPE-HTC 的网络架构, 该架构结合了 Transformer 和 CNN 的特性, 通过分组注意力并行融合模块 (PLMG) 和空间−通道双注意力门控模块 (SCAM) 显著提升了医学图像分

割的精度和鲁棒性。GAPE-HTC 采用双分支编码器结构，在特征提取阶段同时捕捉全局语义信息和局部细节信息，PLMG 模块采用分组注意力有效解决了局部与全局语义割裂及多尺度结构冲突问题，SCAM 模块通过通道注意力和空间注意力解决分割中的特征干扰问题。通过在 Synapse、ACDC 和 AVT 等多个医学图像分割数据集上的实验验证，GAPE-HTC 在 DSC 和 HD95 等关键指标上均优于现有的先进方法，充分验证了其在医学图像分割中的准确性和有效性。在未来的工作中，将重点优化 GAPE-HTC 网络架构以提高其计算效率，同时将其拓展到三维医学图像分割任务中，改进注意力机制和特征融合策略，以提高三维网络的分割精度，使其能够更好地满足复杂临床需求。

参考文献：

[1] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015: 234−241.

[2] ZHOU Zongwei, SIDDIQUEE M M R, TAJBAKHSH N, et al. UNet: redesigning skip connections to exploit multiscale features in image segmentation[J]. IEEE transactions on medical imaging, 2020, 39(6): 1856−1867.

[3] DIAKOGIANNIS F I, WALDNER F, CACCETTA P, et al. ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data[J]. ISPRS journal of photogrammetry and remote sensing, 2020, 162: 94−114.

[4] 邢素霞, 李珂娴, 方俊泽, 等. 深度学习下的医学图像分割综述[J]. 计算机工程与应用, 2025, 61(7): 25−41.
XING Suxia, LI Kexian, FANG Junze, et al. Survey of medical image segmentation in deep learning[J]. Computer engineering and applications, 2025, 61(7): 25−41.

[5] 蒋婷, 李晓宁. 采用多尺度视觉注意力分割腹部 CT 和心脏 MR 图像[J]. 中国图象图形学报, 2024, 29(1): 268−279.
JIANG Ting, LI Xiaoning. Segmentation of abdominal CT and cardiac MR images with multi scale visual attention[J]. Journal of image and graphics, 2024, 29(1): 268−279.

[6] 潘丹, 吕锦, 曾安. 基于自注意力与多视角注意力的医学图像分割方法[J]. 生物医学工程学杂志, 2025, 42(5): 919−927.
PAN Dan, LÜ Jin, ZENG An. Medical image segmentation method based on self-attention and multi-view attention[J]. Journal of biomedical engineering, 2025, 42(5): 919−927.

[7] 赵亮, 刘晨, 王春艳. 位置信息增强的 TransUnet 医学图像分割方法[J]. 计算机科学与探索, 2025, 19(4): 976−988.
ZHAO Liang, LIU Chen, WANG Chunyan. Positional enhancement TransUnet for medical image segmentation[J]. Journal of frontiers of computer science and technology, 2025, 19(4): 976−988.

[8] CHEN Jieneng, LU Yongyi, YU Qihang, et al. TransUNet: transformers make strong encoders for medical image segmentation[EB/OL]. (2021−02−08)[2026−01−06]. https://doi.org/10.48550/arXiv.2102.04306. [LinkOut]

[9] SUN Hongkun, XU Jing, DUAN Yuping. ParaTransCNN: parallelized TransCNN encoder for medical image segmentation[EB/OL]. (2024−01−27)[2026−01−06]. https://doi.org/10.48550/arXiv.2401.15307.

[10] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3431−3440.

[11] 黄晓鸣, 何富运, 唐晓虎, 等. U-Net 及其变体在医学图像分割中的应用研究综述[J]. 中国生物医学工程学报, 2022, 41(5): 567−576.
HUANG Xiaoming, HE Fuyun, TANG Xiaohu, et al. Review on applications of U-Net and its variants in medical image segmentation[J]. Chinese journal of biomedical engineering, 2022, 41(5): 567−576.

[12] HUANG Gao, LIU Zhuang, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2261−2269.

[13] LIAO Weibin, ZHU Yinghao, WANG Xinyuan, et al. LightM-UNet: mamba assists in lightweight UNet for medical image segmentation[EB/OL]. (2024−03−11)[2026−01−06]. https://doi.org/10.48550/arXiv.2403.05246.

[14] XIE Bin, YAN Yan, AGAM G. MM-UNet: meta mamba UNet for medical image segmentation[EB/OL]. (2025−03−21)[2026−01−06]. https://doi.org/10.48550/arXiv.2503.17540.

[15] XING Zhaohu, WAN Liang, FU Huazhu, et al. Diff-UNet: a diffusion embedded network for volumetric segmentation[EB/OL]. (2023−03−18)[2026−01−06]. https://doi.org/10.48550/arXiv.2303.10326.

[16] 彭雨彤, 梁凤梅. 融合 CNN 和 ViT 的乳腺超声图像肿瘤分割方法[J]. 智能系统学报, 2024, 19(3): 556−564.
PENG Yutong, LIANG Fengmei. Tumor segmentation method for breast ultrasound images incorporating CNN and ViT[J]. CAAI transactions on intelligent systems, 2024, 19(3): 556−564.

[17] 傅励瑶, 尹梦晓, 杨锋. 基于 Transformer 的 U 型医学图像分割网络综述[J]. 计算机应用, 2023, 43(5): 1584−1595.
FU Liyao, YIN Mengxiao, YANG Feng. Transformer based U-shaped medical image segmentation network: a survey[J]. Journal of computer applications, 2023, 43(5): 1584−1595.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30: 5998−6008.

[19] 王骁崴, 邢树礼, 毛国君. 多尺度特征融合的轻量化 Transformer 医学图像分割研究[J]. 中国生物医学工程学报, 2025, 44(2): 165−173.
WANG Xiaowei, XING Shuli, MAO Guojun. Research on lightweight transformer medical image segmentation with multi-scale feature fusion[J]. Chinese journal of biomedical engineering, 2025, 44(2): 165−173.

[20] ZHANG Lin, GUO Xinyu, SUN Hongkun, et al. Alternate encoder and dual decoder CNN-Transformer networks for medical image segmentation[J]. Scientific reports, 2025, 15: 8883.

[21] CAO Hu, WANG Yueyue, CHEN J, et al. Swin-unet: unet-like pure transformer for Medical image segmentation[C]//Computer Vision–ECCV 2022 Workshops. Cham: Springer, 2023: 205−218.

[22] HEIDARI M, KAZEROUNI A, SOLTANY M, et al. HiFormer: hierarchical multi-scale representations using transformers for medical image segmentation[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2023: 6191−6201.

[23] ZHENG Fuchen, CHEN Xuhang, LIU Weihuang, et al. SMAFormer: synergistic multi-attention transformer for medical image segmentation[C]//2024 IEEE International Conference on Bioinformatics and Biomedicine. Lisbon IEEE, 2024: 4048−4053.

[24] HATAMIZADEH A, NATH V, TANG Yucheng, et al. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images[C]//Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Cham: Springer, 2022: 272−284.

[25] 袁宝华, 陈佳璐, 王欢. 融合多尺度语义和双分支并行的医学图像分割网络[J]. 计算机应用, 2025, 45(3): 988−995.
YUAN Baohua, CHEN Jialu, WANG Huan. Medical image segmentation network integrating multi-scale semantics and parallel double-branch[J]. Journal of computer applications, 2025, 45(3): 988−995.

[26] 李赫, 刘建军, 肖亮. 融合交叉注意力与双编码器的医学图像分割[J]. 中国图象图形学报, 2024, 29(11): 3462−3475.
LI He, LIU Jianjun, XIAO Liang. Dual-encoder global-local cross-attention network for medical image segmentation[J]. Journal of image and graphics, 2024, 29(11): 3462−3475.

[27] ZHU Xinjun, HAN Zhiqiang, ZHANG Zhizhi, et al. PCT-Net: depth estimation from single structured light image with a parallel CNN-transformer network[J]. Measurement science and technology, 2023, 34(8): 085402.

[28] PAN Dan, LUO Genqiang, ZENG An. Coronary artery segmentation based on transformer and convolutional neural networks dual parallel branch encoder neural network[J]. Journal of biomedical engineering, 2024, 41(6): 1195−1203,1212.

[29] GAO Junbo, HU Jingcheng, SUN Wei. RBSTUNet: a dual-branch parallel network for colorectal polyp segmentation[J]. Biomedical signal processing and control, 2025, 108: 107931.

[30] YE Zhaoyi, HUANG Jin, ZHANG Yimin, et al. DFLNet: Disentangled feature learning network for breast cancer ultrasound image segmentation[J]. Digital signal processing, 2025, 165: 105331.

[31] ZHANG Jing, QIN Qiuge, YE Qi, 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: 106516.

[32] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[33] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2021−06−03)[2026−01−06]. https://doi.org/10.48550/arXiv.2010.11929.

[34] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//International Conference on Machine Learning. Lille, PMLR, 2015: 448−456.

[35] HENDRYCKS D, GIMPEL K. Gaussian error linear units (GELUs)[EB/OL]. (2023−06−06)[2026−01−06]. https://doi.org/10.48550/arXiv.1606.08415.

[36] ZHANG R. Making convolutional networks shift-invariant again[C]//International Conference on Machine Learning. Los Angeles: PMLR, 2019: 7324−7334.

[37] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Computer Vision–ECCV 2018. Cham: Springer, 2018: 3−19.

[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] ZHU Zhiqin, WANG Ziyu, QI Guanqiu, et al. Brain tumor segmentation in MRI with multi-modality spatial information enhancement and boundary shape correction[J]. Pattern recognition, 2024, 153: 110553.

[40] OKTAY O, SCHLEMPER J, LE FOLGOC L, et al. Attention U-Net: learning where to look for the pancreas[EB/OL]. (2018−05−20)[2026−01−06]. https://doi.org/10.48550/arXiv.1804.03999.

[41] AZAD R, HEIDARI M, SHARIATNIA M, et al. Trans-DeepLab: convolution-free transformer-based DeepLab v3+ for Medical image segmentation[C]//Predictive Intelligence in Medicine. Cham: Springer, 2022: 91−102.

[42] HUANG Xiaohong, DENG Zhifang, LI Dandan, et al. MISSFormer: an effective transformer for 2D medical image segmentation[J]. IEEE transactions on medical imaging, 2023, 42(5): 1484−1494.

[43] AZAD R, ARIMOND R, AGHDAM E K, et al. DAE-former: dual attention-guided efficient transformer for Medical image segmentation[C]//Predictive Intelligence in Medicine. Cham: Springer, 2023: 83−95.

[44] FU Liyao, CHEN Yunzhu, JI Wei, et al. SSTrans-Net: smart swin Transformer network for medical image segmentation[J]. Biomedical signal processing and control, 2024, 91: 106071.

[45] RUAN Jiacheng, LI Jincheng, XIANG Suncheng. VM-UNet: vision mamba UNet for medical image segmentation[J]. ACM transactions on multimedia computing, communications, and applications, 2025: 3767748.

[46] ZHU Yun, ZHANG Dong, LIN Yi, et al. Merging context clustering with visual state space models for medical image segmentation[J]. IEEE transactions on medical imaging, 2025, 44(5): 2131−2142.

[47] AZAD R, JIA Yiwei, AGHDAM E K, et al. Enhancing medical image segmentation with TransCeption: a multi-scale feature fusion approach[EB/OL]. (2023−01−25)[2026−01−06]. https://doi.org/10.48550/arXiv.2301.10847.

[48] LI Jing, CHEN Qiaohong, FANG Xian. Sub-pixel multi-scale fusion network for medical image segmentation[J]. Multimedia tools and applications, 2024, 83(41): 89355−89373.

[49] CHEN Nuo, WANG Shaoyu, LU Ran, et al. PHMamba: preheating state space models with context-augmented features for medical image segmentation[C]//ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. Hyderabad: IEEE, 2025: 1−5.

[50] XING Zhaohu, YE Tian, YANG Yijun, et al. SegMamba: long-range sequential modeling Mamba for 3D medical image segmentation[EB/OL]. (2024−09−15)[2026−01−06]. https://doi.org/10.48550/arXiv.2401.1356.

[51] SHEN Longfeng, DIAO Liangjin, PENG Rui, et al. MCT-Net: a multi-branch hybrid CNN-transformer model for medical image segmentation[J]. Pattern analysis and applications, 2025, 28(2): 106.

[52] PENG Ke, HONG Bi, LIU Yiyang, et al. FE-MAANet: a frequency-enhanced multi-scale adaptive attention network for medical image segmentation[J]. Biomedical signal processing and control, 2026, 113: 108884.

[53] JIA Xiaofen, WANG Wenjie, ZHANG Mei, et al. Atten-Nonlocal Unet: attention and non-local Unet for medical image segmentation[J]. Computers in biology and medicine, 2025, 191: 110129.

[54] LIU Jun, LI Kunqi, HUANG Chun, et al. MixFormer: a mixed CNN–transformer backbone for medical image segmentation[J]. IEEE transactions on instrumentation and measurement, 2025, 74: 1−20.

作者简介：

孟祥福，教授，博士，中国计算机学会高级会员。主要研究方向为医学影像分析、时空数据智能、人工智能算法。发表学术论文 46 篇。E-mail：marxi@126.com。

杨雨卓，硕士研究生，主要研究方向为医学图像分析与处理、数据可视化分析。E-mail：lambfishnab@163.com。

张霄雁，副教授，博士，主要研究方向为医学影像、人工智能。E-mail：zhangxiaoyan@lntu.edu.cn。