基于多尺度频域特征与空间注意力的多器官分割网络

doi:10.20269/j.cnki.1674-7100.2026.1010

曾业战[1] 康运成[2]
王 帅[1] 欧阳洪波[1]
钟春良[1] 黄 钊[2]

1. 湖南工业大学
交通与电气工程学院
湖南 株洲 412007
2. 湖南工业大学
生物与医学工程学院
湖南 株洲 412007

摘 要：为解决多器官分割中受大小不一、形状多样、几何结构复杂等因素的影响，提出了一种基于 MFSA-Net（multi-scale frequency spatial attention network）的多器官分割方法。利用多层级、多方向频域分解获取不同尺度的多器官频域特征表达，以有效扩大感受野并提高网络浅层语义特征的辨识性；提出多层次门控注意力机制，实现局部细粒度特征和长距离依赖的融合，聚焦关键目标并抑制背景区域；针对器官的结构差异和多样性，设计了方向增强双分支空间注意力模块，以深度融合边缘像素的空间位置和灰度分布特征，提高模型的空间特性捕获能力。实验结果表明，所提方法可以有效分割尺度差异大、结构复杂的多器官，在 Synapse 和 ACDC 数据集的平均 DSC 分别达到了 81.66% 和 91.61%，优于现有主流方法。

关键词：多器官分割；自注意力机制；频率变换；空间注意力
中图分类号：TP391.41；R318 **文献标志码**：A
文章编号：1674-7100(2026)01-0088-10
引文格式：曾业战，康运成，王 帅，等 . 基于多尺度频域特征与空间注意力的多器官分割网络 [J]. 包装学报，2026，18(1)：88-97.

1 研究背景

在医学图像处理领域，多器官分割是三维重建、计算机辅助诊断和疾病分析的前提，在手术导航、器官移植、术后疗效评估等方面发挥重要作用。临床上，准确的多器官分割一般由经验丰富的医生通过手动反复勾勒来完成。由于 CT 或者 MRI 的切片数量多，手动分割不仅费时费力，而且分割过程易受主观性影响。此外，医学图像中器官弱边界、噪声以及几何形状的多样性，也进一步增加了分割的难度。因此，研究自动、高效的多器官分割方法具有重要意义。

卷积神经网络[1]（CNN）凭借其自动特征提取能力，备受瞩目。以 U-Net[2] 为代表的编码器 - 解码器通过跨层次跳跃连接实现浅层高分辨率特征与深层语义信息的融合，广泛应用于医学图像分割。基于 U-Net 的改进版本，如 U-Net++[3]、Attention U-Net[4]、SAU-Net[5] 以及其他 CNN 方法均获得了不错的性能，并奠定了 CNN 在医学分割中的重要地位。然而，受感受野大小的限制，CNN 方法无法建立长距离依赖，在临床医学的应用受到制约。针对该问题，部分学者提出了大核卷积方法，以拓展感受野，但同时也导致了计算量的增大和细节信息的丢失[6-7]。在此情况下，基于 Transformer[8] 的分割方法逐渐引起了大家的关注。

Transformer 最初用于自然语言处理，通过捕获输入序列的长距离依赖实现对全局上下文建模。随后提出的 Vision Transformer[9] 和 DETR[10] 模型被应用到

收稿日期：2025-10-17
基金项目：湖南省自然科学基金资助项目（2020JJ4276）
作者简介：曾业战，男，讲师，博士，主要研究方向为数字图像处理、人工智能，E-mail：yezhanzeng@126.com
通信作者：钟春良，男，副教授，博士，主要从事微电子器件、光伏电池、微电网研究，E-mail：1769838453@qq.com

二维图像中，并取得了与 CNN 相媲美的性能。为增强多尺度语义表达，Swin-Unet[11] 基于 Swin Transformer[12] 构建 U 型结构，利用滑动窗口注意力获取长距离依赖。MISSFormer[13] 通过引入 ReMix-FFN 来增强 Transfomrer 的上下文桥接模块，以有效提取多尺度特征中的全局依赖与局部上下文。上述网络在建模全局语义方面表现出色，但由于缺乏局部细粒度的细节捕获，在检测小目标、边缘结构等关键区域时面临挑战。为融合 CNN 和 Transformer 两者的优势，许多学者提出了两者的混合架构。Chen J. N. 等 [11] 将 ViT 作为编码器，基于 CNN 构建 U 型解码器，提出了 TransUNet 网络。类似地有，LeViT-Unet[14]、MT-UNet[16]、HiFormer[17] 综合 CNN 和 Transformer 构建网络主干，利用 Transformer 构建注意力机制，实现全局和局部信息的有效提取。此类方法在长距离依赖的特征提取方面有了显著改善，但未充分挖掘图像中丰富的频域信息来强化小目标和边界等重要局部信息。

为深度融合 CNN 和 Transformer，充分挖掘频域和位置信息，本文提出了一种基于多尺度频域特征与空间注意力的多器官分割网络 MFSA-Net（multi-scale frequency spatial attention network）。本文的主要工作如下：

1）针对小尺度器官分割问题，提出基于 Haar 小波的多尺度频域特征增强模块（multi-scale frequency feature enhancement，MFFE）。

2）设计多层级门控注意力（multi-level gated attention，MGA）模块。其在降低计算复杂度的同时，兼顾局部细粒度特征和长距离依赖建模，旨在提高模型的上下文理解能力。

3）提出方向增强双分支空间注意力模块（directional-enhanced dual-branch spatial attention，DEDA）。通过方向梯度编码卷积获取边缘信息，将先验知识融入特征表示；双分支空间注意力模块能深度融合边缘像素的空间位置和灰度分布特征，增强空间邻域相关性。

2 方法

2.1 MFSA-Net 架构

MFSA-Net 架构如图 1 所示，是由编码器、解码器和跳跃连接组成的 U 型结构，包含 MFFE 和 MGA-DEDA 两个基本模块。其中，MFFE 用于扩大感受野和提高不同尺度多器官的检测能力，MGA-DEDA 用于提高模型的空间相关性。

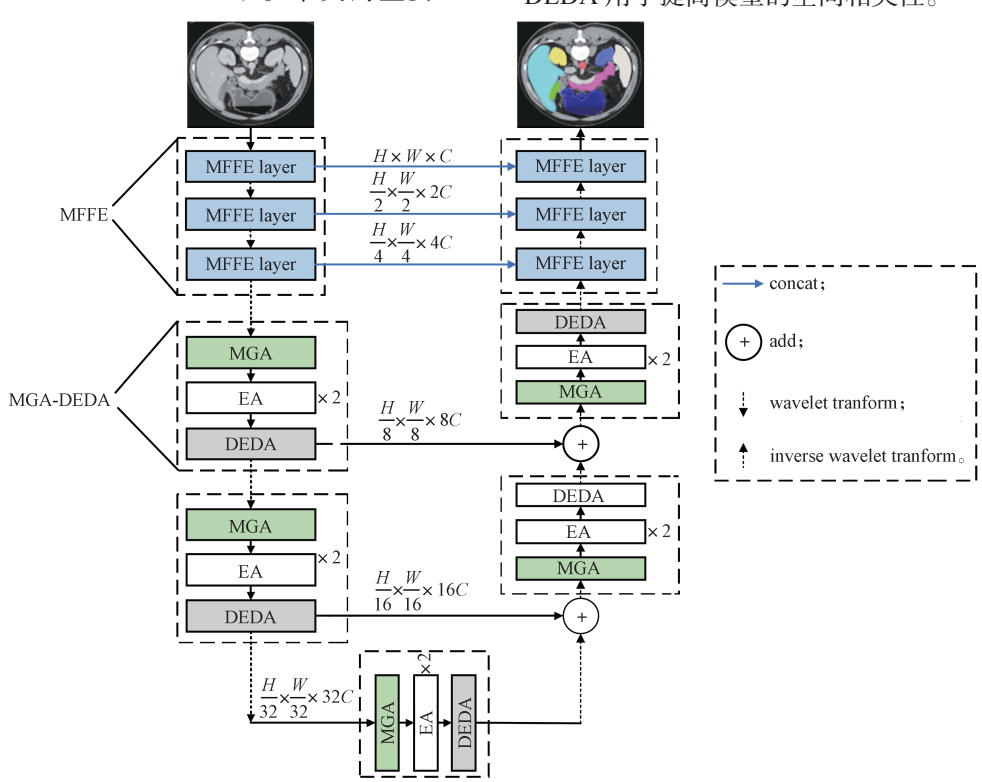

图 1 MFSA-Net 架构

Fig. 1 MFSA-Net overall structure

2.2 MFFE

由于不同器官在图像中存在明显的尺度差异，采用单一尺度的特征提取方法难以全面捕获关键细节和全局信息，且易受图像尺寸或分辨率变化的影响。针对这一问题，利用 Haar 小波多尺度变换[18]的优势，提出了 MFFE 模块（见图 2），以提高不同尺寸器官的检测能力。

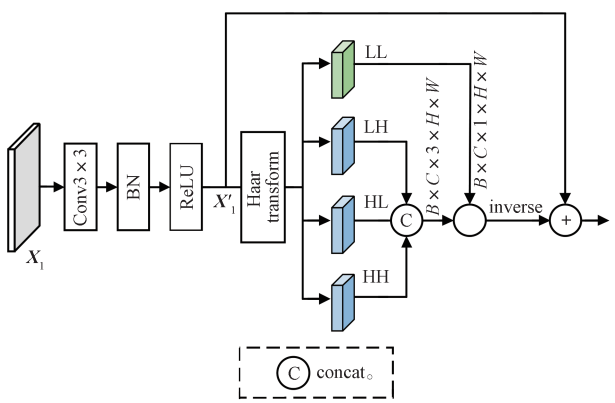

图 2 MFFE 模块

Fig. 2 MFFE block

如图 2 所示，$\boldsymbol{X}_1 \in \mathbb{R}^{B\times C\times H\times W}$ 为输入变量，首先对 $\boldsymbol{X}_1$ 进行 Conv-BN-ReLU 运算得到 $\boldsymbol{X}_1'$，然后，基于 Haar 小波变换将 $\boldsymbol{X}_1'$ 分解为 4 个子带 LL、LH、HL、HH，实现多尺度特征提取和融合。具体步骤如下：

1）沿通道维度对 $\boldsymbol{X}_1'$ 进行水平方向的 Haar 小波变换，即

$$\begin{cases} L_C(i,j)=\dfrac{X_C(i,2j-1)+X_C(i,2j)}{2}, \\ H_C(i,j)=X_C(2i-1,j)-X_C(2i,j)。 \end{cases} \tag{1}$$

式中：$\boldsymbol{X}_C \in \mathbb{R}^{H\times W}$，为 $\boldsymbol{X}_1'$ 在通道 C 的特征矩阵；$L_C(i,j)$ 和 $H_C(i,j)$ 分别为通道 C 的低频系数和高频系数，其中，i 和 j 分别为行和列序号，$1 \leqslant i \leqslant H$，$1 \leqslant j \leqslant W/2$。

2）对 $L_C(i,j)$ 和 $H_C(i,j)$ 进行垂直方向上的 Haar 小波变换，进而获取 4 个特征子带 LL、LH、HL、HH，即

$$\begin{cases} \hat{F}_{\mathrm{LL}}=LL_C(i,j)=\dfrac{L_C(i,2j-1)+L_C(i,2j)}{2}, \\ \hat{F}_{\mathrm{LH}}=LH_C(i,j)=\dfrac{H_C(i,2j-1)+H_C(i,2j)}{2}, \\ \hat{F}_{\mathrm{HL}}=HL_C(i,j)=L_C(i,2j-1)-L_C(i,2j-1), \\ \hat{F}_{\mathrm{HH}}=HH_C(i,j)=H_C(i,2j-1)-H_C(i,2j-1)。 \end{cases} \tag{2}$$

式中：$\hat{F}_{\mathrm{LL}}$、$\hat{F}_{\mathrm{LH}}$、$\hat{F}_{\mathrm{HL}}$、$\hat{F}_{\mathrm{HH}}$ 分别为低频子带、水平高频子带、垂直高频子带以及对角线高频子带系数；$1 \leqslant i \leqslant H/2$；$1 \leqslant j \leqslant W/2$。

3）对 $\hat{F}_{\mathrm{LH}}$、$\hat{F}_{\mathrm{HL}}$、$\hat{F}_{\mathrm{HH}}$ 进行拼接、归一化和 1×1 卷积运算，实现高频子带融合，得到高频向量 $\boldsymbol{X}_{\mathrm{H}}$；对低频子带 $\hat{F}_{\mathrm{LL}}$ 直接进行归一化和 1×1 卷积运算，得到 $\boldsymbol{X}_{\mathrm{L}}$。

$$\begin{cases} \boldsymbol{X}_{\mathrm{H}}=\varphi(conv_{1\times1}(concat(\hat{F}_{\mathrm{LH}},\hat{F}_{\mathrm{HL}},\hat{F}_{\mathrm{HH}}))), \\ \boldsymbol{X}_{\mathrm{L}}=\varphi(conv_{1\times1}(\hat{F}_{\mathrm{LL}}))。 \end{cases} \tag{3}$$

式中，$\varphi(\cdot)$ 表示归一化。

4）为融合高频和低频特征，将 $\boldsymbol{X}_{\mathrm{H}}$ 和 $\boldsymbol{X}_{\mathrm{L}}$ 进行拼接，并基于残差思想将拼接结果与 $\boldsymbol{X}_1'$ 相加，得到 MFEE 的最终输出。

2.3 MGA-DEDA

考虑到位置信息的提取有助于复杂的医学图像分割，为提高算法精度和鲁棒性，基于空间相关性设计了 MGA-DEDA 模块。如图 3 所示，MGA-DEDA 由 MGA、外部注意力[19]（external attention，EA）和方向增强双分支空间注意力（DEDA）3 部分组成。其中，MGA 用于提高模型对关键结构区域的关注力；EA 模块用于突破单一样本视角的局限；DEDA 模块用于通过双分支结构强化模型的空间位置一致性和连续性。

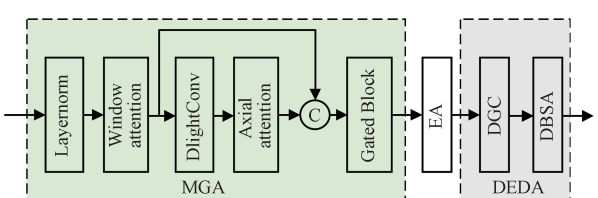

图 3 MGA-DEDA 模块

Fig. 3 MGA-DEDA block

2.3.1 MGA 模块

由于位置毗邻和灰度相似，多器官的浅层特征在语义上高度相似，易出现误分割现象。针对这一问题，综合窗口注意力机制[13]、轴向注意力机制[20]、门控机制[21]设计 MGA 模块，实现从局部细节到全局特征的多层次特征重构与选择。MGA 模块结构如图 4 所示。

为均衡计算效率和局部特征（如边界等）提取，基于窗口注意力机制将尺寸为 $H\times W$ 的特征图划分为 $\dfrac{H}{w}\times\dfrac{W}{w}$ 个窗口，每个窗口大小为 $w\times w$。通过并行算法将窗口注意力的计算复杂度从 $O(n^2)$ 降至 $O(w^2\times n)$。

为构建像素间长距离依赖，引入轴向注意力机

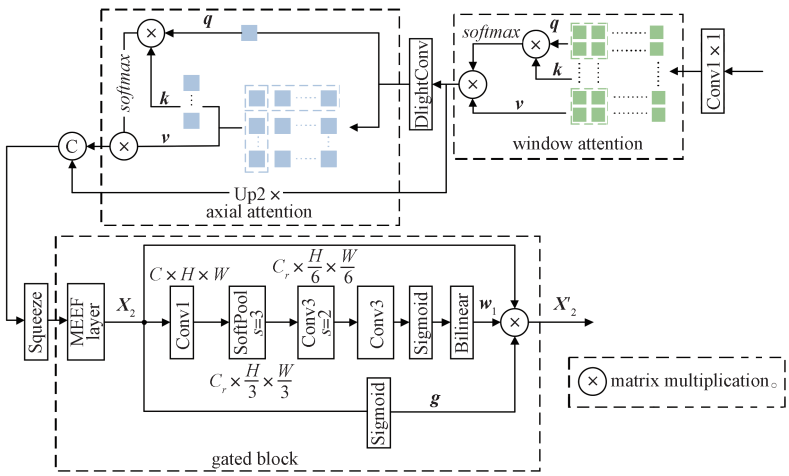

图 4　多层级门控注意力模块

Fig. 4　Multi-level gated attention block

制。沿行和列方向计算注意力，即

$$\begin{cases} A_{\text{row}} = softmax(\boldsymbol{Q}_{\text{row}}\boldsymbol{K}_{\text{row}}^{\text{T}})\boldsymbol{V}_{\text{row}}, \\ A_{\text{col}} = softmax(\boldsymbol{Q}_{\text{col}}\boldsymbol{K}_{\text{col}}^{\text{T}})\boldsymbol{V}_{\text{col}}。 \end{cases} \tag{4}$$

式中：$\boldsymbol{Q}_{\text{row}}$、$\boldsymbol{K}_{\text{row}}$、$\boldsymbol{V}_{\text{row}}$ 分别为行方向展开的查询（query）、键（key）和值（value）向量；$\boldsymbol{Q}_{\text{col}}$、$\boldsymbol{K}_{\text{col}}$、$\boldsymbol{V}_{\text{col}}$ 分别为列方向展开的 query、key 和 value 向量。

鉴于双线性插值和步长卷积易引入伪影 [22]，采用门控模块（gate block）重新校准注意力图，实现背景和干扰信息抑制。门控机制首先通过式(5)和(6)计算注意力特征向量 $\boldsymbol{w}_1$ 和门控信号 g。

$$\boldsymbol{w}_1=\sigma(\psi(conv3(conv2(softmax(Conv1(\boldsymbol{X}_2)))))), \tag{5}$$

$$g=\sigma(\boldsymbol{X}_2), \tag{6}$$

式中：$\boldsymbol{X}_2\in\mathbb{R}^{C\times H\times W}$ 为 MFFE 输出；$\sigma(\cdot)$ 为 Sigmoid 函数；$\psi(\cdot)$ 为双线性插值。

然后，通过非线性权重分配方法聚焦特征图的关键信息 [23]，即

$$\omega_i = \frac{\mathrm{e}^{a_i}}{\sum\limits_{j\in R}\mathrm{e}^{a_j}}, \tag{7}$$

$$\boldsymbol{y}_i = \sum_{k\in R}\sum_{i\in R_k}\frac{\mathrm{e}^{X_i}}{\sum\limits_{j\in R_k}\mathrm{e}^{X_j}}\cdot\omega_k, \tag{8}$$

式中：$\boldsymbol{y}_i$ 为 SoftPool 输出；a_i 为激活因子；$\boldsymbol{X}_i$ 为输入特征图中第 i 个空间位置处的特征向量；$\boldsymbol{X}_j$ 为位于其邻域内的特征向量；R_k 为以 $\boldsymbol{X}_i$ 为中心、大小为 $k\times k$ 的局部空间区域；ω_k 为计算可学习权重；i 和 j 分别为激活区域和像素序号。

最后，将 $\boldsymbol{w}_1$、$\boldsymbol{g}$ 和 $\boldsymbol{X}_2$ 逐元素相乘后，获得门控模块的最终输出，

$$\boldsymbol{X}_2' = \boldsymbol{X}_2\odot\boldsymbol{w}_1\odot\boldsymbol{g}。 \tag{9}$$

2.3.2　EA 模块

自注意力机制主要关注单个样本内各元素之间的关系，忽略了不同样本之间的相关性 [19]。针对这一问题，在 MGA-DEDA 模块中引入可学习的外部键记忆单元 $\boldsymbol{M}_{\text{k}}$ 与外部值记忆单元 $\boldsymbol{M}_{\text{v}}$（$\boldsymbol{M}_{\text{k}}\in\mathbb{R}^{C\times S}$，$\boldsymbol{M}_{\text{v}}\in\mathbb{R}^{S\times C}$，其中 S 为记忆单元数目，$S\ll N$），构建 EA，将数据集的全局结构信息融入特征表示中，如图 5 所示。外部注意力分布矩阵（$\boldsymbol{E}\in\mathbb{R}^{B\times N\times C}$）可表示为

$$\boldsymbol{E}=softmax(\boldsymbol{X}_3\boldsymbol{M}_{\text{k}}), \tag{10}$$

式中，$\boldsymbol{X}_3\in\mathbb{R}^{B\times N\times C}$ 为输入特征。

外部注意力的输出（$\boldsymbol{Y}\in\mathbb{R}^{B\times N\times C}$）可表示为

$$\boldsymbol{Y}=\boldsymbol{E}\boldsymbol{M}_{\text{v}}。 \tag{11}$$

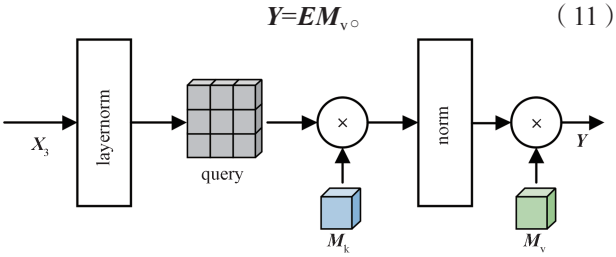

图 5　EA 模块

Fig. 5　EA block

2.3.3　DEDA 模块

为有效分割器官结构变化剧烈的区域，提出基于空间位置一致性的 DEDA 模块，如图 6 所示。DEDA 由方向梯度编码卷积（directional gradient-encoded convolution，DGC）和双分支空间注意力（dual-branch spatial attention，DBSA）组成。

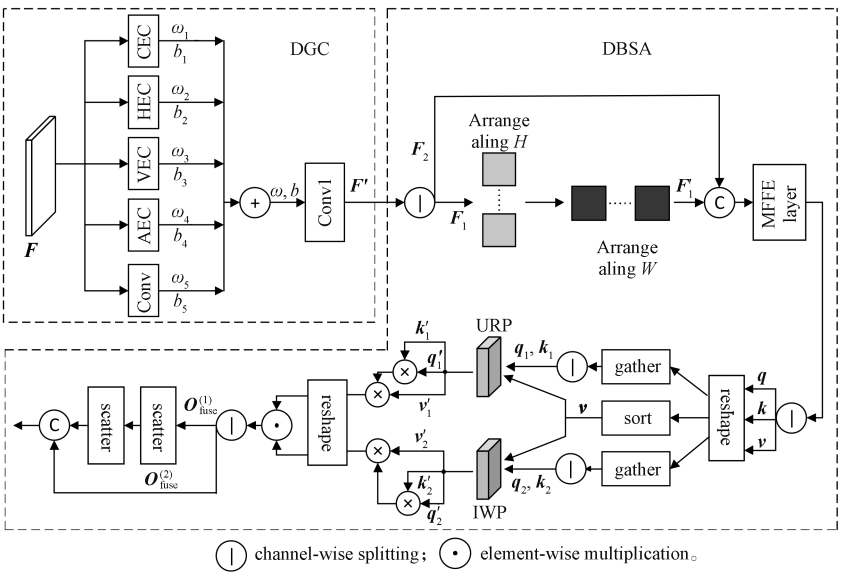

图 6　方向增强双分支空间注意力

Fig. 6　Directional-enhanced dual-branch spatial attention

1）DGC

DGC 主要用于强化局部边缘和空间结构表达，通过在水平、垂直、对角等方向，以及中心梯度编码提取局部像素梯度信息，并将其融入卷积核中，达到增强特征表示和泛化能力的目的。以中心梯度编码为例，设 p_0 为卷积核中心坐标，编码输出为

$$y(p_0)=\sum_{P_n\in R}\omega(p_n)[x(p_0+p_n)-x(p_0)] \tag{12}$$

式中：R 为 p_0 的邻域；$\omega(\cdot)$ 为卷积核权重；$x(\cdot)$ 为灰度值。基于残差思想，将原始特征 $\boldsymbol{F}$ 相加后得到特征 $\boldsymbol{F}'$（$\boldsymbol{F}'\in\mathbb{R}^{B\times C\times H\times W}$），

$$\boldsymbol{F}'=\boldsymbol{F}+\alpha\times DCG(\boldsymbol{F})。 \tag{13}$$

2）DBSA

DBSA 通过等量区域划分分支（uniform region partitioning，URP）和强度相似分支（intensity-wise partitioning，IWP）双分支空间划分策略对输入特征进行子区域划分和自注意力建模，从而提升模型对多器官复杂解剖结构的空间建模能力。为扩大感受野，对特征图 $\boldsymbol{F}'$ 进行拆分（$split(\cdot)$）、排序（$sort(\cdot)$）和拼接（$concat(\cdot)$）运算后[24]，通过 MFFE 层映射为 $\boldsymbol{q}$、$\boldsymbol{k}$、$\boldsymbol{v}$（$\boldsymbol{q}$、$\boldsymbol{k}\in\mathbb{R}^{B\times 2C\times \mathrm{H}\times W}$，$\boldsymbol{v}\in\mathbb{R}^{B\times C\times \mathrm{H}\times W}$），即

$$\begin{cases}\boldsymbol{F}_1,\boldsymbol{F}_2=split(\boldsymbol{F}');\\ \boldsymbol{F}_1'=sort_W(sort_H(\boldsymbol{F}_1));\\ \boldsymbol{q},\boldsymbol{k},\boldsymbol{v}=MFFE(concat(\boldsymbol{F}',\boldsymbol{F}_2))。\end{cases} \tag{14}$$

式中：$\boldsymbol{F}_1$ 为 $\boldsymbol{F}'$ 的前半通道部分；$\boldsymbol{F}_2$ 为 $\boldsymbol{F}'$ 的后半通道部分。

构建查询－键对，即

$$\begin{cases}\boldsymbol{q}_0,\boldsymbol{q}_0'=split(\boldsymbol{q});\\ \boldsymbol{k}_0,\boldsymbol{k}_0'=split(\boldsymbol{k});\\ \boldsymbol{v},\boldsymbol{idx}=sort(\boldsymbol{v});\\ \boldsymbol{q}_1,\boldsymbol{k}_1=chunk(gather((\boldsymbol{q}_0,\boldsymbol{k}_0),idx));\\ \boldsymbol{q}_2,\boldsymbol{k}_2=chunk(gather((\boldsymbol{q}_0',\boldsymbol{k}_0'),idx))。\end{cases} \tag{15}$$

式中：$chunk(\cdot)$ 为均分操作；$gather(\cdot)$ 为张量元素重排运算；idx 为排序索引。

此外，针对多器官边缘的不连续性和弱边界，以及解剖结构位置不一致，采用URP和IWP双分支结构。

URP 分支用于提高空间上分散多器官（如肾脏）的分割性能。首先，将特征图等量划分为 s 个子区域，每个子区域空间大小为（H/s、W/s），并将 $\boldsymbol{q}_1$、$\boldsymbol{k}_1$、$\boldsymbol{v}_1\in\mathbb{R}^{B\times(N_h\times c)\times(s\times h\times w)}$ 映射为 $\boldsymbol{q}_1'$、$\boldsymbol{k}_1'$、$\boldsymbol{v}_1'\in\mathbb{R}^{B\times N_h\times(c\times s)\times(h\times w)}$，其中，$N_h$ 为注意力的头数，c 为单头通道数，s 为子区域数量，h 为子块的高度，w 为子块的宽度。URP 的空间注意力权重 $\boldsymbol{\alpha}_{\mathrm{URP}}$ 和最终输出 $\boldsymbol{O}_{\mathrm{UPR}}$ 可表示为：

$$\boldsymbol{\alpha}_{\mathrm{URP}}=softmax\left(\frac{\boldsymbol{q}_1'(\boldsymbol{k}_1')^{\mathrm{T}}}{\sqrt{d}}\right),\ d=c\times s\ , \tag{16}$$

$$\boldsymbol{O}_{\mathrm{UPR}}=\boldsymbol{\alpha}_{\mathrm{URP}}\boldsymbol{v}_1', \tag{17}$$

式中，d 为注意力缩放因子。

IWP 分支考虑了器官的灰度相似性。先对输入特征中的像素按灰度值进行排序见式（15），从而形成灰度一致的特征子空间；然后，将 $\boldsymbol{q}_2$、$\boldsymbol{k}_2$、$\boldsymbol{v}_2\in\mathbb{R}^{B\times(N_h\times c)\times(h\times w\times s)}$ 映射为 $\boldsymbol{q}_2'$、$\boldsymbol{k}_2'$、$\boldsymbol{v}_2'\in\mathbb{R}^{B\times N_h\times(c\times s)\times(h\times w)}$，便于在每个子区间内独立执行注意力机制。IWP 分支是基于灰度特

征而非空间位置，因此，其能够有效降低解剖结构位置差异或器官的空间位置分散（如肾脏）带来的影响。IWP 分支的空间注意力权重 $\boldsymbol{\alpha}_{\mathrm{IWP}}$ 和最终输出 $\boldsymbol{O}_{\mathrm{IWP}}$ 可表示为：

$$\boldsymbol{\alpha}_{\mathrm{IWP}} = softmax\left(\frac{\boldsymbol{q}_2'(\boldsymbol{k}_2')^{\mathrm{T}}}{\sqrt{d}}\right),\ d = c\times s\ , \tag{18}$$

$$\boldsymbol{O}_{\mathrm{IWP}} = \boldsymbol{\alpha}_{\mathrm{IWP}}\boldsymbol{v}_2'\ 。 \tag{19}$$

为有效融合 URP 和 IWP，首先将 URP 和 IWP 在通道维度上进行重排，并通过逐元素相乘的方式实现空间上下文的交叉和融合；然后，对交叉特征进行均分，生成 $\boldsymbol{O}_{\mathrm{fuse}}^{(1)}$ 和 $\boldsymbol{O}_{\mathrm{fuse}}^{(2)}$，$\boldsymbol{O}_{\mathrm{fuse}}^{(1)}$ 经过重排序还原其原始空间位置，得到 $\boldsymbol{O}_{\mathrm{rep}}$，$\boldsymbol{O}_{\mathrm{fuse}}^{(2)}$ 与 $\boldsymbol{O}_{\mathrm{rep}}$ 拼接后，得到最终的融合结果 $\boldsymbol{O}_{\mathrm{final}}$。

$$\boldsymbol{O}_{\mathrm{rep}} = scatter_W\left(scatter_H\left(\boldsymbol{O}_{\mathrm{fuse}}^{(1)}\right)\right), \tag{20}$$

$$\boldsymbol{O}_{\mathrm{final}} = concat\left(\boldsymbol{O}_{\mathrm{rep}}, \boldsymbol{O}_{\mathrm{fuse}}^{(2)}\right)。 \tag{21}$$

2.4 损失函数

为克服交叉熵损失产生的细节信息缺失[25]，综合交叉熵损失 L_{CE} 和 Dice 损失 L_{D} 构建损失函数，即

$$L = 0.5L_{\mathrm{CE}} + 0.5L_{\mathrm{D}}, \tag{22}$$

$$L_{\mathrm{CE}} = -\frac{1}{P}\sum_{p=1}^{P}\sum_{k=1}^{K} G_{p,k}\cdot\log\left(G_{p,k}'\right), \tag{23}$$

$$L_{\mathrm{D}} = 1-\frac{1}{K}\sum_{k=1}^{K}\frac{2\sum_{p=1}^{P} G_{p,k}G_{p,k}'}{\sum_{p=1}^{P} G_{p,k}+\sum_{p=1}^{P} G_{p,k}'}, \tag{24}$$

式中，$G_{p,k}$ 和 $G'_{p,k}$ 分别为第 p 个像素属于第 k 类的真实标签和预测概率。

3 实验结果与分析

3.1 实验数据集与评估指标

实验用 Synapse、ACDC（Automated Cardiac Diagnosis Challenge）医学图像数据集。其中，Synapse 数据集包含 30 个案例，共 3779 张腹部临床 CT 图像，涵盖主动脉（AO）、胆囊（GA）、左肾（LK）、右肾（RK）、肝脏（LI）、胰腺（PA）、脾脏（SP）和胃（ST）8 个器官，切片层厚为 2.5~5.0 mm，大小为 512×512。此数据库中，选用 18 个案例用于训练，12 个案例用于测试。ACDC 数据集包含 100 例心脏检查数据，涵盖左心室（LV）、心肌（MYO）、右心室（RV）3 个标注。此数据库中，随机选取 70 个训练样本、20 个测试样本和 10 个验证样本。

采用戴斯相似性系数（Dice similarity coefficient，DSC）、95% 豪斯多夫距离（Hausdorff distance，HD95）进行评价。其中，DSC 值越高，则表示预测分割结果与真实标注之间的重叠程度越大；HD95 值越低，则表示分割边界与真实边界之间的最大偏差越小，从而表明模型分割性能越优。

3.2 实验平台与设置

实验是在 NVIDIA Geforce RTX3060、Ubuntu 22.04.2、PyTorch 2.4.0 的开发环境中进行。为了验证所提模型的有效性，在 Synapse 和 ACDC 数据集上将其与多种先进方法进行比较。实验分为训练和测试两个阶段。在训练过程中，输入图像尺寸设置为 224×224，小波变换分解层级数设置为 3，模型的初始学习率和权重衰减均设置为 0.0001，批次数设置为 4，最大训练轮数设置为 120。将训练集中的 CT 或者 MRI 图像输入 MFSA-Net 模型，并通过定义的损失函数计算模型预测结果与真实标注之间的误差；然后，采用反向传播算法将误差信息逐层传递至各网络层，根据其对整体误差的贡献程度调整相应的权重参数，权重更新过程由 Adam 优化器实现。在测试阶段，将已训练好的模型对测试集中的 CT 或者 MRI 图像进行预测，并结合真实标签，采用不同的评估指标来综合评估模型的分割性能。

3.3 不同方法的实验结果分析

3.3.1 不同方法在 Synapse 数据集上的结果对比

表 1 给出了不同方法在 Synapse 数据集中的分割性能比较。由表 1 可以看出，受限于尺寸固定的卷积核，V-Net、DARR、R50-UNet、UNet、AttnUNet 等基于 CNN 模型的分割模型通过跳跃连接缓解梯度消失，但由于无法建立长距离关系，最高的平均 DSC 仅为 77.77%。而基于 Transformer 建立长距离依赖的 TransUNet、TransClaw UNet、MT-Unet 等模型，DSC 有一定上升。本模型平均 DSC 达到 81.66%，优于现有主流算法，且 GA、LK 和 PA 的 DSC 分别为 69.91%、85.85% 和 64.33%，均为最优，这验证了本模型的有效性。

图 7 为不同方法的 CT 图像可视化分割结果。图中，第 1 列为手动分割结果，第 2 列到第 7 列分别为 TransUnet、MissFormer、HiFormer、SelfReg-UNet、ParaTransCNN 和本模型的分割结果。从图 7 第 1 行

表 1　不同方法在 Synapse 数据集上的 DSC 和 HD95 结果

Table 1　DSC and HD95 results of different methods on the Synapse dataset

模型	HD95/mm	平均 DSC/%	DSC/%							
			AO	GA	LK	RK	LI	PA	SP	ST
V-Net[26]		68.81	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR[27]		69.77	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50-UNet[2]	36.87	74.68	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50-AttnUNet[14]	36.97	72.10	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
UNet[2]	39.70	76.85	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
AttnUNet[28]	36.02	77.77	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
R50 ViT[9]	32.87	71.29	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
ViT[9]	39.61	61.50	44.38	39.59	67.46	62.94	89.21	43.14	75.45	69.78
TransUNet[29]	31.69	77.48	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
TransClaw UNet[15]	26.38	78.09	85.87	61.38	84.83	79.36	94.28	57.65	87.74	73.55
LeVit-Unet-384[16]	16.84	78.53	87.33	62.23	84.61	80.25	93.11	59.07	88.86	72.76
MT-UNet[27]	26.59	78.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
Swin-Unet[12]	21.55	79.13	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
HiFormer-B[17]	14.70	80.39	86.21	65.69	85.23	79.77	94.61	59.52	90.99	81.08
MissFormer[13]	17.93	80.43	86.24	67.37	84.68	82.00	93.79	61.01	89.99	78.42
SelfReg-UNet[30]		80.54	86.07	69.65	85.12	82.58	94.18	61.08	87.42	78.22
ParaTransCNN[31]	18.12	80.82	87.83	67.30	84.88	81.49	94.08	63.70	89.00	78.29
本模型	23.11	81.66	88.22	69.91	85.85	82.30	94.22	64.33	89.21	79.24

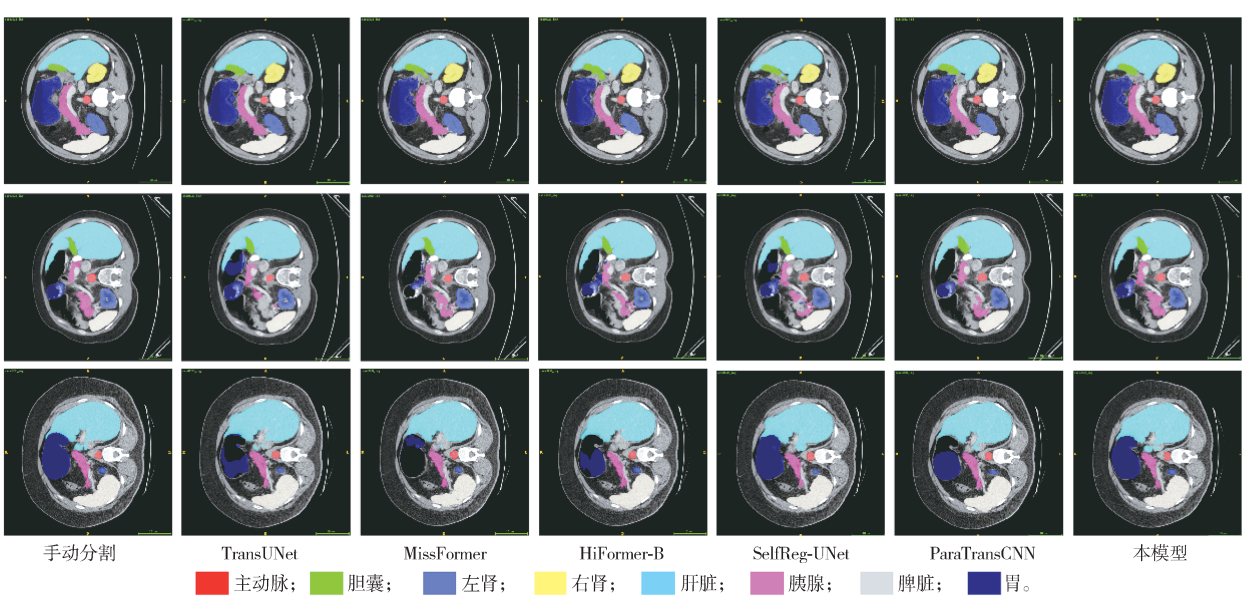

图 7　不同方法的 Synapse 可视化结果

Fig. 7　Visual results of various methods on Synapse

可知，MissFormer、HiFormer 和 SelfReg-UNet 在分割胆囊等小器官时，没有准确捕获边界；TransUNet 将胆囊区域错识别为胰腺；本模型获得了胆囊区域的准确分割。从图 7 第 2 行可知，对于形态差异大且体积较小的狭长形胰腺区域，本模型与手动分割非常接近，而其他方法出现了欠分割。从图 7 第 3 行展示的灰度值低且结构复杂的胃部分割结果可知，本模型能有效分割复杂形状的胃部和肝脏区域，且结果明显优于除 SelfReg-UNet 的其他方法。

以上结果可归因于本模型借助 Haar 小波变换获取多器官在不同尺度、不同方向上的频域特征表征，实现了局部细粒度特征与长距离依赖的有效融合，同

时多层级门控注意力机制对背景区域起到抑制作用。此外，双分支空间注意力结构推动了边缘特征与灰度分布的深度融合，进一步提升了对具有不规则形状和模糊边界器官的识别精度。

3.3.2 不同方法在 ACDC 数据集上的结果对比

表 2 给出了不同方法在 ACDC 数据集中的分割结果。从表 2 可以看出，本模型分割 RV、MYO、LV 的 DSC 分别达到了 89.39%、89.70% 和 95.75%，平均 DSC 为 91.61%。其中，本模型对 MYO、LV 的分割结果为最佳，对 RV 的分割结果，与 LeVit-Unet-384[15] 的接近。这进一步证实了本模型的有效性。

表 2 不同方法在 ACDC 数据集上的 DSC 结果

Table 2 DSC results of different methods on the ACDC dataset %

模型	平均 DSC	DSC		
		RV	MYO	LV
R50-UNet[2]	87.55	87.10	80.63	94.92
R50-AttnUNet[14]	86.75	87.58	79.20	93.47
ViT-CUP[9]	83.41	80.93	78.12	91.17
R50 ViT[9]	86.19	82.51	83.01	93.05
MissFormer[13]	87.90	86.36	85.75	91.59
TransUNet[14]	89.67	86.57	87.27	95.18
LeVit-Unet-384[15]	90.32	89.55	87.64	93.76
Swin-Unet[12]	90.42	88.41	87.71	95.13
MT-Unet[16]	90.43	86.64	89.04	95.62
SegFormer3D[31]	90.96	88.50	88.86	95.53
本模型	91.61	89.39	89.70	95.75

3.4 消融试验

3.4.1 MFFE、MGA、DEDA 的影响分析

表 3 为本模型中 MFFE、MGA 和 DEDA 各模块对分割性能的影响。为保证算法的公平性，保持网络框架和训练参数不变，将验证模块进行替换。

表 3 MFSA-Net 各子模块对分割性能的影响

Table 3 Impacts of MFSA-Net sub-modules on segmentation performance

MFFE	MGA	DEDA	DSC/%	HD95/mm
			78.10	28.84
√			79.28	21.91
√	√		80.53	22.21
√	√	√	81.66	23.11

从表 3 可以看出，MFFE 的增加扩大了感受野并实现了多尺度特征提取，故 DSC 从 78.10% 提高到 79.28%，同时，HD95 由 28.84 mm 降至 21.91 mm。引入 MGA 模块后，增强了局部窗口细节特征和轴向的长距离依赖关系，DSC 上升了 1.25%。在此基础上增加 DEDA，模型对局部边缘细节与空间结构的敏感度提高，DSC 上升至 81.66%。同时还发现，受大噪声、极低对比度的影响，DEDA 对边缘敏感性放大了弱边界处的细微位置差异，导致 HD95 的值略微上升。

3.4.2 小波分解层级数的影响分析

Haar 小波分解层级数决定了频域中被划分的深度，与模型捕获的多尺度结构和边缘特征密切关联。为分析小波分解系数对多器官分割性能的影响，对 MFFE、MGA、DEDA 中不同的小波分解层级数进行了分析比较。MFFE 模块处于网络主干浅层，其主要目的是提取浅层特征，并不需要过高的分解层级。因此，综合考虑频率分解的细粒度和计算成本，将 MFFE 模块的层数设置为 3。位于深层网络的 DEDA 和 MGA，其小波分解层级数对模型整体性能的影响如表 4 所示。

表 4 不同小波分解层级数对模型的影响

Table 4 Impacts of wavelet decomposition levels on model performance

MFFE/ 层	DEDA/ 层	MGA/ 层	DSC/%
3	3	3	80.56
3	3	5	81.66
3	5	5	78.91

由表 4 可知，当 MGA 模块的 Haar 小波分解层级数由 3 增加到 5 时，模型的整体 DSC 由 80.56% 显著提升至 81.66%。这表明 MGA 模块中采用的多层次门控注意力机制能够有效利用更深层次的小波频域信息，捕捉更广泛的语义上下文。因此，将 MGA 的 Haar 小波分解层级数设置为 5。在此基础上，将 DEDA 模块的分解层级数从 3 提高到 5。此时，DSC 出现下滑，其主要原因是过多的小波分解层级数易导致空间信息的冗余和混乱，不利于边界细节信息的准确获取。综上，将 MFFE、MGA、DEDA 的 Haar 小波分解层级数分别设置为 3, 3, 5。

4 结论

针对医学图像多器官分割任务中不同器官差异大、结构形态变化复杂等问题，本文提出了融合多尺度频域增强特征与空间增强注意力机制的高效分割网络——MFSA-Net。该方法设计了 3 个具有互补功

能的关键模块 MFFE、MGA 和 DEDA，分别从频域多尺度建模、语义上下文融合与空间结构增强 3 个角度提升了模型的整体性能。在 Synapse 和 ACDC 两个医学图像多器官分割数据集上开展的实验结果表明，所提方法可以有效分割尺度差异大、结构复杂的多器官，在 Synapse 和 ACDC 数据集的平均 DSC 分别达到了 81.66% 和 91.61%，优于现有主流方法。

参考文献：

[1] LE CUN Y，BOSER B，DENKER J S，et al. Handwritten Digit Recognition with a Back-Propagation Network[C]// Proceedings of the 3rd International Conference on Neural Information Processing Systems. [S. l.]：ACM，1989：396−404.

[2] RONNEBERGER O，FISCHER P，BROX T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]//Medical Image Computing and Computer-Assisted Intervention(MICCAI). Cham：Springer，2015：234−241.

[3] 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，2018：3−11.

[4] OKTAY O，SCHLEMPER J，LE FOLGOC L，et al. Attention U-Net: Learning where to Look for the Pancreas[EB/OL]. 2018：arXiv: 1804.03999. [2024−04−20]. https://arxiv.org/abs/1804.03999.

[5] 张淑军，彭 中，李 辉 . SAU-Net：基于 U-Net 和自注意力机制的医学图像分割方法 [J]. 电子学报，2022，50(10)：2433−2442.

[6] DING X H，ZHANG X Y，ZHOU Y Z，et al. Scaling Up Your Kernels to 31 × 31: Revisiting Large Kernel Design in CNNs[EB/OL]. 2022：arXiv:2203.06717. [2024−05−20]. https://arxiv.org/pdf/2203.06717.

[7] LIU Z，MAO H Y，WU C Y，et al. A Convnet for the 2020s[EB/OL]. 2022：arXiv:2201.03545. [2024−06−10]. https://arxiv.org/pdf/2201.03545.

[8] VASWANI A，SHAZEER N，PARMAR N，et al. Attention Is All You Need[J/OL]. 2017：arXiv:1706.03762. [2024−07−03]. https://arxiv.org/abs/1706.03762.

[9] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale[J/OL]. 2020：arXiv: 2010.11929. [2024−09−11]. https://arxiv.org/abs/2010.11929.

[10] CARION N，MASSA F，SYNNAEVE G，et al. End-to-End Object Detection with Transformers[C]//European Conference on Computer Vision. Cham：Springer，2020：213−229.

[11] CAO H，WANG Y Y，CHEN J，et al. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation[C]//European Conference on Computer Vision. Cham：Springer，2023：205−218.

[12] LIU Z，LIN Y T，CAO Y，et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows[C]//International Conference on Computer Vision (ICCV). Montreal：IEEE，2021：10012−10022.

[13] HUANG X H，DENG Z F，LI D D，et al. MISSFormer: An Effective Medical Image Segmentation Transformer[J/OL]. 2021：arXiv: 2109.07162. [2024−10−18]. https://arxiv.org/abs/2109.07162.

[14] CHEN J N，LU Y Y，YU Q H，et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation[J/OL]. 2021：arXiv: 2102.04306. [2025−03−15]. https://arxiv.org/abs/2102.04306.

[15] XU G P，ZHANG X，HE X W，et al. LeViT-UNet: Make Faster Encoders with Transformer for Medical Image Segmentation[C]//Pattern Recognition and Computer Vision. Singapore：Springer，2024：42−53.

[16] WANG H Y，XIE S，LIN L F，et al. Mixed Transformer U-Net for Medical Image Segmentation[C]//2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S. l.]：IEEE，2022：2390−2394.

[17] 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 (WACV). Waikoloa：IEEE，2023：6191−6201.

[18] FINDER S E，AMOYAL R，TREISTER E，et al. Wavelet Convolutions for Large Receptive Fields[C]// European Conference on Computer Vision(ECCV). Cham：Springer，2025：363−380.

[19] GUO M H，LIU Z N，MU T J，et al. Beyond Self-Attention: External Attention Using Two Linear Layers for Visual Tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2023，45(5)：5436−5447.

[20] HO J，KALCHBRENNER N，WEISSENBORN D，et al. Axial Attention in Multidimensional Transformers[J/OL]. 2019：arXiv: 1912.12180. [2025−01−15]. https://arxiv.org/abs/1912.12180.

[21] 宋艳涛，路云里 . SwinT-Unet：基于双通道自注意力机制的超声图像分割方法 [J]. 电子学报，2024，52(11)：3835−3846.

[22] WANG Y，LI Y S，WANG G，et al. PlainUSR: Chasing Faster ConvNet for Efficient Super-Resolution[C]// Proceedings of the Asian Conference on Computer Vision(ACCV). Singapore：Springer，2025：246−264.
[23] STERGIOU A，POPPE R，KALLIATAKIS G. Refining Activation Downsampling with SoftPool[C]//International Conference on Computer Vision (ICCV). Montreal：IEEE，2022：10337−10346.
[24] SUN S Q，REN W Q，GAO X W，et al. Restoring Images in Adverse Weather Conditions via Histogram Transformer[C]//European Conference on Computer Vision(ECCV). Cham：Springer，2025：111−129.
[25] 王梦溪，雷　涛，姜由涛，等 . 基于空频协同的 CNN-Transformer 多器官分割网络 [J/OL]. 智能系统学报，2025：1−16. https://kns.cnki.net/kcms/detail/23.1538.tp.20250613.1819.002.html.
[26] MILLETARI F，NAVAB N，AHMADI S A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation[C]// 2016 Fourth International Conference on 3D Vision (3DV). Stanford：IEEE，2016：565−571.
[27] FU S H，LU Y Y，WANG Y，et al. Domain Adaptive Relational Reasoning for 3D Multi-Organ Segmentation[C]//Medical Image Computing and Computer Assisted Intervention(MICCAI). Cham：Springer，2020：656−666.
[28] SCHLEMPER J，OKTAY O，SCHAAP M，et al. Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images[J]. Medical Image Analysis，2019，53：197−207.
[29] YAO C，HU M H，LI Q L，et al. Transclaw U-Net: Claw U-Net with Transformers for Medical Image Segmentation[C]//2022 5th International Conference on Information Communication and Signal Processing (ICICSP). Shenzhen：IEEE，2023：280−284.
[30] 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). Cham：Springer，2024：601−611.
[31] PERERA S，NAVARD P，YILMAZ A. SegFormer3D: An Efficient Transformer for 3D Medical Image Segmentation[J/OL]. 2024：arXiv: 2404.10156. [2025−03−25]. https://arxiv.org/abs/2404.10156.

（责任编辑：邓 彬）

A Multi-Organ Segmentation Network Based on Multi-Scale Frequency Features and Spatial Attention

ZENG Yezhan[1]，KANG Yuncheng[2]，WANG Shuai[1]，OUYANG Hongbo[1]，
ZHONG Chunliang[1]，HUANG Zhao[2]

（1. School of Transportation and Electrical Engineering，Hunan University of Technology，Zhuzhou Hunan 412007，China；
2. School of Biological Science and Medical Engineering，Hunan University of Technology，Zhuzhou Hunan 412007，China）

Abstract：To address the influences of factors such as different sizes, diverse shapes, and complex anatomical structures in multi-organ segmentation, a novel multi-organ segmentation method based on MFSA-Net (multi-scale frequency spatial attention network) is proposed. The network utilizes multi-level and multi-directional frequency decomposition to extract frequency-domain features of organs at different scales, which effectively expands the receptive field and enhances the discriminability of shallow semantic features. Moreover, a multi-level gated attention mechanism is introduced to achieve the integration of local fine-grained features and long-range dependencies, enabling the network to focus on critical regions while suppressing background noise. To address the structural diversity and variation among organs, a direction-enhanced dual-branch spatial attention module is designed to deeply integrate the spatial position and gray-scale distribution features of edge pixels, thereby improving the model's spatial representation capability. Experimental results demonstrate that the proposed method can effectively segment multiple organs with large scale variations and complex structures, achieving average Dice similarity coefficient (DSC) scores of 81.66% and 91.61% on the Synapse and ACDC datasets respectively, which outperforms existing mainstream methods.

Keywords：multi-organ segmentation；self-attention mechanism；frequency transformation；spatial attention