智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于空频协同的CNN-Transformer多器官分割网络

王梦溪, 雷涛, 姜由涛, 刘乐, 刘少庆, 王营博

引用本文:
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]. 智能系统学报, 2025, 20(5): 1266-1280.
WANG Mengxi, LEI Tao, JIANG Youtao, et al. CNN-Transformer multiorgan segmentation network based on space-frequency collaboration[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(5): 1266-1280.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202409011

DOI: 10.11992/tis.202409011
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20250613.1819.002

基于空频协同的 CNN-Transformer 多器官分割网络

王梦溪[1,2], 雷涛[1,2], 姜由涛[1,2], 刘乐[1,2], 刘少庆[1,2], 王营博[1,2]
（1. 陕西科技大学 电子信息与人工智能学院, 陕西 西安 710021; 2. 陕西科技大学 陕西省人工智能联合实验室, 陕西 西安 710021）

摘　要: 针对目前主流的医学多器官分割网络未能充分利用卷积神经网络 (convolutional neural network, CNN) 的局部细节提取优势以及 Transformer 的全局信息捕获潜力, 并缺乏空频特征协同建模的问题, 提出了一种基于空频协同的 CNN-Transformer 双分支编解码网络。该网络在局部分支中设计了空频协同注意力, 使网络从频域和空间域捕获到更为丰富的局部细节信息; 在全局分支设计了多视图频域提取器, 该模块通过频谱层和自注意力层联合建模, 提高了模型的空频特征协同建模能力和泛化性能。此外, 设计了局部与全局特征融合模块, 有效整合了 CNN 分支的局部细节信息和 Transformer 分支的全局信息, 解决了网络无法兼顾局部细节和全局感受野的难题。实验结果表明, 该架构克服了医学图像中器官边界模糊导致误分割的问题, 有效提升了多器官分割精度, 同时计算成本更低, 参数量更少。

关键词: 多器官分割; 空频协同; 多视图频域; 注意力机制; CNN; Transformer; 协同注意力; 局部−全局特征融合
中图分类号: TP391.4　　**文献标志码:** A　　**文章编号:** 1673−4785(2025)05−1266−15

中文引用格式: 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]. 智能系统学报, 2025, 20(5): 1266–1280.
英文引用格式: WANG Mengxi, LEI Tao, JIANG Youtao, et al. CNN-Transformer multiorgan segmentation network based on space-frequency collaboration[J]. CAAI transactions on intelligent systems, 2025, 20(5): 1266–1280.

CNN-Transformer multiorgan segmentation network based on space-frequency collaboration

WANG Mengxi[1,2], LEI Tao[1,2], JIANG Youtao[1,2], LIU Le[1,2], LIU Shaoqing[1,2], WANG Yingbo[1,2]
(1. School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China; 2. Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China)

Abstract: Current mainstream medical multi-organ segmentation networks fail to fully exploit the local detail extraction capabilities of convolutional neural network (CNN) and the global information capturing potential of Transformers. Additionally, they lack an effective mechanism for collaboration modeling of spatial and frequency domain features. To address these limitations, we propose a dual-branch encoder-decoder network based on CNN-Transformer with space-frequency collaboration. The network incorporates space-frequency collaborative attention in local branches, allowing the network to capture richer local details from both the frequency and spatial domains. A multi-view frequency domain extractor is designed in the global branch. This module improves the model's ability to jointly model spatial and frequency features and its generalization performance through joint modeling of spectral layers and self-attention layers. In addition, a local and global feature fusion module is designed to effectively integrate the local detail information of the CNN branch and the global information of the Transformer branch, solving the problem that the network cannot balance local details and global receptive fields. Experimental results demonstrate that this architecture effectively addresses the challenges posed by blurred boundary segmentation in medical images, which often leads to mis-segmentation of organs, significantly enhancing the accuracy of multi-organ segmentation while simultaneously reducing the computational costs and the number of parameters required.

Keywords: multiorgan segmentation; space-frequency collaboration; multiview frequency domain; attention mechanism; CNN; Transformer; coattention; local-global feature fusion

收稿日期: 2024−09−06.　**网络出版日期:** 2025−06−16.
基金项目: 国家自然科学基金项目 (62271296,62201334); 陕西省创新能力支撑计划项目 (2025RS-CXTD-012); 陕西高校青年创新团队项目 (23JP014, 23JP022).
通信作者: 雷涛. E-mail: leitao@sust.edu.cn.

腹部多器官分割是医学图像分割的主要难点之一, 它旨在通过像素分类准确捕捉目标器官和组织的形状和体积, 帮助医生更准确地分析、评

估和治疗。目前, 主流的多器官分割网络已取得显著成效, 但仍面临以下挑战。首先, 多器官分割任务不仅面临单器官分割中常见的边缘模糊、对比度低、形变大的问题, 还存在由于其他器官、组织、骨髓阻挡导致的目标器官不连贯问题, 相比单器官分割, 多器官分割需要挖掘更多图像中隐藏的信息。其次, 经典的 U 形主干分割网络受限于有限的感受野, 只能定位那些最具辨别力的局部目标区域。最后, 主流的图像分割方法主要从空间域视角出发探索模型的表达能力, 但在应用于医学多器官分割时, 仍然存在对图像结构理解不足和模型泛化性不足的问题。因此, 应从新的视角出发, 以充分激发深度学习网络模型的潜力。

为了解决上述问题, 本文提出了一种空频协同的 CNN-Transformer 编解码分割网络 (space-frequency coordination CNN-Transformer encoder-decoder network, SFC-Net), 该网络基于交叉注意力, 有效地结合了卷积神经网络 (convolutional neural network, CNN) 分支卓越的局部信息提取能力和 Transformer 分支独特的全局信息提取能力。全局信息帮助模型在整体上理解医学图像不同器官的位置关联; 而局部信息帮助模型在局部上准确分割出器官的轮廓和边界, 从而提高了模型对多器官的分割精度和泛化能力。此外, 为了进一步增强分割网络对图像自身结构信息的理解, 本文提出了一种基于频域和自注意力的多视角频域特征提取器, 实现从频域和空间域视角共同出发充分挖掘图像信息, 从而提高网络模型的特征表达能力和泛化性。本文的主要贡献总结如下:

1) 设计了一种频域空间域协同注意力模块 (frequency domain spatial domain collaborative attention, FSSA), 该模块从两个分支来细化局部信息, 第一个分支学习空间位置之间的关系, 有助于更好地理解和记忆空间维度的结构和布局; 第二个分支利用多频谱有效整合不同通道维度蕴含的器官结构信息。两个分支共同协作, 旨在从通道和空间两方面对特定的组织结构进行准确的定位和分割。

2) 设计了一个基于离散傅里叶变换的多视图频域提取器 (multi-view fequency-domain extractor, MFE), 该提取器能够从多个视角全面深入地捕捉和利用图像中蕴含的丰富频域特征, 并以较小的参数规模有效增强模型对全局信息的感知能力。此外, MFE 还能够从频域和空间域理解样本信息, 从而增加了模型对输入图像变化的适应能力, 提高了模型的泛化性。

3) 提出了基于空频协同的 CNN-Transformer 编解码分割网络模型并应用于多器官分割。该模型成功兼顾并优化了网络对局部与全局信息、频域与空间域信息的理解和处理能力, 进而提升了分割精确度。实验结果显示, 在 3 个权威公开的医学影像多器官分割数据集 Synapse、ACDC 和 AbdomenCT-1K 上, 该模型相较于同类型先进网络表现出了更为优越的性能。

1　相关工作

1.1　基于 CNN 的分割方法

随着深度学习的快速发展, CNN 在密集预测任务中展现出了巨大潜力。由于出色的图像特征提取能力, 其在医学影像分割任务中得到了广泛的应用。U-Net[1] 及其变体[2-4] 在医学图像分割任务中取得了显著的成功, 进一步推动了 CNN 在计算机视觉任务中的发展。然而, 卷积操作对全局上下文关系建模的有限能力会影响分割结果的准确性[5]。此外, 通过多次叠加和下采样以扩大感受野, 可能会导致深度网络的训练出现特征重用问题[6]。除此之外, 空间分辨率的逐渐降低, 可能会导致关键的局部信息的丢失。

1.2　基于 Transformer 的分割方法

近年来, Transformer 在自然语言处理 (natural language processing, NLP) 中取得了优异成就[7], 研究人员也开始尝试将其引入计算机视觉任务中, 以弥补 CNN 的不足。在此背景下, Dosovitskiy 等[8] 引入了第一个用于图像识别的视觉 Transformer(vision-Transformer, ViT), 它完全依赖于自注意力机制。Swin-Unet[9] 提出基于 Transformer 的 U 形网络用于医学图像分割。MS-Unet(multi-scale UNet)[10] 受 UNet++的启发, 改进了 Swin-Unet 的解码器, 创建了一个多尺度嵌套解码器。另一个值得注意的改进是 MISSFormer[11], 它是一个无位置的分层 U 形 Transformer, 使用 Transformer 解决了特征识别约束和规模信息融合问题。然而, Transformer 缺乏 CNN 固有的一些关键特征, 如移位、缩放和失真不变性, 这可能会限制其在视觉处理任务中的有效性[12]。虽然 Transformer 在全局上下文建模方面表现出色, 但其自注意力机制导致缺乏低级特征, 使得它在捕获细粒度细节方面不足, 这可能会影响前景和背景的可区分性, 并对图像分割性能产生不利影响[13-14]。

1.3　基于 CNN-Transformer 混合架构的分割方法

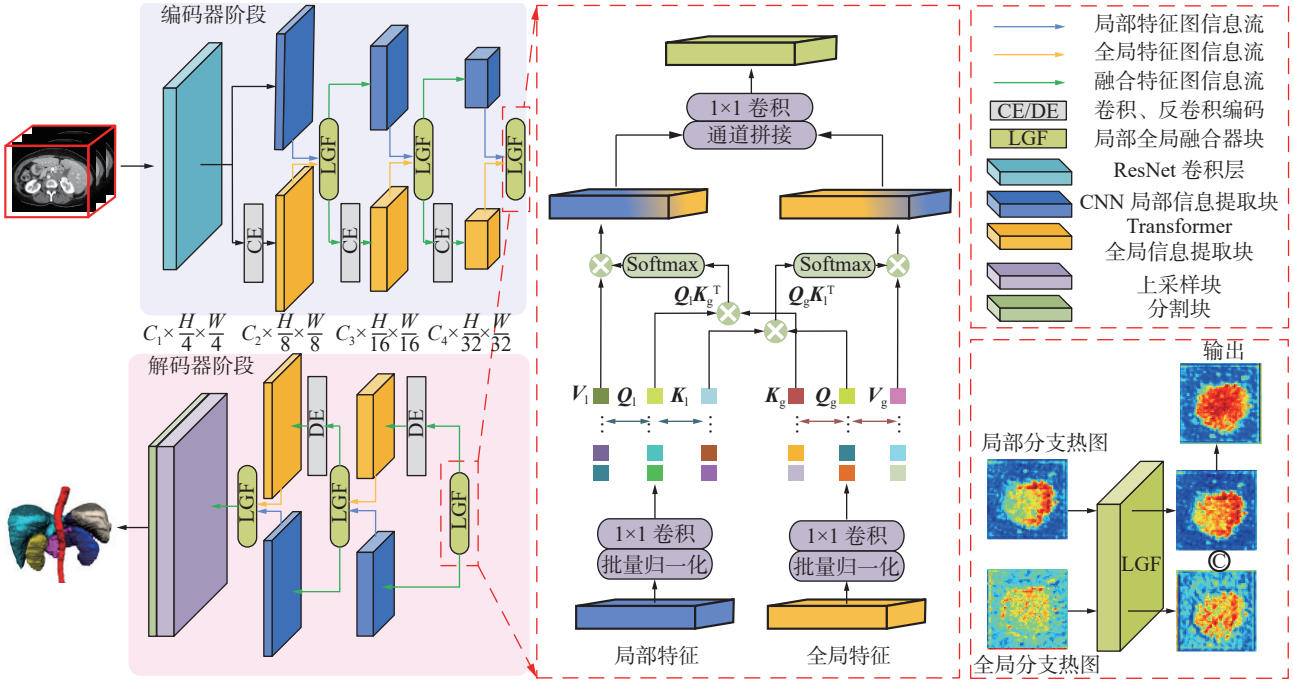
为了克服上述限制, 相关研究者们提出了将

卷积和注意机制结合起来的混合 CNN-Transformer 模型。TransUNet[15] 是第一个用于医学图像分割的 CNN-Transformer 混合模型，通过将 Transformer 层插入到 UNet 的编码器中，与完全使用 Transformer 作为编码器相比，实现了更好的分割效果。LeViT-UNet[16] 用 LeViT 块增强了 TransUNet 的编码器。MT-UNet[17] 引入局部−全局高斯加权自注意力，并将其与外部注意相结合，进一步增强了分割效果。CoTrFuse[18] 使用单独的 CNN 和 Transformer 模块在多个尺度上独立捕获特征，HiFormer[19] 提出了一种新的混合 CNN-Transformer 编码器结构，将 CNN 和 Transformer 结合在网络的浅层，以学习多尺度交互。HiFormer 还引入了双层融合 (double-level fusion, DLF) 模块，通过 Transformer 层对最大和最小尺度层的特征进行进一步编码，并通过全局平均池化和交叉关注机制进行融合。这些方法在医学影像分割任务中表现出良好的性能，然而，大多数方法仅考虑了空间域信息，忽略了图像中丰富的频域信息，对网络模型的分割能力探索仍然不足。通过结合空间域和频域信息，有望进一步提升医学分割模型的性能。

2　本文方法

2.1　网络结构概述

为了获得更好的医学图像分割效果，研究人员从多尺度特征融合、更有效的卷积设计和全局与局部信息融合等角度对网络进行优化，并获得了更佳的性能。然而，这些方法大多仅考虑了空间域信息，忽略了图像中丰富的频域信息。频域信息可以有效捕捉图像中的纹理、边缘等特征，有助于提高模型对图像结构的理解。因此，研究如何有效结合图像的频域特征与空间域特征成为实现医学影像多器官精准分割的重要内容。本文提出了一种空频协同的 CNN-Transformer 编解码分割网络模型 (SFC-Net) 并将其应用于医学影像多器官分割任务，整体架构如图 1 所示。



(a) SFC-Net 分割网络整体架构　(b) LGF 模块细节原理图　(c) LGF 模块特征融合注意力热图

图 1　SFC-Net 分割网络整体架构

Fig. 1　Overall architecture of the SFC-Net segmentation network

为了充分利用图像的局部和全局信息，SFC-Net 采用了 CNN 和 Transformer 并联的网络架构，并通过局部与全局特征融合模块 (local global fusion, LGF) 有效地交互和整合 CNN 的归纳偏置以及 Transformer 的长远距离依赖。基于此，SFC-Net 不仅具有较大的模型容量且不需要大量训练数据就能获得归纳偏置，从而提高了模型的表征能力。这种设计有利于适应输入图像的变换，从而提高了多器官的分割精度。

SFC-Net 编解码网络共有 7 层，其中中间的 5 层采用了融合 CNN 和 Transformer 分支的结构，这部分是提升分割性能的关键。自注意力作为提取全局上下文信息的重要机制，其参数量和计算量随着输入特征图的增大呈指数级增长，为了减

少输入特征图的大小, SFC-Net 在网络的第一层使用了 ResNet50 中的卷积块, 将输入尺寸减小为原来的 1/4。这不仅有效降低了自注意力计算量和参数量, 还初步获取了图像的浅层特征。此外, 模型的最后一层使用上采样和分割头来获取最终的分割掩码, 这种设计在有效减小计算复杂度的同时, 提高了分割精度。

2.2　频域空间协同注意力模块

过度拟合是医学多器官分割任务中的一个常见问题。为了克服这一问题, 在分割网络中引入注意力机制是常用的方法, 它可以提升网络模型对输入数据的关注和理解能力, 有利于提升网络的泛化性。相较于计算资源受限的自注意力, 通道注意力和空间注意力被广泛使用。然而, 后来的研究表明, 将注意力同时应用于通道和空间两个维度上效果更佳, 例如混合注意力 CBAM(convolutional block attention module)[20]。由于特征图的通道和空间都蕴含着丰富的信息, 显然 CBAM 是更高效的方法。此前的研究大多在空间域获取信息而忽略了频域的重要性。对于空间域, 分割对象和背景之间的边界往往不清晰, 而在频域, 对象和背景处于不同的频率, 是易区分的[21]。因此, 本文提出了全新的频域空间域协同注意力 (FSSA) 并将其应用于基于 CNN 的局部信息提取分支, 以增强模型对输入图像信息的挖掘能力, 进一步利用获取的丰富信息来适应输入变换。

FSSA 的结构如图 2 所示, 该方法由并行的空间域分析和频域分析组成, 与 CBAM 的串行结构具有本质的差异。从空间域分析图像, 本质上是对图像中不同位置的特征根据网络学习到的重要程度进行加权整合, 使得模型更加关注有益特征, 从而提高图像分割的准确性。而从频域分析图像, 本质上是分析图像的频率成分、频谱密度等特征, 这可以提供空间域没有的多频谱信息, 有助于识别并理解图像中的目标物体。在 FSSA 中, 最后一步是对两个分支的注意力系数相乘并分别激活, 然后使用卷积进行信息整合, 这样可以有效融合频域特征和空间域特征, 进而得到更丰富、更全面的特征表达, 从而提高模型对输入数据的理解能力。

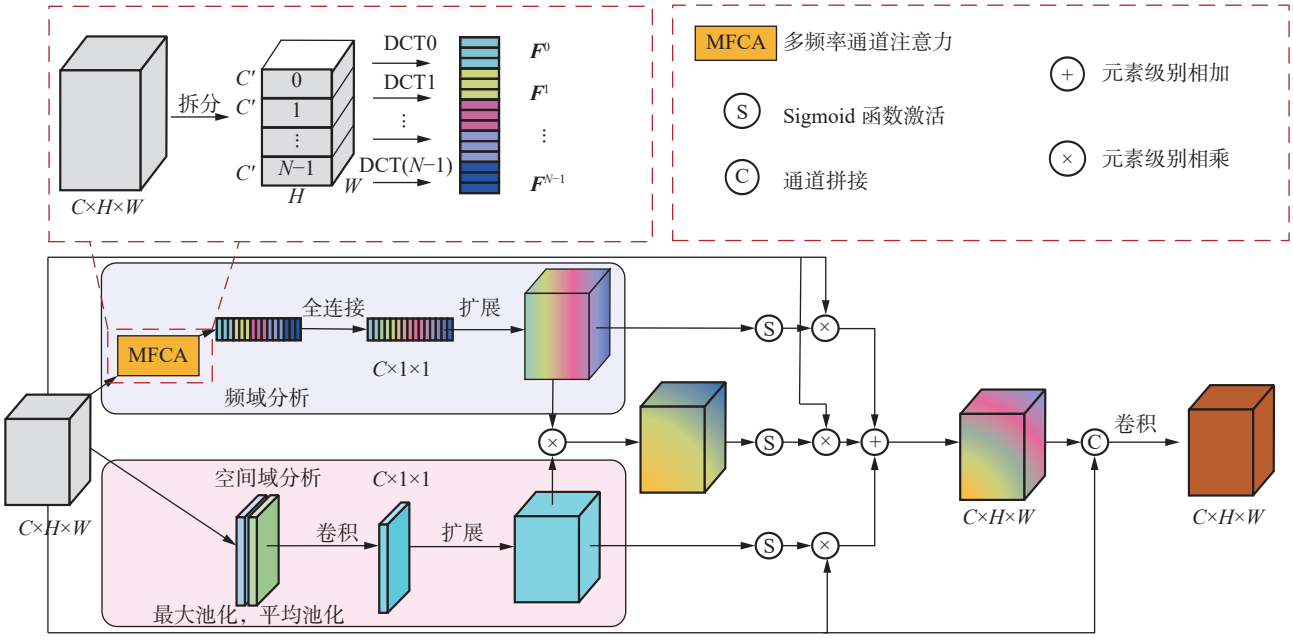


图 2　频域空间域协同注意力 (FSSA)

Fig. 2　Frequency domain spatial domain collaborative attention (FSSA)

如图 2 所示, 在 FSSA 中, 给定一个特征图 $\boldsymbol{X}_{\text{in}} \in \mathbb{R}^{C\times H\times W}$ 作为输入, C 代表输入通道数, $H\times W$ 代表输入特征图的分辨率, 特征图将分别进入空间域分支和频域分支。

2.2.1　空间域分析

对图像空间域的分析旨在研究空间中各像素区域的重要程度, 从而区分有用特征和无用特征。为了平衡计算量和性能, 本文使用空间注意力进行空间域分析:

$$\text{Att}_{\text{s}}(\boldsymbol{X}_{\text{in}}) = \text{Sigmoid}(\text{Con}_{7\times 7}[\text{AvgPool}(\boldsymbol{X}_{\text{in}});\text{MaxPool}(\boldsymbol{X}_{\text{in}})]) \tag{1}$$

$$\boldsymbol{F}^{1} = \text{Att}_{\text{s}}(\boldsymbol{X}_{\text{in}})\otimes \boldsymbol{X}_{\text{in}} \tag{2}$$

具体细节如式 (1)、(2) 所示, 首先, 输入特征通过跨通道的平均自适应池化和最大池化操作对特征图的空间信息进行聚合, 生成两个不同的空间上下文描述量, 即 $\boldsymbol{F}_{\text{avg}} = \text{AvgPool}(\boldsymbol{X}_{\text{in}}) \in \mathbb{R}^{1\times H\times W}$,

$\boldsymbol{F}_{\max}=\mathrm{MaxPool}(\boldsymbol{X}_{\mathrm{in}})\in\mathbb{R}^{1\times H\times W}$，它们分别表示平均池化特征和最大池化特征。其次，将两个描述量按通道维度进行拼接，并对其进行卷积核大小为7×7的卷积操作来压缩通道，生成包含空间信息的映射图，即 $\boldsymbol{F}_{\mathrm{spatial}}=\mathrm{Con}_{7\times7}([\boldsymbol{F}_{\mathrm{avg}};\boldsymbol{F}_{\max}])\in\mathbb{R}^{1\times H\times W}$。最后，对空间映射图 $\boldsymbol{F}_{\mathrm{spatial}}$ 进行Sigmoid激活生成空间注意力图 $\mathrm{Att}_{\mathrm{s}}(\boldsymbol{X}_{\mathrm{in}})\in\mathbb{R}^{1\times H\times W}$，并与输入特征进行逐元素相乘⊗得到经过空间域分析的特征图 $\boldsymbol{F}^1\in\mathbb{R}^{C\times H\times W}$。

2.2.2　频率域分析

通道注意力机制的核心是为每个通道赋予一个注意力系数，该注意力系数本质上是个标量。全局平均池化 (global average pooling, GAP) 由于其简单和高效被广泛使用。然而，GAP 的简单性使得通道注意力机制难以捕获各种输入的复杂信息。一些方法如 CBAM[20] 和 SRM(style-based recalibration module)[22] 进一步使用全局最大值池化和全局标准差池化来提高 GAP 的性能。

为了在有限的计算开销下对标量信道进行压缩，同时尽可能地保持整个信道的表示能力，本文使用离散余弦变换 (discrete cosine transform, DCT) 对通道注意机制中的通道进行压缩。首先，DCT 是在数字图像和视频处理中广泛使用的数据压缩方法，其具有很强的能量紧凑性[23]，因此可以获得高质量的数据压缩结果。其次，离散余弦变换可以通过乘法实现并且是可微的，很容易集成到 CNN 中。最后，GAP 在 SE-Net(squeeze-and-excitation networks) 中的有效性仅等同于 DCT 的最低频率分量，而许多其他潜在有用的频率分量被遗弃。通过 DCT，可以保留更多的频率信息，从而提高通道表示能力。

因此，为了更好地捕捉输入特征图蕴含的丰富频率信息，本文使用基于二维离散余弦变换的多频谱通道注意力 (multi-frequency channel attention, MFCA)。首先，将输入特征图 $\boldsymbol{X}_{\mathrm{in}}\in\mathbb{R}^{C\times H\times W}$ 沿着通道维度分为 N 组，表示为 $[\boldsymbol{X}^0,\boldsymbol{X}^1,\cdots,\boldsymbol{X}^{N-1}]$，其中 $\boldsymbol{X}^i\in\mathbb{R}^{C'\times H\times W}$，$i\in\{0,1,\cdots,N-1\}$，$C'=C/N$。在多频谱通道注意力中，二维 (2D)DCT 为每组通道分配对应的频率分量，并且 2D DCT 的结果可以作为通道压缩结果，具体细节用公式表示为

$$\boldsymbol{F}^i=\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\boldsymbol{X}^i_{:,h,w}f^{u_i,v_i}_{h,w}$$

式中：$[u_i,v_i]$为 $\boldsymbol{X}^i$对应的2D频率分量指数，$\boldsymbol{F}^i\in\mathbb{R}^{C'\times1\times1}$ 为压缩后的 C' 维向量，$f^{u_i,v_i}_{h,w}=\cos(\pi h/H\cdot(u_i+1/2))\cdot\cos(\pi w/W\cdot(v_i+1/2))$为 2D DCT 常用的基函数。

整个输入的压缩通道标量可以通过拼接得到，公式表示为

$$\boldsymbol{F}(\boldsymbol{X}_{\mathrm{in}})=\mathrm{Concat}([\boldsymbol{F}^0,\boldsymbol{F}^1,\cdots,\boldsymbol{F}^{N-1}])$$

$$\mathrm{Att}_{\mathrm{c}}(\boldsymbol{X}_{\mathrm{in}})=\mathrm{Sigmoid}(\mathrm{FC}(\boldsymbol{F}(X_{\mathrm{in}})))$$

$$\boldsymbol{F}^2=\mathrm{Att}_{\mathrm{c}}(\boldsymbol{X}_{\mathrm{in}})\otimes\boldsymbol{X}_{\mathrm{in}}$$

式中 $\boldsymbol{F}\in\mathbb{R}^{C\times1\times1}$ 为所得到的多频谱向量，由此可以看到，通过将原始的 GAP 单频谱方法推广到具有多个频率分量的通道注意力，有效地丰富了通道信息以提升模型的表示学习能力。

2.2.3　空间域信息融合

输入特征图经过空间域和频域分析后，需要进行有效融合。首先，将频域和空间域分析所得到的权重相乘并激活，然后再与输入特征图相乘，以此获得同时包含空间域信息和频域信息的特征图 $\boldsymbol{F}^3$。随后，将经过空域分析的特征图 $\boldsymbol{F}^1$、经过频域分析的特征图 $\boldsymbol{F}^2$ 以及 $\boldsymbol{F}^3$ 相加后，与输入特征图在通道维度进行拼接，并通过卷积操作融合特征。整个频域空间域协同注意力的框架表示为

$$\boldsymbol{F}^3=\mathrm{Sigmoid}(\mathrm{expand}(\boldsymbol{F}(\boldsymbol{X}_{\mathrm{in}}))\times\mathrm{expand}(\boldsymbol{F}_{\mathrm{spatial}}))\otimes\boldsymbol{X}_{\mathrm{in}}$$

$$\boldsymbol{F}^{\mathrm{FSSA}}(\boldsymbol{X}_{\mathrm{in}})=\mathrm{Conv}\{[(\boldsymbol{F}^3+\boldsymbol{F}^2+\boldsymbol{F}^1),\boldsymbol{X}_{\mathrm{in}}]\}$$

值得注意是，输出结果 $\boldsymbol{F}^{\mathrm{FSSA}}\in\mathbb{R}^{C\times H\times W}$ 与输入 $\boldsymbol{X}_{\mathrm{in}}\in\mathbb{R}^{C\times H\times W}$ 具有相同的维度，因此，频域空间域协同注意力可以作为高效注意力组件快速迁移到其他分割网络。

2.3　多视图频域提取器

在医学图像中，Lee-Thorp 等的一项研究 FNet (fourier transforms)[24] 提出了频域信息有助于捕捉图像中的纹理、边缘等特征，进而增强模型对图像结构的理解频谱层，并证明频谱层通过提供丰富的特征表示和全局视角来提升模型精度。最近的工作 SpectFormer[25] 将频谱层引入 Transformer，以弥补传统 Transformer 仅关注空域信息的不足，从而使 Transformer 模型能够更全面地理解图像的特征。值得注意的是 SpectFormer 证明了频域层和多头自注意力层在图像处理中均具有重要作用。

虽然 SpectFormer 方法在 Transformer 引入了频域层来提取全局信息，但它仅考虑了 $x-y$ 视图的频域信息，导致获得的信息存在片面性。为了解决这一问题，本文提出基于二维傅里叶变换的多视图频域提取器 (MFE)，并将其应用于基于 Transformer 的全局信息提取分支，从而在频域获得更全面的上下文信息，其主要结构如图 3 所示。

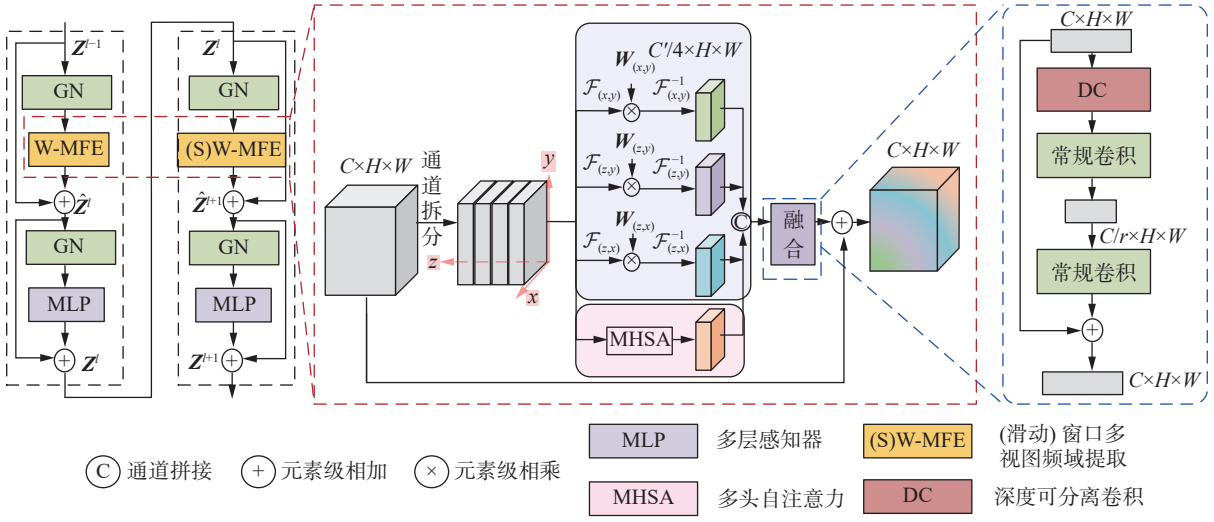


图 3　多视图频域提取器 (MFE)

Fig. 3　Multi-view frequency domain extractor (MFE)

对于输入$\boldsymbol{I}\in\mathbb{R}^{C\times H\times W}$, C代表输入通道数, $H\times W$代表输入特征图的分辨率。首先, 将输入按通道均分为 4 部分, 表示为$[\boldsymbol{I}^0,\boldsymbol{I}^1,\boldsymbol{I}^2,\boldsymbol{I}^3]$。其中$[\boldsymbol{I}^0,\boldsymbol{I}^1,\boldsymbol{I}^2]$ 3 部分输入到频域层进行多视角的全局频域特征提取:

$$\boldsymbol{I}^0_{(x,y)}=\boldsymbol{W}_{(x,y)}\otimes\mathcal{F}_{(x,y)}[\boldsymbol{I}^0]$$

$$\boldsymbol{I}^{0'}_{(x,y)}=\mathcal{F}^{-1}_{(x,y)}[\boldsymbol{I}^0_{(x,y)}]$$

式中: $\boldsymbol{W}_{(x,y)}$、$\boldsymbol{W}_{(z,y)}$、$\boldsymbol{W}_{(z,x)}$表示可学习的全局滤波器, 用来确定每个频率分量的权重, 滤波器可以使用网络的反向传播进行参数更新; $\mathcal{F}_{(x,y)}[\cdot]$、$\mathcal{F}_{(z,y)}[\cdot]$、$\mathcal{F}_{(z,x)}[\cdot]$表示对应视图的二维傅里叶变换; $\mathcal{F}^{-1}_{(x,y)}$、$\mathcal{F}^{-1}_{(z,y)}$、$\mathcal{F}^{-1}_{(z,x)}$表示逆傅里叶变换。这里以第一个分支为例进行详细说明, 其他分支同理。

首先, 对输入$\boldsymbol{I}^0$的$x-y$视图进行傅里叶变换, 得到$\mathcal{F}_{(x,y)}[\boldsymbol{I}^0]\in\mathbb{R}^{C\times H\times W}$, 它是一个复数形式的张量, 代表输入$I^0$的频谱。然后, 将频谱与可学习的滤波器$\boldsymbol{W}_{(x,y)}\in\mathbb{R}^{C\times H\times W}$进行逐元素相乘 (⊗) 以实现对频谱的调制。可学习的滤波器$\boldsymbol{W}_{(x,y)}$被称为全局滤波器, 因为它与输入的维度相同, 可以表示为频域中任意滤波器的线性组合。最后, 使用二维逆傅里叶变换$\mathcal{F}^{-1}_{(x,y)}$将学习到的全局频谱$\boldsymbol{I}^0_{(x,y)}$变换到空间域$\boldsymbol{I}^{0'}_{(x,y)}$。同理, 对于第 2、3 个分支, 分别在$z-y$和$z-x$视图上执行上述操作, 细节为

$$\boldsymbol{I}^1_{(z,y)}=W_{(z,y)}\otimes\mathcal{F}_{(z,y)}[\boldsymbol{I}^1]$$

$$\boldsymbol{I}^{1'}_{(z,y)}=\mathcal{F}^{-1}_{(z,y)}[\boldsymbol{I}^1_{(z,y)}]$$

$$\boldsymbol{I}^2_{(z,x)}=\boldsymbol{W}_{(z,x)}\otimes\mathcal{F}_{(z,x)}[\boldsymbol{I}^2]$$

$$\boldsymbol{I}^{2'}_{(z,x)}=\mathcal{F}^{-1}_{(z,x)}[\boldsymbol{I}^2_{(z,x)}]$$

此外, 空间域信息在多器官分割任务中同样重要。因此, 将剩余部分$\boldsymbol{I}^3$输入到多头自注意力层, 以提取全局空间域特征。对于第 4 个分支, 利用自注意力机制来获得空间域的全局信息:

$$\mathrm{SA}(\boldsymbol{I}^3)=\mathrm{Softmax}\left(\frac{\boldsymbol{q}\boldsymbol{k}^{\mathrm{T}}}{\sqrt{D_k}}+\boldsymbol{B}\right)\boldsymbol{v}\tag{3}$$

$$\boldsymbol{Y}=\mathrm{Concat}(\boldsymbol{I}^{0'}_{(x,y)},\boldsymbol{I}^{1'}_{(z,y)},\boldsymbol{I}^{2'}_{(z,x)},\mathrm{SA}(\boldsymbol{I}^3))$$

式 (3) 中: 相对位置偏差$\boldsymbol{B}\in\mathbb{R}^{M^2\times M^2}$; $\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}=\boldsymbol{I}^3\boldsymbol{W}_{(q,k,v)}\in\mathbb{R}^{M^2\times\frac{C}{4}}$分别代表查询向量 (query)、线索向量 (key) 和值向量 (value) 矩阵, 其中M^2则代表图像嵌入的数量; D_k表示 key 的维度。

最后, 对输入的特征图进行残差连接, 得到输出:

$$\boldsymbol{Y}'=\mathrm{Conv}_1\left(\mathrm{DWConv}_{(3,1)}(\boldsymbol{Y})\right)+\boldsymbol{Y}$$

全局分支的整体架构如图 3 所示, 对于某层特征图输入$\boldsymbol{Z}^{l-1}$, 全局分支的整体处理过程为

$$\hat{\boldsymbol{Z}}^l=\text{W-MFE}(\mathrm{GN}(\boldsymbol{Z}^{l-1}))+\boldsymbol{Z}^{l-1}$$

$$\boldsymbol{Z}^l=\mathrm{MLP}\left(\mathrm{GN}\left(\hat{\boldsymbol{Z}}^l\right)\right)+\hat{\boldsymbol{Z}}^l$$

$$\hat{\boldsymbol{Z}}^{l+1}=\text{(S)W-MFE}(\mathrm{GN}(\boldsymbol{Z}^l))+\boldsymbol{Z}^l$$

$$\boldsymbol{Z}^{l+1}=\mathrm{MLP}\left(\mathrm{GN}\left(\hat{\boldsymbol{Z}}^{l+1}\right)\right)+\hat{\boldsymbol{Z}}^{l+1}$$

式中: $\hat{\boldsymbol{Z}}^l$和$\hat{\boldsymbol{Z}}^{l+1}$分别表示 W-MFE 和 (S)W-MFE 的输出特征; "W"和"(S)W"分别代表窗口和滑动窗口, 与 Swin Transformer 中的操作保持一致; GN(·)表示组归一化 (group normalization), 实验证明相较于 Swin Transformer 中的层归一化 (layer normalization, LN), GN 更适合所提出的模型。因此, 通过多视图的频域运算可以学习到更丰富的全局信息。需要注意的是, 在具体实现中, 使用 PyTorch 库中支持的快速傅里叶变换来加速运算。

不同于以往的自注意力机制, (S)W-MFE 能充分利用医学图像的频域特征和空间域特征进行长程依赖建模, 弥补了常规自注意力仅关注图像

的空间域信息而忽略图像中丰富频谱特征的缺陷。这不仅提高了对医学影像中的多目标分割的特征提取能力，还提升了分割网络对医学图像的整体特征表达能力。

2.4　局部与全局特征融合模块

为了有效地融合 CNN 和 Transformers 的编码特征，本文提出了一个新的局部与全局特征融合模块 (LGF)，如图 1(b) 所示。首先，将 CNN 分支的局部特征图 $\boldsymbol{F}_{\mathrm{L}} \in \mathbb{R}^{N\times C}$ 和 Transformers 分支的全局特征图 $\boldsymbol{F}_{\mathrm{G}} \in \mathbb{R}^{N\times C}$ 进行批量归一化和线性映射，得到 $\boldsymbol{Q}_{\mathrm{L}},\boldsymbol{K}_{\mathrm{L}},\boldsymbol{V}_{\mathrm{L}} \in \mathbb{R}^{N\times C}$，其中 $N = H\times W$，公式表示为

$$\boldsymbol{Q}_{\mathrm{L}} = \boldsymbol{K}_{\mathrm{L}} = \boldsymbol{V}_{\mathrm{L}} = \mathrm{Conv}_{1\times 1}(\mathrm{BN}(\boldsymbol{F}_{\mathrm{L}})) \in \mathbb{R}^{N\times C}$$

$$\boldsymbol{Q}_{\mathrm{G}} = \boldsymbol{K}_{\mathrm{G}} = \boldsymbol{V}_{\mathrm{G}} = \mathrm{Conv}_{1\times 1}(\mathrm{BN}(\boldsymbol{F}_{\mathrm{G}})) \in \mathbb{R}^{N\times C}$$

对于全局信息提取部分，将 Transformer 分支的查询向量 $\boldsymbol{Q}_{\mathrm{G}}$ 与 CNN 分支的线索向量 $\boldsymbol{K}_{\mathrm{L}}$ 进行点乘，并经过 Softmax 后与值向量 $\boldsymbol{V}_{\mathrm{G}}$ 相乘，以引导 Transformer 分支对局部信息的学习。对于局部信息提取部分，将 CNN 分支的查询向量 $\boldsymbol{Q}_{\mathrm{L}}$ 与 Transformer 分支的线索向量 $\boldsymbol{K}_{\mathrm{G}}$ 进行点乘并经过 Softmax 后与值向量 $\boldsymbol{V}_{\mathrm{L}}$ 相乘来引导 CNN 分支对全局信息的学习，公式表示为

$$\boldsymbol{S}_{\mathrm{G}} = \mathrm{Softmax}(\boldsymbol{Q}_{\mathrm{G}}\boldsymbol{K}_{\mathrm{L}}^{\mathrm{T}}/\sqrt{d_{\boldsymbol{K}_{\mathrm{L}}}})\boldsymbol{V}_{\mathrm{G}}$$

$$\boldsymbol{S}_{\mathrm{L}} = \mathrm{Softmax}(\boldsymbol{Q}_{\mathrm{L}}\boldsymbol{K}_{\mathrm{G}}^{\mathrm{T}}/\sqrt{d_{\boldsymbol{K}_{\mathrm{G}}}})\boldsymbol{V}_{\mathrm{L}}$$

然后将新的 $\boldsymbol{S}_{\mathrm{G}}$ 与 $\boldsymbol{S}_{\mathrm{L}}$ 按通道进行拼接，并通过卷积进行信息融合，得到最终输出 $\boldsymbol{F}_{\mathrm{LGF}} \in \mathbb{R}^{N\times C}$ 进入下一层的学习：

$$\boldsymbol{F}_{\mathrm{LGF}} = \mathrm{Conv}_{1\times 1}([\boldsymbol{S}_{\mathrm{G}},\boldsymbol{S}_{\mathrm{L}}])$$

为了减少注意力机制的计算复杂度，在 LGF 中，无论是局部分支还是全局分支，当生成线索向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$ 时，都插入了尺度比例缩小模块：

$$\boldsymbol{K}_r = \mathrm{Reshape}\left(\boldsymbol{K},\left(\frac{N}{r},rC\right)\right)\mathrm{Proj}(rC,r)$$

式中 $\boldsymbol{K}_r$ 表示新的线索向量，其通过应用空间缩减率 r 来将像素序列的长度从 N 转换为 N/r。相应地，通道尺寸从 C 增加到 rC。然后，线性投影层 $\mathrm{Proj}(\cdot)$ 将中间特征层的通道深度从 rC 恢复到 C。$\boldsymbol{V}_r$ 以相同的方式生成：

$$\boldsymbol{V}_r = \mathrm{Reshape}\left(\boldsymbol{V},\left(\frac{N}{r},rC\right)\right)\mathrm{Proj}(rC,r)$$

因此，自注意力的表达式变为

$$\boldsymbol{S}_{\mathrm{G}} = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{G}}\boldsymbol{K}_{\mathrm{L}r}^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{K}_{\mathrm{L}r}}}}\right)\boldsymbol{V}_{\mathrm{G}r}$$

$$\boldsymbol{S}_{\mathrm{L}} = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{L}}\boldsymbol{K}_{\mathrm{G}r}^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{K}_{\mathrm{G}r}}}}\right)\boldsymbol{V}_{\mathrm{L}r}$$

观察这些公式可以发现自注意力操作的计算成本减少为原来的 $1/r$，从而提高了自注意机制的效率。局部与全局特征融合模块基于自注意和跨模式融合机制，整合了局部分支和全局分支的信息，其效果可视化如图 1(c) 所示。可以看出，全局分支特征图经过 LGF 后补充了细节信息，局部分支特征图经过 LGF 后获取了长远距离像素点的关系，两个分支的特征图经过拼接融合后获得了更为完整的分割结果。

3　实验结果与分析

3.1　实验数据集及预处理

Synapse 数据集[26] 包括 30 个腹部 CT(computed tomography) 扫描，共包含 3 779 个轴向对比增强的临床 CT 图像。每个体积样本由 85~198 个切片组成，每个切片大小为 512 像素×512 像素。本实验对 8 个腹部器官：主动脉、胆囊、左肾、右肾、肝脏、胰腺、脾脏和胃进行评估。本工作将 Synapse 数据集随机分为 18 个训练样本 (2 212 个轴位切片)、6 个验证样本和 6 个测试样本。

心脏自动诊断挑战数据集[27](automated cardiac diagnosis challenge, ACDC) 包含从法国第戎大学医院收集的 100 名患者的心脏短轴影像 MRI (magnetic resonance imaging) 数据，包括健康患者、既往心肌梗死患者、扩张型心肌病、肥厚型心肌病和右室异常患者，每组 20 例。MRI 影像采用屏气扫描，一系列短轴切片覆盖心脏，切片厚度为 5~8 mm。每个 3D 体数据空间分辨率为 0.70~1.92 像素/mm^2，28~40 个切片覆盖全部心动周期。每一个患者的 MRI 影像都手动标注出左心室、右心室和心肌。为了有效比较本文对 ACDC 数据集的分割效果，本文遵循相关系列的工作[4,15]，将 3 个器官的 Dice 相似系数 (Dice similarity coefficient, DSC) 作为最终的评价指标，并将整个数据集随机分为 70 个训练用例 (1 930 个轴位切片)、10 个验证用例和 20 个测试用例。

AbdomenCT-1K 数据集是目前最大的腹部 CT 多器官分割数据集，包含来自 12 个医疗中心的 1 112 个 CT 扫描，包括多阶段、多供应商和多疾病病例。标注共包括 4 446 个器官，器官种类分为 4 类，分别是肝、肾、脾和胰腺，这些器官比现有的腹部器官分割数据集要大得多。数据集包含整合后数据集中所有病例的 4 个器官的注释，为了保证注释的专业性，首先使用训练好的单器官模型来推断每个病例，然后，15 名初级注释员使用 ITK-SNAP3.6 在 2 名委员会认证的放射科医生的监督下手动细化分割结果，最后，一位拥有 10 余年经验的放射科医生验证并完善了注释。

3.2　实验环境及超参数设置

本实验算法运行的平台如下: GPU 为显存为 24 GB 的 NVIDIA GeForce RTX 3090, 编程语言包括 Python 和 C++, 并使用 PyTorch 1.7.0 框架训练。输入图像大小设置为 224 像素×224 像素, 训练时将批量大小和初始学习率分别设置为 12 和 0.01。此外, 本文使用在 ImageNet 上为 CNN 和 Swin Transformer 模块预先训练的权重来初始化它们的参数。所设计的模型使用 SGD 优化器进行优化, 动量为 0.9, 权重衰减为 0.000 1。在训练过程中, 还采用了翻转和旋转等数据增强技术, 以提高数据多样性。

为了克服多器官分割网络在使用交叉熵损失函数 (cross-entropy loss, CE-Loss) 时出现的部分细节信息遗失问题, 本研究采取了一种融合策略, 将 CE-Loss 与面向类别不均衡的 Dice 损失函数 (Dice-Loss) 结合, 作为联合优化的损失函数。Dice-Loss 的数学表达式定义为

$$L_{\text{dice}} = 1 - \frac{2 < p, \hat{p} >}{\| p\|_1 + \| p\|_2}$$

式中 $< p, \hat{p} >$ 表示每个通道的标签结果与预测结果矩阵点乘。Dice-Loss 能减少过拟合情况的出现, 但会影响反向传播导致训练误差曲线混乱等问题, 与此相反, CE-Loss 在梯度表现方面展现了更优的特性。CE-Loss 的定义为

$$L_{\text{cross}} = -(p \cdot \log(\hat{p}) + (1-p) \cdot \log(1-\hat{p}))$$

式中 p、$\hat{p}$ 分别表示标签结果和预测结果。当分割前景数量远远小于背景像素数量时, 会导致模型偏向于背景。本文采用联合损失函数, 保留了 CE-Loss 优异的梯度形式, 同时引入 Dice-Loss 缓解前景目标和背景不平衡的问题, 联合损失的定义为

$$L_{\text{loss}} = \alpha L_{\text{cross}} + (1-\alpha) L_{\text{dice}}$$

式中: $\alpha \in [0,1]$ 是可学习的参数, 用来自适应控制 CE-Loss 所占总损失的权重比; $(1-\alpha)$ 则表示 Dice-Loss 所占的权重比。

3.3　对比实验

为了验证本文方法 SFC-Net 的有效性, 在 Synapse、ACDC 以及 AbdomenCT-1K 3 个多器官数据集上与其他方法进行了比较, 观察和评估本文提出的方法在不同模式和不同目标解剖结构下的多器官分割性能。

3.3.1　Synapse 数据集分割

为了确保公正无偏的比较分析, 本文将 SFC-Net 与 CNN 和基于 Transformer 的方法, 以及将两者结合形成的方法模型进行了对比。CNN 模型包括 U-Net[1]、Attention-Unet[2], 基于 Transformer 的方法包括 Swin-Unet[9]、MISSFormer[11]、MEW-UNet[4] 和 Dae-Former[28], 以及两者合并时形成的模型包括 Trans-Unet[15]、LeViT-Unet[16]、AFTer-Unet[29]、MT-Unet[17]、CoTr[30]、DENTC[31]、STA-Former[32]。在对 8 个腹部器官进行图像分割的任务中, SFC-Net 与主流方法相比, 在 DSC 和豪斯多夫距离 (Hausdorff distance, HD) 这两大关键指标上取得了显著提升, 实验结果见表 1, 各项最佳结果以粗体显示。

表 1　在 Synapse 数据集上不同方法的定量评估和比较

Table 1　Quantitative evaluation and comparison of different methods on the Synapse dataset

对比方法	平均值		DSC/%↑							
	DSC/%↑	HD/mm↓	主动脉	胆囊	左肾	右肾	肝脏	胰腺	脾脏	胃
U-net[1][2015]	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
Att-Unet[2][2018]	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
Swim-Unet[9][2022]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
Trans-UNet[15][2021]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	76.62
LeViT-Unet[16][2023]	78.53	16.84	78.53	62.23	84.61	80.25	93.11	59.07	88.86	72.76
MISSFormer[11][2022]	80.74	19.65	85.31	66.47	83.37	81.65	94.52	63.49	91.51	79.63
CoTr[30][2021]	78.46	22.35	87.06	63.65	82.64	78.69	94.06	57.86	87.95	75.74
AFTer-UNet[29][2022]	81.02	18.96	**90.91**	64.81	87.90	85.30	92.20	63.54	90.99	72.48
Dae-Former[28][2023]	82.43	17.43	88.96	**72.30**	86.08	80.88	94.98	65.12	91.94	79.19
MT-UNet[17][2022]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
MEW-UNet[4][2022]	78.92	16.44	86.68	65.32	82.87	80.02	93.63	58.36	90.19	74.26
STA-Former[32][2024]	82.98	15.73	86.51	71.34	86.56	82.26	94.91	68.20	91.29	82.77
DENTC[31][2023]	80.68	25.46	86.98	70.07	84.40	81.20	94.34	61.06	89.41	77.96
SFC-Net(本文方法)	**83.97**	**13.73**	87.55	69.20	**88.61**	**84.85**	**95.36**	**69.61**	**92.41**	**84.18**

注: 加粗表示最佳结果。

从 3 个方面分析 SFC-Net 的有效性。首先，将该方法与 CNN 模型进行比较，SFC-Net 相较于 2 种 CNN 方法，DSC 分别提高了 7.12 和 6.2 百分点，HD 分别降低了 65.42%，61.88%。其次，将该方法与 Transformer 架构进行比较分析。SFC-Net 与 Swin-Unet 相比，DSC 提高了 4.84 百分点，HD 降低了 36.29%；与次优的 Dae-Former 相比，SFC-Net 的 DSC 提高了 1.54 百分点，HD 降低了 21.23%。最后，将该方法与 CNN-Transformer 混合网络架构进行比较分析。SFC-Net 与 AFTer-UNet 相比，SFC-Net 的 DSC 提高了 2.95 百分点，HD 降低了 27.58%；与 STA-Former 相比，SFC-Net 的 DSC 提高了 1 百分点，HD 降低了 12.71%。这些结果表明，在多器官分割任务上，Transformer 架构比 CNN 架构拥有更优异的特征表达能力，并且混合网络架构以及空频协同建模是具有潜力的研究方向。综合来看，SFC-Net 在 DSC 和 HD 这 2 个评估指标上都展现出了优越的性能。具体而言，SFC-Net 在 6 个器官的分割任务中稳步超过之前的工作，尤其在胰腺、肾脏和胃的分割任务中表现显著。图 4 给出了部分方法在 ITK-SNAP 中的可视化分割结果，可以观察到，相较于其他网络，SFC-Net 对胃、胆囊和胰腺等复杂结构的整体和边缘分割较好，这表明空频协同建模有利于模型对输入图像结构的理解，从而产生更完整和精细的分割结果。

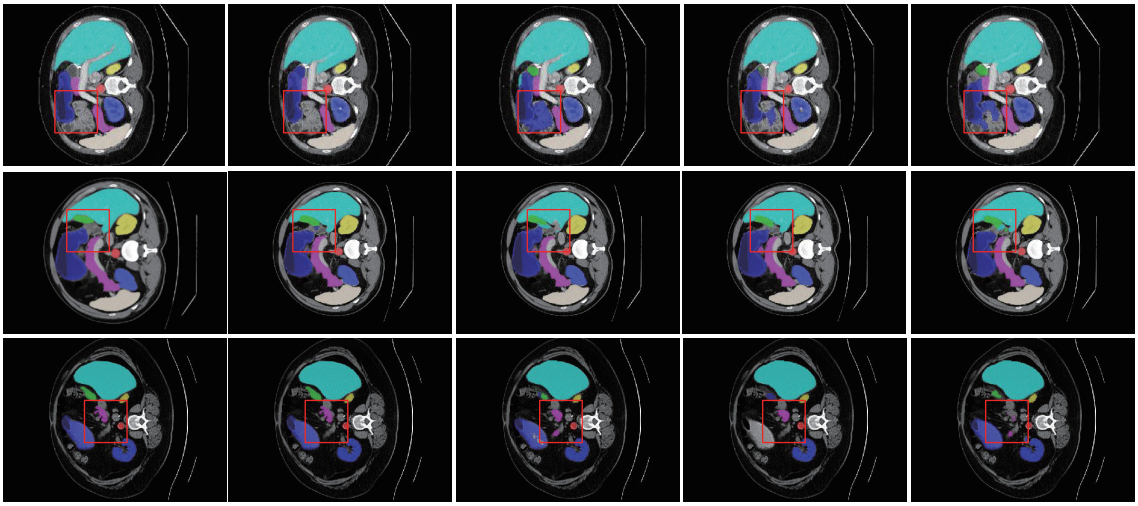


图 4　在 Synapse 数据集上使用不同方法的腹部多器官分割结果

Fig. 4　Abdominal multi-organ segmentation results using different methods on the Synapse dataset

3.3.2　ACDC 数据集分割

为了验证 SFC-Net 在不同模态的多器官数据下的有效性，本文在 MRI 多器官影像 ACDC 数据集上进行了扩展实验，并将 SFC-Net 与 4 个 CNN 网络 U-Net[1]、Attention-Unet[2]、UNet++[3]、ResUNet[33] 以及 5 个 Transformer 的网络或混合网络 TransUNet[15]、SwinUNet[9]、LeViT-Unet[16]、UTNet[34]、MT-UNet[17]、MEW-Unet[4]、DENTC[31] 进行了比较。定量结果如表 2 所示，各项最佳结果以粗体显示。

表 2　在 ACDC 数据集上不同方法的定量评估

Table 2　Quantitative evaluation of different methods on the ACDC dataset

对比方法	平均值		右心房		心肌		左心房	
	DSC/%↑	HD/mm↓	DSC/%↑	HD/mm↓	DSC/%↑	HD/mm↓	DSC/%↑	HD/mm↓
Unet[1][2015]	88.55	3.78	87.10	5.91	86.86	2.49	91.69	2.94
UNet++[3][2019]	90.21	4.06	89.02	6.91	87.31	**2.42**	94.30	2.83
ResUNet[33][2020]	89.75	3.71	89.44	5.53	88.13	2.53	91.68	3.08
Att-UNet[2][2018]	86.75	3.85	87.58	**5.48**	79.20	2.82	93.47	3.25
TransUNet[15][2021]	89.71	6.46	88.86	7.61	84.53	5.03	95.73	6.75
SwinUNet[9][2022]	90.30	5.07	81.81	5.75	81.18	3.78	90.01	5.70
UTNet[34][2021]	91.32	3.69	**90.41**	5.59	89.15	2.54	94.39	2.96

续表 2

对比方法	平均值		右心房		心肌		左心房	
	DSC/%↑	HD/mm↓	DSC/%↑	HD/mm↓	DSC/%↑	HD/mm↓	DSC/%↑	HD/mm↓
LeViT-Unet[16][2023]	90.32	—	89.55	—	87.64	—	93.76	—
MT-Unet[17][2022]	90.43	3.27	86.64	5.68	89.04	3.02	95.62	3.05
MEW-Unet[4][2022]	91.00	3.65	88.82	5.87	88.61	2.68	95.56	5.68
DENTC[31][2023]	91.12	—	90.34	—	88.35	—	94.69	—
SFC-Net[本文方法]	**92.03**	**3.22**	90.21	5.97	**89.94**	2.57	**95.94**	**2.12**

注: 加粗表示最佳结果。

分析表 2 可以发现, 所提出的方法 SFC-Net 在 3 个器官的平均分割精度以及轮廓分割准确度上都得到具有竞争力的结果。具体地, SFC-Net 在 3 个器官的分割任务中, 心肌和左心房的分割取得了最优结果。SFC-Net 相较于完全使用 Transformer 的 SwinUnet 模型, 其分割精度 DSC 提升了 1.73 百分点, HD 降低了 36.49%, 这表明 CNN-Transformer 混合网络在多器官分割任务中相较于完全使用 Transformer 的架构可能更具优势; SFC-Net 相较于同类架构的混合分割网络 TransUNet、UTNet、MT-Unet 分割精度平均提升了 2.2 百分点, 这说明 SFC-Net 设计了更好的局部全局融合策略; SFC-Net 相较于仅使用频域层构建的 MEW-Unet 模型, DSC 提升了 1.03 百分点, 这进一步验证了 SFC-Net 中的空频协同策略的有效性, 是挖掘图像信息的更优选择。SFC-Net 在 CT 模态数据集 Synapse 和 MRI 模态数据集 ACDC 上均达到了最佳的 DSC 指标和 HD 指标, 这些结果验证了 SFC-Net 在不同模态数据集上的泛化性。

此外, 图 5 给出了 SFC-Net 与其他方法在 ACDC 数据集上对心脏 MRI 影像分割的可视化结果。图 5(a) 表示人工标注结果, 将 SFC-Net 的分割结果与前 6 个方法的分割结果进行对比, 可以发现, SFC-Net 可以实现对右心房、心肌和左心房更完整的分割, 这再次说明了 SFC-Net 可以对兼顾图像的全局和细节信息的明显优势。相比之下, 其他方法都存在对右心房的不完整分割、对左心房的错误分割以及对心肌的过度分割的问题。

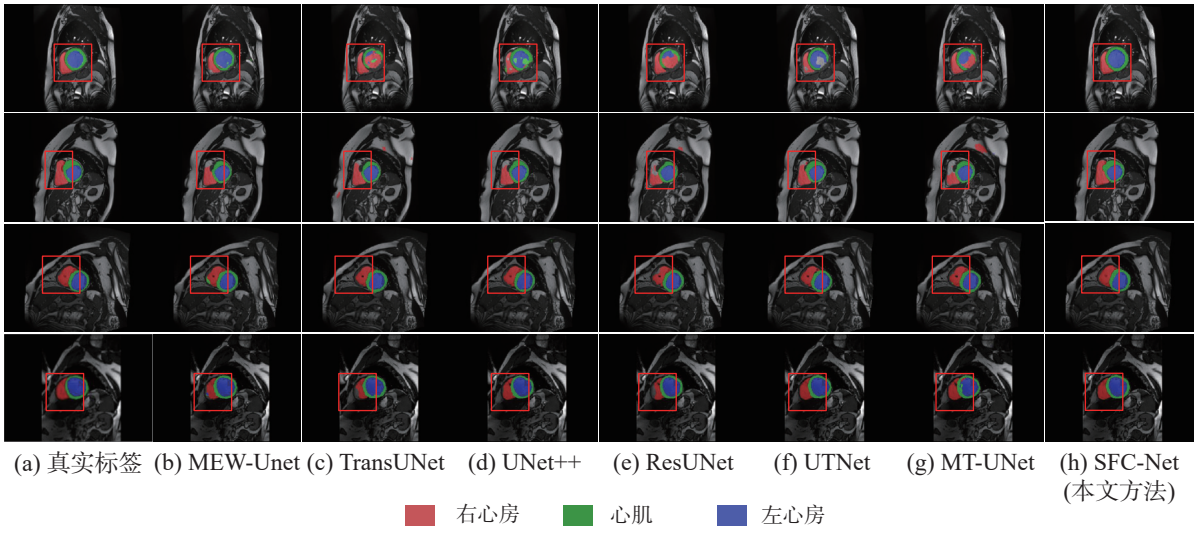


图 5　不同方法在 ACDC 数据集上对心脏 MRI 影像分割的可视化结果

Fig. 5　Visualization results of cardiac MRI image segmentation using different methods on the ACDC dataset

3.3.3　AbdomenCT-1K 数据集分割

为了进一步验证 SFC-Net 在多器官分割任务上的有效性, 本文在数据种类丰富的 AbdomenCT-1K 数据集上进行了扩展实验, 并将 SFC-Net 与目前在该数据集上表现最好的 3 个网络 nnUNet[35]、TransUNet[15]、CoTr[30] 进行了比较, 定量分析结果汇总于表 3, 各项最佳结果以粗体显示。观察结果发现, 在涉及 4 种器官肝脏、肾脏、胰腺与脾脏的分割任务中, SFC-Net 在易漏分割的小器官以及形状和位置易变的胰腺上展现出了卓越的性能。此外, 相较于同类混合式网络 TransUNet, 其 DSC 分别提高了 3.5 和 9.5 百分点。这一结果突显了 SFC-Net 在处理多器官形态变化问题上的有效性。在总体分割性能上, SFC-Net 以 86.4% 的 DSC 超越了其他对比方法, 并且比次优的 CoTr 网络高出了 1.6 百分点。

表 3　在 AbdomenCT-1K 数据集上先进方法对腹部多器官的定量评估结果

Table 3　Advanced methods' quantitative evaluation results for abdominal multi-organ segmentation on the AbdomenCT-1K dataset

对比方法	DSC/%↑				
	平均	肝脏	肾脏	脾脏	胰腺
NnUNet[35][2021]	83.7±14.8	**95.8±6.0**	84.1±14.8	89.9±15.5	65.0±22.7
TransUNet[15][2021]	83.3±13.9	95.3±3.5	83.9±17.0	**91.7±11.9**	62.2±23.1
CoTr[30][2021]	84.8±13.9	95.6±5.7	83.5±16.8	90.6±12.1	69.5±21.2
SFC-Net[本文方法]	**86.4±11.4**	95.3±4.7	**87.4±11.4**	91.0±15.2	**71.7±14.2**

注: 加粗表示最佳结果。

图 6 给出了在 AbdomenCT-1K 数据集上各种方法分割效果的可视化对比。从图中第 1 行和第 4 行可以观察到, SFC-Net 在处理肝脏形态变化较大区域以及分割胰腺器官时, 相比其他方法能够产生更为精细的分割轮廓。根据第 2 行和第 3 行, 在区分具有相似解剖结构的脾脏和邻近器官 (如肝脏或肾脏) 时, 其他方法容易产生误分割, 而 SFC-Net 能通过频域信息辨识它们的差异, 实现更精确的分割。SFC-Net 在上述 3 个数据集上卓越的性能证明了其有效性, 表明 SFC-Net 可以为医学图像多器官分割领域提供有力的经验证据和数据支持。

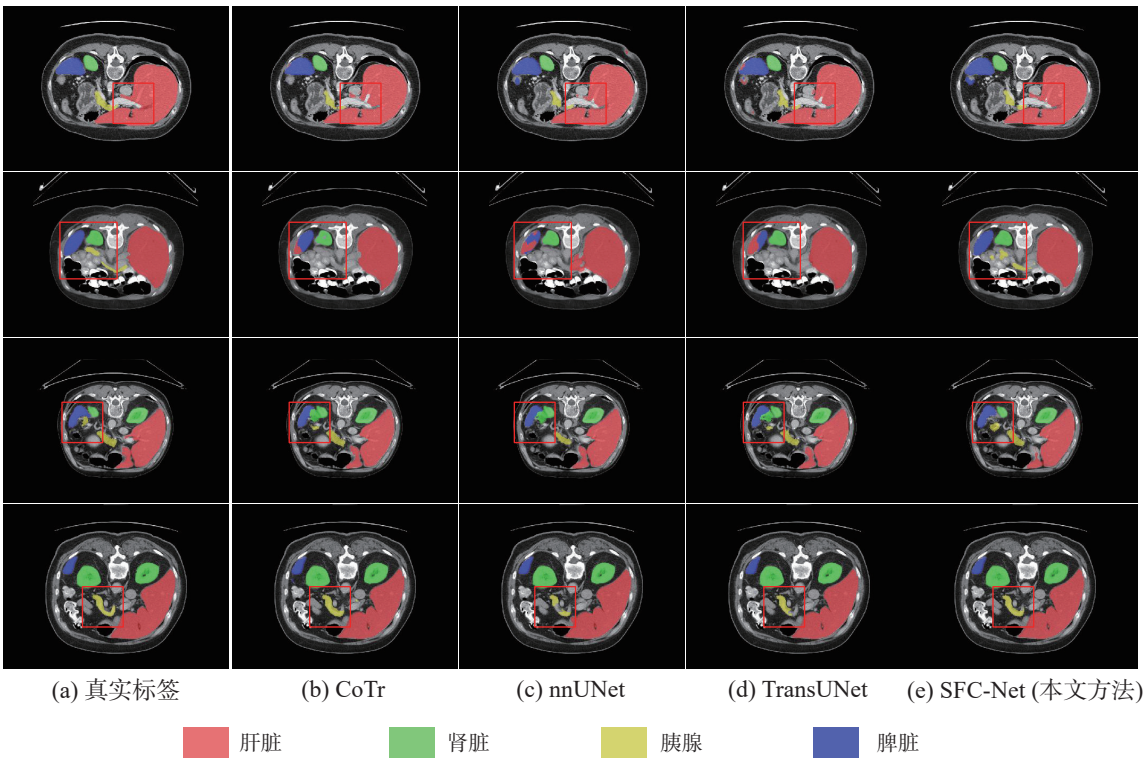


图 6　在 AbdomenCT-1K 数据集上使用不同方法对腹部多器官分割的定性结果

Fig. 6　Qualitative results of abdominal multi-organ segmentation using different methods on the AbdomenCT-1K dataset

3.3.4　模型计算成本比较

SFC-Net 在参数规模与计算复杂度上与其他方法的对比见表 4, 最佳结果以粗体显示。为了确保比较的客观性, 选取了多种不同网络架构进行比较。分析结果显示, SFC-Net 在参数规模上相比完全基于 CNN 的模型 Att-Unet 降低了 50%, 计算复杂度减少了 85.90%; 与完全基于 Transformer 的模型 Swin-Unet 相比, 参数数量降低了 59.97%, 计算复杂度降低了 19.34%; 与轻量级混合模型 LeViT-Unet 相比, 参数数量减少了 68.23%, 计算复杂度减少了 71.89%; 与 STA-Former 相比, 参数量减少了 52.19%, 计算复杂度减少了 51.49%。这归功于 SFC-Net 中全局分支的优化, 该分支采用资源效率更高的多视图频域层替代多头自注意力机制, 在不牺牲模型性能的前提下显著减少了参数和计算量。总体而言, SFC-Net 在显著降低参数和计算需求的同时, 保持了高效性能, 这进一步促进了其在移动设备上的应用以及在工程领域的实际部署。

表 4　在 Synapse 数据集上不同方法的模型参数比较

Table 4　Comparison of model parameters for different methods on the Synapse dataset

对比方法	参数量/10^6↓	计算量/10^9↓	DSC/%↑	HD/mm↓
U-Net[1][2015]	34.52	65.39	76.85	39.70
Att-Unet[2][2018]	34.88	66.57	77.77	36.02
Swin-Unet[9][2022]	41.40	11.63	79.13	21.55
Trans-Unet[15][2021]	105.30	15.21	77.48	31.69
LeViT-Unet[16][2023]	52.17	33.25	78.53	16.84
Dae-Former[28][2023]	59.70	26.16	82.43	17.43
MEW-Unet[4][2022]	138.89	31.04	78.92	16.44
STA-Former[32][2024]	34.66	19.34	82.98	18.73
SFC-Net[本文方法]	**16.57**	**9.38**	**83.97**	**13.73**

注: 加粗表示最佳结果。

3.4　消融实验

本文的主要贡献在于充分利用图像频域空间的多频谱信息, 并结合空间域信息来提高网络的分割精度。具体来说, 本文提出了 3 个模块: 频域空间域协同注意力模块 (FSSA)、多视图频域提取器 (W-MFE) 以及局部与全局特征融合模块 (LGF)。为了充分证明本文所提出模块的有效性, 本文以 HiFormer 作为主干网络, 在 Synapse 数据集上进行了一系列消融实验。需要注意的是, 实验中的 DSC 是 8 个器官的平均 DSC。

首先, 验证所设计的模块可以提升模型的整体分割精度。如表 5 所示, 各项最佳结果以粗体显示。在主干网络的基础上, 选择在编码器部分使用 CBAM 时模型分割精度提升不大。Synapse 数据集本身具有较高的复杂性, 器官具有不同的形状、大小和纹理特征。由于 HiFormer 提取的原始特征已经足够丰富且针对性强, 直接加入 CBAM 带来的额外增益相对较小。另外在编码器中加入 CBAM 的方式可能不适用于 Synapse 数据集, 因此本文设计了频域空间域协同注意力模块 (FSSA)。

表 5　在 Synapse 数据集上各模块功能的消融实验结果

Table 5　Ablation study results of various module functionalities on the Synapse dataset

行标	CBAM	FSSA	(S)W-MFE	LGF	DSC/%↑	HD/mm↓
A					80.39	14.70
B	√				80.74	15.13
C	√		√		81.27	14.96
D	√		√	√	82.06	13.89
E		√			81.37	14.06
F			√		82.23	13.98
G				√	80.79	15.02
H		√	√		83.53	13.85
I			√	√	82.65	13.87
J		√	√	√	**83.97**	**13.73**

注: 加粗表示最佳结果。

在主干网络的基础上, 仅使用设计的频域空间域协同注意力模块 (FSSA) 在多器官分割精度方面平均提升了 0.98 百分点, 仅使用设计的多视图频域提取器 (W-MFE) 模块提升了 1.84 百分点, 而将两者组合使用分割精度则提升了 3.14 百分点, 证明空频特征协同建模有利于提升模型的分割精度。

其次, 边缘分割准确度在临床应用中是极其重要的。HD 作为模型边缘分割精确度的衡量标准被广泛使用, 其值越低表明模型分割效果越好。如表 5 所示, 在主干网络的基础上, 仅使用设计的频域空间域协同注意力模块 (FSSA) 在多器官边缘分割准确度方面提升了 4.35 百分点, 仅使用设计的多视图频域提取器 (W-MFE) 模块提升了 4.9 百分点, 将两者组合使用提升了 5.78 百分点。这些结果表明, 充分利用频域信息和空间域

信息可以提升多器官的平均分割精度和边缘分割准确度。

本文设计的局部与全局特征融合模块 (LGF) 是有效的。从 D 行和 A 行的对比以及 G 行和 E 行的对比可以发现，仅使用 LGF 可以提升大约 0.5 百分点的整体分割精度。为了进一步证明 LGF 的功能，本文对该模块的每一步进行可视化，其热图如图 7 所示。可以发现，全局分支经过 LGF 后，补充了局部细节信息，而局部分支经过 LGF 后，分割效果更全面。最后将两条支路进行融合，得到了分割范围更广、精度更高的最终结果。

为了充分利用图像的细节信息，网络浅层使用卷积神经网络作为前置信息提取层。为了研究不同 CNN 骨干网络的贡献，本文使用 ResNet[36] 和 DenseNet[37] 两种基本方法进行消融实验，结果如表 6 所示，各项最佳结果以粗体显示。从 C 行可知，利用 ResNet50 主干可获得最佳性能。此外，从 C 行和 D 行可以发现，参数量更多的 CNN 主干并不一定会带来性能提升，因此本文使用 ResNet50 作为前置信息提取层。

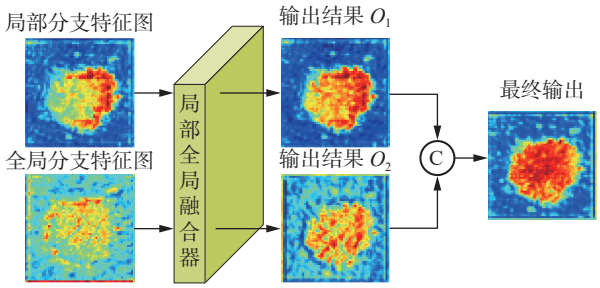


图 7　局部与全局特征融合模块注意力热图可视化

Fig. 7　Attention heatmap visualization of local-global fusion module

表 6　在 Synapse 数据集上使用不同大小的 ResNet 的消融实验结果

Table 6　Ablation study results using different sizes of ResNet on the Synapse dataset

行标	模型	参数/10^6↓	DSC/%↑	HD/mm↓
A	SFC-Net+ResNet18	10.06	80.85	15.86
B	SFC-Net+ResNet34	15.81	82.73	21.76
C	SFC-Net+ResNet50	**16.57**	**83.97**	**13.73**
D	SFC-Net+ResNet101	35.56	82.95	16.05
E	SFC-Net+DenseNet121	14.98	82.15	15.26
F	SFC-Net+DenseNet169	20.61	82.32	14.93
G	SFC-Net+DenseNet201	26.42	82.67	20.26

注：加粗表示最佳结果。

多视图频域提取器 (W-MFE) 的核心理念是多轴运算。因此，本文逐一增加多轴运算的数量，以验证所提出模块的有效性，其消融实验结果如表 7 所示，各项最佳结果以粗体显示。分析表 7 可知，$x-y$视图的频域运算在仅使用空间域的全局多头注意力方法上可以提升 0.47 百分点，证明频域信息的利用可以有效提升分割精度。$z-y$视图的频域运算可以在运用$x-y$视图运算后提升 0.63 百分点，$z-x$视图可以在之前的两视图运算上提升 0.65 百分点。此外，对比 D 行和 E 行数据可以看出，引入 MHSA 可以提升 0.62 百分点。因此，多视图频域提取器模块的多视图运算获取的信息是互不冗余的，其设计有助于对更全面的频域信息和全局知识进行建模。

表 7　在 Synapse 数据集上多视图频域提取器的消融实验结果

Table 7　Ablation study results of the multi-view frequency domain extractor on the Synapse dataset

行标	MHSA	$\boldsymbol{I}^{0'}_{(x,y)}$	$\boldsymbol{I}^{1'}_{(z,y)}$	$\boldsymbol{I}^{2'}_{(z,x)}$	DSC/%↑	HD/mm↓
A	√				82.22	15.62
B	√	√			82.69	15.08
C	√	√	√		83.32	14.68
D		√	√	√	83.35	14.51
E	√	√	√	√	**83.97**	**13.73**

注：加粗表示最佳结果。

4　结束语

本文提出了一种基于空频协同的 CNN-Transformer 编解码多器官分割网络 SFC-Net，SFC-Net 通过交叉注意力将 CNN 分支良好的局部信息提取能力和 Transformer 分支良好的全局信息提取

能力有效结合, 设计了空频协同注意力模块 (FSSA)、基于傅里叶变换的多视频域提取器 (MFE) 以及一个新的局部与全局特征融合模块 (LGF)。SFC-Net 优化了网络对局部与全局信息、空间域与频域信息的理解处理能力, 解决了当前多器官分割任务的主要问题。在 3 个权威公开的医学影像多器官分割数据集上的实验结果表明, SFC-Net 能够在减少计算量和参数量的同时, 取得更先进的分割结果。在后续的研究中, 团队计划将 SFC-Net 中提出的各个模块推广到更广泛的数据集上, 并通过调整或引入新模块来解决更复杂的任务和达到更高精度的分割需求, 从而进一步提高多器官分割的性能。

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 International Publishing, 2015: 234−241.

[2] OKTAY O, SCHLEMPER J, LE FOLGOC L, et al. Attention U-Net: learning where to look for the pancreas [EB/OL]. (2018−05−20)[2024−09−06]. https://arxiv.org/abs/1804.03999v3.

[3] LEI Tao, SUN Rui, DU Xiaogang, et al. SGU-Net: shape-guided ultralight network for abdominal image segmentation[J]. IEEE journal of biomedical and health informatics, 2023, 27(3): 1431−1442.

[4] RUAN Jiacheng, XIE Mingye, XIANG Suncheng, et al. MEW-UNet: Multi-axis representation learning in frequency domain for medical image segmentation[EB/OL]. (2022−10−25)[2024−09−06]. https://arxiv.org/abs/2210.14007v1.

[5] 刘万军, 姜岚, 曲海成, 等. 融合 CNN 与 Transformer 的 MRI 脑肿瘤图像分割[J]. 智能系统学报, 2024, 19(4): 1007−1015.
LIU Wanjun, JIANG Lan, QU Haicheng, et al. MRI brain tumor image segmentation by fusing CNN and Transformer[J]. CAAI transactions on intelligent systems, 2024, 19(4): 1007−1015.

[6] 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.
ZHANG Shujun, PENG Zhong, LI Hui. SAU-Net: medical image segmentation method based on U-Net and self-attention[J]. Acta electronica sinica, 2022, 50(10): 2433−2442.

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

[8] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]//International Conference on Learning Representations. NewOrleans: ICLR, 2021: 1−22.

[9] 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 Nature Switzerland, 2023: 205−218.

[10] KUSHNURE D T, TALBAR S N. MS-UNet: a multi-scale UNet with feature recalibration approach for automatic liver and tumor segmentation in CT images[J]. Computerized medical imaging and graphics, 2021, 89: 101885.

[11] HUANG Xiaohong, DENG Zhifang, LI Dandan, et al. MISSFormer: an effective medical image segmentation Transformer[EB/OL]. (2021−12−19)[2024−09−06]. https://arxiv.org/abs/2109.07162v2.

[12] 雷涛, 张峻铭, 杜晓刚, 等. 基于混洗特征编码与门控解码的医学图像分割网络[J]. 电子学报, 2024, 52(12): 4142−4152.
LEI Tao, ZHANG Junming, DU Xiaogang, et al. Medical image segmentation network based on shuffled feature encoding and gated decoding[J]. Acta electronica sinica, 2024, 52(12): 4142−4152.

[13] 周新民, 熊智谋, 史长发, 等. 基于多尺度卷积调制的医学图像分割[J]. 电子学报, 2024, 52(9): 3159−3171.
ZHOU Xinmin, XIONG Zhimou, SHI Changfa, et al. Medical image segmentation based on multi-scale convolution modulation[J]. Acta electronica sinica, 2024, 52(9): 3159−3171.

[14] 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.

[15] CHEN Jieneng, LU Yongyi, YU Qihang, et al. TransUNet: Transformers make strong encoders for medical image segmentation[EB/OL]. (2021−02−08)[2024−09−06]. https://arxiv.org/abs/2102.04306v1.

[16] XU Guoping, ZHANG Xuan, HE Xinwei, et al. LeViT-UNet: make faster encoders with Transformer for medical image segmentation[C]//Pattern Recognition and Computer Vision. Singapore: Springer Nature Singapore, 2023: 42−53.

[17] JHA A, KUMAR A, PANDE S, et al. MT-UNET: a novel U-Net based multi-task architecture for visual scene understanding[C]//2020 IEEE International Conference on Image Processing. Abu Dhabi: IEEE, 2020: 2191−2195.

[18] CHEN Yuanbin, WANG Tao, TANG Hui, et al. CoTr-Fuse: a novel framework by fusing CNN and Transformer for medical image segmentation[J]. Physics in medicine & biology, 2023, 68(17): 175027.

[19] 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 Com-

puter Vision. Waikoloa: IEEE, 2023: 6191−6201.
[20] 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.
[21] 王婷, 宣士斌, 周建亭. 融合小波变换和编解码注意力的异常检测[J]. 计算机应用研究, 2023, 40(7): 2229−2234, 2240.
WANG Ting, XUAN Shibin, ZHOU Jianting. Anomaly detection fusing wavelet transform and encoder-decoder attention[J]. Application research of computers, 2023, 40(7): 2229−2234, 2240.
[22] LEE H, KIM H E, NAM H. SRM: a style-based recalibration module for convolutional neural networks[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 1854−1862.
[23] RAO K R, YIP P C. Discrete cosine transform - algorithms, advantages, applications[J]. Academic press, 2014, 34(4): 6315−6322.
[24] LEE-THORP J, AINSLIE J, ECKSTEIN I, et al. FNet: mixing tokens with Fourier transforms[EB/OL]. (2022−05−26)[2024−09−06]. https://arxiv.org/abs/2105.03824v4.
[25] PATRO B N, NAMBOODIRI V P, AGNEESWARAN V S. SpectFormer: frequency and attention is what you need in a vision Transformer[EB/OL]. (2023−04−14)[2024−09−06]. https://arxiv.org/abs/2304.06446v2.
[26] 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.
[27] 叶晋豫, 李娇, 邓红霞, 等. SwinEA: 融合边缘感知的医学图像分割网络[J]. 计算机工程与设计, 2024, 45(4): 1149−1156.
YE Jinyu, LI Jiao, DENG Hongxia, et al. SwinEA: Medical image segmentation network fused with edge-aware[J]. Computer engineering and design, 2024, 45(4): 1149−1156.
[28] 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 Nature Switzerland, 2023: 83−95.
[29] YAN Xiangyi, TANG Hao, SUN Shanlin, et al. AFTer-UNet: axial fusion Transformer UNet for medical image segmentation[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2022: 3270−3280.
[30] XIE Yutong, ZHANG Jianpeng, SHEN Chunhua, et al. CoTr: efficiently bridging CNN and Transformer for 3D medical image segmentation[C]//Medical Image Computing and Computer Assisted Intervention. Cham: Springer International Publishing, 2021: 171−180.
[31] HONG Zhifang, CHEN Mingzhi, HU Weijie, et al. Dual encoder network with Transformer-CNN for multi-organ segmentation[J]. Medical & biological engineering & computing, 2023, 61(3): 661−671.
[32] LIU Yuzhao, HAN Liming, YAO Bin, et al. STA-Former: enhancing medical image segmentation with Shrinkage Triplet Attention in a hybrid CNN-Transformer model[J]. Signal, image and video processing, 2024, 18(2): 1901−1910.
[33] 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.
[34] GAO Yunhe, ZHOU Mu, METAXAS D N. UTNet: a hybrid Transformer architecture for medical image segmentation[C]//Medical Image Computing and Computer Assisted Intervention. Cham: Springer International Publishing, 2021: 61−71.
[35] ISENSEE F, JAEGER P F, KOHL S A A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation[J]. Nature methods, 2021, 18(2): 203−211.
[36] 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.
[37] IANDOLA F, MOSKEWICZ M, KARAYEV S, et al. DenseNet: implementing efficient ConvNet descriptor Pyramids[EB/OL]. (2014−04−07)[2024−09−06]. https://arxiv.org/abs/1404.1869v1.

作者简介：

王梦溪，硕士研究生，主要研究方向为计算机视觉、机器学习。E-mail：202007020606@sust.edu.cn。

雷涛，教授，博士生导师，陕西科技大学电子信息与人工智能学院副院长，IEEE 高级会员。主要研究方向为计算机视觉、机器学习。发表学术论文 90 余篇。E-mail：leitao@sust.edu.cn。

姜由涛，硕士研究生，主要研究方向为计算机视觉、机器学习。E-mail：2819423992@qq.com。