基于 Transformer 与多层注意力机制的膀胱医学图像分割算法研究

高博艺，丁学明，丁雪峰，胡鸿翔

（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘　要：深度学习技术目前已经广泛应用于医学图像分割领域，然而，膀胱图像的精准分割依然面临挑战。本研究针对膀胱 MRI 图像分割问题，提出了一种基于特征融合以及注意力机制的改进算法（MIV-Net），通过引入了多尺度核心特征融合器（Multi-scale Core Feature Fusionizer，MCF），以增强模型图像特征提取能力，视觉特征解析与转换（Vision Transformer，VIT）网络，优化了全局信息处理能力。此外，在解码过程中通过多层次注意力融合（Multi-level Attention Fusion，MAF）模块来增强对上下文信息的利用。实验结果表明：MIV-Net 模型在膀胱壁和膀胱肿瘤的分割中展现出优异的性能，其 IoU、Precision、Dice 系数等关键评价指标均显著超越当前主流方法，在两个分割任务上分别达到了 83.16%、91.66%、90.81% 以及 82.49%、90.50%、90.40%，该模型不仅在技术上具有创新性，更在实际应用中有望显著提升膀胱癌诊断的准确性。

关键词：膀胱图像分割；融合注意力机制；视觉特征解析与转换网络；多尺度核心特征融合器；U-Net

中图分类号：TP391　　**文献标识码：**A

Research on Bladder Medical Image Segmentation Algorithm Based on Transformer and Multi-Level Attention Mechanism

Gao Boyi,，Ding Xueming，Ding Xuefeng，Hu Hongxiang

（School of Optical-Electrical & Computer Engineering, University of Shanghai for Science & Technology, Shanghai 200093,China）

Abstract: Deep learning technology has been widely applied in the field of medical image segmentation; however, precise segmentation of bladder images remains challenging. This study addresses the issue of bladder MRI image segmentation by proposing an improved algorithm based on feature fusion and attention mechanisms (MIV-Net). The algorithm introduces a Multi-scale Core Feature Fusionizer (MCF) to enhance the model's image feature extraction capability, and a Vision Transformer (VIT) Network to optimize global information processing. Additionally, the utilization of contextual information is enhanced during the decoding process through the Multilevel Attention Fusion (MAF) module. Experimental results demonstrate that the MIV-Net model exhibits superior performance in the segmentation of the bladder wall and bladder tumors. Key evaluation metrics such as IoU, Precision, and Dice coefficients significantly surpass current mainstream methods, achieving 83.16%, 91.66%, and 90.81% for bladder wall segmentation, and 82.49%, 90.50%, and 90.40% for bladder tumor segmentation, respectively. This model not only shows technical innovation but also holds promise for significantly improving the accuracy of bladder cancer diagnosis in practical applications.

Keywords: Bladder image segmentation; attention mechanism integration; Vision Transformer; Multi-scale Core Feature Fusionizer; U-Net

0　引　言

膀胱癌是全国范围内十大最常见的恶性肿瘤之一，发病率在所有恶性肿瘤中位第九。据统计 2022 年约有 61.4 万新病例和 22 万例死亡[1]。根据癌症统计报告，2023 年癌症的总体死亡率自 1991 年以来下降了 33%，预计避免了 380 万人死亡，其中，膀胱癌的发病率从 2015 年开始每年下降 0.6%，2019 年之后下降幅度加快至 1.8%[2]。然而，膀胱癌的早期诊断和治疗仍然面临着巨大挑战。

目前，膀胱癌的诊断主要依赖于计算机断层扫描（Computed Tomography，CT）、磁共振成像（Magnetic Resonance Imaging，MRI）等医学图像技术[3]。尽管这些技术为疾病的辅助诊断提供了重要支持，但由于膀胱肿瘤和正常组织的边界较为模糊，医学图像分割仍然是一个难点。即使是经验丰富的放射科医生，对于一些分布特征相似或边界模糊的影像，依然可能出现诊断结果不一致的情况。因此，如何通过计算机视觉技术实现医学图像中膀胱肿瘤的精准分割成为解决这一问题的关键。

精准分割医学图像中的病灶对医生探寻病因和制定诊疗方案起关键作用，计算机视觉技术的发展促使深度学习在医学图像分割领域衍生出多种模型架构。其中，U-Net 架构以其巧妙的跳跃连接、易于优化的特性，成为医学图像分割领域的主流基准模型[5]。本研究提出了一种基于特征融合以及注意力机制的改进算法(MIV-Net)，通过对膀胱 MRI 图像的高精度分割，能够具体了解和分析肿瘤的情况，以此来确定适合病人情况的医疗方法。

基金项目：国家自然科学基金资助项目(11502145)

论文主要贡献如下：1）提出了一种多尺度的辅助分割模型，增强模型捕捉输入图像局部细节的能力。2）通过视觉特征解析与转换网络，优化整体模型对全局信息的整合，使得模型能够有效融合更多的上下文信息。3）开发的 MIV-Net 模型在膀胱肿瘤以及膀胱壁的分割任务上显示出优于其他主流模型的性能，验证了该模型在实际的医疗场景中所具备的潜力。

1 相关工作

近年来，深度学习在医学图像领域的进步显著，特别是膀胱肿瘤分割和膀胱壁分割方法的研究也越加深入。传统的图像处理肿瘤分割算法有基于阈值[4]的分割方法、基于区域的分割方法和基于边界的分割方法。随后的技术进步，出现了全卷积网络（Fully Convolutional Network，FCN），FCN 是由 Jonathan Long 等人[6]于 2015 年提出的深度学习模型，开创了医学图像的像素级语义分割。相比于传统的卷积神经网络（Convolutional Neural Network，CNN）只能输出单一的标签或分类，FCN 能够输出与输入图像相同尺寸的像素级别的预测结果，从而实现了端到端的像素级别语义分割，其在语义分割领域取得了显著成果，成为了后续很多语义模型的基础。

2015 年欧洲核子研究组织（CERN）的研究人员 Ronneberger[7]等在 2015 年生物医学成像国际挑战赛提出 U-Net 网络，该网络特别为生物医学图像设计，通过其独特的跳跃连接架构，提高了生物医学图像的精准分割能力。然而，U-Net 在细节和局部特征的感知能力上仍然有限。在某些情况下，低层特征和高层特征通过跳跃连接进行融合时，可能并未完全发挥其优势。基于 U-Net 的进一步研究，Ozan Oktay[8]等人在 U-Net 网络的基础上提出了一种新的注意力门（Attention Gate，AG）模型用于医学图像分割，AGs 训练模型通过学习抑制输入图像中的无关区域，同时突显出有用的显著特征，从而增强 U-Net 的感知能力和特征提取能力。但是，尽管 Attention U-Net 通过注意力机制能够增强对重要区域的关注，但对于一些极小的目标物体，注意力机制可能未能完全捕捉到细节信息，导致微小病变部位在分割结果中表现不佳。

最近，Nabilibtehaz 等人[9]提出了 Transunet，将 Transformer 的全局自注意力机制与 U-Net 的局部特征提取相结合，在多器官分割和心脏分割等任务中取得了卓越的性能。此外，Jie Hu 等人[10]提出了 Squeeze-and-Excitation (SE)模块，通过显式建模通道之间的相互依赖关系，自适应地重新校准通道级特征响应，为现有的 CNN 带来了显著的性能提升。Xi Guan 等人[11]采用组合模型将 SE 模块和 AG 模块集成到 AGSE-VNet 模型中，用于 3D MRI 胶质瘤脑肿瘤的图像分割，来提高分割性能，从结果来看，模型在增强肿瘤（ET）和肿瘤核心（TC）区域的分割效果不如整个肿瘤（WT）区域，表明模型在处理小目标和边界不清晰的区域时可能存在不足。Wang Y 等人[12]提出了 LCANet 网络，基于 U-Net 网络分别加入了 ResUNet Block、SA 等，对膀胱肿瘤进行分割取得了较好的结果。Qi Zhang 等人[13]提出了一种基于注意力机制的膀胱镜图像分割（ACS）模型，在编码器-解码器中加入了通道注意力模块和空间注意力模块，在跳跃连接部分加入了融合关注模块，增加了初始注意模块，最终 ACS 模型的肿瘤分割性能优于对比模型。吴港等人**[14]**介绍了一种用于结肠息肉分割的新型模型 FTDC-Net，它结合了 Transformer 编码模块和空洞卷积模块，利用 ResNet50 作为编码器提取深层次图像特征。该模型在 Kvasir-SEG 和 ETIS-LARIBPOLYPDB 数据集上进行了实验验证，结果表明其在各项评价指标上均优于主流息肉分割模型。近年来，DEA-Net 网络[15]的提出，通过设计一个细节增强注意力块（DEAB），包括细节增强卷积（DEConv）和内容引导注意力（CGA），来提升特征学习能力并改善去雾性能。在一些极端的雾条件下，例如非常浓密的雾或雾的分布非常不均匀的情况，DEA-Net 的去雾效果可能不如在中等雾条件下显著。这是因为在这种情况下，图像的细节信息可能被严重丢失，而模型可能难以完全恢复这些丢失的细节。VM-UNet 网络结构[16]引入了视觉状态空间（VSS）块作为基础块来捕获广泛的上下文信息，并构建了一个不对称的编码器-解码器结构，以减少卷积层的数量。论文中的实验表明，随着输入图像分辨率的增加，VM-UNet 的性能并没有达到预期。Hou 等人[17]提出了一种基于注意力机制和特征融合的方法，应用于主动脉和肺的图像分割。

现有的医学图像分割模型，如 U-Net 及其衍生的 Attention U-Net 等网络，虽然在大部分场景中表现优异，但在处理小目标、模糊边界和细节恢复方面仍存在不足。特别是在复杂环境或低分辨率图像中，这些模型的分割效果不理想。为了解决上述问题，本研究提出了一种基于 Transformer 和多层注意力机制的膀胱医学图像分割算法。Transformer 能够有效捕捉图像中的全局上下文信息，而多层注意力机制可以在不同尺度上强化对细节和小目标的关注，从而提高模型的分割精度，特别是在处理小目标和细节丰富的区域时。现阶段通过深度学习进行膀胱癌医学图像分割领域的研究相对较少，且传统方法对于膀胱肿瘤的 MRI 图像的分割效果并不理想，针对上述问题，文章提出 MIV-Net 模型旨在提供一种有效的解决方案，并进行深入探索。

2 MIV-Net 模型结构

U-Net 网络模型是深度学习在医学图像分割领域的一个重要应用，其结构可分为编码层、解码层。编码层主要是负责提取图像的特征，通过一系列的 3×3 卷积层和 2×2 最大池化层构成下采样模块。在编码网络中一共包含四层下采样模块，每一层之间逐层连接，将提取到的特征通过特征图的形式保存并向下一层进行传递，有效地在多个尺度上提取到高维特征。解

码层主要负责还原特征图尺寸和具体的特征信息，在解码部分中一共包括了四个上采样模块，每个模块都将得到的特征图进行上采样，直到还原至原始图像的尺寸。

在进行上采样的过程中，每级都添加了跳跃连接层，跳跃连接层将编码层的特征图和上采样得到的特征图在通道维度上进行拼接，并进行卷积操作，从而整合特征通道。通过结合编码层的特征图像信息，可以更好地恢复原始图像的信息。

本研究在 U-Net 网络模型的基础上，MIV-Net 引入了多层次融合注意力（Multi-level Attention Fusion，MAF）模块，多尺度核心特征融合器（Multiscale Core Feature Fusionizer，MCF）和视觉特征解析与转换网络（Vision Transformer，VIT），增强了模型对于图像特征的学习能力。通过在跳跃连接层中加入 MCF 和 VIT，使得模型能够在多个尺度上更有效地捕捉图像的全局和局部特征，进一步提升学习能力。

在解码部分，MAF 模块替代了传统的拼接操作。采用双路经特征融合模块（Dual Path Fusion Block，DPFB）、通道注意力（Channel Attention，CA）模块和空间注意力（Spatial Attention，SA）模块，每层通道数减半，尺寸翻倍，模型能够有针对性地关注输入数据的不同通道和特征图的各个区域，从而实现选择性注意，减少无关的信息的干扰，最终提高了图像分割的精确性，特别是在处理一些具有复杂背景的医学图像。模型的整体框架图如图 1 所示，提供了对于 MIV-Net 模型架构的直观理解。

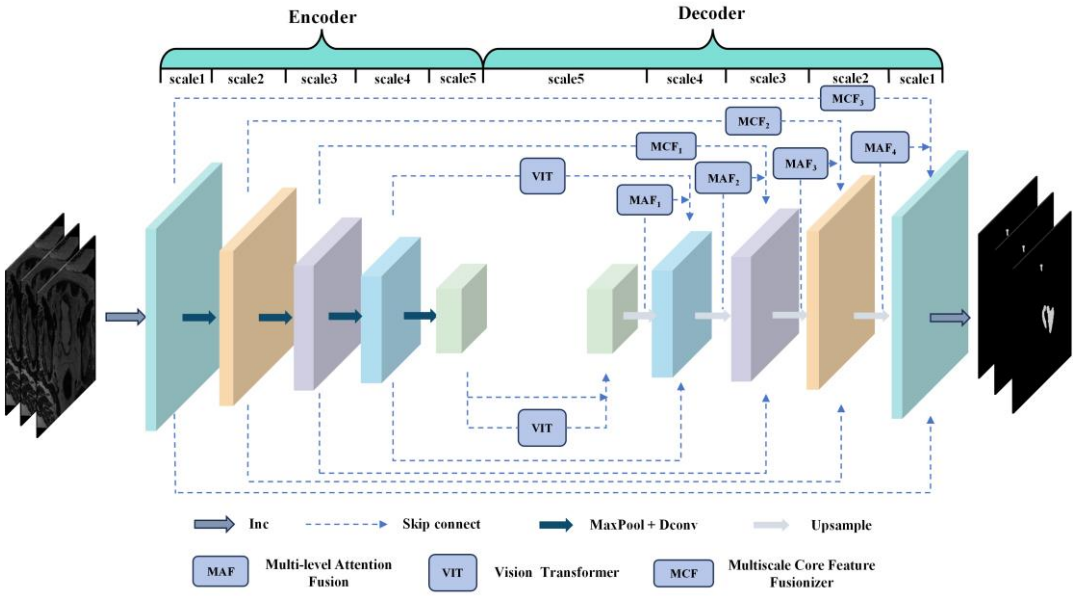

图 1 基于 Transformer 与多层注意力机制的膀胱肿瘤分割模型

Fig. 1 The Bladder Tumor Segmentation Model Based on Transformer and Multi-level Attention Mechanism

2.1 多尺度核心特征融合器

MCF 在 U-Net 网络模型的跳跃连接层中实现了一种动态融合机制，通过采用不同尺寸的卷积核处理图像，MCF 能够计算每个特征图片的权重并进行融合，从而优化图像特征提取过程。MCF 主要由以下三个部分组成，模型结构图如图 2 所示。

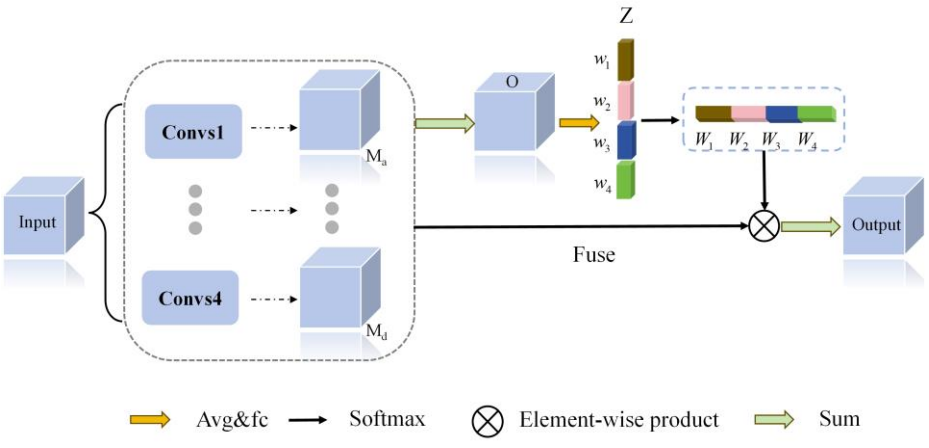

图 2 多尺度核心特征融合器模型结构图

Fig. 2 MCF model structure diagram

1)多尺度卷积分支：输入图像 $X \in R^{H\times W\times C}$，首先经过四个并行的卷积分支，如式（1）所示，这些分支使用不同大小的卷积核（1×1、3×3、5×5、7×7）进行不同程度的特征提取，以捕获不同的图像特征信息，之后经过标准化模块和 ReLU 激活函数，网络深度保持不变。这样的设计允许模型在保留局部细节的同时，也能够考虑到更广泛的上下文信息，从而提高特征提取的效率和准确性。得到特征图分别为 M_a、M_b、M_c、M_d，特征图的尺寸以及通道数和输入特征图保持一致。

$$M_i = Fcon_i(B(\theta(X))) \tag{1}$$

式(1)中，$Fcon_i$ 表示卷积操作，B 表示标准化模块，θ 表示 ReLU 激活函数。

2)特征图融合和重标准化：在经过初步的卷积操作后，将四个不同卷积核尺寸提取的特征图进行融合，通过 Sum 操作整合为一个特征图 O，如式（2）所示。确保了来自多个分支的特征信息能够有效地传递至下一层神经网络，从而为后续的特征学习提供更全面的上下文信息。

$$O = M_a + M_b + M_c + M_d \tag{2}$$

对特征图 O 进行全局平均池化（Global Average Pooling，GAP）处理得到 Q，Q 中第 i 个通道的元素表达式如式（3）所示。通过全局平局池化操作可以减少特征图 O 的空间维度、减少参数数量、计算量和过拟合同时保留了重要的通道特征信息。

$$Q_i = F_{avg}(O_i) = \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W} O_{i,h,w} \tag{3}$$

之后使用多个全连接层进行线性映射得到向量 Z，计算各分支的权重 $Z = F_{fc}(Q)$，H 和 W 分别是特征图的高度和宽度。

3)加权特征融合：通过 Softmax 函数将计算所得到的权重归一化，得到最终的注意力权重，然后将应用回各自的特征图（M_a、M_b、M_c、M_d），并将它们加权合并，生成最终的融合特征图。这种方法确保了模型能够根据图像内容的不同区域和特征，动态调整每个尺寸特征的权重。

面对不同形状和复杂图像分割任务时，使用相同卷积核的卷积方法可能会导致一些细节被忽略，从而影响分割的精度。特别是在处理具有复杂形状和细节的医学图像时，这一问题尤为突出。为了解决这一挑战，本文设计了 MCF 模型，能够有效地处理具有复杂形状和细节的医学图像，特别是在执行精确图像分割任务时，能够显著提高分割精度。

MCF 通过结合多尺度特征和动态权重的方法，使得模型能够更好地理解和重建图像的重要部分，从而减少与任务无关信息的干扰。同时，MCF 允许模型在不同的尺度上处理图像，捕捉到更多的细节和上下文信息。动态权重则能够根据图像的具体特征信息，自动调节卷积核的权重，使得整体模型更加的灵活、适应性更强。

2.2 多层次注意力融合 MAF 模块

2.2.1 双路径特征融合 DPFB 模块

医学图像分割任务中，旨在将医学图像中有意义的区域划分出来。CNN 通常难以同时将全局上下文和局部细节同时捕捉。针对以上问题，本文提出了 DPFB 模块，通过整合方向卷积和注意力机制，能够将图像特征进行更有效地融合，DPFB 模型结构图如图 3 所示。

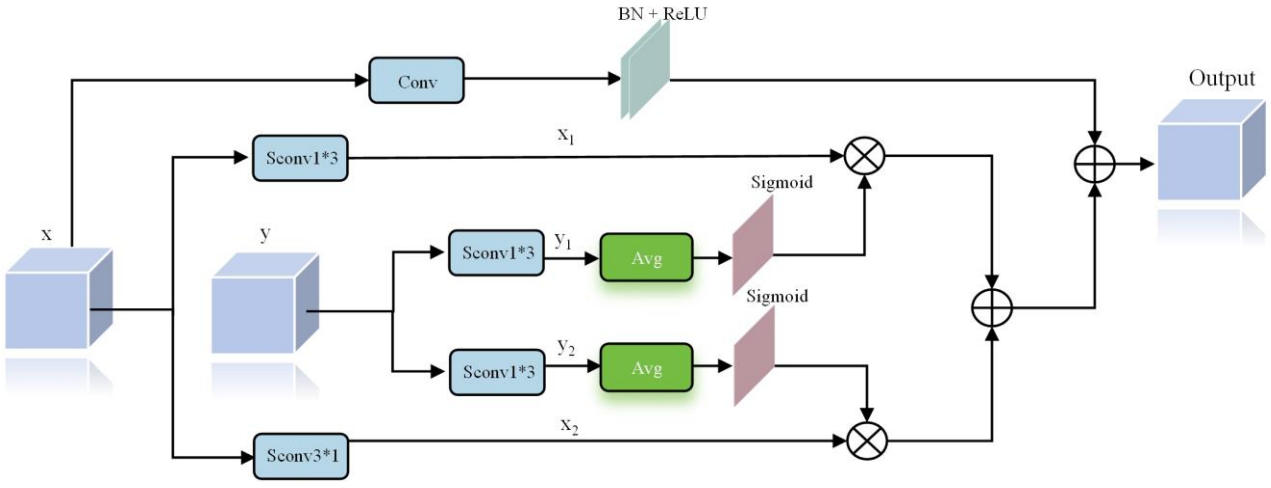

图 3 双路径特征融合模块模型结构图

Fig. 3 Dual Path Fusion Block Structure Diagram

在 DPFB 模块中，首先对当前层特征图 y 以及上一层特征图 x 分别进行方向卷积（1×3 和 3×1），从而生成特征图 y_1、y_2 和 x_1、x_2。针对 y_1 和 y_2，通过平均池化操作来降低维度，并使用 Sigmoid 激活函数来生成它们的注意力权重图。然后利用注意力图加权融合 x_1 和 x_2。同时对特征图 x 进行 3×3 卷积操作，随后进行归一化和 ReLU 激活函数处理，进一步特征提取。最终，将经过卷积处理的特征图和加权融合后的特征图相加，从而生成最终的输出特征图。通过这一系列操作的目的是为了通过多方向的卷积和注意力机制，使得不同层次的特征在融合过程中能够有效互补，从而提升模型的整体特征表达能力。

在 DPFB 模块中，对当前层的特征图 y 在医学图像分割任务中主要作用是增强特征融合，通过使用不同方向的卷积核，DPFB 可以更全面地捕捉图像中的结构信息。通过自适应池化操作和 Sigmoid 激活函数，可以动态地调整特征图的权重，从而突出特征图的有效信息。DPFB 模块能够更好地帮助网络区分图像中的细节和边界，从而提高分割精度。

2.2.2 通道注意力 CA 模块

在深度学习模型中，特别是在进行图像特征处理过程中，每个通道的信息往往具有不同的代表性和重要性。为了有效利用通道之间的依赖关系，论文重点分析了每个通道传递的信息。每个通道在处理时会经过一个局部感受野的滤波器。但经过此操作之后，每个单元只能利用局部的感受野进行信息提取，无法使用局部区域以外的信息。

为了解决这个问题，本研究采用了 CA 模块来优化这一过程。该模块能够识别并增强有用的特征通道，同时抑制不相关的通道，具体模型结构如图 4 所示。

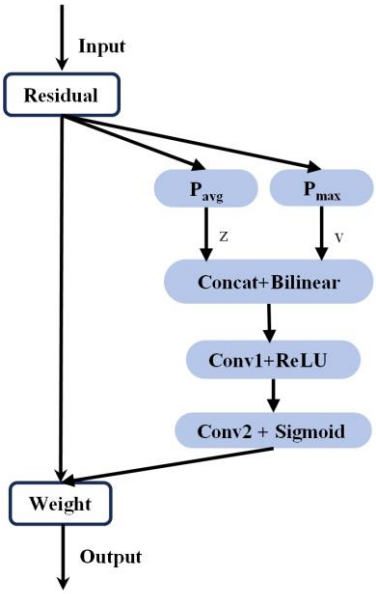

图 4 通道注意力模型结构图

Fig. 4 Channel Attention model structure diagram

通过全局平均池化操作将全局的空间信息压缩成一个通道描述符，在网络中能够更好的理解整体上的特征，而不会出现因为局部感受野而无法利用局部区域以外的信息的困扰，可以使得网络能够更好地理解数据整体结构，从而提高了网络的表达能力。z 是通过维度为 $C\times H\times W$ 的输入 U 进行全局平均池化操作产生的，z 中第 i 个通道的元素表达式如公式(4)所示。

$$z_i = P_{avg}(U_i) = \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}U_{i,h,w} \tag{4}$$

v 是通过维度为 $C\times H\times W$ 的 U 进行最大池化（Max Pooling）操作产生的，v 中的第 j 个通道的元素表达式如公式(5)所示。

$$v_j = P_{\max}(U_j) = \max_{h=1}^{H}\max_{w=1}^{W}U_{j,h,w} \tag{5}$$

为了能够更好地帮助网络在进行特征提取时更加有效地利用通道之间的信息，通过一系列卷积核激活函数操作，特征图进行通道注意权重的生成，将这些生成的权重加权输入到原本的特征图像的通道中去，对通道中重要的特征信息进行突出。

$$\mathbf{Z} = F_{con_1}(\theta(F_{con_2}(z\ \mathrm{e}\ v))) \tag{6}$$

式（6）中 e 表示拼接操作，F_{con1}和F_{con2}分别指的是两个1×1的卷积层，F_{con1}对每个通道中的特征进行降维，缩小输入特征图的通道数，以减少计算复杂度和避免过拟合。F_{con2}对特征进行重建，将特征图的通道数恢复到原始的通道数，以便之后得到的特征权重能够和输入特征图相乘。特征 **Z** 经过Sigmoid激活函数，最终得到特征图权重系数，之后与输入U相乘得到最终输出S。

$$S = X \otimes \mathrm{l}(\mathbf{Z}) \tag{7}$$

式(7)中，⊗ 表示逐元素相乘，l 是sigmoid函数。经过(4)、(5)的处理后，网络可以学习到全局信息，使得网络对于整个信息特征有了更好的理解，提升了网络的性能和表达能力，进而在各个图像分割任务中实现了更精确的性能。

2.2.3 空间注意力 SA 模块

在深度学习模型中，特别是图像处理应用中，SA 是通过处理空间信息动态地调整输入数据不同空间位置的重要性，来增强有用信息并抑制无用信息。这一机制通过处理空间信息，能够有效提升模型性能。在实现空间注意力机制时，首先计算输入的特征图在通道维度上的平均值和最大值，分别得到两个不同的特征图，代表每个位置上的平均信息和最大信息，具体计算过程如公式(8)、(9)所示。

$$X_{Avg} = \frac{1}{C}\sum_{c=1}^{C} X(c,h,w) \tag{8}$$

$$X_{Max} = \max_{c=1}^{C}(X(c,h,w)) \tag{9}$$

式(8)、(9)中，c，h，w分别代表特征图像对应的通道数、高、宽。如图5中SA模块所示，将输入特征图像X分别经过Avg_out_c和Max_out_c得到X_{Avg}和X_{Max}，由此分别得到了特征图在通道维度上每个位置的平均值信息和最大值信息，将X_{Avg}和X_{Max}进行拼接，通过1×1的卷积层和Sigmoid函数，得到每个位置的注意力权重，将输出的特征图的每个像素值都在0-1之间，用来表示该位置的信息的重要程度，公式如式（10）所示。

$$X_C = \mathrm{l}(F_{con}(X_{Avg}\ \mathrm{e}\ X_{Max})) \tag{10}$$

通过这种机制，模型能够更好理解和利用图像中每个位置的信息，从而在处理复杂的视觉任务时提高准确性。

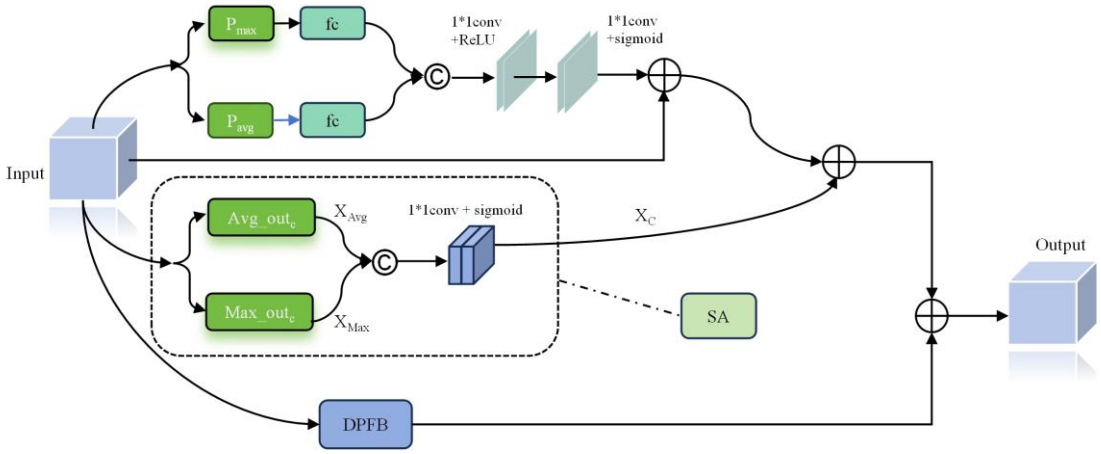

图 5 多层次注意力融合模块模型结构图

Fig. 5 MAF Model Structure Diagram

提出的 MAF 模块模型结构图如图 5 所示，将特征图分别输入到 CA 模块、SA 模块和本文提出的 DPFB 模块中进行细节特征提取，之后逐级拼接到一起，这种方法可以充分利用特征图的通道和空间信息并增强特征融合能力，进而提高模型的表征能力和分割性能。

2.3 视觉特征解析与转换网络（VIT）

在图像处理领域，VIT 将原本用于处理自然语言任务的 Transformer 架构[18]进行了创新性适配，VIT 通过将输入的图像划分为固定大小的图块（patches），然后将这些图块展平为固定长度的向量，引入到经典的 Transformer 结构中。通过这种方法，模型能够利用 Transformer 的强大能力去处理图像中的全局信息，VIT 的模型结构图如图 6 所示。

在自然语言处理（NLP）中时，Transformer 通过多头注意力机制来分析输入序列中不同位置之间的关系，计算每个位置相对于其他位置的重要性，这种操作可以使得模型在理解上下文信息时不受位置顺序的限制。在进行图像处理时，VIT 采用相似的策略，将图像划分为小的图块，将这些图块扁平化成一个序列，之后将图块和位置信息嵌入到解码器当中，通过计算 Query (Q)、Key (K)、和 Value (V)来处理这些图块的关系。在这个过程中，每个图块的信息都被全面的进行分析，通过利用子注意力机制来捕获图块之间的关系，确保模型能够捕捉到图像的细节和上下文信息。

综上所述，VIT 创新性地将 Transformer 架构应用于图像处理领域，利用多头注意力机制和图块划分策略，有效地处理了图像中的全局和局部信息。VIT 展现出在捕捉图像细节和上下文信息方面的强大能力，显著提升了图像处理任务的效果。

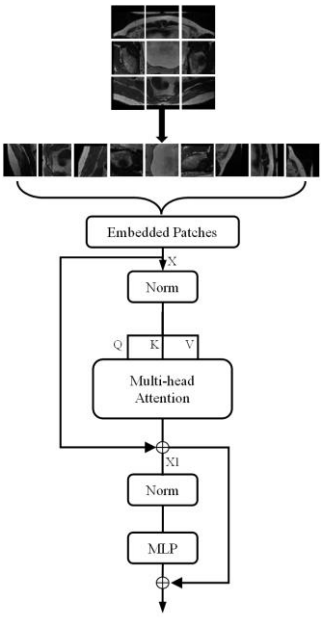

图 6 视觉特征解析与转换网络模型结构图

Fig. 6 Vision transformer model structure diagram

具体来说，输入 X 分别经过三个线性层（q、k、v）得到对应的 Q、K、V，然后进行多头注意力计算，其中注意力分数是通过 Q 与 K 的点积并应用 Softmax 函数获得的，如式（11）所示，然后与输入 X 相加得到 $X1$，进而实现了残差连接[19]，将输入直接传递到网络的后续层中，进而将最初的信息得以保留。将相加的结果经过两个线性层之后，再与 $X1$ 进行残差连接，最终输出结果。

$$Attention(Q,K,V)=softmax\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{11}$$

通过引入 VIT 模块，加强了模型对小分辨率的语义信息的理解，并采用残差连接来保留最初的信息，从而提高网络性能。相较于 CNN，VIT 通过多头注意力机制来捕捉图像的全局特征，更好地处理高分辨率图像和复杂的图像内容。

3 实验结果和分析

3.1 数据集

数据来源于 ISICDM 2019 临床数据分析挑战赛的基于磁共振成像的膀胱内外壁分割与肿瘤检测数据集，该数据采集自空军军医大学（第四军医大学），已经分别对肿瘤区域和膀胱壁区域进行了标记，标签使用 one-hot 编码进行处理。

3.2 实验设置

实验是在 Ubuntu22.04 操作系统上进行的，使用深度学习框架 Pytorch。GPU 配置为单卡的 NVIDIA GeForce GTX 4090，CUDA 的版本为 11.8。学习率为 0.0005，并设置了衰减权重，衰减率为 0.0001，每次训练批次（Batch）大小设置为 8，优化器选取了 Adam，并进行了 2000 个 Epoch 的训练周期。

3.3 评价指标

采用 IoU、Dice 系数、VOE、Precision 指标对模型分割性能进行评价。

$$Dice=\frac{2TP}{FP+2TP+FN} \tag{12}$$

$$IoU = \frac{TP}{FP + TP + FN} \tag{13}$$

$$VOE = 1 - \frac{2 \times |A \cap B|}{|A| + |B|} \tag{14}$$

$$Presion = \frac{TP}{FP + TP} \tag{15}$$

其中 TP（True Positive）表示模型预测为正类别且实际上就是正类别的像素数量，FP（False Positives）表示模型未能正确预测负类别的像素数量，FN（False Negatives）表示模型未能正确预测为正类别的像素数量。

Dice 指标的高低表示了模型的预测结果和真实结果的重叠程度，一定程度上反映了模型分割的性能。

IoU 表示预测区域与实际区域之间的重叠程度，取值范围在 0 到 1 之间，值越高表示模型预测与真实标签的重叠越好，模型的性能越强。

VOE 是用来衡量分割结果和真实标签之间重叠程度的指标。VOE 越低表示分割结果和真实标签之间吻合度越高。

Precision 是指分割结果中正确分割的像素数量占分割结果总像素数量的比例。它表示了分割结果中真正属于目标类别的像素的准确率。

3.4 模型性能对比分析

为了验证提出的 MIV-Net 网络的图像分割性能，将模型与当今主流的图像分割算法模型进行比较，结果如表 1 和表 2 所示。从表 1 可以看出，模型在进行膀胱肿瘤分割时 IoU、Precision、Dice 系数分别达到了 83.16%、91.66%、90.91%，相较于 U-Net 分别提高了 8.18%、10.5%、5.21%，取得了较大的提升，VOE 为 16.84%，比 U-Net 下降了 8.18%。

表 1 不同分割算法对膀胱肿瘤的分割效果

Table 1 The segmentation performance of different segmentation algorithms on bladder tumors.

方法	IoU(%)	VOE(%)	Precision(%)	Dice(%)
Unet(2015)	74.98	25.02	81.16	85.70
Unet3+(2020)	78.34	21.66	85.98	87.86
Attention_Unet(2018)	80.64	19.36	87.41	89.28
Unet++(2018)	66.93	33.07	87.18	80.19
Transunet(2021)	76.02	23.98	86.88	86.83
AGSE-VNet(2022)	81.92	18.08	88.64	90.06
DEA-Net(2024)	79.34	20.66	89.35	88.48
VM-UNet(2024)	80.25	19.75	89.42	89.04
MIV-Net	**83.16**	**16.84**	**91.66**	**90.91**

表 2 不同分割算法对膀胱壁的分割效果

Table 2 The segmentation performance of different segmentation algorithms on the bladder wall

方法	IoU(%)	VOE(%)	Precision(%)	Dice(%)
Unet(2015)	75.34	24.66	82.15	85.93
Unet3+(2020)	74.63	25.37	79.29	85.47
Attention_Unet(2018)	79.58	20.42	86.86	88.63
Unet++(2018)	66.93	33.07	87.18	80.19
Transunet(2021)	73.25	26.75	83.99	84.56
AGSE-VNet(2022)	81.62	18.38	88.70	89.88
DEA-Net(2024)	77.81	22.19	89.12	87.52
VM-UNet(2024)‘	77.37	22.63	89.42	89.04
MIV-Net	**82.66**	**17.34**	**90.43**	**90.51**

从表 2 可以看出，MIV-Net 模型在进行膀胱壁的图像分割时评价指标 IoU、Precision、Dice 系数分别达到了 82.66%、90.43%、90.51%，相比于基础模型 U-Net 提高了 7.32%、8.28%、4.58%，VOE 的值为 17.34%，比 U-Net 下降了 7.32%。Dice 系数比 Unet3+提高 5.04%，比 Attention_Unet 高 1.88%，比 Unet++高 10.32%，比 Transunet 高 5.95%，比 AGSE-VNet 高 0.63%，

比 DEA-Net 高 2.43%，比 VM-UNet 高 1.87%，可以看出，本研究所提出的模型分割性能在膀胱壁数据集中提升性能显著，优于与之对比的八个模型。

此外，由表 1 和表 2 可以看出，提出的 MIV-Net 模型相比于现在的主流模型 AGSE-VNet[11]、Unet3+[20]、Attention_Unet[8]、DEA-Net[15]、VM-UNet[16]、Unet++[21]、Transunet[22]等分割效果均有良好的表现，由此证明 MIV-Net 对于图像的特征提取和融合以及分割等性能都有着显著的优化效果。不同分割算法对膀胱肿瘤和膀胱壁的分割效果可视化对比如图 7 所示。

膀胱肿瘤的特点在于形态多样、边界模糊[23]，在进行分割时分割难度会大大增加，在图 7 的第一行和第二行可以看出 AGSE-VNet、Unet3+、Attention_Unet、DEA-Net、VM-UNet、Unet++、Transunet 分割模型在预测膀胱肿瘤边界以及形态多样大小不一的膀胱肿瘤时，出现了溢出和误判，MIV-Net 能更精准地预测肿瘤的边界和形态，更接近于真实的标签。

在图 7 的第三行和第四行可以看出 Unet3+、Attention_Unet、Unet++、Transunet、DEA-Net 分割模型在预测膀胱壁时，出现了误判，将部分膀胱壁组织预测成了膀胱肿瘤；在第四行中 Transunet、Attention_Unet、VM-UNet 模型在预测膀胱壁时，均出现了缺失的情况；膀胱壁在一些部分突然变薄时，其他模型对于边缘的预测结果均不理想，存在大量的误判；MIV-Net 模型在面对较为复杂情况时，仍然能够精准分割出了膀胱壁的形状和位置。

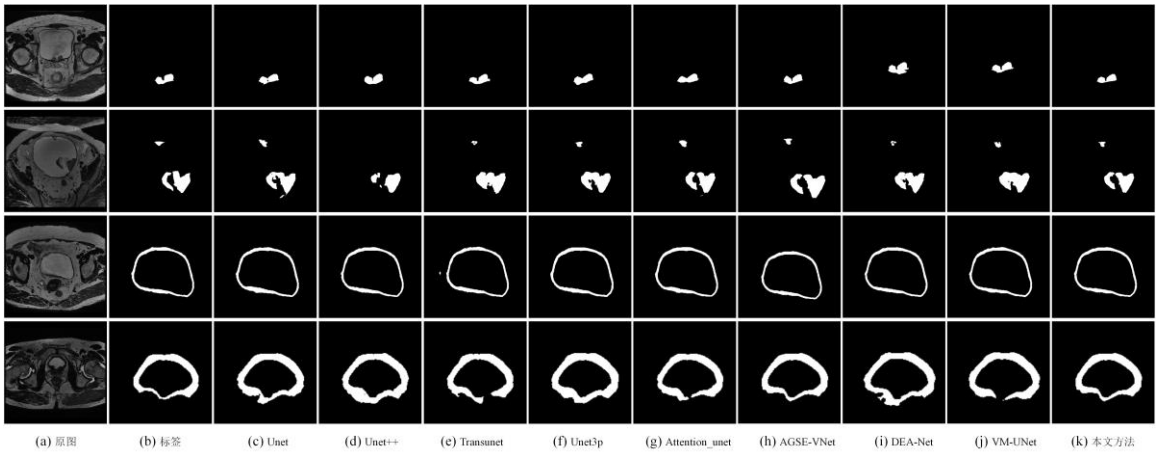

图 7 不同分割算法对膀胱肿瘤和膀胱壁的分割效果可视化对比

Fig. 7 Visualization of the segmentation performance comparison of different segmentation algorithms on bladder tumors and bladder walls.

3.5 消融实验

为了具体研究加入的各个模块对于膀胱肿瘤和膀胱壁分割效果的影响，在实验条件相同的情况下，依次将多层次注意力融合模块、CNN 模块、多尺度核心特征融合器模块、VIT 模块加入到基础网络中进行消融实验。

膀胱肿瘤消融实验的具体指标结果如表 3 所示，膀胱壁的消融实验的具体指标结果如表 4 所示。通过分析两组消融数据，在加入了多层次注意力融合模块、CNN 模块、多尺度核心特征融合器模块、VIT 模块，IoU、Precision、Dice 系数有 4.44%-5.87%的提高。多层次注意力融合模块、多尺度核心特征融合器模块、VIT 模块三种模块组合在整体上的分割效果是最好的，如图表 3、4 所示，在加入了三种模块之后，网络模型在 IoU、Precision、Dice 系数以及 VOE 系数中均表现为最佳，三种模块的组合在总体上提供了最佳的分割效果。

表 3 加入不同模块对膀胱肿瘤的分割效果

Table 3 Visualization of the segmentation performance comparison of bladder tumors with the incorporation of different modules

方法	IoU(%)	VOE(%)	Precision(%)	Dice(%)
U-Net	74.98	25.02	81.16	85.70
+CNN	78.72	20.28	85.79	87.63
+InA	81.82	18.18	89.05	90.0
+MFC	81.92	18.08	88.64	90.06
+MFC_MAF	82.24	17.76	88.88	90.25
+MFC_VIT	82.47	17.53	90.82	90.4
MIV-Net	**83.16**	**16.84**	**91.66**	**90.91**

表 4 加入不同模块对膀胱壁的分割效果

Table 4 Visualization of the segmentation performance comparison of bladder tumors with the incorporation of different modules

方法	IoU(%)	VOE(%)	Precision(%)	Dice(%)
Unet	75.34	24.66	82.15	85.93
+CNN	78.29	21.71	84.61	87.82
+ InA	80.16	19.84	88.96	88.99
+MFC	81.62	18.38	88.70	89.88
+ MFC_MAF	81.74	18.16	89.67	89.95
+MFC_VIT	81.93	18.07	90.12	90.07
MIV-Net	**82.66**	**17.34**	**90.43**	**90.51**

膀胱壁和膀胱肿瘤的可视化消融实验对比图如图 8 所示，在第二行中可以看出在基础模型 U-Net 中加入了 InA、MFC、VIT 等模型之后，在进行膀胱肿瘤分割预测时，膀胱肿瘤的可视化图像轮廓更加清晰，大幅减少了误判以及边界浑浊的现象。相较于 U-Net 模型和加入了 CNN 模型的可视化图均有较好的提升，在同时加入了三个模块之后，本文方法图像分割的模糊混浊边界得到了改善,分割预测出了轮廓更加平滑、清晰，更接近标签值(groundtruth,，GT)的分割效果。

在图 7 中第三行和第四行的可视化图中，U-Net 模型出现了误判以及边界浑浊等现象，在加入了 CNN 模块之后，可视化图像的表现依旧较差，在加入了 InA、MFC、VIT 等模型时候，出现了小部分缺失，误判。在结合三个模型在整体上的表现和标签值更加接近，表现优异。

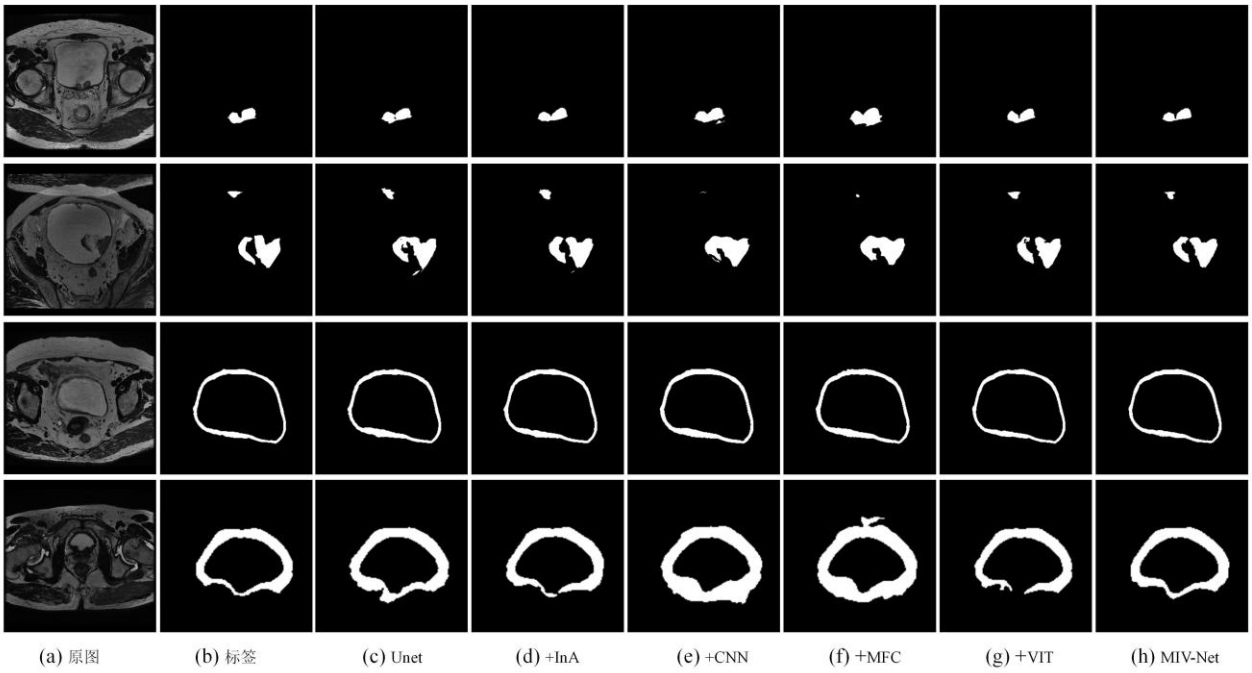

图 8 消融实验可视化对比图

Fig. 8 Visualization of the comparison between ablation experiments

4 结束语

本研究提出一种基于 Transformer 与多层注意力机制的膀胱医学图像分割预测模型 MIV-Net，创新性地加入了多尺度核心特征融合器模块、视觉特征解析与转换网络，有效地融合了图像的全局信息和局部特征信息。此外，在解码部分加入了多层次注意力融合模块增强了模型的边缘特征能力以及提取重要特性信息的能力，同时有效地抑制了非关键信息，从而在膀胱肿瘤和膀胱壁数据集上展现出卓越的分割性能。该算法通过广泛的对比试验和消融实验，在多个性能指标上均表现优异，验证了其在实际的医学图像分割场景中的应用潜力。下一步将优化模型的参数量，降低计算复杂度提高效率，将先进的图像分割技术推广到更广泛的医学图像处理应用中。

参考文献:

[1] Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA: a cancer journal for clinicians, 2024, 74(3): 229-263.

[2] Siegel R L, Miller K D, Wagle N S, et al. Cancer statistics, 2023[J]. Ca Cancer J Clin, 2023, 73(1): 17-48.

[3] Gu Z, Cheng J, Fu H, et al. Ce-net: Context encoder network for 2d medical image segmentation[J]. IEEE transactions on medical imaging, 2019, 38(10): 2281-2292.

[4] Moghadas-Dastjerdi H, Ahmadzadeh M, Samani A. Towards computer based lung disease diagnosis using accurate lung air segmentation of CT images in exhalation and inhalation phases[J]. Expert Systems with Applications, 2017, 71: 396-403.

[5] 张玮智, 于谦, 苏金善, 等. 从 U-Net 到 Transformer: 深度模型在医学图像分割中的应用综述[J]. 计算机应用, 2023: 0.
ZHANG W J,YU Q,SU J S,et al. Application review of deep models in medical image segmentation: from U-Net to Transformer[J]. Journal of Computer Applications. 2023: 0.

[6] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[7] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.

[8] Oktay O, Schlemper J, Folgoc L L, et al. Attention u-net: Learning where to look for the pancreas[J]. arXiv preprint arXiv:1804.03999, 2018.

[9] Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation[J]. arXiv preprint arXiv:2102.04306, 2021.

[10] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[11] Guan X, Yang G, Ye J, et al. 3D AGSE-VNet: an automatic brain tumor MRI data segmentation framework[J]. BMC medical imaging, 2022, 22(1): 1-18.

[12] Wang Y, Li X, Ye X. LCANet: A Lightweight Context-Aware Network for Bladder Tumor Segmentation in MRI Images[J]. Mathematics, 2023, 11(10): 2357.

[13] Zhang Q, Liang Y, Zhang Y, et al. A comparative study of attention mechanism based deep learning methods for bladder tumor segmentation[J]. International Journal of Medical Informatics, 2023, 171: 104984.

[14] 吴港,全海燕.一种基于特征融合的息肉分割双解码模型[J].数据采集与处理,2024,39(4):954-966
WU Gang,QUAN Haiyan, A Double-Decoding Model for Polyp Segmentation Based on Feature Fusion[J]. Journal of Data Acquisition and Processing, 2024,39(4):954-966

[15] Chen Z, He Z, Lu Z M. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention[J]. IEEE Transactions on Image Processing, 2024.

[16] Ruan J, Xiang S. Vm-unet: Vision mamba unet for medical image segmentation[J]. arXiv preprint arXiv:2402.02491, 2024.

[17] Hou G, Qin J, Xiang X, et al. Af-net: A medical image segmentation network based on attention mechanism and feature fusion[J]. Computers, Materials & Continua, 2021, 69(2): 1877-1891.

[18] Vaswani A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017.

[19] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[20] Huang H, Lin L, Tong R, et al. Unet 3+: A full-scale connected unet for medical image segmentation[C]//ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2020: 1055-1059.

[21] Zhou Z, 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, 2019, 39(6): 1856-1867.

[22] Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation[J]. arXiv preprint arXiv:2102.04306, 2021.

[23] 陶惜婷, 叶青. 融合 CNN 和 Transformer 的并行双分支皮肤病灶图像分割 [J]. 计算机应用研究, 2024, 41 (8).
TAO X T,YE Q. Parallel dual-branch image segmentation of skin lesions fusing CNN and Transformer[J]. Application Research of Computers,2024,41 (8).

作者简介：

高博艺(2000-)，男，硕士研究生，研究方向：深度学习，E-mail:gby17839488691@163.com。

丁学明(1971-)，**通讯作者**，男，博士，副教授，硕士生导师，研究方向：深度学习，E-mail：xuemingding@usst.edu.cn。

丁雪峰(1999-)，男，硕士研究生，研究方向：深度学习，E-mail:dingfaded@icloud.com。

胡鸿翔(2000-)，男，硕士研究生，研究方向：深度学习，E-mail:1906760471@qq.com。