DOI: 10.13382/j.jemi.B2508933

基于改进 U-Net 网络的遥感图像分割算法*

赵亚伟[1] 霍翊茗[2] 葛 坤[3] 冉 宁[1,4]

(1. 河北大学电子信息工程学院 保定 071002;2. 河北大学国际学院 保定 071002;
3. 河北大学化学与材料科学学院 保定 071002;4. 河北大学节能技术研发中心 保定 071002)

摘 要:针对遥感图像存在目标尺度差异大、对象的稀疏分布和易混淆类别之间的高度相似性等问题,构建了一种基于改进 U-Net 的遥感图像语义分割方法 DKTU-Net,在 U-Net 框架基础上,对编码器与特征融合结构进行了针对性改进。首先,将原有卷积编码器替换为融合卷积与 Shunted Transformer 的混合型编码结构,在保持局部细节敏感性的同时,能够有效捕获遥感图像中跨尺度、跨区域的全局语义依赖;然后,采用细节保持上下文融合模块(detail-preserving contextual fusion, DPCF)实现跳跃连接与上采样特征融合,通过可学习的融合策略自适应地组合多尺度特征,改善了目标尺度差异大、对象的稀疏分布问题;最后,引入核选择融合注意力模块(kernel selective fusion attention, KSFA),对上下文融合后的特征进行特征增强,该模块能根据不同类别需求自适应地调整感受野,缓解了易混淆类别之间的高度相似性的问题。在 ISPRS Vaihingen 与 ISPRS Potsdam 两个高分辨率遥感数据集上进行了实验,结果表明,提出的 DKTU-Net 相比于原 U-Net 网络模型,在 ISPRS Vaihingen 数据集上的 MIoU 和 m-F1 分别提高了 6.52%和 4.87%;在 ISPRS Potsdam 数据集上的 MIoU 和 m-F1 分别提高了 3.54%和 2.4%,改善了遥感图像分割存在目标尺度差异大、对象的稀疏分布和易混淆类别之间的高度相似性等问题。

关键词: 遥感图像;语义分割;多尺度特征融合;注意力机制

中图分类号: TP391.41;TN911.73 **文献标识码**: A **国家标准学科分类代码**: 520.6040

Remote sensing image segmentation algorithm based on the improved U-Net network

Zhao Yawei[1] Huo Yiming[2] Ge Kun[3] Ran Ning[1,4]

(1. College of Electronic & Information Engineering, Hebei University, Baoding 071002, China; 2. International College, Hebei University, Baoding 071002, China; 3. College of Chemistry and Materials Science, Hebei University, Baoding 071002, China; 4. Laboratory of Energy-saving Technology, Hebei University, Baoding 071002, China)

Abstract: To address the issues of significant scale differences in targets, sparse distribution of objects, and high similarity between certain classes in remote sensing images, an improved semantic segmentation method DKTU-Net for remote sensing image based on U-Net structure is proposed. Based on the U-Net framework, this method makes targeted improvements to the encoder and feature fusion structure. Firstly, the original convolutional encoder is replaced by a hybrid encoding structure that fuses convolution and Shunted Transformer, which can effectively capture cross-scale and cross-region global semantic dependencies while maintaining local detail sensitivity. Then, the detail-preserving contextual fusion module (DPCF) is adopted to fuse skip connections and upsampled features, and through a learnable fusion strategy, multi-scale features are adaptively combined to improve the problems of significant scale differences and sparse distribution of objects. Finally, the kernel selective fusion attention module (KSFA) is introduced to enhance the features after context fusion. This module can adaptively adjust the receptive field according to the needs of different classes, alleviating the problem of high similarity between easily confused classes. Experiments were conducted on the two high-resolution remote sensing datasets, ISPRS Vaihingen and ISPRS Potsdam. The results show that the proposed DKTU-Net, compared with the original U-Net

收稿日期: 2025-12-03 Received Date: 2025-12-03

*基金项目:国家自然科学基金(62373132)、河北省自然科学基金(F2025201023)、石家庄市驻冀高校基础研究项目(241791367A)、河北大学优秀青年科研创新团队建设项目(QNTD202411)、河北大学多学科交叉研究计划资助项目(DXK202409)

network model, The MIoU and M-F1 on the ISPRS Vaihingen dataset increased by 6.52% and 4.87% respectively, and the MIoU and M-F1 on the ISPRS Potsdam dataset increased by 3.54% and 2.4% respectively It has effectively improved the problems existing in remote sensing image segmentation, such as large differences in target scale, sparse distribution of objects, and high similarity between easily confused categories.

Keywords: remote sensing image; semantic segmentation; multi-scale feature fusion; attention mechanism

0 引　言

遥感图像包含丰富的地表信息,广泛应用于环境监测、交通监控、土地普查、海洋测绘等方面[1],在国民经济发展中发挥着重要作用。遥感图像分割是对遥感图像中的每个像素进行精确分类并分配特定的语义标签,以实现对建筑物、道路等物体的精确识别。随着卫星技术和无人机遥感技术的高速发展,遥感影像的空间分辨率已经达到厘米级[2-4],这使得地物细节更加丰富,但也随之出现了遥感图像阴影干扰严重、地物类型繁多、易混淆类别之间高度相似、尺度变化大等问题[5-8]。K-means 聚类算法、基于图形法、马尔可夫随机场以及阈值分割等传统图像分割方法在单一的场景下能够取得较好的分割效果,但在高分辨率、复杂结构的城市遥感影像中难以实现精细化分割的要求。

随着深度学习的不断发展,卷积神经网络(convolutional neural network, CNN)凭借其强大的特征提取能力和非线性拟合能力被广泛应用于语义分割领域[9]。Long 等[10]提出的全卷积网络(fully convolutional network, FCN),用卷积层替代传统的全连接层,实现了端到端像素级分割。Ronneberger 等[11]提出的 U-Net 使用 U 型的设计来获取不同尺度信息,并对称连接不同尺度信息,进一步提高了分割效果。Badrinarayanan 等[12]提出了 SegNet,引入了池化索引技术,在解码器中使用索引进行上采样,可以省去其上采样过程的学习,在计算内存与效率上得到了充分提升。Zhou 等[13]提出的 U-Net++采用密集连接,缓解了高低级语义信息在连接时存在的语义差距。但是,这些模型依靠固定的卷积核,在感受野方面是非常有限的,在处理遥感图像时难以提取存在长距离依赖关系的特征,这使得分割会出现不够连续、目标边界模糊等问题。为了解决感受野有限的问题,Chen 等[14-16]提出了空洞金字塔卷积(atrous spatial pyramid pooling, ASPP)来分割多个尺度的对象。利用不同空洞率的空洞卷积并进行不同尺度的特征融合,从而获得多个尺度的对象以及图像上下文。Zhao 等[17]提出的 PSPNet,其核心模块是金字塔池化模块,通过不同尺寸的卷积核收集了不同感受野的特征层,增强了全局上下文特征感知。虽然这些模型能够提取丰富的上下文信息,但是金字塔池化模块固定的尺度设计会导致模型在处理多尺度变化情况下的适应性有限,导致在小目标识别和边缘细节恢复方面表现不足。

注意力机制出现使模型能够将更重要的特征集中在特定区域,提高了恢复图像细节的能力和语义分割的精度。Woo 等[18]引入了空间通道注意力机制,旨在增强任何基础网络中信道和空间维度的特征信息,从而增强模型的性能。尽管 CNN 在添加注意力机制后在遥感图像分割方面取得了良好的效果,但是,在进行卷积运算时,每个卷积主要强调局部信息,这使得卷积运算存在局限性,难以将上下文信息与全局信息相结合。Transformer[19]为解决这个问题提供了一个新思路,自视觉变换器(vision transformer, ViT)[20]提出以来,基于多头自注意力机制的 Transformer 能够熟练地捕获全面的全局上下文信息,有效弥补了卷积结构的局限性。随后 Liu 等提出 Swin-Transformer[21]方法加入了分层特征提取机制,这使得 Transformer 能够适应高分辨率视觉任务。虽然基于 Transformer 的语义分割模型在全局依赖建模和语义一致性方面表现突出,但是,单纯利用 Transformer 结构在图像的局部信息处理中因为缺乏空间归纳偏差,使其捕获细节能力受到限制。难以精准刻画地物边界和小目标结构。同时 Transformer 模型的高计算复杂度和对数据规模的依赖,也限制了其在实际遥感任务中的广泛应用。

针对以上问题,研究人员逐渐意识到 CNN 与 Transformer 各自的优势具有互补性:CNN 善于提取局部纹理与空间细节,而 Transformer 擅长捕获长程依赖与全局语义。为了充分利用两者的互补特性,人们通过在网络的编码阶段引入 Transformer 模块[22],或者在卷积特征上施加多尺度注意力机制,在保持分辨率与边界敏感度的同时,提升模型的全局特征表达能力。

本文在 U-Net 网络框架基础上,构建了一种基于改进 U-Net 的遥感图像语义分割方法 DKTU-Net。通过对现有特征建模与融合模块进行合理组合与结构优化,保留了 U-Net 经典的“编码-解码”对称结构及跳跃连接机制。在编码阶段,引入融合卷积特征与 Shunted Transformer 的混合型编码结构,在保持局部细节建模能力的同时,有助于增强了模型对遥感影像中跨尺度、跨区域语义信息的建模能力;在特征融合阶段,引入细节保持上下文融合模块(detail-preserving contextual fusion, DPCF)实现跳跃连接与上采样特征融合,通过可学习的融合策略自适应地整合多尺度特征,以更好地适应目标

尺度差异较大、对象分布稀疏的遥感场景;在特征增强阶段,采用核选择融合注意力模块(kernel selective fusion attention, KSFA),对上下文融合后的特征进行进一步增强,在一定程度上缓解了易混淆类别之间的高度相似性的问题。实验结果表明,提出的改进模型 DKTU-Net 在遥感图像上表现出良好的分割性能。

1　U-Net 算法

1.1　U-Net 概述

U-Net 是 Ronneberger 等[11]提出一种用于生物医学图像分割的卷积网络,该网络采用了对称的编码器-解码器网络结构实现端到端的语义分割。跳跃连接方式使得解码器能够直接访问编码器中高分辨率的特征信息。在编码器阶段,通过连续的下采样操作逐步提取图像特征,使网络能够获得更深层次语义信息;在解码器阶段,通过上采样恢复图像分辨率,并利用跳跃连接将编码器的低层特征直接传递给解码器所对应的层。通过跳跃连接,实现特征融合,保证解码网络能利用更多的信息恢复图像。因 U-Net 网络具有结构简单、容易训练等特点,被广泛地应用在医学影像分析、自然图像分割等领域。

1.2　U-Net 的不足与局限

U-Net 网络被用于医学影像的分割,但遥感影像与医学影像不同,其背景复杂、地物的类型丰富,基础的 U-Net 网络无法很好地适应。遥感图像中存在地物目标的尺度差异巨大问题,如车辆、植被斑块、大型建筑同时存在,U-Net 是基于传统卷积神经的网络,其感受野有限,难以同时有效捕获大目标和小目标的全局上下文信息与局部细节。跳跃连接在一定程度上能够融合细节信息,但缺乏有效的特征选择机制,可能将未处理的噪声信息直接传递给解码器,导致解码过程受到干扰。U-Net 的核心操作是局部卷积,捕获的远程依赖关系是有限的。这导致 U-Net 在处理遥感影像这类复杂场景时,难以凭借远距离像素之间的语义关系来消除歧义,容易将高度相似性地物混淆,导致分割错误。因此,后续的许多研究工作都致力于在 U-Net 的基础上进行改进,以克服这些局限性。

2　DKTU-Net 模型框架

2.1　网络结构

DKTU-Net 使用 U-Net 作为基本网络架构,总体网络框架如图 1 所示。该网络整体采用编码器-解码器架构,由 Shunted Transformer 编码器、DPCF 模块、KSFA 模块以及多级上采样解码器组成。在编码器端,用混合型 Shunted Transformer 作为 DKTU-Net 网络的特征提取部分,通过 4 个阶段下采样,提取出不同层级的特征,分辨率依次为原图像的 1/4、1/8、1/16、1/32。该编码器通过分层卷积嵌入机制,以此来实现全局上下文与局部空间特征的联合编码,提取出不同尺度的特征表示。在解码端,DPCF 模块将来自于编码端的浅层特征和与之对应的解码端的深层语义特征进行融合,将高层特征和低层特征在通道维度上划分为多个子块,通过自适应融合单元动态的调整高层特征和低层特征的权重比例,以此来充分的保留空间和语义信息。在融合阶段后引入 KSFA 模块,该模块通过两个不同的膨胀深度卷积提取不同尺度的特征,将得到的两个分支的特征分别在空间和光谱维度上进行处理,从而实现空间与光谱的联合建模,这一设计进一步增强了目标边界识别和光谱相似辨别的能力。最后,通过恢复模块逐步恢复特征分辨率,生成与原始图像相同尺寸的分割结果。

2.2　编码器结构

传统 Transformer 的多头自注意力在固定尺度的 token 表示上计算注意力,每个 token 只代表单一尺度的特征,这限制了每个自注意层捕捉多尺度特征的能力,导致模型在捕获小目标对象时效率降低。本文采用 Shunted Transformer[23]编码器,该编码器引入了一种新颖且通用的分流自注意力机制(shunted self-attention, SSA),其设计使同一层内的不同注意力头能够分别建模粗粒度与细粒度特征,以实现多尺度信息的有效表征。与传统通过合并太多 token 导致无法获得小对象的方法不同,SSA 通过在同一层内的多组注意力头并行的方法来提取不同尺度的目标,使得在保证计算效率的同时也能对细粒度信息的充分保留。

Shunted Self-Attention 模块是 Shunted Transformer 编码器的核心组成部分,如图 2 所示。该模块将每个分支在不同尺度上分别加入自注意力机制,以此来获取不同层级的空间依赖关系,同时将这些特征在通道维上进行加权融合,以获得综合特征。Shunted Self-Attention 模块使网络具备了在单个层之间同时获得局部纹理信息和全局结构语义的能力,从而在遥感图像分割任务中表现出更强的特征表达能力和目标识别能力。DKTU-Net 编码器的特征提取过程由 4 个阶段构成,各个阶段包含重叠卷积嵌入、Shunted Self-Attention 机制以及前馈网络。每个阶段输出的特征图的空间分辨率逐渐降低,分辨率约为原始图像的 1/4、1/8、1/16、1/32,语义表达能力逐步增强。通过这种层级式的特征提取操作,DKTU-Net 算法在获取细微局部特征的同时也能兼顾全局语义信息,使模型在遥感图像这类复杂场景下能维持良好的分割精度。

2.3　DPCF 细节保持上下文融合模块

在遥感图像中,不同目标之间往往存在显著的尺度

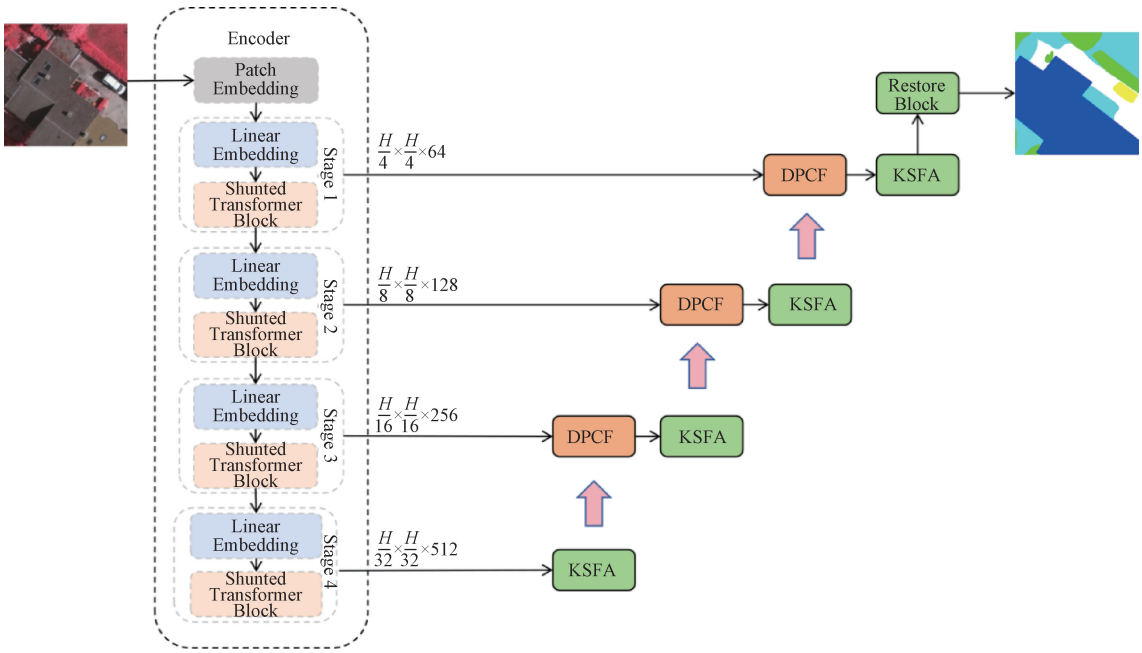

图1　DKTU-Net结构

Fig. 1　DKTU-Net architecture diagram

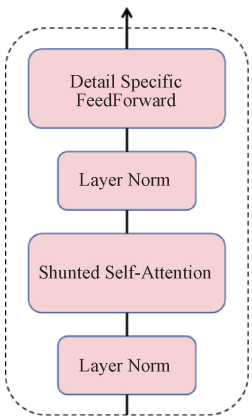

图2　Shunted Self-Attention模块

Fig. 2　Shunted Self-Attention module

差异,例如地面区域与车辆等小尺度目标。在解码阶段进行多尺度特征融合时,尺度较小目标所对应的语义特征容易被大尺度特征所淹没,从而导致小目标信息的稀释甚至丢失。为缓解上述问题,引入DPCF细节保持上下文融合模块[24],该模块通过一种自适应融合机制动态调整高、低分辨率特征的贡献比例,使模型能够根据每个像素的上下文判断该保留哪些细节、接受哪些语义信息,从而最大程度上避免细节丢失。

DPCF模块的结构如图3所示。该模块将低分辨率特征图 $\boldsymbol{F}_l$ 和高分辨率特征图 $\boldsymbol{F}_h$ 沿通道维度划分为4个相等的特征段,引入一个单一的可学习参数 $\alpha'\in\boldsymbol{R}^{1\times1\times1}$,该参数在空间维度和通道维度上进行扩展,以匹配特征段的形状,从而得到 $\alpha'\in\boldsymbol{R}^{H\times W\times\frac{C}{4}}$,然后将这个扩展后的参数 α 通过sigmoid函数处理,以生成用于每个空间位置和通道段的门控权重 $\beta\in[0,1]$,如式(1)所示,权重 β 代表低分辨率特征的贡献度,而 $(1-\beta)$ 代表高分辨率特征对每个特征段的贡献度。自适应融合通过每个段的加权和来实现,如式(2)所示,自适应融合后,4个特征段沿着通道维度重新拼接成特征图 $\boldsymbol{F}$,然后通过一个卷积块,通常是3×3的卷积组成,然后是批归一化(batch normalization, BN)和一个ReLU激活函数,对融合进行细化,得到最终的输出。

$$\beta=sigmoid(\alpha),\beta\in\boldsymbol{R}^{H\times W\times\frac{C}{4}} \tag{1}$$

$$o_i=\beta\odot l_i+(1-\beta)\odot h_i \tag{2}$$

式中:$\odot$ 表示逐元素乘法;o_i 代表第 i 个分段的选择性聚合特征图;l_i 代表第 i 个分段的低分辨率特征图;h_i 代表第 i 个分段的高分辨率特征图。

通过上述自适应融合方式,在多尺度特征融合时,网络在保持全局语义一致性的同时,能够更有效地保留遥感影像中小尺度目标等细节信息。

2.4　KSFA核选择注意力模块

在遥感影像中,不同地物类别在光谱响应和纹理特征上往往具有较高相似性,例如道路与屋顶、低植被与树木等,容易产生误分类问题。为此,引入KSFA核选择注意力模块[25],该模块通过核选择机制对不同尺度的感受野进行自适应调节,使模型能够根据不同类别的空间结构特性选择更合适的上下文范围,从而增强对易混淆类

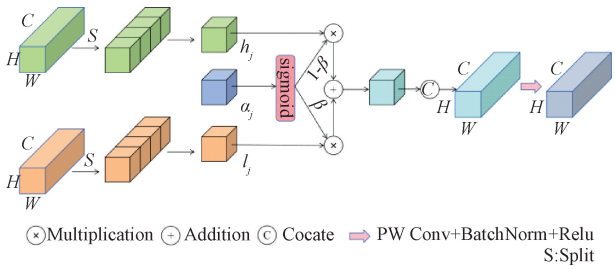

图 3　DPCF 模块

Fig. 3　DPCF module

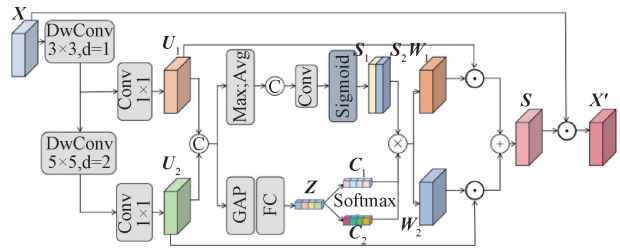

图 4　KSFA 模块

Fig. 4　KSFA module

别的判别能力。

KSFA 模块主要由两部分构成，如图 4 所示，第 1 部分是通过 2 个不同的膨胀深度卷积提取不同尺度的特征，接着通过 1×1 卷积统一通道数，得到 2 个不同感受野的特征图 $\boldsymbol{U}_1$、$\boldsymbol{U}_2$。第 2 部分是空间-光谱选择机制，包括空间选择和光谱选择。在空间选择方面，将多尺度特征 $\boldsymbol{U}=[\boldsymbol{U}_1;\boldsymbol{U}_2]$ 在通道维度拼接后，分别施加通道维度的平均池化与最大池化，得到 2 个空间描述符。将二者拼接并通过一个 1×1 卷积层，再经过 Sigmoid 激活函数，生成与卷积核数量对应的空间选择掩码 $\boldsymbol{S}_i$。在光谱选择方面，对拼接特征 $\boldsymbol{U}$ 进行空间全局平均池化，压缩为光谱特征描述符，再通过一个全连接层进行非线性变换。随后，通过一个 Softmax 函数在 2 个可学习的向量间进行计算，生成光谱选择权重，如式(3)所示。

$$c_1^{[j]}=\frac{e^{\boldsymbol{M}_j\boldsymbol{Z}_j}}{e^{\boldsymbol{M}_j\boldsymbol{Z}_j}+e^{\boldsymbol{N}_j\boldsymbol{Z}_j}},c_2^{[j]}=\frac{e^{\boldsymbol{N}_j\boldsymbol{Z}_j}}{e^{\boldsymbol{M}_j\boldsymbol{Z}_j}+e^{\boldsymbol{N}_j\boldsymbol{Z}_j}} \tag{3}$$

式中：$c_1^{[j]}$ 和 $c_2^{[j]}$ 是第 j 个光谱通道的注意力值，分别对应 2 个不同的感受野的光谱权重；$\boldsymbol{M}_j$ 和 $\boldsymbol{N}_j$ 是第 j 个光谱通道对应的两个可学习的向量；$\boldsymbol{Z}_j$ 是通过全局平均池化和全连接层得到的特征向量 $\boldsymbol{Z}$ 的第 j 个元素。

接着将得到的光谱权重 $\boldsymbol{C}_i$ 与空间掩码 $\boldsymbol{S}_i$ 进行矩阵相乘，得到每个尺度特征对应的复合空间-光谱权重 $\boldsymbol{W}_i$。此权重同时考虑了空间位置和光谱通道的重要性。随后，各尺度特征图 U_i 分别与对应的权重 $\boldsymbol{W}_i$ 进行加权求和，并通过一个 1×1 卷积融合，得到增强后的注意力特征 $\boldsymbol{S}$。最后将输入特征 $\boldsymbol{X}$ 与学习到的注意力特征 $\boldsymbol{S}$ 进行逐元素相乘得到最后的输出 $\boldsymbol{X}'$。

3　实验数据与结果分析

3.1　实验环境与超参数设置

实验环境如表 1 所示，训练使用了权重衰减为 0.001 的 Ranger 优化器。学习率初始化为 0.01，使用 1/epoch 的自定义调整方案进行学习率调度，训练轮数设置为 150，训练批大小为 24，图像大小为设置为 256×256 pixels，其余超参数均采用默认值。

表 1　实验环境

Table 1　Experimental environment

名称	配置信息
操作系统	Ubuntu 22.04
开发语言	Python3.10
深度学习框架	Cuda12.1+Pytorch2.1.0
GPU	RTX 3090(24G)*4
CPU	Intel Xeon Gold 6226R*2

3.2　数据集与评价指标

使用国际摄影测量与遥感学会(ISPRS)提供的 2 个高分辨率遥感公开数据集 ISPRS Vaihingen 和 ISPRS Potsdam[26] 对 DKTU-Net 算法进行评估。这 2 个数据集包括 6 种最常见的土地覆盖类别：不透水地表、建筑物、低矮植被、树木、车辆和背景。ISPRS Vaihingen 数据集包含 33 张高分辨率航拍影像，该影像采用近红外、红、绿(IRRG)三通道，空间分辨率为 9 cm/pixel。选择 11 张遥感影像进行训练(图像 ID：1、3、5、7、13、17、21、23、26、32 和 37)，5 张遥感影像进行测试(图像 ID：11，15，28，30 和 34)。ISPRS Potsdam 数据集包含 38 张高分辨率航拍影像，采用红、绿、蓝、近红外四通道以及相应的 DSM。相比于 ISPRS Vaihingen 数据集，空间分辨率更高，达到 5 cm/pixel。该数据集包含大量遮挡和阴影及多样化建筑形态，场景结构更复杂，对模型的泛化能力和鲁棒性提出了更高要求。选择 26 张遥感影像进行训练，其他 12 张遥感影像进行测试(图像 ID：2_13、2_14、3_13、3_14、4_13、4_15、5_13、5_14、5_15、6_14、6_15、7_13)。将用到的影像通过滑动窗口裁剪将它们裁剪为 256×256 pixels 的大小。

为了评估模型的性能，本实验采用了在语义分割领域常用的评估标准，包括交并比(intersection over union，IoU)、平均交并比(mean IoU，MIoU)以及平均 F1 分数(mean F1-score，m-F1)。IoU 是衡量单一类别预测区域与真实区域重叠程度的指标，而 MIoU 是在所有类别上计算 IoU，然后将结果取平均，反映模型对所有类别的

平均性能,能避免模型只在主要类上表现好而忽视小类的问题,F1 分数是精确率(precision)与召回率(recall)的调和平均,而 m-F1 是各类别 F1 平均值。计算公式如式(4)~(7)所示。

$$IoU = \frac{TP}{TP + FP + FN} \tag{4}$$

$$MIoU = \frac{1}{N}\sum_{i=1}^{N} IoU_i \tag{5}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{6}$$

$$m\text{-}F1 = \frac{1}{N}\sum_{i=1}^{N} F1_i \tag{7}$$

式中:TP(true positive)表示实际是正类,模型预测为正类的数量;FP(false positive)表示实际是负类,模型预测为正类的数量;FN(false negative)表示实际是正类,模型预测为负类的数量;C 表示数据集中类的总数。

3.3　消融实验

为了验证每个组件的有效性,在 Vaihingen 数据集上进行了消融实验,结果如表 2 所示,其中 Baseline 为 U-Net 网络、ST 为 Shunted Transformer 编码器。

表 2　ISPRS Vaihingen 数据集消融实验

Table 2　Ablation experiments on the ISPRS Vaihingen dataset　　(%)

模型	IoU					MIoU	m-F1
	不透水地表	建筑	低矮植被	树木	车辆		
Baseline	73.78	82.51	59.54	71.15	47.65	66.93	79.53
Baseline+ST	76.70	85.76	59.59	73.41	63.73	71.84	83.27
Baseline+ST+KSFA	77.44	**86.68**	61.24	**73.84**	63.58	72.56	83.76
Baseline+ST+KSFA+DPCF	**78.56**	86.62	**61.48**	73.36	**67.23**	**73.45**	**84.40**

由表 2 可知,将 U-Net 模型的编码部分替换为混合型 Shunted Transformer 编码器后,模型的分割性能显著提升。相比于原模型的 MIoU 和 m-F1 分别提高了 4.91%和 3.74%,特别是对于不透水表面、建筑、树和车辆,在 IoU 上分别提高了 2.92%、3.25%、2.26%和 16.08%,其中车辆类别的识别性能提升最为显著,这表明引入混合型 Shunted Transformer 后,模型在小尺度及稀疏分布目标上的特征表征能力得到了明显增强。在添加 KSFA 核选择注意力模块后,对于低矮植被、树木类光谱与纹理特征高度相似的类别在 IoU 上分别提高了 1.70%和 2.69%,这表明该模块通过自适应感受野选择增强了模型对上下文差异的建模能力,有助于提升模型对易混淆类别的判别能力。在此基础上引入 DPCF 模块后,MIoU 和 m-F1 分别提高了 0.64%和 0.89%,同时,对不透水地表和车辆这类尺度差异较大的目标上,IoU 分别提高了 1.12%和 3.65%。结果表明,DPCF 通过自适应调节高、低分辨率特征的融合权重,在保持语义一致性的同时有效保留细节信息,从而在一定程度上缓解了遥感影像中目标尺度差异大和小目标易被忽略的问题。

3.4　对比实验

为了评估所提模型 DKTU-Net 的有效性,在相同的训练条件下,选取了近年来在语义分割领域具有代表性的经典模型 U-Net[11]、UNet++[13]、DeepLabv3+[27]、PSPNet[17]、Swin-Unet[28]进行了对比实验。该实验在 ISPRS Vaihingen 和 ISPRS Potsdam 两个数据集上进行训练和测试,实验结果如表 3、4 所示。

表 3　不同网络架构在 ISPRS Vaihingen 数据集上的分割结果

Table 3　The segmentation results of different network architectures on the ISPRS Vaihingen dataset　　(%)

模型	IoU					MIoU	m-F1
	不透水地表	建筑	低矮植被	树木	车辆		
U-Net	73.78	82.51	59.54	71.15	47.65	66.93	79.53
UNet++	76.09	84.56	59.34	73.03	61.78	70.96	82.67
DeepLabv3+	71.54	81.17	52.23	68.76	39.76	62.69	76.01
PSPNet	75.55	83.52	58.59	72.18	50.16	68.00	80.33
Swin-Unet	73.19	79.81	57.41	71.23	43.93	65.12	78.10
DKTU-Net	**78.56**	**86.62**	**61.48**	**73.36**	**67.23**	**73.45**	**84.40**

由表 3 可知,在 ISPRS Vaihingen 数据集上,DKTU-Net 算法在综合评估指标上都取得了优异的性能,对于基线模型 U-Net,MIoU 和 M-F1 分别提高了 6.52%和 4.87%。PSPNet 通过采用金字塔池化模块进行多尺度全局池化,从而实现全局场景语义的建模,与基线模型相比评估指标虽然有所提高,但是效果不是很好。UNet++结合了密集的跳跃连接与中间层的深层监督,有助于细化目标边界,处理小目标和边缘复杂的图像。例如,在处

理车辆这类小目标上 UNet++相比于基线模型 IoU 提高了 14.1%,但是在处理低矮植被和树木 2 个可混淆类别的任务中稍有欠佳。本文提出的 DKTU-Net 算法的低矮植被和树木的 IoU 分别提高了 1.94%和 2.21%,是所有比较模型中表现最优的,在处理车辆这类小目标上 IoU 提高了 19.6%。由表 4 可知,DKTU-Net 算法在 ISPRS Potsdam 数据集上的分割结果相比于其他模型也是最好的。上述结果表明,所引入的自适应感受野选择与细节保持多尺度融合机制能够在复杂遥感场景中有效增强模型对易混淆类别和小尺度目标的判别与表征能力。

表 4　不同网络架构在 ISPRS Potsdam 数据集上的分割结果

Table 4　The segmentation results of different network architectures on the ISPRS Potsdam dataset　　(%)

模型	IoU					MIoU	m-F1
	不透水地表	建筑	低矮植被	树木	车辆		
U-Net	75.92	80.97	64.38	62.33	75.29	71.78	83.37
UNet++	77.33	83.57	66.29	66.74	76.30	74.05	84.92
DeepLabv3+	71.78	77.94	58.77	54.54	69.02	66.41	79.49
PSPNet	76.15	83.52	65.30	64.60	69.94	71.90	83.46
Swin-Unet	74.80	81.10	64.88	61.84	72.12	70.95	82.81
DKTU-Net	**77.92**	**84.54**	**67.66**	**68.32**	**78.17**	**75.32**	**85.77**

3.5　可视化结果分析

为了验证 DKTU-Net 算法的有效性和优越性,从 ISPRS Vaihingen 数据集的测试集中选取了不同类型的图像作为检测对象,对 DKTU-Net 算法的分割结果进行可视化,并与 U-Net、UNet++、DeepLabv3+、PSPNet、Swin-Unet 算法进行可视化结果对比,对比结果如图 5 所示。从检测结果对比可以观察到,DKTU-Net 的结果与标签值是最为接近的,由前 3 组对比可以看出,传统,算法在建筑物、不透水地表与车辆等边界区域存在一定的混淆,容易出现边缘模糊和小目标漏检的情况。相比之下,DKTU-Net 算法能够很好地分辨不同地物类别的边界,在建筑物、车辆与不透水地表等类别上表现出更加清晰的轮廓和连贯的区域分布,这在一定程度上反映出 DPCF 模块在多尺度特征融合及空间细节保持方面的优势。在车辆这类小尺寸目标上,DKTU-Net 能够很好的保留目标结构并减少漏检现象,这与其编码阶段引入混合型 Shunted Transformer 后对长程依赖关系和局部细节的联合建模能力相关。此外,在遥感图像中树木和低矮植被的光谱和纹理特征高度相似,传统算法在识别这些类别时往往会出现混淆的情况。从最后一组对比可以看出,DKTU-Net 能够很好地区分低矮植被和树木这 2 个类别。这表明,引入 KSFA 模块后,模型在建模上下文差异方面具有一定的优势,有助于缓解高度相似性类别之间的混淆问题。综上所述,可视化结果从定性角度验证了 DKTU-Net 在处理多尺度目标、对象稀疏分布以及高相似度类别方面的优势。

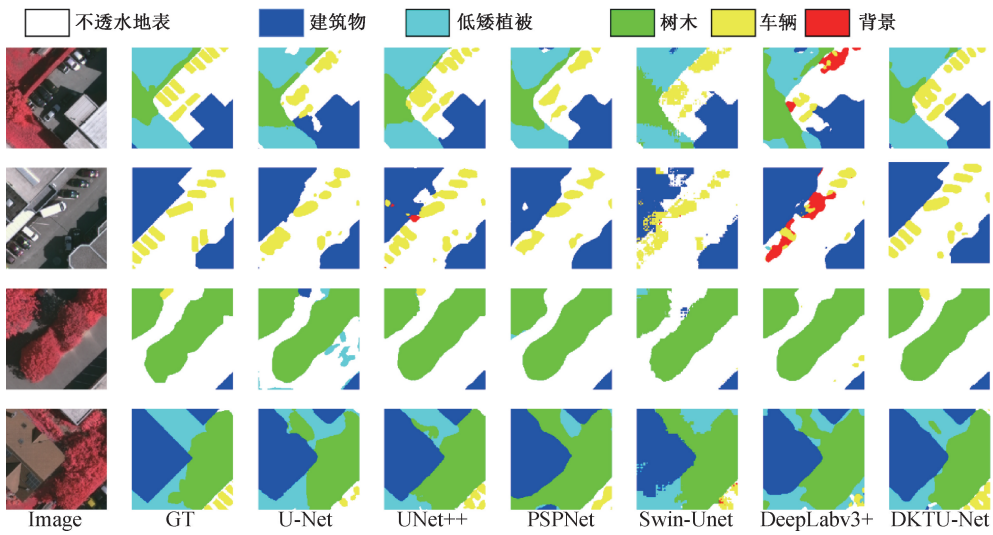

图 5　不同算法在 ISPRS Vaihingen 数据集上的效果图对比

Fig. 5　Comparison of different algorithms on the ISPRS Vaihingen dataset

4 结 论

针对遥感图像中存在目标尺度差异大、对象稀疏分布和易混淆类之间的高度相似性等问题,在 U-Net 框架的基础上,提出了一种结构改进型遥感图像分割网络 DKTU-Net。通过对现有特征建模与融合模块进行合理组合与结构优化,提升了模型在复杂遥感场景下的分割性能。在编码阶段,引入混合型 Shunted Transformer 编码结构,以增强对局部细节与长程依赖关系的联合建模能力。在解码阶段,采用 DPCF 上下文融合模块实现跳跃连接与上采样特征进行自适应融合,在保证高层语义一致性的同时保留高分辨率特征中空间信息。在特征增强阶段,引入 KSFA 核选择融合注意力模块,对融合后的特征进行进一步的增强,从而提升对光谱与纹理特征高度相似地物类别的区分能力。实验结果表明,与其他主流的语义分割算法相比,DKTU-Net 网络在 ISPRS Vaihingen 与 ISPRS Potsdam 数据集上具有较好的分割效果。后续工作将进一步研究模型轻量化与计算效率优化方法,以提高其在资源受限环境下的实际应用能力。

参考文献

[1] 梁静桦, 梁杰文. 基于 DeepLabV3+的遥感图像语义分割方法[J]. 北京测绘, 2023, 37(12): 1596-1600.
Liang Jingye, Liang Jiewen. Semantic segmentation method for remote sensing images based on DeepLabV3+[J]. Beijing Surveying and Mapping, 2023, 37(12): 1596-1600.

[2] 潘伟豪, 徐赛博, 郭弘扬, 等. 基于 D-S 证据理论的高分遥感影像建筑建筑物变化检测[J]. 电子测量与仪器学报, 2022, 36(8): 194-203.
Pan Weihao, Xu Saibo, Guo Hongyang, et al. Building change detection in high-resolution remote sensing images based on D-S evidence theory[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(8): 194-203.

[3] 王超, 石爱业, 顾爱华, 等. 一种结合阴影补偿的城市高分遥感影像分割方法[J]. 电子测量与仪器学报, 2017, 31(10): 1687-1692.
Wang Chao, Shi Aiye, Gu Aihua, et al. An urban high-resolution remote sensing image segmentation method combined with shadow compensation[J]. Journal of Electronic Measurement and Instrumentation, 2017, 31(10): 1687-1692.

[4] 柯淇晟, 花强, 张博. 基于改进 YOLOv8 的遥感图像目标检测[J]. 河北大学学报(自然科学版), 2025, 45(3): 309-316.
Ke Qisheng, Hua Qiang, Zhang Bo. Remote sensing image object detection based on improved YOLOv8[J]. Journal of Hebei University (Natural Science Edition), 2025, 45(3): 309-316.

[5] 黄聪, 杨珺, 刘毅, 等. 基于改进 DeeplabV3+的遥感图像分割算法[J]. 电子测量技术, 2022, 45(21): 148-155.
Huang Cong, Yang Jun, Liu Yi, et al. Remote sensing image segmentation algorithm based on improved DeepLabV3+[J]. Electronic Measurement Technology, 2022, 45(21): 148-155.

[6] 闫钧华, 张琨, 施天俊, 等. 融合多层级特征的遥感图像地面弱小目标检测[J]. 仪器仪表学报, 2022, 43(3): 221-229.
Yan Junhua, Zhang Kun, Shi Tianjun, et al. Weak ground target detection in remote sensing images using multi-level feature fusion[J]. Chinese Journal of Scientific Instrument, 2022, 43(3): 221-229.

[7] 李子茂, 于舒, 郑禄, 等. 基于融合注意力机制的小样本遥感场景分类方法[J]. 国外电子测量技术, 2023, 42(7): 59-67.
Li Zimao, Yu Shu, Zheng Lu, et al. Few-shot remote sensing scene classification method based on fused attention mechanism[J]. Foreign Electronic Measurement Technology, 2023, 42(7): 59-67.

[8] 简萌, 陈旭凤, 鲁军, 等. 一种改进的 K-PCA 与 PNN 结合的快速高光谱遥感分类算法[J]. 河北大学学报(自然科学版), 2025, 45(4): 343-351.
Jian Meng, Chen Xufeng, Lu Jun, et al. A fast hyperspectral remote sensing classification algorithm combining improved K-PCA and PNN[J]. Journal of Hebei University (Natural Science Edition), 2025, 45(4): 343-351.

[9] 孟月波, 王静, 刘光辉. 多层级特征增强聚合的遥感图像细小水体提取[J]. 探测与控制学报, 2023, 45(6): 123-134.
Meng Yuebo, Wang Jing, Liu Guanghui. Small water body extraction in remote sensing images using multi-level feature enhancement and aggregation[J]. Journal of Detection & Control, 2023, 45(6): 123-134.

[10] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Visionand Pattern Recognition, 2015: 3431-3440.

[11] Ronneberge O, Fische 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.
[12] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.
[13] Zhou Z, Siddiquee M M R, Tajbakhsh N, et al. UNet++: A nested U-Net architecture for medical image segmentation [C]//International Workshop on Deep Learning in Medical Image Analysis. Cham: Springer International Publishing, 2018: 3-11.
[14] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs [C]//International Conference on Learning Representations (ICLR). 2015.
[15] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.
[16] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 1248-1257.
[17] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 2881-2890.
[18] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2018: 3-19.
[19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.
[20] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. ArXiv preprint arXiv: 2010.11929, 2020.
[21] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 10012-10022.
[22] He X, Zhou Y, Zhao J, et al. Swin Transformer embedding UNet for remote sensing image semantic segmentation[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-15.
[23] Ren S, Zhou D, He S, et al. Shunted self-attention via multi-scale token aggregation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10853-10862.
[24] Xu W, Zheng S, Wang C, et al. SAMamba: Adaptive state space modeling with hierarchical vision for infrared small target detection [J]. ScienceDirect Information Fusion, 2025, 124: 103338.
[25] Xu Y, Wang D, Zhang L, et al. Dual selective fusion transformer network for hyperspectral image classification[J]. Neural Networks, 2025, 187: 107311.
[26] Rottensteiner F, Sohn G, Jung J, et al. The ISPRS benchmark on urban object classification and 3D building reconstruction [J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2012, I-3(1): 293-298.
[27] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 801-818.
[28] Cao H, Wang Y, Chen J, et al. Swin-Unet: U-Net-like pure transformer for medical image segmentation[C]// European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 205-218.

作者简介

赵亚伟,2024 年于四川轻化工大学获得学士学位,现为河北大学硕士研究生,主要研究方向为深度学习、语义分割。
E-mail:2199463409@qq.com

Zhao Yawei received his B. Sc. degree from Sichuan University of Science & Engineering in 2024. he is currently a M. Sc. candidate at Hebei University. His main research interests include deep learning and Semantic Segmentation.

冉宁(通信作者),2017 年于浙江大学获博士学位,现为河北大学电子信息工程学院教授,主要研究方向为人工智能、图像处理等。
E-mail:ranning87@hotmail.com

Ran Ning (Corresponding author) received his Ph. D. degree from Zhejiang University in 2017. He is currently a professor with the College of Electronic and Information Engineering, Hebei University. His current research interests include artificial intelligence and image processing.