E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP18;TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)04-1272-13

论文引用格式:Pan Z Z, Gao F, Gong C Z, Gan Y H and Dong J Y. 2026. Remote sensing image semantic segmentation with selective attention and directional feature enhancement. Journal of Image and Graphics, 31(4):1272-1284(潘子哲, 高峰, 宫传政, 甘言海, 董军宇. 2026. 选择注意力与方向特征增强的遥感图像语义分割. 中国图象图形学报, 31(4):1272-1284)[DOI:10.11834/jig.250317]

选择注意力与方向特征增强的遥感图像语义分割

潘子哲,高峰*,宫传政,甘言海,董军宇

1. 中国海洋大学海洋动力—物理环境与智能感知全国重点实验室,青岛　266100;
2. 中国海洋大学计算机科学与技术学院,青岛　266100

摘　要:**目的**　遥感图像语义分割是遥感解译的核心任务,但现有模型普遍面临两大挑战:一是不同尺度特征融合时存在信息不平衡,高层语义与低层细节的有效融合不足;二是传统卷积难以有效提取道路、河流等具有强方向性的线性地物特征,导致分割结果边缘模糊、结构不连续。为解决上述问题,提出一种基于选择注意力与方向特征增强的遥感图像语义分割模型。**方法**　首先,构建了一种新颖的选择性交叉注意力机制,该机制采用跨层级查询与Top-k选择策略,使高层语义特征能够主动地从低层特征中高效筛选并融合最相关的细节信息,有效缓解了多尺度信息不平衡问题并提升了计算效率;其次,设计了一个精巧的方向性特征增强模块,该模块采用两级并行架构,在多个并行的多尺度分支内部,进一步通过并行的水平与垂直一维卷积独立地提取并自适应融合方向性特征,显著增强了模型对线性地物结构的感知能力。**结果**　在公开的ISPRS Vaihingen和Potsdam基准数据集上进行了实验。在Vaihingen数据集上,所提模型的平均交并比达到84.68%,相比于性能第2的CMTFNet(CNN and multiscale Transformer fusion network)模型提升了0.94%;在Potsdam数据集上,平均交并比达到86.84%。特别是在线性地物(如汽车、建筑)和细粒度类别(如低矮植被)的分割上,精度和边界完整性均显著优于现有主流方法。消融实验也验证了所提出的选择性注意力和方向性增强两个核心模块的有效性,二者协同作用使模型性能相较于基线提升了3.43%。**结论**　所提出的模型通过创新的选择性注意力和方向性特征增强设计,协同解决了遥感图像分割中的多尺度信息融合不平衡和方向性特征提取不足的核心难题,在提升分割精度的同时,改善了线性地物的连续性和小目标的辨识度,为复杂场景下的遥感图像精细化解译提供了一种鲁棒且高效的解决方案。

关键词:遥感图像处理;图像语义分割;选择性注意力;方向性特征增强;注意力机制;多尺度特征融合;卷积神经网络(CNN);编码器—解码器

Remote sensing image semantic segmentation with selective attention and directional feature enhancement

Pan Zizhe, Gao Feng*, Gong Chuanzheng, Gan Yanhai, Dong Junyu

1. State Key Laboratory of Physical Oceanography, Ocean University of China, Qingdao　266100, China;
2. School of Computer Science and Technology, Ocean University of China, Qingdao　266100, China

Abstract: **Objective** Semantic segmentation, which aims to perform dense, pixel-level classification, has emerged as a

收稿日期:2025-07-15;**修回日期**:2025-10-31;**预印本日期**:2025-11-07

*通信作者:高峰　gaofeng@ouc.edu.cn

基金项目:科技创新2030——新一代人工智能重大项目(2022ZD0117202);山东省自然科学基金项目(ZR2024MF020)

Supported by: Science and Technology Innovation 2030 —"New Generation Artificial Intelligence" Major Project(2022ZD0117202); Natural Science Foundation of Shandong Province, China(ZR2024MF020)

pivotal technology in computer vision. Often termed remote sensing image semantic segmentation, this task is fundamental for interpreting vast amounts of geospatial data within the domain of Earth observation. Its applications are wide-ranging and critical, including land-cover mapping, urban development monitoring, change detection, and environmental surveillance. Early approaches to this problem rely on low-level features and classic machine learning, but they struggle with the immense complexity found in very-high-resolution imagery. The advent of deep learning, particularly convolutional neural networks, revolutionized the field. Models based on the fully convolutional network and U-Net architectures redefine semantic segmentation as an end-to-end pixel-labeling problem. These models excel at learning hierarchical feature representations directly from data. However, despite their progress, two persistent and critical challenges hinder their performance. The first challenge is the information imbalance in multiscale feature fusion. The encoder-decoder structure generates high-level features that are rich in semantic context but spatially coarse and low-level features that are spatially precise but semantically weak. Standard fusion strategies, such as symmetric skip-connections, treat these features equally. This approach leads to suboptimal feature fusion where fine details can be diluted by overly smooth semantic information or to that where high-level context is corrupted by irrelevant low-level texture. The second challenge is the difficulty in extracting directional features. Man-made and natural structures such as roads, rivers, and building boundaries are inherently linear and exhibit strong directional anisotropy. Standard square convolutional kernels are isotropic and ill-suited for capturing these long, continuous structures, which often lead to segmented outputs with discontinuous lines and fragmented object boundaries. This study aims to develop a novel deep learning architecture that intelligently fuses multiscale features while explicitly enhancing the network's ability to perceive directional information, thereby producing accurate and structurally coherent segmentation maps. In this way, the dual challenges can be overcome. **Method** In this study, we propose a new network architecture, the selective attention with directional feature enhancement network (SADENet), which is constructed on a robust encoder-decoder framework. We utilize a ResNet backbone to serve as the feature encoder. The core innovations are encapsulated within two new modules integrated into the decoder path: the top-k cross-attention (TCA) module and the directional feature enhancement module (DFEM). The TCA module is specifically designed to facilitate a highly intelligent fusion of features. It employs an asymmetric cross-attention mechanism where the high-level feature acts as the query, whereas the low-level feature provides the key and value. This approach allows for a context-aware selection process where high-level semantic information actively guides the search for the most relevant fine-grained details. A top-k selection strategy is incorporated to maintain computational tractability. This strategy prunes the attention matrix to consider only the most significant feature interactions, thereby improving efficiency. DFEM is specifically engineered to address the challenge of linear feature segmentation. It consists of a multibranch parallel structure where each branch utilizes a pair of asymmetric 1D convolutions——one horizontal (e. g. , $1 \times k$) and one vertical ($k \times 1$) ——to capture features along cardinal orientations explicitly and independently. The module can model directional structures across a spectrum of scales by using multiple branches with varying kernel sizes ($k = 1, 3, 5, 7$). Then, the features from these parallel branches are adaptively fused using a position enhancement module. The entire network is implemented using the PyTorch deep learning framework. As detailed in the paper, the model training utilized the Lookahead optimizer wrapping AdamW, with a cosine annealing learning rate scheduler and an initial learning rate of 6×10^{-4}. The models were trained for 105 epochs with a batch size of 8. We applied extensive data augmentation, including random cropping, flipping, rotation, and mosaic augmentation. All experiments were conducted on a workstation equipped with an NVIDIA RTX 4090 GPU. **Result** We evaluated our proposed SADENet against a suite of five representative state-of-the-art methods——U-Net, DeepLabv3+, BANet, UNetFormer, and CNN and multiscale Transformer fusion network(CMTFNet)——on two public benchmark datasets: ISPRS Vaihingen and ISPRS Potsdam. On the Vaihingen dataset, our model achieved a mean intersection-over-union (mIoU) of 84. 68% and an overall accuracy (OA) of 93. 55%, outperforming the second-best method, CMTFNet, by a notable margin of 0. 94% in mIoU and 2. 4% in OA. This superior performance was observed across all six land-cover classes. On the more challenging Potsdam dataset, SADENet achieved an mIoU of 86. 84% and a mean F1-score of 92. 84%, thereby demonstrating the best performance among all compared methods. We conducted comprehensive ablation experiments on the Vaihingen dataset to validate the effectiveness of our proposed components. Starting from a U-Net baseline scoring 81. 25% mIoU, the integration of the TCA module alone increased the mIoU by 1. 99 percentage points to

83.24%. Similarly, integrating only the DFEM module improved the baseline by a margin of 2.31 percentage points to 83.56%, which is even greater than the improvement gained from integrating the TCA module alone. This finding underscores the powerful impact of the integration of only the DFEM module on feature representation. The complete SADENet model, which combines both modules, achieved the final mIoU of 84.68%, representing a total performance gain of 3.43 percentage points over the baseline. This finding confirms that both modules are effective and contribute synergistically. Visual comparisons of the segmentation maps further illustrated our model's advantages, thereby showing visibly sharp building boundaries, continuous and correctly delineated road networks, and a superior ability to distinguish small, densely packed objects, such as cars, where other methods tended to produce blurred or merged results. This finding confirms the model's practical utility. **Conclusion** In this study, we proposed a novel network, SADENet, which integrates a selective attention mechanism and a directional feature enhancement module for the task of remote sensing image semantic segmentation. The experimental results conclusively show that our model significantly outperforms several state-of-the-art approaches on challenging and widely used benchmark datasets. The designed modules effectively address the key issues of multiscale feature imbalance and poor linear structure representation, thereby leading to highly accurate and structurally coherent segmentation results. SADENet provides a powerful new tool for geospatial analysis by fostering a highly targeted fusion of features and enhancing the perception of anisotropic structures. The work demonstrates that substantial improvements in segmentation quality can be achieved by carefully designing architectural components tailored to specific data challenges, thereby paving the way for highly reliable automated interpretation of complex remote sensing imagery.

Key words: remote sensing image processing; image semantic segmentation; selective attention; directional feature enhancement; attention mechanism; multi-scale feature fusion; convolutional neural network (CNN); encoder-decoder

0 引　言

遥感图像语义分割作为遥感领域的基础性任务,对地物识别、变化检测、城市规划和环境监测等应用具有重要意义。随着遥感技术的发展,高分辨率遥感影像数据已广泛应用于国民经济、社会生活与国家安全等各个方面。然而,遥感图像具有分辨率高、地物类别多样和边界模糊(Zhang等,2018)等特点,使得传统语义分割方法面临诸多挑战。

遥感图像语义分割经历了从传统方法到深度学习方法的发展过程。传统方法主要基于图像的低级特征,如颜色、纹理等,包括分水岭算法(Beucher,1994)、马尔可夫随机场(Kato和Pong,2006)等。近年来,深度学习特别是基于卷积神经网络(convolutional neural network, CNN)和Transformer的方法在遥感图像语义分割领域取得了显著的进展。Shelhamer等人(2017)提出的全卷积网络(fully convolutional network, FCN)通过将分类网络改造成全卷积结构,利用反卷积操作融合不同层次的特征图,从而在保持高效推理的同时,提升分割的准确性和细节。Ronneberger等人(2015)提出的U-Net通过编码器—解码器结构和跳跃连接,有效融合高层语义信息和低层空间细节,广泛应用于遥感图像语义分割。He等人(2016)提出了深度残差网络,通过残差连接和多尺度特征融合提高了分割精度。Chen等人(2018)设计的DeepLabv3+(encoder-decoder semantic segmentation network with atrous separable convolution)引入了空洞卷积和特征金字塔,增强了对多尺度地物的感知能力。然而,这些方法在处理高分辨率遥感图像时,仍面临着地物边界模糊、小目标识别困难等问题。

随着深度学习,特别是Transformer架构的发展,遥感图像语义分割领域涌现了大量先进模型。例如,SegFormer(simple and efficient design for semantic segmentation with Transformers)(Xie等,2021)和SETR(segmentation Transformers)(Zheng等,2021)将语义分割视为序列预测任务,利用Transformer强大的全局建模能力实现了高效分割(刘思涌和赵毅力,2025)。然而,这些方法在取得显著进展的同时,仍有两个核心挑战未能得到充分解决,这限制了其在复杂遥感场景下的应用效果。

第1个核心挑战是跨层级特征融合中的信息不平衡问题。针对这一问题,研究者们引入了各种注意力机制(李林娟 等,2024)。例如,DANet(dual attention network)(Fu等,2019)通过双重注意力模块

整合全局依赖，而 UNetFormer（Wang 等，2022）则将自注意力与卷积结合以自适应地融合特征。这些方法与现有主流的注意力变体（如 CBAM（convolutional block attention module）、SKNet（selective kernel network））一样，大多遵循“软选择（soft selection）”的思路，即通过学习可微的权重对所有特征进行加权。这种方式虽然灵活，却无法从根本上剔除无关低层纹理的干扰。值得注意的是，尽管类似 top-k 的“硬选择”策略已在其他视觉任务中有所探索，例如 Xiao 等人（2024）在 TTST（top-k token selective Transformer）中将其应用于超分辨率任务的同层级特征优选，但其设计并未针对语义分割中核心的跨层级信息融合（陶超 等，2025）不平衡问题。

第 2 个核心挑战是方向性地物特征的提取不足问题。 遥感图像中的道路、河流等线性地物具有强方向性，但传统方形卷积核难以有效捕捉。尽管有工作尝试通过特定设计来强化方向性感知，如 Xu 等人（2022）提出的对象校准框架，但大多数主流分割模型仍普遍缺乏对这一问题的专门建模，导致线性地物分割结果不连续、边缘模糊。

综上所述，尽管上述方法在特定方面取得了进展，但在遥感图像语义分割领域仍存在以下亟待解决的核心难题：1）不同尺度特征信息不平衡：现有的“软选择”注意力机制未能高效地解决高层语义与低层细节的融合冲突；2）方向性地物特征提取不足：模型普遍缺乏对线性结构的专门感知能力，影响分割精度；3）复杂背景干扰严重：在多样化的地物背景下，模型需要更精确的特征表达来进行区分。

针对这些问题，本文提出了一种基于选择注意力与方向特征增强的新型遥感图像语义分割框架（selective attention with directional feature enhancement network，SADENet），主要贡献如下：1）提出选择性交叉注意力机制，通过跨层级特征交互与 top-k 选择注意力策略，优先融合高层语义特征与最相关的低层细节特征，有效缓解多尺度信息不平衡问题并提升计算效率；2）设计了方向性特征增强模块，其采用了一种精巧的两级并行架构：在多个并行的多尺度分支内部，进一步通过并行的水平与垂直卷积来独立提取并自适应融合方向性特征，显著增强了模型对道路、河流等线性地物的结构感知能力；3）在 ISPRS Vaihingen 和 Potsdam 两个公开遥感数据集上的实验结果表明，所提方法显著优于现有的语义分割方法，尤其在线性地物分割方面表现出色。

1 本文方法

1.1 网络架构概述

如图 1 所示，本文方法采用编码器—解码器架构，以 ResNet（residual network）为骨干网络。网络核心包含两个创新模块：1）选择性交叉注意力机制，用于解决不同尺度特征间的信息不平衡问题；2）方向性特征增强模块，用于提高对线性地物的识别能力。

编码器提取的 res1—res4 多尺度特征经过解码器逐层融合。选择性交叉注意力机制位于特征传递路径上（res4→res3→res2→res1），选择性地融合低层特征细节与高层语义信息；解码器每个阶段后的方向性特征增强模块通过多方向卷积增强线性地物特征表达。

通过空间降采样、通道压缩和 top-k 选择，TCA（top-k cross attention）模块自身的计算成本被大幅降低。与一个在原始分辨率特征图上进行的全尺寸标准交叉注意力相比，处理 1 024 × 1 024 像素的图像时，注意力部分的计算量可减少约 75%。两个模块协同工作，选择性交叉注意力机制负责跨层次特征交互，方向性特征增强模块增强方向性特征表达，使网络在保持全局语义理解的同时提高线性地物分割精度。

1.2 选择性交叉注意力机制

传统的交叉注意力机制在处理不同尺度特征时，往往采用统一的操作方式，未能充分考虑特征间的信息差异性，导致特征融合效率不高。为解决这一问题，本文提出选择性交叉注意力机制（TCA），如图 2 所示。

给定高层语义特征 $\boldsymbol{F}_{\mathrm{h}} \in \mathbf{R}^{C\times H\times W}$ 和低层详细特征 $\boldsymbol{F}_{\mathrm{l}} \in \mathbf{R}^{C\times H\times W}$，TCA 模块首先通过空间降采样将特征图尺寸减半，显著降低后续注意力计算的内存开销，具体为

$$\boldsymbol{F}_{\mathrm{h}}^{\mathrm{down}} = AvgPool2D(\boldsymbol{F}_{\mathrm{h}}) \tag{1}$$

$$\boldsymbol{F}_{\mathrm{l}}^{\mathrm{down}} = AvgPool2D(\boldsymbol{F}_{\mathrm{l}}) \tag{2}$$

式中，$AvgPool2D(\cdot)$ 代表步长为 2 的二维平均池化操作。$\boldsymbol{F}_{\mathrm{h}}^{\mathrm{down}}$ 和 $\boldsymbol{F}_{\mathrm{l}}^{\mathrm{down}}$ 分别表示经过空间下采样后的高层特征和低层特征。经过此操作，它们的空间维度

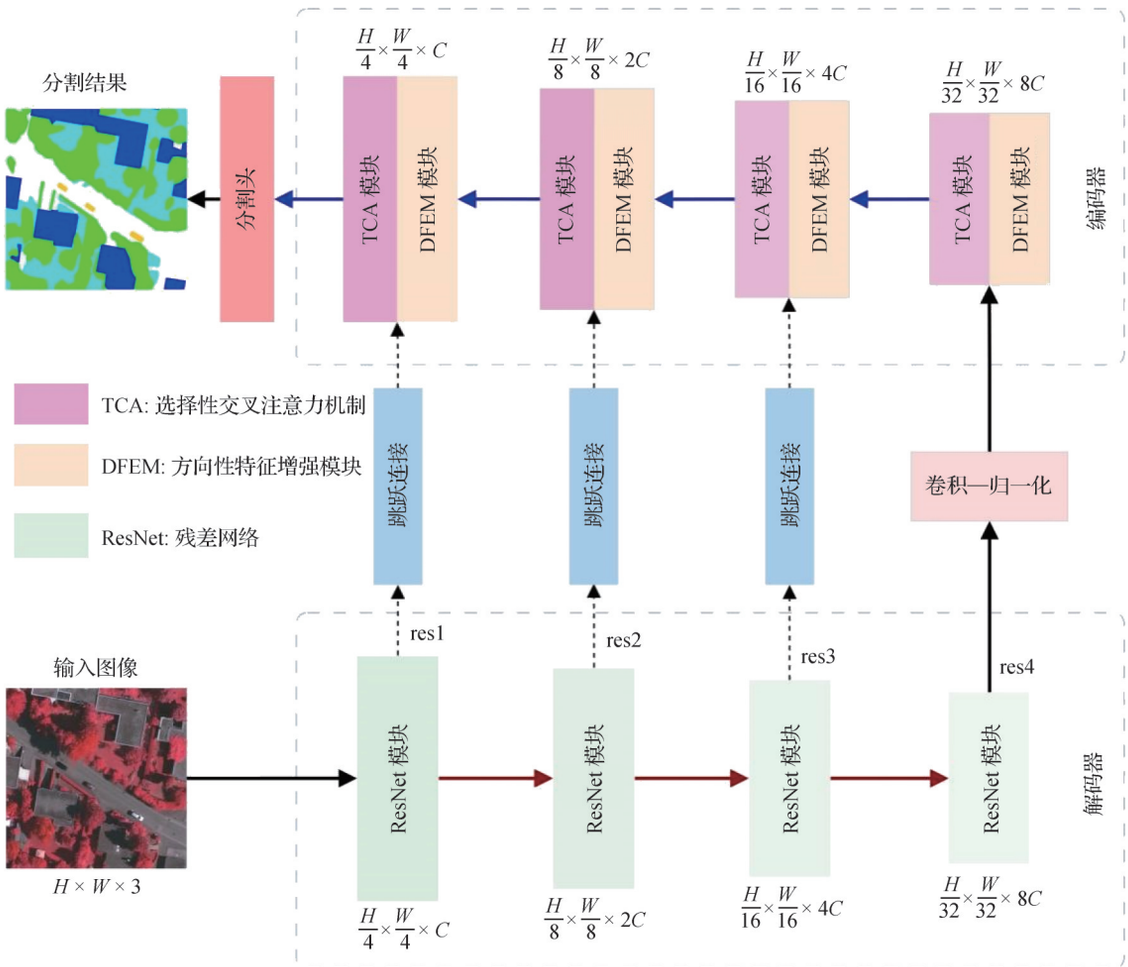

图1　基于选择注意力与方向性特征增强的遥感图像语义分割网络架构图

Fig. 1　The architecture of the proposed semantic segmentation network for remote sensing images based on selective attention and directional feature enhancement

(高度H和宽度W)均减小为原来的一半,而通道数保持不变。

然后,通过1×1卷积降低通道维度,进一步提升计算效率。具体为

$$\boldsymbol{Q}_{\mathrm{h}}=\boldsymbol{W}_{\mathrm{q}}^{\mathrm{h}}\boldsymbol{F}_{\mathrm{h}}^{\mathrm{down}},\ \boldsymbol{Q}_{\mathrm{h}}\in\mathbf{R}^{B\times C/2\times HW/4} \tag{3}$$

$$\boldsymbol{K}_{\mathrm{l}}=\boldsymbol{W}_{\mathrm{k}}^{\mathrm{l}}\boldsymbol{F}_{\mathrm{l}}^{\mathrm{down}},\ \boldsymbol{K}_{\mathrm{l}}\in\mathbf{R}^{B\times C/2\times HW/4} \tag{4}$$

$$\boldsymbol{V}_{\mathrm{l}}=\boldsymbol{W}_{\mathrm{v}}^{\mathrm{l}}\boldsymbol{F}_{\mathrm{l}}^{\mathrm{down}},\ \boldsymbol{V}_{\mathrm{l}}\in\mathbf{R}^{B\times C/2\times HW/4} \tag{5}$$

式中,$\boldsymbol{W}_{\mathrm{q}}^{\mathrm{h}}$,$\boldsymbol{W}_{\mathrm{k}}^{\mathrm{l}}$和$\boldsymbol{W}_{\mathrm{v}}^{\mathrm{l}}$是3组独立的可学习权重矩阵,通过1×1卷积实现。$\boldsymbol{W}_{\mathrm{q}}^{\mathrm{h}}$专门用于将高层特征$\boldsymbol{F}_{\mathrm{h}}^{\mathrm{down}}$投影为查询(query),而$\boldsymbol{W}_{\mathrm{k}}^{\mathrm{l}}$和$\boldsymbol{W}_{\mathrm{v}}^{\mathrm{l}}$则专门用于将低层特征$\boldsymbol{F}_{\mathrm{l}}^{\mathrm{down}}$分别投影为键(key)和值(value)。在注意力计算前,将特征重塑为序列形式,便于执行矩阵乘法。

查询矩阵$\boldsymbol{Q}_{\mathrm{h}}$源自高层语义特征$\boldsymbol{F}_{\mathrm{h}}$,而键值矩阵$\boldsymbol{K}_{\mathrm{l}}$和$\boldsymbol{V}_{\mathrm{l}}$则由低层细节特征$\boldsymbol{F}_{\mathrm{l}}$生成。该跨层注意力机制使高层特征能够主动查询并融合低层信息,而非仅进行自注意力计算。这种交叉查询方式充分结合了高层语义和低层细节,尤其适用于遥感图像分割任务,能有效提升边缘定位与语义判别能力。

完成注意力分数矩阵计算后,引入选择性top-k策略(Chen等,2019),只保留每行查询位置最相关的K个键值对,显著降低计算复杂度的同时保留重要信息。具体为

$$\boldsymbol{A}_{\mathrm{hl}}=top\text{-}k(\boldsymbol{Q}_{\mathrm{h}}\boldsymbol{K}_{\mathrm{l}}^{\mathrm{T}}/\sqrt{d_k},k) \tag{6}$$

式中,$top\text{-}k$操作保留每行(即每个查询位置)最大的k个值,其余置零,并对保留的值进行softmax归一化。在实现中,选择$k=3$。如图3所示,对于注意力矩阵的每一行(即每一个查询位置),首先计算其与所有键位置的相似度得分,然后仅保留得分最高的3个键值对进行后续计算,而将其他位置的权重置零。这种“硬选择”机制能够有效过滤掉冗余或无关的低层细节,强制模型关注最相关的特征,从而在保留关

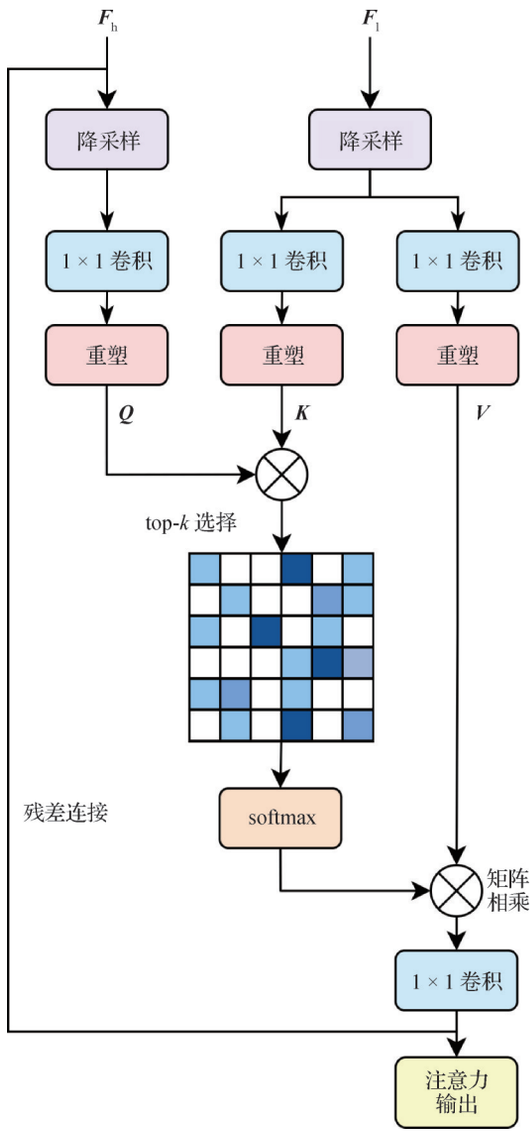

图 2　选择性交叉注意力机制(TCA)结构

Fig. 2　Architecture of TCA module

键信息的同时,显著降低了后续计算的复杂度。

基于选择性top-k注意力,计算加权特征并通过残差连接融合。具体为

$$\boldsymbol{F}_{\text{out}} = \boldsymbol{F}_{\text{h}} + Conv\left(Norm\left(UpSample\left(\boldsymbol{A}_{\text{hl}}\boldsymbol{V}_{\text{l}}\right)\right)\right) \quad (7)$$

使用BatchNorm进行特征归一化,GELU(Gaussian error linear unit)作为激活函数,并通过双线性插值上采样将特征恢复到原始分辨率。与标准交叉注意力相比,TCA通过跨层级查询实现了高层语义对低层细节的选择性引导,并通过top-k选择注意力策略显著降低了计算复杂度,即使对降采样后的高分辨率遥感图像(注意力矩阵可达[$\boldsymbol{B}$, 4 096, 4 096]),B为批大小(batch size),计算量变小的同时也提升了模型的效率和适用性。

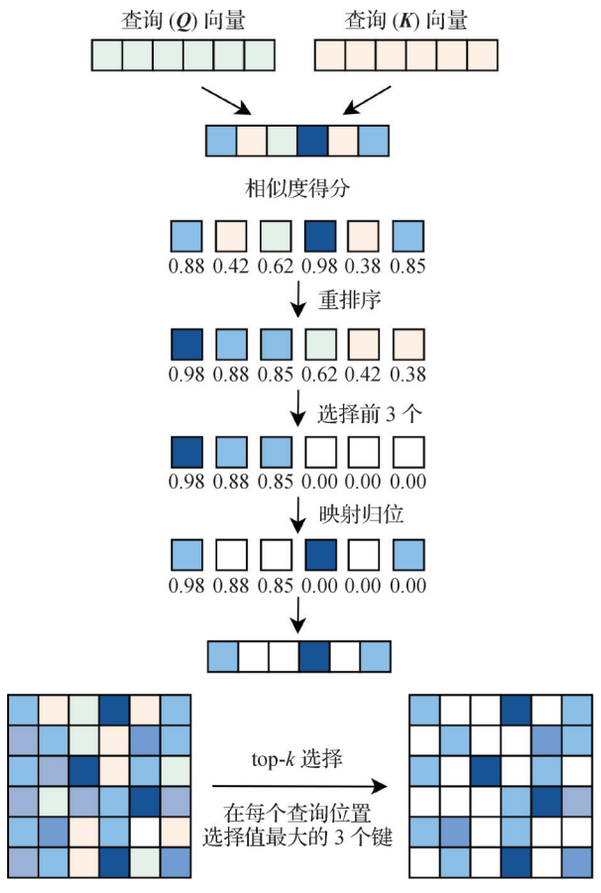

图 3　top-k选择策略过程示意

Fig. 3　Illustration of the top-k selection process

1.3　方向性特征增强模块

遥感图像中的线性地物(如道路、河流和建筑边界)具有显著的方向各向异性特征。然而,传统卷积操作受限于正方形卷积核的各向同性特性(Liu等,2024),难以有效建模长距离线性结构,导致细节信息丢失与分割边缘断裂。为此,本文提出了方向性特征增强模块(directional feature enhancement module,DFEM),如图4所示。

DFEM模块由两个主要部分组成:全局上下文建模分支和方向性卷积增强分支。全局上下文建模分支的核心是global-local attention机制,目的是捕捉特征图中任意两个像素之间的长距离依赖关系,从而建立全局上下文。具体而言,该分支采用了标准的Transformer编码器模块结构,主要由一个多头自注意力(multi-head self-attention, MHSA)层和一个前馈网络(feed-forward network, FFN)构成。输入特征首先经过MHSA层捕捉空间依赖关系,然后通过残差连接和层归一化,再送入由两个全连接层构成的FFN进行非线性特征变换。这种结构能够高效

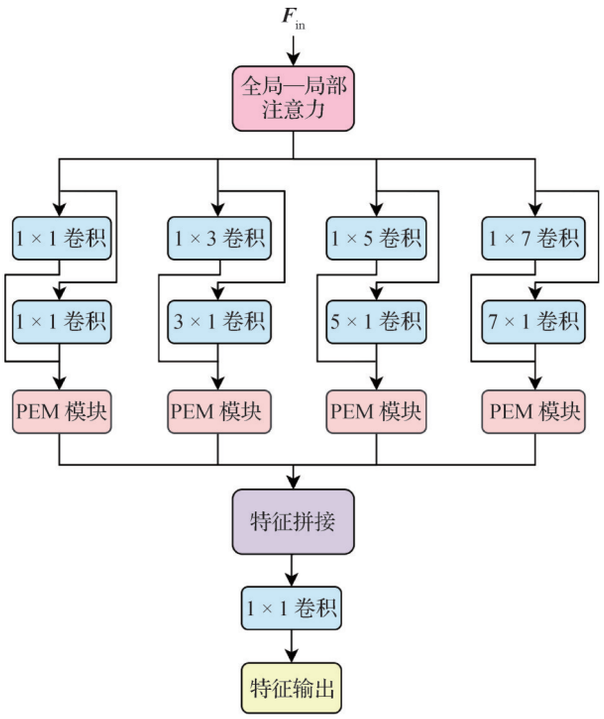

图4　方向性特征增强模块(DFEM)结构图

Fig. 4　Architecture of DFEM

地建模全局信息,为后续的方向性卷积分支提供丰富的语义上下文,是增强模型可复现性的关键。方向性卷积增强分支则专注于方向性特征的提取与增强,是DFEM模块的核心创新点。

方向性卷积增强分支采用多个平行的多尺度分支,在每个分支内部,通过并行的方向卷积独立捕获水平和垂直的结构特征。以核大小为k的分支为例,其特征$\boldsymbol{F}_k$的生成过程为

$$\boldsymbol{F}_k = Conv_{1\times k}(\boldsymbol{F}_{in'}) + Conv_{k\times 1}(\boldsymbol{F}_{in'}) \tag{8}$$

式中,$\boldsymbol{F}_{in'}$是来自global-local attention的输出特征,$Conv_{1\times k}$和$Conv_{k\times 1}$分别代表核形状为$(1, k)$的水平卷积与$(k, 1)$的垂直卷积。本文将这两个并行方向卷积的输出通过逐元素相加进行融合,以捕捉双轴方向的综合特征。同时,并行设置了$k \in \{1, 3, 5, 7\}$的4个分支,以捕获从点状到长条状结构的多尺度信息。

每个多尺度方向分支输出的特征$\boldsymbol{F}_k$,通过位置增强模块(position enhancement module, PEM)进行自适应加权,具体为

$$\boldsymbol{F}_{k'} = \boldsymbol{F}_k \cdot PEM(\boldsymbol{F}_k), \quad k \in \{1, 3, 5, 7\} \tag{9}$$

PEM模块通过全局平均池化和通道维度上的softmax操作,生成通道注意力权重,具体为

$$PEM(\boldsymbol{F}) = \boldsymbol{F} \cdot f_{\text{softmax}}(AvgPool(\boldsymbol{F})) \tag{10}$$

这种设计允许模型根据输入特征的内容自适应地调整不同尺度与方向组合特征的重要性,抑制噪声响应并增强有效结构特征,使模型能够更好地适应不同方向和尺度的线性地物。

最后,将4个并行分支的增强特征拼接(Concat)并融合(Fusion)(Zhang等,2019),具体为

$$\boldsymbol{F}_{\text{out}} = f_{\text{Fusion}}\left(f_{\text{Concat}}\left[\boldsymbol{F}_{1'}, \boldsymbol{F}_{3'}, \boldsymbol{F}_{5'}, \boldsymbol{F}_{7'}\right]\right) \tag{11}$$

DFEM模块的创新之处在于其定制化的两级并行设计。全局上下文建模部分提供了特征的整体语义理解,而方向性卷积增强分支则通过解耦并融合多尺度与多方向信息,专注于线性地物的特征增强,两者相辅相成。

2　实验结果与分析

2.1　数据集与评价指标

本文在两个公开遥感图像数据集上评估所提方法。

1)ISPRS Vaihingen(ISPRS, 2012)。源自德国Vaihingen地区的高分辨率遥感影像,包含33幅正射影像。每幅影像平均尺寸为2 494 × 2 064像素,空间分辨率达到9 cm(地面采样距离(ground sampling distance, GSD))。该数据集提供了3个光谱波段组合(近红外、红光和绿光),并配套完整的数字表面模型(digital surface model, DSM)和归一化数字表面模型(nDSM)数据。所有影像均经过精细标注,涵盖6类地物:不透水表面、建筑物、低矮植被、树木、汽车以及杂乱区域/背景。

2)ISPRS Potsdam(ISPRS, 2012)。采集于德国波茨坦城市区域,包含38幅超高分辨率正射影像。每幅影像尺寸为6 000 × 6 000像素,空间分辨率为5 cm(GSD),优于Vaihingen数据集。该数据集提供4个光谱波段(红、绿、蓝及近红外),同时附带数字表面模型(DSM)和真实正射影像(true orthophoto, TOP)数据。

评价指标采用平均交并比(mean intersection over union, mIoU)、F1分数、总体精度(overall accuracy, OA)、参数量(parameters, Params)和计算复杂度(floating-point operations, FLOPs)。

2.2　实现细节

本文提出的SADENet模型基于PyTorch实现,

以 ResNet-18 作为骨干网络。模型训练采用 Lookahead 包装的 AdamW 优化器，初始学习率设为 6×10^{-4}，使用余弦退火调度策略（初始周期 T_0 = 15，周期乘子 T_mult = 2）。训练总共进行 105 轮，批处理大小为 8。数据增强包括随机裁剪、翻转、旋转以及 Mosaic 混合（比例 0.25）。所有实验在配备 NVIDIA RTX 4090 GPU 的工作站上进行。测试阶段采用滑动窗口策略处理大尺寸图像，并使用测试时增强（test-time augmentation，TTA）进一步提高模型性能。

2.3 与现有方法的比较

本实验选取了 5 种代表性方法进行对比：经典的编码器—解码器结构 U-Net（Ronneberger 等，2015）、引入空洞卷积的 DeepLabv3+（Chen 等，2018）、结合 Transformer 和卷积的 BANet（bilateral awareness network）（Wang 等，2021）、针对遥感图像优化的 UNetFormer（Wang 等，2022）以及最新的 CNN 与多尺度 Transformer 融合网络 CMTFNet（CNN and multiscale Transformer fusion network）（Wu 等，2023）。表 1 和表 2 展示了本文方法与主流语义分割方法在 Vaihingen 和 Potsdam 数据集上的详细性能比较。结果表明，所提 SADENet 方法在两个数据集上均达到了最优性能。

表 1 Vaihingen 数据集上不同方法的实验结果

Table 1 Experimental results of different methods on the Vaihingen dataset

/%

类别	F1						mF1	mIoU	OA
	不透水表面	建筑	低矮植被	树木	汽车	背景			
U-Net	96.29	94.75	82.87	88.99	84.04	60.18	89.39	81.25	90.21
DeepLabv3+	96.64	95.40	83.16	89.14	85.15	60.56	89.90	82.08	90.78
BANet	96.58	95.32	83.27	89.54	88.29	61.12	90.60	83.18	90.83
UNetFormer	96.80	95.49	83.61	89.36	87.62	61.09	90.57	83.14	90.99
CMTFNet	96.74	95.70	84.23	89.76	88.31	61.18	90.95	83.74	91.15
SADENet（本文）	**96.95**	**96.00**	**84.99**	**90.32**	**89.34**	**61.20**	**91.52**	**84.68**	**93.55**

注：加粗字体表示各列最优结果。

表 2 Potsdam 数据集上不同方法的实验结果

Table 2 Experimental results of different methods on the Potsdam dataset

/%

类别	F1						mF1	mIoU	OA
	不透水表面	建筑	低矮植被	树木	汽车	背景			
UNet	92.66	94.89	86.17	87.71	95.36	57.57	91.35	84.36	89.94
DeepLabv3+	93.15	95.43	86.45	87.78	95.21	57.86	91.4	84.42	90.07
BANet	93.54	96.23	87.06	88.45	95.45	58.70	92.15	85.62	90.89
UNetFormer	93.78	96.20	87.52	88.44	96.17	58.65	92.42	86.14	91.21
CMTFNet	91.21	96.62	87.32	88.39	96.21	58.78	92.31	86.35	91.36
SADENet（本文）	**94.57**	**96.86**	**88.17**	**89.78**	**96.83**	**58.89**	**92.84**	**86.84**	**91.54**

注：加粗字体表示各列最优结果。

在 Vaihingen 数据集上，SADENet 的 mIoU 达到 84.68%，比次优方法 CMTFNet 高出 0.94%；总体精度（OA）达到 93.55%，比 CMTFNet 高出 2.4%，表现出显著优势。从各类别结果来看，SADENet 在所有 6 个类别上均取得最佳性能，特别是在低矮植被类别上达到 84.99%，比其他方法有明显提升；在汽车类别上达到 89.34%，比次优方法 CMTFNet 高出 1.03%，表明了本文方法对小目标的识别能力。值

得注意的是，在传统方法普遍表现较好的不透水表面和建筑类别上，SADENet仍有提升，分别达到96.95%和96.00%，表明方法的全面性。

在Potsdam数据集上，SADENet同样取得了最佳性能，mIoU达到86.84%，虽然仅比CMTFNet高出0.49%，但在关键类别上展现出更明显的优势。具体而言，在树木类别上，SADENet达到89.78%，比其他方法高出至少1.33%；在低矮植被类别上达到88.17%，比最接近的UNetFormer高出0.65%。这些结果表明了SADENet在处理具有复杂纹理和形状的地物时具有显著优势。

在总体指标上，SADENet的mF1分别在Vaihingen和Potsdam数据集上达到91.52%和92.84%，明显高于其他方法。这表明所提方法能够在保持高精度(precision)的同时保证高召回率(recall)，实现了更平衡的分割性能。

与现有方法相比，SADENet不仅在整体性能上取得了提升，还在各类地物上均表现出色，特别是在具有复杂纹理和边界的类别(如低矮植被、树木和汽车)上具有明显的优势，表明了TCA和DFEM模块的有效性。图5和图6分别展示了Vaihingen和Potsdam数据集上的可视化结果。从图中可以清晰地看

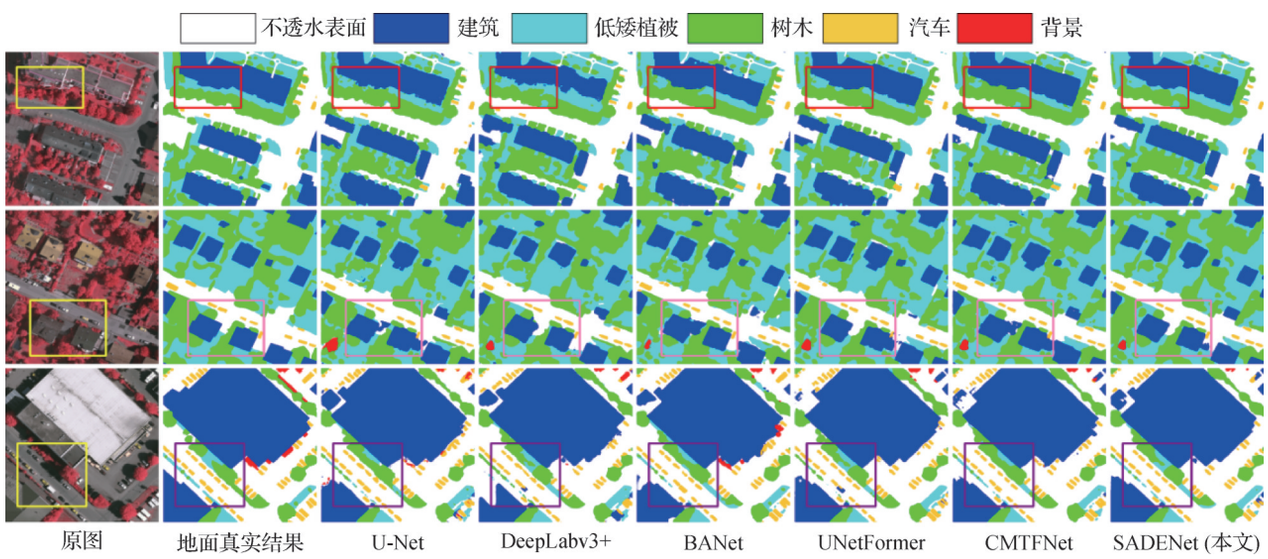

图5 Vaihingen数据集上的分割结果可视化

Fig. 5 Visualization of the segmentation results on the Vaihingen dataset

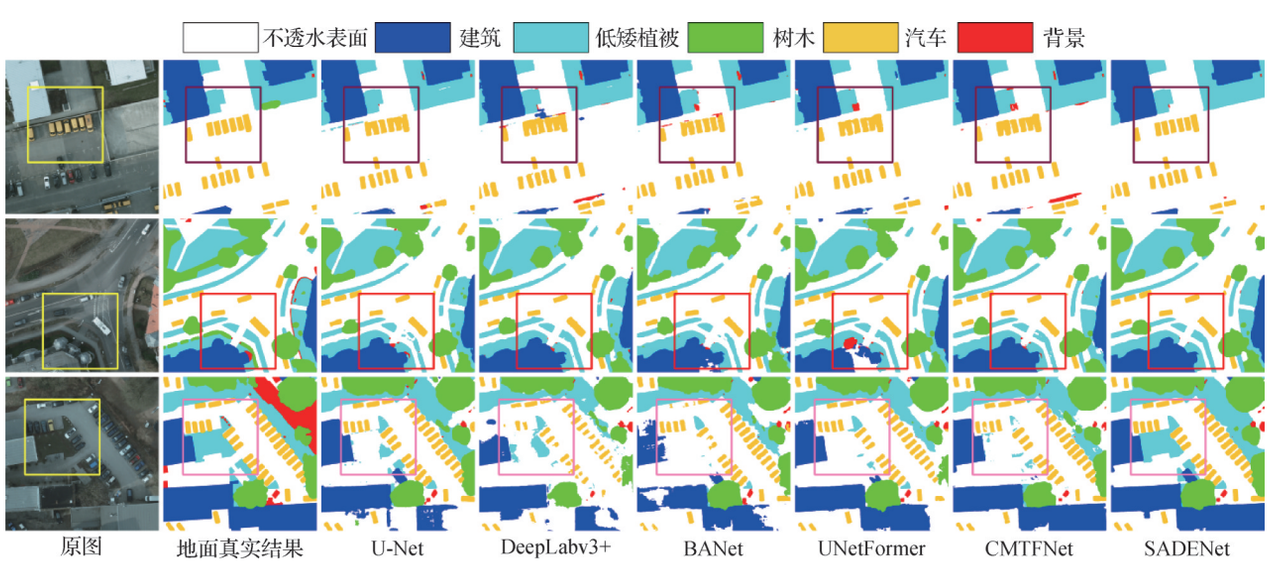

图6 Potsdam数据集上的分割结果可视化

Fig. 6 Visualization of the segmentation results on the Potsdam dataset

出，相比于其他主流方法，SADENet在线性地物（如道路、建筑边界）的分割上表现出更高的精确度和连续性。在Vaihingen数据集中，标注的区域显示SADENet能够更好地保持道路的连续性；在Potsdam数据集中，建筑物边界更加清晰，细节保留更为完整。这些可视化结果进一步验证了定量分析中所体现的优势，特别是DFEM模块对方向性特征的增强效果。

为了更直观地展示SADENet在处理复杂遥感场景时的实际优势，图7提供了在3个典型挑战场景下与多种主流方法的定性比较结果。1）倾斜的线性结构（墙面斜线）（图7第1行）；2）密集排布的细小目标（汽车）（图7第2行）；3）复杂的物体边界（房屋拐角）（图7第3行）。高亮区域的对比结果直观地证明了SADENet在保持结构连续性、辨识小目标和刻画精确边界方面的卓越性能。这些场景分别代表了线性结构、细小目标和复杂边界的分割难题。

在线性结构处理上（图7第1行），对于具有挑战性的倾斜墙面，多数对比方法的结果中出现了明显的断裂、空洞或边缘噪声，无法保持建筑立面的完整性。相比之下，得益于DFEM模块对方向性特征的强大感知能力，SADENet能够生成最连续、最平滑的分割结果，精准地还原了线性结构。

在细小目标辨识上（图7第2行），在密集停放的停车场景中，其他方法普遍存在严重的物体粘连问题，难以区分单个车辆，导致形态模糊。而SADENet则展现出卓越的细节分辨能力，这归功于TCA模块通过“硬选择”高效融合了关键的低层细节，使得每一辆车都能被清晰地独立分割出来。

在复杂边界刻画上（图7第3行），对于房屋拐角这类需要精确边界定义的场景，其他模型往往产生过于平滑或错误的边界。SADENet则能够最精确地贴合真实轮廓，生成了最为锐利的房屋拐角。

综上所述，这些可视化结果与定量分析高度一致，均证明了SADENet通过其创新的选择性注意力和方向性特征增强设计，在解决遥感图像分割的关键挑战上，相较于现有方法具有显著的优越性。

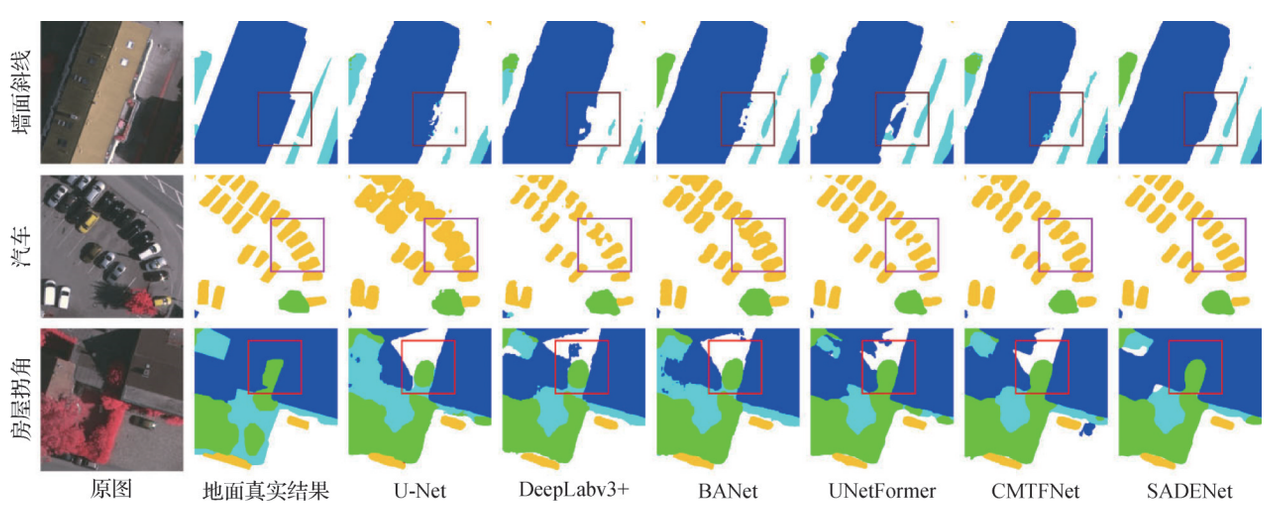

图7　不同方法在典型挑战场景下的分割结果可视化对比

Fig. 7　Visual comparison of segmentation results using different methods on typical challenging scenarios

不同方法的参数量和计算复杂度对比如表3所示。对表3中模型效率指标的深入剖析，揭示了一个值得关注的现象：尽管SADENet架构中的TCA模块旨在提升计算效率，其最终的FLOPs（53.5 G）却高于部分轻量化对标模型。这一看似矛盾的结果，并非设计的疏漏，而是一个经过深思熟虑的架构权衡（architectural trade-off）。

SADENet计算复杂度的主要来源可归因于其方向性特征增强模块（DFEM）。为了追求在线性地物分割上无与伦比的精度，DFEM采用了强大的、但计算密集型的全局注意力与多尺度并行卷积组合。该设计遵循了一种“以计算换精度”（compute-for-accuracy）的范式，即在对性能提升起决定性作用的关键模块上进行策略性的计算资源投入。

此外，对模型效率的全面评估不应局限于FLOPs这一单一维度。从模型参数量（Params）的角

表3　不同方法的参数量和计算复杂度对比

Table 3　Comparison of the number of parameters and computational complexity of different methods

方法	Params/M	FLOPs/G
UNet	31.0	**28.5**
DeepLabv3+	41.1	55.7
BANet	14.0	38.5
UNetFormer	13.7	47.0
CMTFNet	30.1	33.1
SADENet (本文)	**12.3**	53.5

注：加粗字体表示各列最优结果。

度审视，SADENet的优越性则毋庸置疑。凭借仅12.3 M的参数量，它在所有对比方法中最为精简，这一特性直接转化为更小的模型存储需求和内存占用，显著增强了其在资源受限环境下的部署潜力。

因此，SADENet的架构哲学并非单纯地最小化FLOPs，而是在模型紧凑性（最低参数量）、计算策略（高价值FLOPs）和分割性能（最高精度）三者之间，实现了一种精妙的战略平衡。它通过在DFEM模块上进行靶向性的高强度计算，有力地证明了其设计的先进性——即将计算预算以最高的效率，精准地分配给能够产生最大性能收益的体系结构组件。

在处理包含倾斜线性结构与建筑边界的场景时（图8第1行），多种基线方法在分割细长的非正交结构时，出现了严重的粘连与形状扭曲。相比之下，本文提出的SADENet能够生成轮廓清晰且独立的分割结果，这主要得益于DFEM模块，该模块通过并行的多方向卷积，有效解耦并感知了不同方向的特征，从而精确捕捉了倾斜结构。

同时，在处理包含密集停放车辆的停车场场景时（图8第2行），其他方法难以区分紧凑的小目标，导致形态模糊。而SADENet则能精确地还原每一个独立的车辆轮廓，展现了其卓越的精细结构分辨能力。这一优势凸显了本文模型中TCA模块提供的全局上下文与DFEM模块的多尺度设计（特别是小感受野分支）协同工作的有效性。

综上，这些定性结果直观地证实了SADENet能够显著提升分割结果的边界清晰度、线性结构完整性以及对小目标的辨别能力，从而在复杂的遥感解译任务中展现出更高的可靠性。

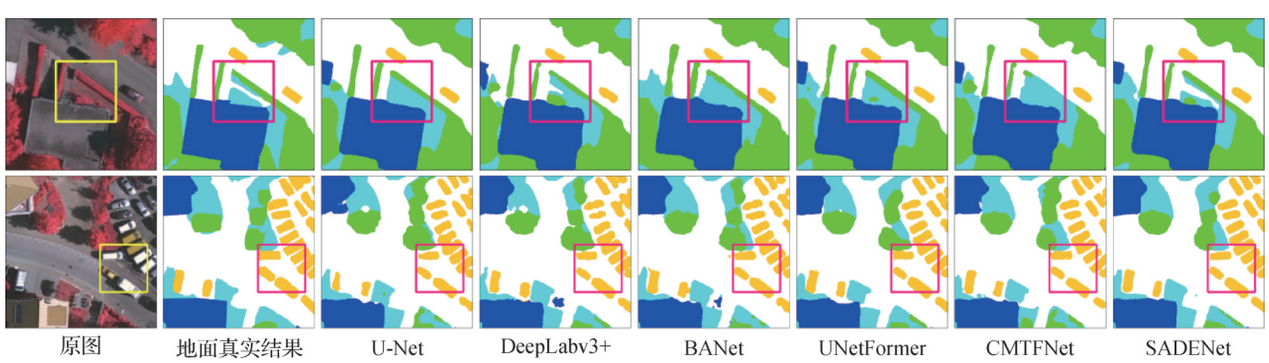

图8　不同方法对线性地物与边界的分割结果对比

Fig. 8　Comparison of segmentation results for linear features and boundaries by different methods

2.4　消融实验

为验证所提方法的有效性，进行了详细的消融实验，结果如表4所示。基线模型采用标准的U-Net结构，分别加入TCA和DFEM模块验证其效果。实验结果表明：1）引入TCA模块将基线模型的mIoU从81.25%提升至83.24%，提高了1.99%，表明选择性交叉注意力机制能有效提升特征融合效果，解决不同尺度特征间的信息不平衡问题；2）引入DFEM模块将mIoU提升至83.56%，提高了2.31%，表明方向性特征增强对遥感图像分割具有显著促进作用，特别是对线性地物的分割精度有明显改善；3）两个模块的协同作用（SADENet）使模型性能达到84.68%，相比基线模型提高了3.43%，表明了TCA和DFEM模块的互补性。尽管各模块单独使用已能带来性能提升，联合使用后分割精度进一步提高，充分证明了所提方法的有效性。

为了进一步探究TCA模块内部超参数的影响，对其核心的top-k参数进行了实验，结果如图9所示。

表 4　不同模块在 Vaihingen 数据集的消融实验结果

Table 4　Ablation study results for different modules on the Vaihingen dataset

模型配置	mIoU/%
Baseline (U-Net)	81.25
Baseline + TCA	83.24
Baseline + DFEM	83.56
Baseline + TCA + DFEM (SADENet)	84.68

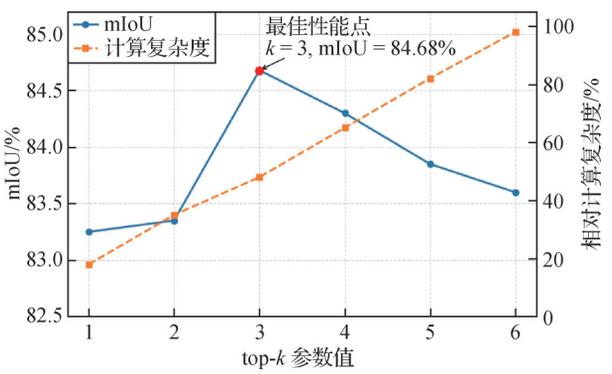

图 9　top-k 参数对模型性能和计算复杂度的影响

Fig. 9　The effect of the top-k parameter on model performance and computational complexity

实验表明，k = 3 时模型性能最佳，既保证了关键信息的保留，又显著降低了计算复杂度。

3　结　论

本文针对遥感图像语义分割中多尺度特征信息不平衡与方向性地物提取困难的挑战，提出了一种基于选择注意力与方向特征增强的新型网络框架 SADENet。核心贡献在于设计了两个关键模块：选择性交叉注意力机制（TCA）和方向性特征增强模块（DFEM）。TCA 通过跨层级特征交互与 top-k 选择注意力策略，高效融合高层语义与最相关的关键低层细节，有效缓解了信息不平衡问题；DFEM 则通过其精巧的两级并行架构，在多尺度分支内部独立地提取并融合水平与垂直方向的特征，显著增强了模型对道路、建筑边界等线性结构的感知能力。在 Vaihingen 和 Potsdam 两个公开数据集上的大量实验表明，SADENet 在 mIoU、mF1 和 OA 等关键指标上均展现出优异性能，优于当前主流方法。更重要的是，详尽的定量结果与针对性的可视化分析共同验证了本文方法在提升分割精度，特别是在改善线性地物（如道路和建筑边界）的分割连续性与边缘清晰度方面的显著有效性。此外，效率分析表明，SADENet 在实现最优轻量化设计（最低参数量）的同时，通过“高价值计算”达成了性能与效率的最佳权衡。

尽管 SADENet 表现优异，但仍存在局限。首先，其在复杂光照和阴影遮挡条件下的鲁棒性尚待进一步验证。其次，DFEM 模块虽然在增强正交结构上表现优异，但其基于水平和垂直卷积的设计，在处理任意角度的细长地物（如倾斜的道路）时能力有限。其固定的轴对齐特性难以适应非正交方向，这为未来引入旋转等变卷积等更通用的方向建模方法，提供了明确的研究方向。最后，top-k 策略虽然提升了效率，但在超高分辨率或实时性要求极高的场景中，其性能仍有优化空间。

未来，可将 SADENet 的核心设计思想扩展至变化检测、目标提取等遥感下游任务，并通过持续的结构优化与更先进的方向建模策略，提升其在复杂场景下的适应性和实用性。

参考文献（References）

Beucher S. 1994. Watershed, hierarchical segmentation and waterfall algorithm//Serra J, Soille P, eds. Mathematical Morphology and Its Applications to Image Processing. Dordrecht, Netherlands: Springer: 69-76 [DOI: 10.1007/978-94-011-1040-2_10]

Chen L C, Zhu Y K, Papandreou G, Schroff F and Adam H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 833-851 [DOI: 10.1007/978-3-030-01234-2_49]

Chen M, Beutel A, Covington P, Jain S, Belletti F and Chi E H. 2019. Top-*K* off-policy correction for a REINFORCE recommender system//Proceedings of the 12th ACM International Conference on Web Search and Data Mining. Melbourne, Australia: ACM: 456-464 [DOI: 10.1145/3289600.3290999]

Fu J, Liu J, Tian H J, Li Y, Bao Y J, Fang Z W, et al. 2019. Dual attention network for scene segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 3141-3149 [DOI: 10.1109/CVPR.2019.00326]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

ISPRS. 2012. ISPRS 2D semantic labeling contest—— Vaihingen and

Potsdam datasets [EB/OL]. [2024-05-21]. https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx

Kato Z and Pong T C. 2006. A Markov random field image segmentation model for color textured images. Image and Vision Computing, 24(10): 1103-1114 [DOI: 10.1016/j.imavis.2006.03.005]

Li L J, He Y, Xie G, Zhang H X and Bai Y H. 2024. Cross-layer detail perception and group attention-guided semantic segmentation network for remote sensing images. Journal of Image and Graphics, 29(5): 1277-1290 (李林娟, 贺赟, 谢刚, 张浩雪, 柏艳红. 2024. 跨层细节感知和分组注意力引导的遥感图像语义分割. 中国图象图形学报, 29(5): 1277-1290) [DOI: 10.11834/jig.230653]

Liu H J, Zhou X Y, Wang C L, Chen S T and Kong H. 2024. Fourier-deformable convolution network for road segmentation from remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 62: #4415117 [DOI: 10.1109/TGRS.2024.3476087]

Liu S Y and Zhao Y L. 2025. DP-SAM: efficient semantic segmentation of remote sensing images by fine-tuning SAM. Journal of Image and Graphics, 30(8): 2884-2896 (刘思涌, 赵毅力. 2025. 微调SAM的遥感图像高效语义分割模型DP-SAM. 中国图象图形学报, 30(8): 2884-2896) [DOI: 10.11834/jig.240540]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Shelhamer E, Long J and Darrell T. 2017. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4): 640-651 [DOI: 10.1109/TPAMI.2016.2572683]

Tao C, Guo X, Hu K Y, Shen Y X and Wang H. 2025. Language-guided cross-spatiotemporal domain adaptation for remote sensing image semantic segmentation. Journal of Image and Graphics, 30(9): 3153-3170 (陶超, 郭鑫, 胡柯彦, 沈羽翔, 王昊. 2025. 以语言为媒介的遥感图像跨时空领域自适应语义分割. 中国图象图形学报, 30(9): 3153-3170) [DOI: 10.11834/jig.240640]

Wang L B, Li R, Wang D Z, Duan C X, Wang T and Meng X L. 2021. Transformer meets convolution: a bilateral awareness network for semantic segmentation of very fine resolution urban scene images. Remote Sensing, 13(16): #3065 [DOI: 10.3390/rs13163065]

Wang L B, Li R, Zhang C, Fang S H, Duan C X, Meng X L, et al. 2022. UNetFormer: a UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 190: 196-214 [DOI: 10.1016/j.isprsjprs.2022.06.008]

Wu H, Huang P, Zhang M, Tang W L and Yu X Y. 2023. CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 61: #2004612 [DOI: 10.1109/TGRS.2023.3314641]

Xiao Y, Yuan Q Q, Jiang K, He J, Lin C W and Zhang L P. 2024. TTST: a top-k token selective transformer for remote sensing image super-resolution. IEEE Transactions on Image Processing, 33: 738-752 [DOI: 10.1109/TIP.2023.3340058]

Xie E Z, Wang W H, Yu Z D, Anandkumar A, Alvarez J M and Luo P. 2021. SegFormer: simple and efficient design for semantic segmentation with transformers//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: ACM: #924

Xu X H, Wang J L, Ming X and Lu Y. 2022. Towards robust video object segmentation with adaptive object calibration//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal: ACM: 2709-2718 [DOI: 10.1145/3503161.3547824]

Zhang H, Dana K, Shi J P, Zhang Z Y, Wang X G, Tyagi A, et al. 2018. Context encoding for semantic segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7151-7160 [DOI: 10.1109/CVPR.2018.00747]

Zhang Y, Huynh C P and Ngan K N. 2019. Feature fusion with predictive weighting for spectral image classification and segmentation. IEEE Transactions on Geoscience and Remote Sensing, 57(9): 6792-6807 [DOI: 10.1109/TGRS.2019.2908679]

Zheng S X, Lu J C, Zhao H S, Zhu X T, Luo Z K, Wang Y B, et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 6877-6886 [DOI: 10.1109/CVPR46437.2021.00681]

作者简介

潘子哲,男,硕士研究生,主要研究方向为遥感图像分割与数字图像处理。E-mail: 1091084984@qq.com

高峰,通信作者,男,副教授,主要研究方向为人工智能海洋学与高光谱图像处理。E-mail:gaofeng@ouc.edu.cn

宫传政,男,硕士研究生,主要研究方向为人工智能海洋学与高光谱图像处理。E-mail:2350967118@qq.com

甘言海,男,讲师,主要研究方向为人工智能海洋学与数字图像处理。E-mail:ganyanhai@ouc.edu.cn

董军宇,男,教授,主要研究方向为海洋大数据与水下图像分析。E-mail: dongjunyu@ouc.edu.cn