http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2025.0356

上下文注意力引导的航拍图像语义分割算法

刘露露[1,*]，谢荣磊[2]，姬嘉驰[3]

(1. 北京航空航天大学杭州创新研究院，杭州 310051；2. 中国航天科工飞航技术研究院，北京 100074；
3. 北京航空航天大学 航空科学与工程学院，北京 100191)

摘 要：航拍图像语义分割算法在土地利用、环境监测、资源调查、城市规划等领域具有重要的研究和应用价值。针对高分辨率航拍图像中地物关系复杂、目标尺度多变、类内差异大、类间相似高的问题，提出上下文注意力引导的语义分割算法。在模型的编码阶段引入协同注意力模块，采用空间与通道多维度特征提取的方式，对全局的上下文特征信息进行挖掘，提升模型的特征表征能力。在模型的解码阶段构建全局-局部特征融合模块，利用全局与局部平行支路的方式，实现语义信息和细节特征的提取，并借助十字语义交互模块进行特征融合，提升模型对类别边缘的辨别能力。实验结果表明：所提算法在UAVID数据集上的平均交并比（MIoU）达到了71.0%，在速度与精度上显著优于对比算法。

关 键 词：语义分割；航拍图像；注意力机制；深度学习；特征融合
中图分类号：V279+.2
文献标志码：A **文章编号**：1001-5965(2026)11-3918-11

随着传感器技术的不断进步，在全球范围内拍摄了越来越多的高分辨率航拍图像，这些图像包含了丰富的空间细节和潜在的语义内容，在资源勘探[1]、环境保护[2]、城市规划[3]等领域发挥着重要作用。语义分割作为计算机视觉领域的重要研究方向，可以对图像中的类别进行像素级分类，是满足海量航拍图像快速解译需求的关键算法。然而，航拍图像普遍具有杂波干扰强、地物关系多、目标尺度与方向多变的特点，如何有效提取高分辨率航拍图像中目标间的语义关系、准确预测类别标签仍然是一个亟待解决的问题。

近些年来，随着以深度学习为代表的人工智能技术的发展，卷积神经网络(convolutional neural network, CNN)逐渐成为解决语义分割算法所面临困难的关键技术。Long等[4]提出了全卷积网络(fully convolutional network, FCN)的语义分割算法。该算法对现有的CNN架构进行修改，将所有的全连接层替换为卷积层，从而解决网络输入输出图像尺寸大小不固定的问题。随后，Noh等[5]将转置卷积引入到图像语义分割任务中，构建了一个含有普通卷积层特征编码器和转置卷积层特征解码器的全新语义分割架构。Badrinarayanan等[6]也提出了一种用于图像语义分割的卷积编码器-解码器的结构，被称为SegNet。SegNet的创新之处是其解码器采用的是上采样操作来增加输入特征的分辨率，并利用稀疏映射生成密集的特征映射来完成图像的分割。Ronneberger等[7]在前述基础上提出了U-Net网络，其由2部分组成，一部分用来捕获上下文信息，另一部分用来恢复特征的尺度。尽管这些结构在自然图像上表现良好，但对于高分辨率航拍图像存在两方面不足：①航拍图像中目标尺度跨度极大，单一分辨率的编码-解码难以对极小目标与大尺度地物同时做出精确分割；②高分辨率下的内存与计算开销显著增加，直接将这些网络放大以提升

收稿日期：2025-06-09；录用日期：2025-08-06；网络出版时间：2025-08-29 16:00
网络出版地址：link.cnki.net/urlid/11.2625.V.20250829.1511.001
基金项目：江淮前沿技术协同创新中心追梦基金 (2023-ZM01X008)
*通信作者. E-mail：liu_lulu@buaa.edu.cn

引用格式：刘露露，谢荣磊，姬嘉驰. 上下文注意力引导的航拍图像语义分割算法[J]. 北京航空航天大学学报，2026，52(11)：3918-3928.
Liu L L，Xie R L，Ji J C. Semantic segmentation algorithms of aerial images via context-aware attention guidance[J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52(11)：3918-3928 (in Chinese).

分辨率常常不可行。为解决 FCN 全局上下文信息利用不充分的问题, Liu 等[8] 提出了名为 ParseNet 的模型。ParseNet 首先利用全局平均池化与归一化操作完成上下文特征的提取; 然后, 采用反池化操作将其恢复至原始特征图的尺寸并与原始特征图组合在一起, 形成一个含有上下文全局特征信息的新特征图, 为后续语义分割结果的生成提供丰富的特征信息。为更好地学习一个场景的全局上下文特征, Zhao 等[9] 实现了一个多尺度的网络, 称为金字塔场景解析网络, 其可以克服卷积特征映射的表征空间分辨率较低的问题, 在高维特征表征上提取重要的亚像素定位信息, 以多尺度分析实现较好的分割效果。多尺度与全局上下文的引入确实提升了对场景结构的理解, 但在航拍图像中仍面临 2 类问题: ①局部细节(尤其是极小目标)可能在多尺度汇聚过程中被“平滑”或淹没, 导致定位不准; ②航拍图像常伴随类别不均衡(少数类别像素占比极低), 单纯的多尺度融合无法有效缓解类别偏倚对模型训练的影响。除此之外, Chen 等[10] 发现来自 FCN 最后一层的深层特征图中没有足够的细化信息来进行精确的目标分割, 因此, 提出了一种基于 CNN 和全连接条件随机场的语义分割算法。该算法通过插值对 CNN 的粗分割结果进行上采样, 并将粗分割结果输入给一个全连接条件随机场来细化分割的结果, 克服深层 CNN 的不良定位问题。Schwing 等[11] 提出了一种可以联合训练 CNN 和条件随机场的全连接深度神经网络, 并在数据实验中取得了良好的性能。Zheng 等[12] 受此启发也提出了一种端到端联合训练的 CNN 和条件随机场分割方法。Lin 等[13] 提出了一种基于上下文的条件随机场语义分割方法, 探索了图像中“区域+区域”和“区域+背景”的上下文关系, 并由此改善分割效果。条件随机场类的方法在边界细化上效果显著, 但其对噪声敏感、参数调整复杂, 并且作为后处理往往增加了推理流程的复杂性。在航拍图像中, 由于背景复杂且纹理相似, 简单的基于像素/邻域的平滑约束容易导致小目标被误融合或消失; 另外, 端到端联合训练的模型在大分辨率输入下的计算开销亦不容忽视。虽然上述分割算法在自然图像分割中表现优异, 但却缺乏对航拍图像语义分割任务特有问题的针对性分析(如地物尺度跨度大、类别分布不均衡、复杂背景干扰), 无法将传统视角图像上表现出色的性能直接迁移到高分辨率航拍图像中。

为解决航拍图像语义分割任务中类别分布不均衡、类内差异大、类间相似高导致分割精度低的问题, 本文提出了一种上下文注意力引导的航拍图像语义分割(简称 CAGNet)算法。其核心思想是通过编码端和解码端的协同注意力机制, 分别强化多尺度特征的语义关联与上下文感知能力。在编码阶段, 考虑到高分辨率航拍图像空间尺度大、类别目标尺寸差异大的特点, 本文结合注意力机制提出了协同注意力模块(collaborative attention module, CAM)用于提升编码阶段全局上下文特征信息的提取能力; 在解码阶段, 模型引入全局-局部特征融合模块(global-local feature fusion module, GLFFM), 在局部细节特征和全局语义特征上对编码器提取特征进行融合, 增强预测标签的语义辨别能力。

1　本文算法

1.1　网络结构

本文 CAGNet 算法总体框架如图 1 所示。该框架主要包含特征编码模块(feature encoding module, FEM)、CAM 与 GLFFM, 其中, FEM 利用 ResNet-18 提取多尺度的特征信息, CAM 从通道和空间 2 个维度对关键特征进行挖掘, GLFFM 以多尺度特征信息为输入建立特征的长距离依赖关系, 以增强算法在实际场景中的适用能力。

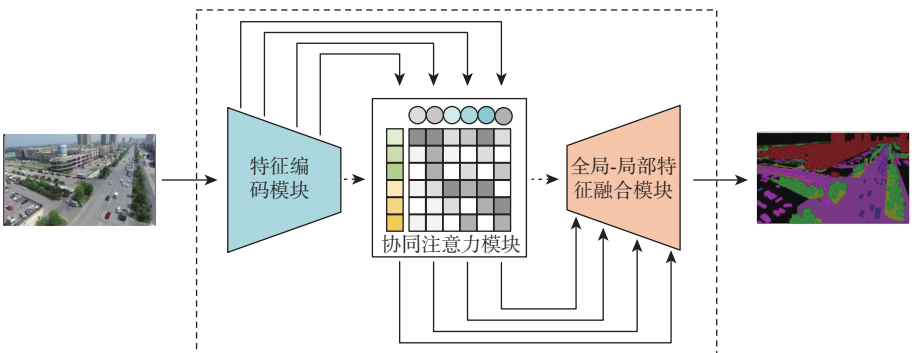


图 1　本文算法框架

Fig. 1　Framework of the proposed algorithm

1.2　协同注意力模块

考虑到航拍图像空间尺度大、各类目标尺寸差异大的特点, 本文结合注意力机制在计算机视觉任务中的出色表现, 提出了协同注意力模块, 如图 2

所示，协同注意力模块主要由协同通道注意力机制和协同空间注意力机制组成，其可以在空间和通道2个维度上对全局的上下文特征信息进行挖掘，从而提升模型特征信息的丰富程度，为后续的语义类别预测做准备。

在协同通道注意力机制中，如图3所示，输入的特征图$\boldsymbol{x}\in\mathbf{R}^{C\times H\times W}$，首先，经过池化操作和维度调整操作，从而转变成维度为$C\times H\times 1$和$C\times 1\times H$的特征图，其中，C为特征通道维度，H为特征高度，W为特征宽度；随后，将其输入一维卷积操作以学习相邻通道的依赖关系，再采用维度调整操作将它们的维度变成$C\times 1\times 1$；然后，分别采用Sigmoid函数对2个特征图进行激活，从而获得高度和宽度2个维度的通道权重；最后，对2个通道权重与原始特征图进行加权与融合操作，从而输出包含通道信息的特征图。

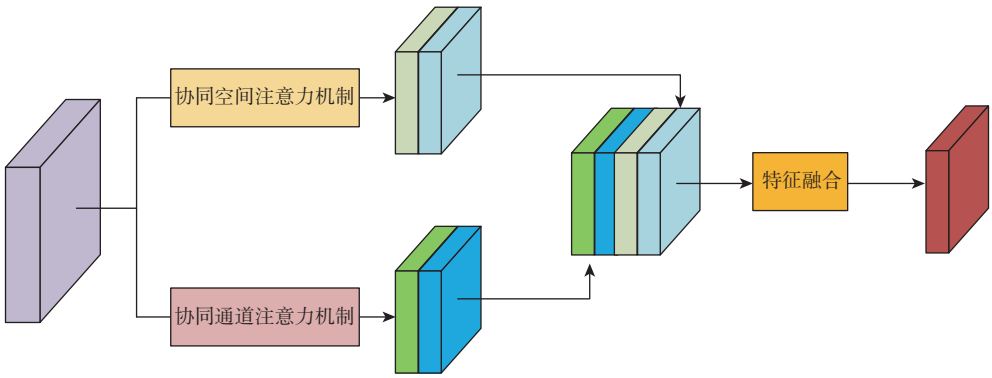


图2 协同注意力模块

Fig. 2 Collaborative attention module

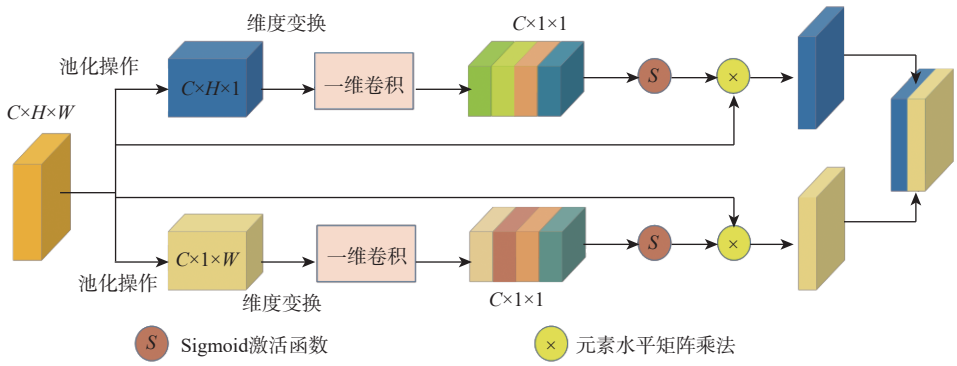


图3 协同通道注意力机制

Fig. 3 Collaborative channel attention mechanism

在协同空间注意力机制中，如图4所示，其中，GAP表示全局平均池化，GMP表示全局最大池化。首先，采用通道平均池化和通道最大池化来压缩特征图的通道信息；随后，利用拼接与卷积操作

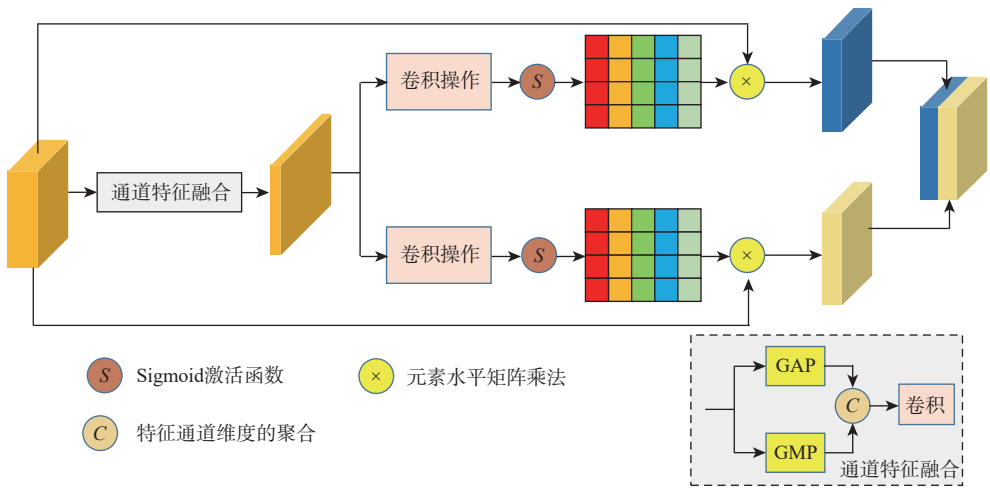


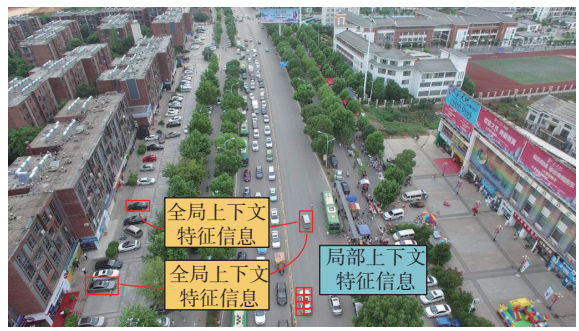
图4 协同空间注意力机制

Fig. 4 Collaborative spatial attention mechanism

对特征进行融合; 然后, 协同空间注意力分别采用与输入特征维度相对应高宽的卷积核来提取特征图的空间信息; 再利用Sigmoid函数获取特征图在宽度和高度 2 个维度的空间权重, 并最终将获得的 2 个空间权重与原始输入特征图进行加权, 得到 2 个空间特征加权的特征图; 最后, 所生成的融合特征图采用拼接和卷积操作进行最终的特征融合, 从而实现空间粒度信息的有效挖掘。

1.3　全局-局部特征融合模块

对于航拍图像语义分割任务, 如图 5 所示, 如果算法仅对局部信息进行建模, 则会导致每个像素的分类通常是模糊的, 而全局上下文信息则可以有效地解决这一问题使每个像素的分类变得更加清晰准确。此外, 复杂的人造物体经常出现在航拍图像中, 如果没有全局语义信息, 就很难实现精确的语义分割。基于上述分析, 如图 6 所示, 在全局-局部特征融合模块中, 全局与局部注意力机制构造了 2 个平行的分支从而实现对全局和局部上下文特征的融合。作为一种相对浅层的结构, 局部注意力分支使用了 2 个平行的卷积层(其卷积核大小分别为 3 和 1)来提取局部的上下文特征信息。随后, 在最终的十字形语义交互操作之前添加了 2 个批归一化操作, 以实现批次特征的融合。



京审字(2026)G第2978号

图 5　全局与局部上下文信息示意

Fig. 5　Illustration of global and local contextual information

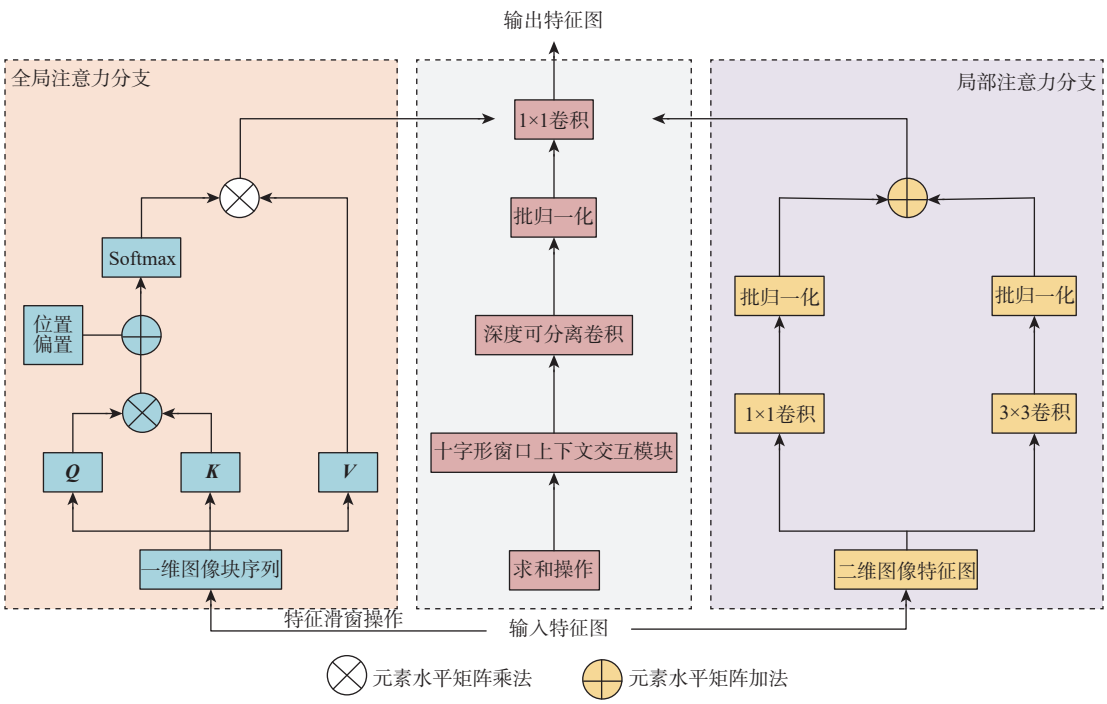


图 6　全局-局部特征融合模块

Fig. 6　Global-local feature fusion module

在全局特征分支中, GLFFM 利用基于窗口的多头自注意力来实现全局上下文信息捕获的能力。如图 7 所示, 首先, 使用一个标准的卷积将输入的二维特征图($\mathbf{R}^{B\times C\times H\times W}$)的通道维度扩展 3 倍。随后, 利用窗口分区操作将一维的序列($\mathbf{R}^{\left(3\times B\times \frac{H}{w}\times \frac{W}{w}\times h\right)\times(w\times w)\times \frac{C}{h}}$)切分为向量序列($\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$), 其中, B 为特征批次尺寸, w 为窗口的尺寸, h 为多头注意力的数量, $\boldsymbol{Q}$ 为注意力机制的 Query 向量, $\boldsymbol{K}$ 为注意力机制的 Key 向量, $\boldsymbol{V}$ 为注意力机制的 Value 向量。具体而言, 这里将通道维度统一设置为 64, 滑窗尺寸和多头注意力的数量都设置为 8。然后, 全局特征分支利用多头自注意力机制捕获全局语义特征信息。

在常用的特征融合方式中通常引入求和或乘积操作来融合全局支路与局部支路所提取的特征信息, 这种方式虽然简单且不会增加模型的复杂度, 但却缺少对遥感图像空间一致性的考虑, 容易导致最终语义分割精度的恶化。由此, 在全局语义信息与局部细节特征融合的过程中, GLFFM 额外引入了十字语义交互模块。如图 8 所示, 十字语义

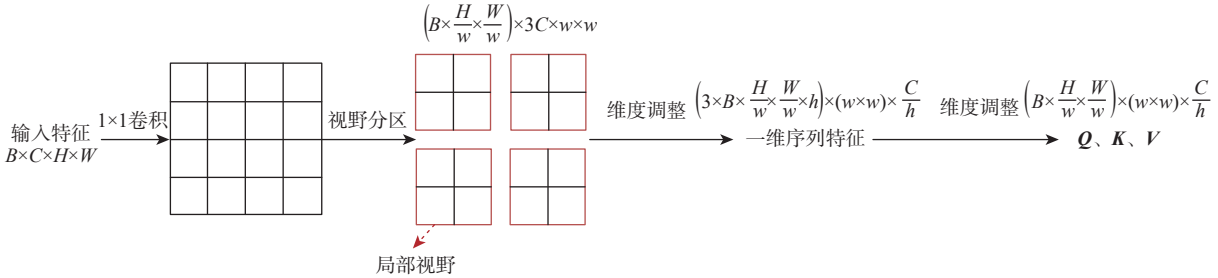


图 7 特征滑窗操作

Fig. 7 Feature sliding window operation

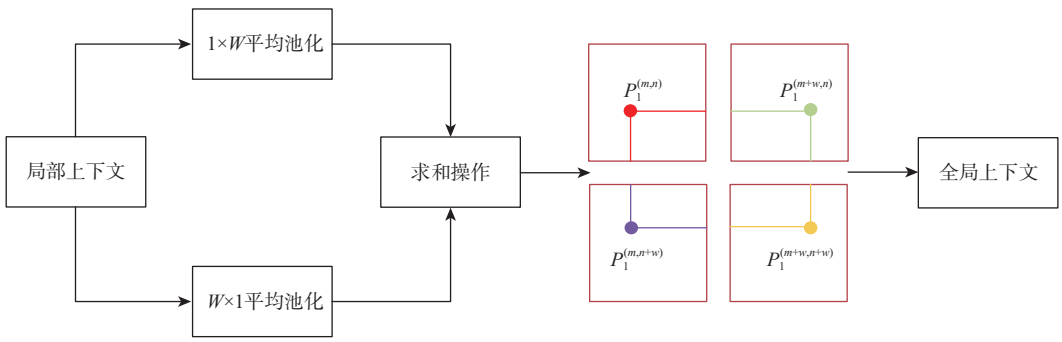


图 8 十字语义交互模块

Fig. 8 Crosswise semantic interaction module

交互模块首先通过一个水平的平均池化层和一个垂直的平均池化层分别生成 2 个特征图，随后，将 2 个特征图相加从而捕获跨视野的全局特征信息。其中，水平的平均池化层建立全局-局部之间水平方向关系的过程如式 (1) 所示。具体而言，对于视野 1 中的任意一点，其与视野 2 中点的依赖关系可以建模为式 (2)～式 (5)：

$$W_{\text{in1}} = H_{\text{p}}(W_{\text{in2}}) \tag{1}$$

$$P_1^{(m,n)} = \left(\sum_{i=0}^{w-m-1} P_1^{(m+i,n)} + \sum_{j=0}^{m} P_2^{(m+2-j,n)}\right)/w \tag{2}$$

$$P_1^{(m+i,n)} = D_i(P_1^{(m,n)}) \tag{3}$$

$$P_2^{(m+w-j,n)} = D_j(P_2^{(m+w,n)}) \tag{4}$$

$$P_1^{(m,n)} = \left(\sum_{i=0}^{w-m-1} D_i\left(P_1^{(m,n)}\right) + \sum_{j=0}^{m} D_2\left(P_2^{(m+w,n)}\right)\right)/w \tag{5}$$

式中：$W_{\text{in}i}$ 为第 i 个视野的窗口特征；H_{p} 为水平池化操作；D表示自注意力操作，其可以建立一个局部视野内像素对之间的依赖关系。因此，对于任何一个在视野 1 中的其他点 $P_1^{(m+i,n)}$ 而言，其与点 $P_1^{(m,n)}$ 的跨视野交互关系都可以用式 (3) 进行建模。同样的，对于任何一个在视野 2 中的其他点 $P_2^{(m+w-j,n)}$，其与点 $P_2^{(m+w,n)}$ 的交互关系可以用式 (4) 建模。由此，式 (2) 也可以用式 (5) 来建模点 $P_1^{(m,n)}$ 与点 $P_2^{(m+w,n)}$ 的跨视野交互关系。实际上，上述跨视野像素点之间的交互也就是局部视野 1 与局部视野 2 在水平方向上的语义信息交互过程。与此类似，视野 1 和视野 3 或视野 2 与视野 4 之间的交互过程也可以通过垂直方向的池化层来实现。采用这种方式，十字语义交互模块可以在不同的视野上对全局与局部的语义信息进行建模，以捕获完整的上下文信息。

2 实验结果与分析

2.1 实验配置

本文选用深度学习框架 Pytorch V1.7.0 搭建航拍图像语义分割算法模型，操作系统为 Windows10，并行计算框架为 CUDA V10.1、CUDNN V8.0.3.33。硬件平台配置参数如下：CPU 为 Intel Core i7-9700 K，GPU 为 NVIDIA GTX 3090，内存 32 GB，显存 24 GB。

在算法训练过程中，本文构建了一个额外的辅助模块来优化各阶段 GLFFM 的参数训练。如图 1 所示，这种多头分割架构实际上在早前的研究工作中已经被证实是非常有效的。基于这种多头分割架构，本文采用一个损失函数和一个辅助损失函数共同训练语义分割算法模型。语义分割训练主损失函数 L_{p} 由一个交叉熵损失函数 L_{ce} 和一个 DICE 损失函数 L_{dice} 组成，具体如下：

$$L_{\text{ce}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^{(n)} \ln \tilde{y}_k^{(n)} \tag{6}$$

$$L_{\text{dice}} = 1 - \frac{2}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{\tilde{y}_k^{(n)} y_k^{(n)}}{\tilde{y}_k^{(n)} y_k^{(n)}} \tag{7}$$

$$L_{\text{p}} = L_{\text{ce}} + L_{\text{dice}} \tag{8}$$

式中: N和K分别为样本数量和类别数量; $y^{(n)}$和$\tilde{y}^{(n)}$分别为真实语义标签的独立热编码和网络相应输出结果的Softmax(·)激活值; 而$\tilde{y}_k^{(n)}$则是样本对应类别的置信度值。对于辅助损失L_{aux}, 本文直接使用交叉熵损失函数进行计算, 如式 (6) 所示。具体而言, 辅助模块以 3 个阶段的 GLFFM 加权融合后的结果为输入, 利用一个3×3的卷积层、批归一化层、Relu激活函数, 以及一个1×1的卷积层和上采样层进行处理, 从而得到最终的辅助预测结果。此外, 为实现最优的损失组合方式, 还额外引入了一个值为 0.44 的尺度加权因子α。整个算法模型训练过程的损失计算式为

$$L = L_{\text{p}} + \alpha L_{\text{aux}} \tag{9}$$

航拍图像语义分割算法训练与性能评估的数据采用的是无人机语义分割数据集(unmanned aerial vehicle semantic segmentation dataset, UAVID)。该数据集具有空间分辨率高、空间差异性大、分类模糊、场景复杂的特点, 符合本文研究的主要内容场景。UAVID 共包含 42 个飞行序列的 420 张高分辨率图像(图像像素分辨率为 3 820 像素×2 160 像素与 4 096 像素×2 160 像素), 按照官方的分配方式, 其中, 200 张图像用于训练, 70 张图像用于验证, 150 张无标签的图像用于测试。数据格式上, 本文将图像统一填充并裁剪为 8 张像素分辨率为 1 024 像素×1 024 像素的图像。在训练过程中, 统一按照 0.5 的概率对图像进行随机水平翻转、随机亮度调整以增强数据分布。本文选用的优化器为 AdamW, 初始学习率设置为6×10^{-4}, 并使用余弦策略对学习率进行衰减, 整个训练周期为 30 个轮次, 每批次训练图像的数量 Batch Size 为 8。

2.2　评估指标

语义分割任务的结果可以分为: 真正例(true positive, TP)表示真实值与预测值相同且为正样本; 真负例(true negative, TN)表示真实值与预测值相同且为负样本; 假正例(false positive, FP)表示预测值为正样本, 而实际值为负样本; 假负例(false negative, FN)表示预测值为负样本, 而实际值为正样本。MIoU 是指平均交并比(mean intersection-over-union), 用于衡量模型在各个类别上的预测准确度。该指标将每个类别的交集和并集之比的平均值作为模型的评估指标。对于每个类别, 将预测的分割结果与真实标签进行比较, 计算出该类别的 IoU 值, 然后, 将所有类别的 IoU 值求平均便可得到平均交并比, 具体的计算如式 (10) 所示。此外, MIoU 的值越高, 则表示模型的性能越好, 表示模型能够更加准确地分割各个类别。

$$M_{\text{IoU}} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{T_{\text{P}}}{F_{\text{N}}+F_{\text{P}}+T_{\text{P}}} \tag{10}$$

2.3　结果分析

本文选择最高有效位(most significant digit, MSD)[14]、交叉注意力网络(cross attention network, CANet)[15]、双注意力网络(dual attention network, DANet)[16]、SwiftNet[17]、BiSeNet[18]、多尺度注意力网络(multi-scale attention network, MANet)[19]、ABCNet[20]、SegFormer[21]、双边注意力网络(bilateral attention network, BANet)[22]、BoTNet[23]、CoaT[24]、Segmenter[25]、快速分割卷积神经网络(fast-segmentation convolutional neural network, Fast-SCNN)[26]、特征注意力网络(feature attention network, FANet)[27]、ShelfNet[28] 这 13 种算法作为对照算法, 在 UAVID 数据集上对本文 CAGNet 算法进行速度和精度 2 个维度的性能验证。如表 1 所示, 与速度最快、结构最轻量的算法 Fast-SCNN 相比, CAGNet 算法在 MIoU 指标上提高了 25.1%。同时, 与同样参数量的

表 1　不同算法的速度与精度对比

Table 1　Comparison of speed and accuracy for different algorithms

算法	内存量	参数量	复杂度	推理速度/(帧·s^{-1})	MIoU/%
Fast-SCNN[26]	619.4×10^6	1.1×10^6	3.4×10^9	222.7	45.9
Segmenter[25]	828.6×10^6	6.7×10^6	26.8×10^9	14.7	58.7
BiSeNet[18]	970.6×10^6	12.9×10^6	51.8×10^9	121.9	61.5
DANet[16]	611.1×10^6	12.6×10^6	39.6×10^9	189.4	60.6
FANet[27]	971.9×10^6	13.6×10^6	86.8×10^9	94.9	
ShelfNet[28]	579.0×10^6	14.6×10^6	46.7×10^9	141.4	47.0
SwiftNet[17]	835.8×10^6	11.8×10^6	51.6×10^9	138.7	61.1
MANet[19]	1 169.2×10^6	12.0×10^6	51.7×10^9	75.6	62.6
ABCNet[20]	1 105.1×10^6	14.0×10^6	62.9×10^9	102.2	63.8
Segformer[21]	933.2×10^6	13.7×10^6	63.3×10^9	31.3	66.0
BoTNet[23]	710.5×10^6	17.6×10^6	49.9×10^9	135.0	63.2
CoaT[24]	3 133.8×10^6	11.1×10^6	104.8×10^9	10.6	65.8
CAGNet	1 203.7×10^6	12.4×10^6	49.3×10^9	110.3	71.0

其他CNN算法相比，CAGNet实现了110.3帧/s的推理速度，且在MIoU指标上超过了其他算法7.2%以上。值得注意的是，CAGNet算法超过了最先进的基于混合空间变换网络的CoaT算法5.2%的MIoU，同时实现了10倍的速度提升。此外，本文CAGNet算法比使用Transformer网络的其他语义分割算法在MIoU指标上高出了12.3%，速度上提升了近7倍。基于以上分析，可以得出以下结论：无论是综合考虑航拍图像语义分割算法速度还是精度上的表现，本文CAGNet算法都是非常出色的。

为充分验证CAGNet算法的先进性，本文还找了12个实用有效的语义分割算法，并在UAVID官方提供的测试数据上进行了详细的比较。如表2所示，CAGNet算法不仅实现了最佳的MIoU性能71.0%，还在每个类别IoU上实现了出色的优势性能。具体而言，CAGNet算法不仅在MIoU指标上比最出色的CNN算法高出了7.2%的性能，还比基于混合空间变换网络的BANet算法和BoTNet算法分别高出了6.4%和7.8%。此外，尽管“人类”这一类别在航拍图像中属于极小的目标，且由于航拍角度的问题也会存在尺度变化比过大的情况，但是CAGNet算法却实现了34.8%的IoU值，远远高于其他算法。由此，可以充分说明，CAGNet算法在解决高分辨率航拍图像语义分割问题时具有出色的有效性与先进性。

表2 不同算型的类别精度对比

Table 2 Comparison of class-wise accuracy for different algorithms

算法	IoU/%								MIoU/%
	杂波	建筑	道路	树木	植被	移动汽车	静止汽车	人类	
MSD[14]	57.0	79.8	74.0	74.5	55.9	62.9	32.1	19.7	57.0
CANet[15]	66.0	86.6	62.1	79.3	78.1	47.8	68.3	19.9	63.5
DANet[16]	64.9	85.9	77.9	78.3	61.5	59.6	47.4	9.1	60.6
SwiftNet[17]	64.1	85.3	61.5	78.3	76.4	51.1	62.1	15.7	61.1
BiSeNet[18]	64.7	85.7	61.1	78.3	77.3	48.6	63.4	17.5	61.5
MANet[19]	64.5	85.4	77.8	77.0	60.3	67.2	53.6	14.9	62.6
ABCNet[20]	67.4	86.4	81.2	79.9	63.1	69.8	48.4	13.9	63.8
Segmenter[25]	64.2	84.4	79.8	76.1	57.6	59.2	34.5	14.2	58.7
SegFormer[21]	66.6	86.3	80.1	79.6	62.3	72.5	52.5	28.5	66.0
BANet[22]	66.7	85.4	80.7	78.9	62.1	69.3	52.8	21.0	64.6
BoTNet[23]	64.5	84.9	78.6	77.4	60.5	65.8	51.9	22.4	63.2
CoaT[24]	69.0	88.5	80.0	79.3	62.0	70.0	59.1	18.9	65.8
CAGNet	72.2	89.5	84.7	82.1	67.9	76.8	59.6	34.8	71.0

为更全面的对CAGNet算法效果进行全面的评估，本文将算法在UAVID数据集的验证集与测试集上的预测结果进行了可视化，通过可视化图像对模型的先进性与有效性进行了定性的评价。

CAGNet算法在UAVID数据集验证集上的可视化结果如图9所示，本文选取对航拍场景更具有代表性的三类图像作为输入图像，经过CAGNet算法推理后，将语义分割预测结果与真实标签采用相同的方式进行可视化。无论是对微小的地面车辆还是对极小的路边行人，又或者对尺度较大的地面植被和建筑物，CAGNet算法的预测结果较真值标签而言，都具有很好的识别与分割效果，唯一的瑕疵在于对形状不规则的植被覆盖区域的边界分割不够精确，容易出现错误的判定，但从全图的维度来衡量算法效果，已经能够充分证明，CAGNet算法在UAVID数据集所对应的航拍场景下可以十分完美地完成语义分割任务。

此外，CAGNet算法在UAVID数据集测试集上也能够在图像尺度对各困难类别进行准确的判定与分割。如图10所示，选取城市、乡村、野外等不同场景的图像作为测试样本，可以看出，CAGNet在如此复杂的航拍场景下依然实现了出色的语义分割效果。图10中，白框所框选的区域不仅存在复杂的地物关系还含有大量的杂波干扰，导致算法在这些地方的分割效果较真值标签有一定差异，但是在航拍图像智能分析任务中，可视化图像的分割效果与当前的技术要求相比已基本满足实际场景的任务需求。这一结论既证明了CAGNet算法在高分辨率航拍遥感图像语义分割场景下的出色能力，也反映出当下的算法可能在鲁棒性和领域泛化能力上存在不足，因而出现测试集未知遥感场景中分割精度有所下降的情况。

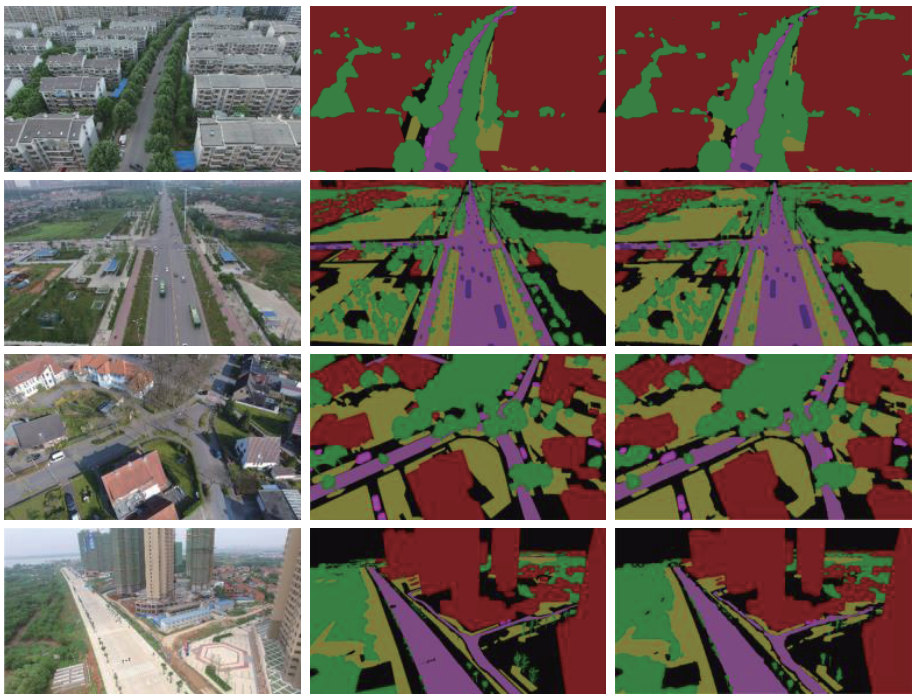


图 9　验证集可视化结果

Fig. 9　Visualization results of validation set

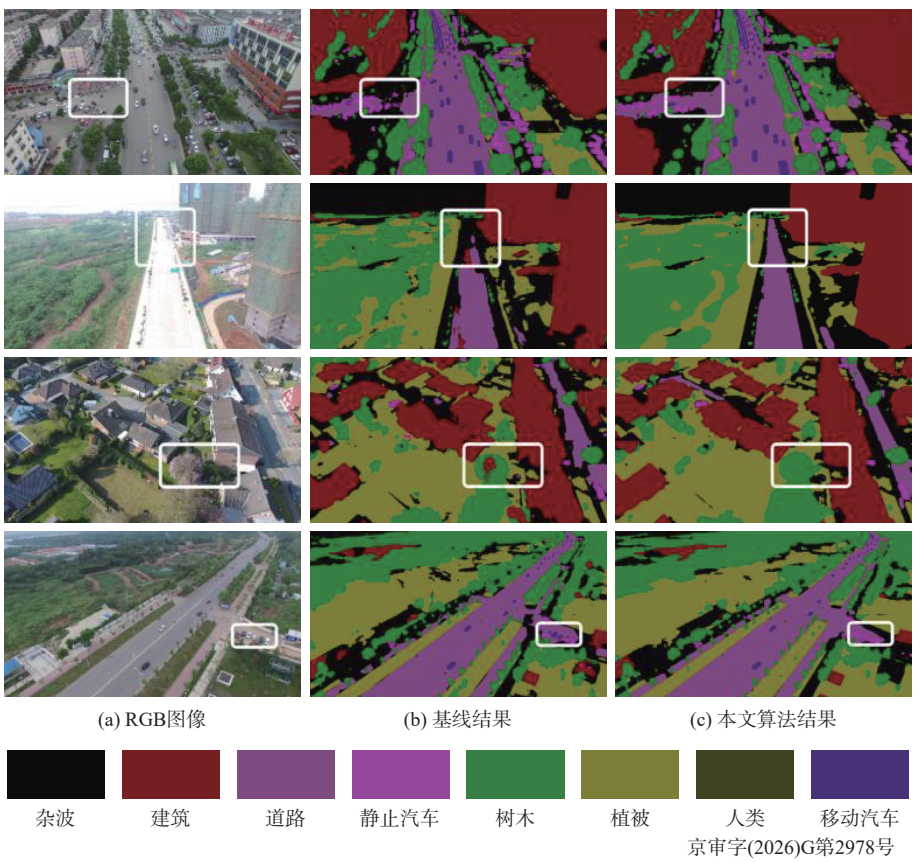


图 10　测试集可视化结果

Fig. 10　Visualization results of test set

如图 9 和图 10 所示，尽管 CAGNet 算法整体分割效果良好，但在形状不规则且纹理复杂的植被覆盖区域，边界精细化仍存在不足，这主要源于多尺度上下文融合过程中细粒度边缘信息在深层特征中被弱化的问题；为此计划在后续工作中引入边缘感知模块、基于梯度的边界损失或轻量级边界解码器以专门提升边界精度，进而改善本文 CAGNet 算法在复杂区域边界分割不精确的问题。

2.4 消融实验

为验证本文算法的有效性，利用 UAVID 数据集的训练集与测试集进行消融实验。在其他参数配置相同的情况下，共进行了 A、B、C、D 这 4 个方案的消融实验。其中，方案 A 使用 ResNet-18 作为特征编码模块，利用上采样模块将最深层特征图上采样成输入图像的大小，得到语义分割的结果；方案 B 在方案 A 的基础上引入 CAM 对高维语义特征进行挖掘，得到语义分割结果；方案 C 在方案 B 的基础上引入 GLFFM 对全局与局部特征进行交互融合，得到语义分割结果；方案 D 则是本文 CAGNet 算法。

如图 11 所示，与方案 A 相比，方案 B 引入了 CAM，在一定程度上增强 FEM 对航拍图像地物关系的建模能力，故 MIoU 提升了 0.7%。其次，方案 C 与方案 A 和方案 B 相比，方案 C 额外增加了 GLFFM，使得算法的分割性能和 MIoU 分别提升了 3.4% 和 2.7%。这组对比可以看出，语义分割模型的特征解码功能与编码提取功能同样重要，仅仅提升模型的特征编码能力无法显著提升其分割效果。随后，方案 D 与方案 C 相比，方案 D 的分割性能 MIoU 较方案 C 提升了 2.2%，可见，语义分割任务是逐像素判别的精确任务，同时提升模型远距离建模与语义交互能力会影响最终的分割效果。综上所述，本文提出的 CAM 与 GLFFM 在提升模型的分割性能上具有显著作用。

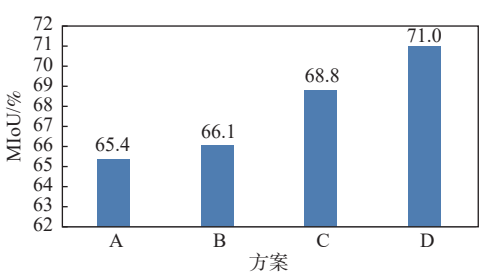


图 11 不同方案的消融实验

Fig. 11 Ablation study on different proposal

3 结 论

1）本文针对航拍图像地物关系复杂、目标尺度多变、类别分布失衡、杂波干扰较多的问题，提出了一种上下文注意力引导的航拍图像语义分割算法。

2）本文提出的协同注意力模块通过在空间和通道 2 个维度上对上下文特征信息进行挖掘，提升算法对不同类别间相似目标的判别能力。

3）本文提出的全局-局部特征融合模块利用全局与局部双分支网络，实现大感受野与小感受野之间语义信息的交互与融合，提升语义分割的精确度。

4）在真实航拍场景中采集的 UAVID 数据上开展的对比实验与消融实验的结果表明，本文 CAGNet 算法在速度与精度上均显著高于其他 15 种对比算法，并且提高了相同类别目标在各种尺寸下的分割精度，表现出算法在航拍图像语义分割任务上的先进性与有效性。

参考文献（References）

[1] Zhong H F, Sun H M, Han D N, et al. Lake water body extraction of optical remote sensing images based on semantic segmentation[J]. Applied Intelligence, 2022, 52(15): 17974-17989.

[2] Yuan Q Q, Shen H F, Li T W, et al. Deep learning in environmental remote sensing: achievements and challenges[J]. Remote Sensing of Environment, 2020, 241: 111716.

[3] Zhang Q L, Seto K C. Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data[J]. Remote Sensing of Environment, 2011, 115(9): 2320-2329.

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

[5] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2016: 1520-1528.

[6] 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.

[7] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[M]//Medical Image Computing and Computer-Assisted Intervention-MICCAI. Berlin: Springer, 2015: 234-241.

[8] Liu W, Rabinovich A, Berg A C. ParseNet: looking wider to see better[EB/OL]. (2015-06-15)[2025-06-02]. https://arxiv.org/abs/1506.04579.

[9] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6230-6239.

[10] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[EB/OL]. (2014-12-22)[2025-06-02]. https://arxiv.org/abs/1412.7062.

[11] Schwing A G, Urtasun R. Fully connected deep structured net-

works[EB/OL]. (2015-03-09)[2025-06-02]. https://arxiv.org/abs/1503.02351.

[12] Zheng S, Jayasumana S, Romera-paredes B, et al. Conditional random fields as recurrent neural networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2016: 1529-1537.

[13] Lin G S, Shen C H, Van Den Hengel A, et al. Efficient piecewise training of deep structured models for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 3194-3203.

[14] Lyu Y, Vosselman G, Xia G S, et al. UAVid: a semantic segmentation dataset for UAV imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 165: 108-119.

[15] Yang M Y, Kumaar S, Lyu Y, et al. Real-time semantic segmentation with context aggregation network[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 178: 124-134.

[16] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 3141-3149.

[17] Oršić M, Šegvić S. Efficient semantic segmentation with pyramidal fusion[J]. Pattern Recognition, 2021, 110: 107611.

[18] Yu C Q, Wang J B, Peng C, et al. BiSeNet: bilateral segmentation network for real-time semantic segmentation[C]//Proceedings of the European Conference on Computer Vision-ECCV. Berlin: Springer, 2018: 334-349.

[19] Li R, Zheng S Y, Zhang C, et al. Multiattention network for semantic segmentation of fine-resolution remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5607713.

[20] Li R, Zheng S Y, Zhang C, et al. ABCNet: attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 181: 84-98.

[21] Xie E Z, Wang W H, Yu Z D, et al. SegFormer: simple and efficient design for semantic segmentation with transformers[J]. Advances in neural information processing systems, 2021, 34: 12077-12090.

[22] Wang L B, Li R, Wang D Z, et al. Transformer meets convolution: a bilateral awareness network for semantic segmentation of very fine resolution urban scene images[J]. Remote Sensing, 2021, 13(16): 3065.

[23] Srinivas A, Lin T Y, Parmar N, et al. Bottleneck transformers for visual recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 16514-16524.

[24] Xu W J, Xu Y F, Chang T, et al. Co-scale conv-attentional image transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 9961-9970.

[25] Strudel R, Garcia R, Laptev I, et al. Segmenter: transformer for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 7242-7252.

[26] Poudel R P K, Liwicki S, Cipolla R. Fast-SCNN: fast semantic segmentation network[EB/OL]. (2019-02-12)[2025-06-03]. https://arxiv.org/abs/1902.04502.

[27] Hu P, Perazzi F, Heilbron F C, et al. Real-time semantic segmentation with fast attention[J]. IEEE Robotics and Automation Letters, 2021, 6(1): 263-270.

[28] Zhuang J T, Yang J L, Gu L, et al. ShelfNet for fast semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop. Piscataway: IEEE Press, 2020: 847-856.

Semantic segmentation algorithms of aerial images via context-aware attention guidance

LIU Lulu[1,*], XIE Ronglei[2], JI Jiachi[3]

(1. Hangzhou Innovation Institute of Beihang University, Hangzhou 310051, China;
2. HIWING Technology Academy of CASIC, Beijing 100074, China;
3. School of Aeronautic Science and Engineering, Beihang University, Beijing 100191, China)

Abstract: Semantic segmentation algorithms for aerial images have significant research and application value in fields such as land use classification, environmental monitoring, resource surveying, and urban planning. To address the challenges posed by high-resolution aerial images—such as complex spatial relationships among objects, varying target scales, large intra-class variations, and high inter-class similarities—this paper proposes a context attention-guided semantic segmentation algorithm. In order to improve the network's ability to represent features, a collaborative attention module is added to the encoder stage of the model. This module pulls features from both spatial and channel dimensions. To efficiently extract both semantic information and detailed features, a global-local feature fusion module is built in the decoder stage using concurrent global and local branches. Furthermore, a crisscross semantic interaction module is employed to fuse multi-level features, improving the model's ability to distinguish object boundaries. Experimental results demonstrate that the proposed algorithm achieves a mean intersection-over-union(MIoU)of 71.0% on the UAVID dataset, significantly outperforming state-of-the-art algorithms in both accuracy and inference speed.

Keywords: semantic segmentation; aerial images; attention mechanism; deep learning; feature fusion

Received: 2025-06-09; **Accepted**: 2025-08-06; **Published Online**: 2025-08-29 16:00

URL: link.cnki.net/urlid/11.2625.V.20250829.1511.001

Foundation item: Jianghuai Frontier Technologies Collaborative Innovation Center Dream Pursuit Fund (2023-ZM01X008)

* **Corresponding author**. E-mail: liu_lulu@buaa.edu.cn