DOI: 10.13382/j.jemi.B2508546

基于 MFFISNet 的低空地物识别方法研究*

张 堃[1,2] 白浚哲[1] 欧阳鹏[1] 徐浚哲[3] 张培建[1] 费敏锐[4] 华 亮[1]

(1. 南通大学电气与自动化学院 南通 226019;2. 南通市智能计算与智能控制重点实验室 南通 226019;
3. 南通电力设计院有限公司 南通 226019;4. 上海大学机电工程与自动化学院 上海 210053)

摘 要:随着无人机航拍技术的快速发展,低空场景中对基础设施等目标的精确识别需求日益增强。然而,传统目标检测与语义分割方法在边界刻画及同类实例区分方面仍存在不足。针对此问题,提出一种改进的多模态特征融合实例分割网络,以提升无人机遥感影像中目标分割的精细化程度与鲁棒性。所得方法的创新点:构建双路径输入结构,同时利用 RGB 影像与 DSM 信息,丰富多模态特征表达;面向 DSM 分支引入 HWF-LM 与 DMSCA,显著增强模型对高程与结构信息的表征能力;提出 FGCA 机制,实现跨模态特征的高效融合,从而提升复杂场景下的实例分割精度。在 Drone-OrthoSeg 数据集上,所提方法的边界框 mAP 和掩模 mAP 分别达到 42.63% 和 42.69%;在 NWPU VHR-10 数据集上分别为 77.86% 和 72.59%;在雾天海上场景的 FoggyShipInsseg 数据集上亦取得 63.86%与 59.69%的良好表现。实验结果表明,该方法在准确性与鲁棒性方面均优于现有先进方法,可为低空场景下的基础设施自动化检测和测量提供高效可靠的技术支持。

关键词:实例分割;多模态特征融合;卷积神经网络;无人机图像;地理空间物体

中图分类号:TP391.41;TN307 **文献标识码**:A **国家标准学科分类代码**:510.4050

Research on low-altitude object recognition method based on MFFISNet

Zhang Kun[1,2] Bai Junzhe[1] Ouyang Peng[1] Xu Junzhe[3] Zhang Peijian[1] Fei Minrui[4] Hua Liang[1]

(1. School of Electrical and Automation Engineering, Nantong University, Nantong 226019, China; 2. Nantong Key Laboratory of Intelligent Control and Intelligent Computing, Nantong 226019, China; 3. Nantong Electric Power Design Institute Co., Ltd., Nantong 226019, China; 4. School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 210053, China)

Abstract: With the rapid development of drone aerial photography technology, the demand for precise recognition of targets such as infrastructure in low-altitude scenarios has been increasingly growing. However, traditional object detection and semantic segmentation methods still have shortcomings in boundary delineation and distinguishing between similar instances. To address these issues, this paper proposes an improved multimodal feature fusion instance segmentation network, MFFISNet, to enhance the fineness and robustness of target segmentation in drone remote sensing images. The method in this paper includes three main innovations: a dual-path input structure is constructed, utilizing both RGB images and DSM information to enrich multimodal feature representation; for the DSM branch, HWF-LM and DMSCA are introduced, significantly enhancing the model's ability to represent elevation and structural information; FGCA mechanism is proposed to achieve efficient fusion of cross-modal features, thereby improving instance segmentation accuracy in complex scenarios. On the Drone-OrthoSeg dataset, MFFISNet achieved a bounding box mAP of 42.63% and mask mAP of 42.69%; on the NWPU VHR-10 dataset, the results were 77.86% and 72.59%, respectively; and on the foggy maritime scenarios of the FoggyShipInsseg dataset, it also achieved good performance of 63.86% and 59.69%. Experimental results indicate that the proposed method outperforms existing advanced methods in both accuracy and robustness, providing efficient and reliable technical support for automated detection and measurement of infrastructure in low-altitude scenarios.

Keywords: instance segmentation; multimodal feature fusion; CNNs; UAV imagery; geospatial object recognition

收稿日期:2025-07-09 Received Date: 2025-07-09

*基金项目:国家自然科学基金(62473216)项目资助

0　引　言

实例分割技术[1-3]在无人机航拍领域已成为一个关键的研究方向。随着低空经济产业成为目前国家重点推动的新兴产业形态之一,其核心目标之一正是依托无人机、低空通航器等平台开展物流运输、巡检勘测、应急救援、城市治理等场景的智能化应用。在低空区域的规划布局中,各类地表物体与建筑坐落于复杂地形结构之上,其空间配置需考虑不同地形因素带来的影响。如何通过无人机图像使用目标实例分割算法对低空场景下的周围物体进行类别识别并精确定位已成为解决低空地物勘探线路巡检工程的关键问题之一。

依托无人机低空航拍技术的高分辨率遥感影像在基础设施巡检、海上目标勘测和工业维护等工程场景中被广泛应用。然而,低空遥感影像受光照变化、遮挡复杂、尺度差异大与纹理变化明显等因素影响,使高精度实例分割面临较大挑战。仅依赖 RGB 的传统影像(只覆盖可见光 3 个波段,光谱表达能力有限)难以准确刻画不同目标的物理属性。与此同时,频谱特征通过离散余弦变换(discrete cosine transform,DCT)获得的纹理、结构与高频信息虽能补充 RGB 的空间细节,但仍不足以提供三维几何语义。数字表面模型(digital surface model,DSM)[4]是测绘学中用于数字化表达地表及地物表面三维形态术语,包含自然地形、建筑物、植被等顶部的高程信息,可为遥感图像中的目标区分与立体分析提供有效补充。RGB 与 DSM 的多模态融合构成当前高精度遥感实例分割的重要技术基础。

在低空勘测工程中,现有目标检测方法虽然能提供目标位置,但无法描绘完整边界,因此不满足实际测量与结构分析需求;而语义分割虽能对图像进行像素级分割,却无法区分同类目标的独立实例。在多种分割模型中,U-Net[5]通过编码器-解码器结构实现了精细边界提取,DeepLabv3+[6-7]通过空洞卷积与 ASPP 模块增强上下文语义,但二者均无法实现实例分割。实例分割模型由此成为更适合工程场景的解决思路,其主流架构包括两阶段模型和单阶段模型。以 Mask R-CNN[8]为代表的两阶段方法通过先检测后分割获得较高精度,但在小目标检测方面容易受背景噪声的干扰,且推理速度较慢。单阶段模型如 YOLACT[9]则通过并行预测原型掩膜与系数,在速度与精度间取得较好平衡,但仍有局限,在复杂遥感场景中表现仍不够稳定。现有方法在低空航拍实例分割任务中仍存在明显的不足。

本研究受 YOLO 的目标检测技术[10-12]思想的启发,并借鉴 YOLACT 的基本思想,提出了一种增强型多模态特征融合实例分割网络(multimodal feature-fused instance segmentation net,MFFISNet),该网络可用于对低空场景下的目标精准识别任务。

本文提出了一种提高实例分割网络准确率的方法,即整体结构部分采用 YOLACT 框架设计,并额外加入双路径输入结构以增强网络识别特征多样性的能力;为实现目标识别任务的具体特点,改进了 DSM 路径特征提取架构,再引入 Haar 小波特征学习模块(haar wavelet feature learning module,HWF-LM)和基于 DCT 的多光谱通道注意力机制,可通过变换特征提取域来增强模型特征提取能力;为了处理主干网络提取的多模态特征,提出一种采用特征引导方式的通道注意力(feature-guided channel attention, FGCA)机制,通过采用渐进式细化策略给每个特征通道生成专用的空间重要性特殊的图策略(spatial importance map, SIM)。本方法融合了通道和空间注意力,可有效促进全面的信息交互。

1　相关工作

1.1　实例分割方法与研究

实例分割方法通常采用“预训练+微调”范式,通过监督或自监督方式对主干网络进行通用的特征预训练,并结合特定的颈部结构以增强多尺度特征整合能力,从而适配不同下游任务。早期方法依赖候选区域与后处理步骤,为缓解这一问题,实例分割方法 Mask R-CNN 在 Faster R-CNN[13-14]框架上增设掩码预测分支,将检测与分割解耦,使模型在保持较高速度的同时显著提升分割精度,成为该领域的典型代表。

后续研究在精度与效率方向持续演进。一类方法聚焦于提升边界质量,如 PointRend[15]将分割视为逐点渲染过程,通过对困难边界点进行自适应细化,有效克服低分辨率上采样导致的边缘过度平滑问题,并在 COCO[16]与 Cityscapes[17]上显著改善掩码质量。另一类研究则致力于结构轻量化,如 QueryInst[18]将实例分割建模为查询机制,利用端到端的查询共享与多阶段并行监督,实现检测与分割信息协同,在多项基准上展示出优越性能与较高推理效率。RTMDet[19]通过兼容的主干网络和颈部匹配融合测试,实现检测与分割协同进行,展示出优越性能。Mask2Former[20]提出一种掩蔽注意力机制,通过改变预测掩蔽区域进行局部提取特征,进而有效地缓解性能瓶颈引发的一系列问题。

1.2　多模态融合技术研究

多模态融合技术[21]通过整合不同传感源数据,能够获取更丰富、更具判别性的特征,是提升目标识别与遥感分割精度的重要途径。随着遥感与视觉算法的发展,如何更高效利用多源互补模态[22]已成为关键挑战。多模

态特征融合可弥补单一模态信息不足,通过协同增强机制提升模型的精准度与鲁棒性。

CMGFNet[23]提出了由 RGB 与 DSM 双编码器组成的架构,通过跨模态门控融合模块实现可控特征融合,并采用自上而下的多层级特征交互,以增强对建筑物边界与细粒结构的判别能力。ScasMNet[24]进一步采用轻量化可分离卷积与级联式多尺度集成策略,将 RGB 与 DSM 作为双分支输入,从颜色纹理与高度信息两个维度补充区域语义,并结合 DenseCRF 优化边界,使遥感语义分割的细节表达更为精确。与前两者侧重层次化与空间信息集成不同,HAFNet[25]从冗余抑制和自适应融合角度出发,引入注意力感知融合模块,以动态选择显著跨模态特征,减少简单级联带来的冗余与信息丢失。

在更细粒度的显著性场景中,DSNet[26]提出动态特征选择机制,以跨模态全局上下文模块获取粗语义线索,并通过动态补偿策略实现 RGB 与深度特征的自适应互补。该网络利用跨尺度细化与边界增强生成清晰显著图,显著图中的掩码可作为实例分割中的重要先验,不仅为实例定位与边界推断提供高置信区域,还能有效减少误检与欠分割现象,从而提升下游实例分割模型的结构感知能力与表达质量。

2 MFFISNet 网络架构与方法设计

本文的模型架构 MFFISNet 是一种基于 YOLACT 框架的“检测+分割”并行的单阶段实例分割方法,该网络由主干网络、颈部网络、边界框预测头和掩码生成头组成。该模型可将无人机光学图像与 DSM 信息有效结合,以融合多种特征模态来提高实例分割精度。此外,本文对网络架构部分对特征提取主干网络做出改进,采用增强型颈部模块融合策略优化头部架构,获得更细分的分割结果。这些改进都可使该模型适合于多种线路建设勘测工程中或其他以目标识别任务为主的相关实例分割任务。提出的 MFFISNet 思路遵循着一个递进设计的流程:首先通过引入 RGB+DSM 双分支结构提高地形几何与纹理表达能力,并利用小波 Haar 变换与频域注意力方法优化其结构表征能力,最终通过 FGCA 增强模态间的融合能力,实现对图像中边界部位分割。

2.1 多模态特征提取网络

本方法通过拓展双分支多模态架构 YOLACT 框架达到增强特征表示的多样性,旨在提高高分辨率无人机采样图像中的物体识别和分割的精度。改进后的特征提取网络完整结构如图 1 所示。

MFFISNet 的双路径多模态融合网络架构同时处理光学图像(图 1 上半部分路径)和 DSM 数据(图 1 下半部分路径)。光学图像路径是指输入图像依次经过初始卷积模块和 3 个由 BottleNeck 模块构成的层级多尺度特征提取。随后,特征通过 SPPF 金字塔池化模块以增强感受野。DSM 路径是由 DSM 数据首先通过一个特征编码器,随后进入包含 Haar 小波变换的特征学习模块进行频域的特征提取,并最终通过基于 DCT 的多光谱通道注意力模块(DCT-based multi-spectral channel attention_spatial pyramid pooling fast, DMSCA_SPPF)进行频域信息增强。特征融合阶段将两个路径产生的同尺度特征通过特征导向跨模态注意力融合模块进行深度融合,有效结合了光学图像的外观信息和 DSM 的高程信息。预测头部分是融合后的多尺度特征被送入模型的颈部和头部网络。

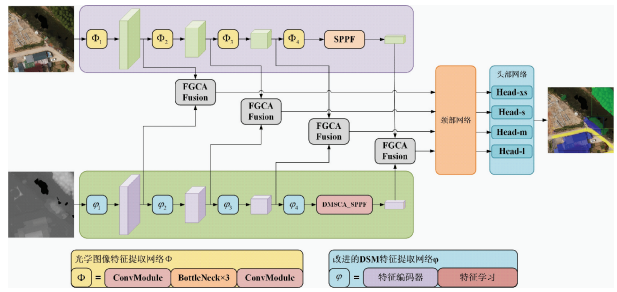

图 1 改进型实例分割网络 MFFISNet 结构

Fig. 1 Improved instance segmentation network MFFISNet structure

2.2 改进的 DSM 特征提取网络

MFFISNet 中使用的多模态特征提取网络由光学图像特征提取网络和 DSM 特征提取网络两个主要部分组成。

光学图像特征提取网络主要基于 YOLOv8[27]的主干结构设计。YOLOv8 的主干结构在设计上继承了 YOLOv3[28]所奠定的多尺度特征金字塔与逐步下采样的核心范式。YOLOv8 在 YOLOv5[29]的特征提取块 C3 基础上改进了 C2f 模块作为其主干网络基本构建单元,该模块通过优化梯度流有效提升了训练效率与性能。

DSM 特征提取网络是 YOLOv8 主干网络的改进版本,用 HWF-LM 取代了原有的下采样模块,并在特征图输入到 SPPF 模块之前引入了 DMSCA,模型结构如图 2 所示。

改进的特征提取网络中引入的 HWF-LM 由无损特征编码和特征表示学习两个模块组成。无损特征编码器的功能是转换特征并进行空间维度的下采样。这通过使用 Haar 小波变换[30-32]来实现,该方法保留完整且重要信息的同时,可有效缩小特征表示空间尺寸。特征表示学习模块由典型的卷积层、归一化操作和激活函数 ReLU 组成。为了在最小化空间分辨率的同时保留关键信息,无损编码器应用了基于 Haar 小波的变换。这种方法降低了特征图的空间尺度,并保持了关键数据的完整性。

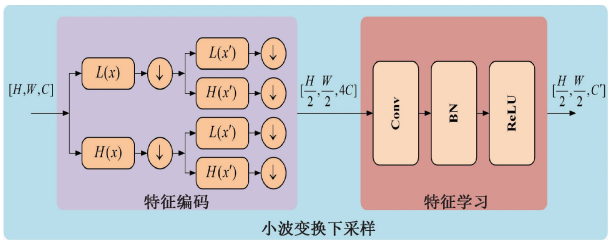

图 2　改进的 DSM 特征提取网络

Fig. 2　improved DSM feature extraction network

一阶一维 Haar 变换的基础小波及其相应的尺度函数如下：

$$\begin{cases}\phi_1(x)=\dfrac{1}{\sqrt{2}}\phi_{1,0}(x)+\dfrac{1}{\sqrt{2}}\phi_{1,1}(x)\\ \psi_1(x)=\dfrac{1}{\sqrt{2}}\phi_{1,0}(x)-\dfrac{1}{\sqrt{2}}\phi_{1,1}(x)\end{cases}\tag{1}$$

式中：$\phi_{j,k}(x)$ 的定义如式(2)所示。

$$\phi_{j,k}(x)=\sqrt{2^j}\phi(2^jx-k),\quad k=0,1,\cdots,2^j-1\tag{2}$$

式中：j 和 k 分别表示 Haar 基函数的阶数(或尺度)和索引(在二维图像中对应方向)。尺度函数 $\phi_{0,0}(x)$ 如式(3)所示。

$$\phi_{0,0}(x)=\phi_0(x)=\begin{cases}0, & x<0\\ 1, & 0\leqslant x<1\\ 0, & x\geqslant 1\end{cases}\tag{3}$$

因此，一级 Haar 变换可通过零级 Haar 基函数表示，如式(4)所示。

$$\begin{cases}\phi_1(x)=\phi_0(2x)+\phi_0(2x-1)\\ \psi_1(x)=\phi_0(2x)-\phi_0(2x-1)\end{cases}\tag{4}$$

长度为 L 的信号可被分解为两部分长度为 $L/2$ 的信号，分别对应低通和高通滤波分量。当 Haar 小波变换应用于二维信号时，可生成 4 个分量，每个分量的空间分辨率均为原始信号的 1/2。

图 2 所示为如何对分辨率为 $H\times W$ 的图像进行 Haar 小波变换分解。其中，L 和 H 分别代表低通和高通滤波器，低通滤波器用于提取图像的近似信息，而高通滤波器对应细节信息。

采用 Haar 小波变换可将输入特征分解为水平(H)、垂直(V)、对角(D)3 类细节分量及近似分量 4 个子带。尽管空间尺度保持不变，但通道数扩展 4 倍，使有效空间信息在不损失内容的情况下被映射至通道维度[33]，从而提升后续卷积网络对关键结构与高频细节的提取能力。其后的特征表示模块由 1×1 卷积、批量归一化和 ReLU 构成，以完成通道对齐与非线性映射抑制冗余特征，以增强后续网络的判别性特征表达。DCT[34] 可将频域特征映射为不同频率的余弦系数，而全局平均池化(GAP)操作[35]本质上等价于二维 DCT 的最低频分量，因此可视为一种特殊的频域特征提取方式，为构建频域注意力模块提供理论支撑。GAP 操作表达式为：

$$f_{0,0}^{2d}=\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}x_{i,j}^{2d}\cos\left(\frac{0}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{0}{W}\left(j+\frac{1}{2}\right)\right)=\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}x_{i,j}^{2d}=GAP(x_{i,j}^{2d})HW\tag{5}$$

式中：$f_{0,0}^{2d}$ 代表二维 DCT 部分的最低频率分量，该计算分量可与 GAP 操作的结果成比例。基于上述公式，在 DMSCA 中采用 GAP 操作仅保留最低频率信息，而其他代表通道的重要模式频率分量则被忽略。从实验结果的角度出发，这些信息对通道建模仍具有重要意义的，不应舍弃。

DMSCA_SPPF 的结构如图 3 所示。该方法可将传统的 GAP 方法扩展至包含多个频率分量的框架，从而引入更多频谱信息，丰富了压缩后的通道特征，从而提升注意力机制的表达能力。

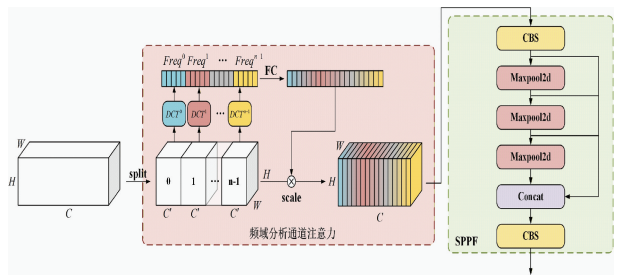

图 3　DMSCA_SPPF 模块结构

Fig. 3　DMSCA_SPPF module structure

首先，将输入 X 沿通道维度划分为多个部分。设划分后的部分为 $[X_0,X_1,\cdots,X_{n-1}]$，其中 $X^i\in\boldsymbol{R}^{C'\times H\times W}$，$i\in\{0,1,\cdots,n-1\}$，$C'=C/n$，且 C 可被 n 整除。对每个部分分配一个对应的二维 DCT 频率分量，并利用其变换结果作为 DMSCA 的压缩信息，如式(6)所示。

$$Freq^i=2DDCT^{u_i,v_i}(X^i)=\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}X_{:,h,w}^iB_{h,w}^{u_i,v_i}\quad i\in\{0,1,\cdots,n-1\}\tag{6}$$

式中：$[u_i,v_i]$ 表示对应于 X^i 频率分量的二维索引；而 $Freq^i\in\boldsymbol{R}^{C'}$ 代表压缩后的 C' 维向量。最终，通过拼接所有部分的压缩向量，可以得到完整的压缩向量，如式(7)所示。

$$Freq=compress(X)=cat([Freq^0,Freq^1,\cdots,Freq^{n-1}])\tag{7}$$

式中：$Freq\in\boldsymbol{R}^C$，该部分最终计算所得多频谱向量。整个多频谱通道注意力机制，即 DMSCA 可由式(8)表示。

$$DMSCA=\sigma(fc(Freq))\tag{8}$$

2.3　改进的融合网络用于提取主干网络的特征

骨干网络的改良若直接采用拼接方式，则无法充分

发挥 RGB 与 DSM 数据间的交互优势。为解决深层交互缺失问题，本文引入特征注意力模块（feature attention module，FAM），以增强在特征融合阶段光学影像与 DSM 数据之间的关联性与交互性。

FAM 由通道注意力机制和空间注意力机制组成，这两种机制依次作用于特征图，分别计算通道维度和空间维度上的注意力权重。通道注意力机制生成通道权重向量，记作 $W_c \in \boldsymbol{R}^{C\times1\times1}$，用于重新校准特征映射。SIM 记作 $W_s \in \boldsymbol{R}^{H\times W}$，自适应地表示特征图中不同区域的重要性。

然而，FAM 中的空间注意力结构部分仅能通过特征图存在不均匀结构时进行建模，而无法解决特征图存在不均匀性的问题。FAM 的通道注意力仅建模了通道之间存在的差异，也尚未能充分考虑上下文信息对特征融合。随着特征通道数增加，不同图像中存在的分布信息被统一编码，不同通道在特征空间中具有不同语义，即对于每个特殊的特征通道，其在空间维度上的数据分布结果是非均匀的，因此需要设计一个通道专属的 SIM。FAM 还存在另一个局限性，即其通道注意力权重 W_c 和空间注意力权重 W_s 是按顺序计算的，且二者增强特征独立，之间不存在有效的信息交互。为全面解决上述存在的问题，本文又提出了 FGCA 为每个输入特征通道注入独立的 SIM，确保在计算过程中通道部分充分通道注意力权重和空间注意力权重，以确保二者信息的有效交互。FGCA 还通过为每个通道生成独立的空间重要性图，使得模型能够根据不同模态中通道的语义差异动态调整关注区域，从机制上提高了对边界细节、小物体和复杂纹理区域的刻画能力。FGCA 模块结构如图 4 所示。

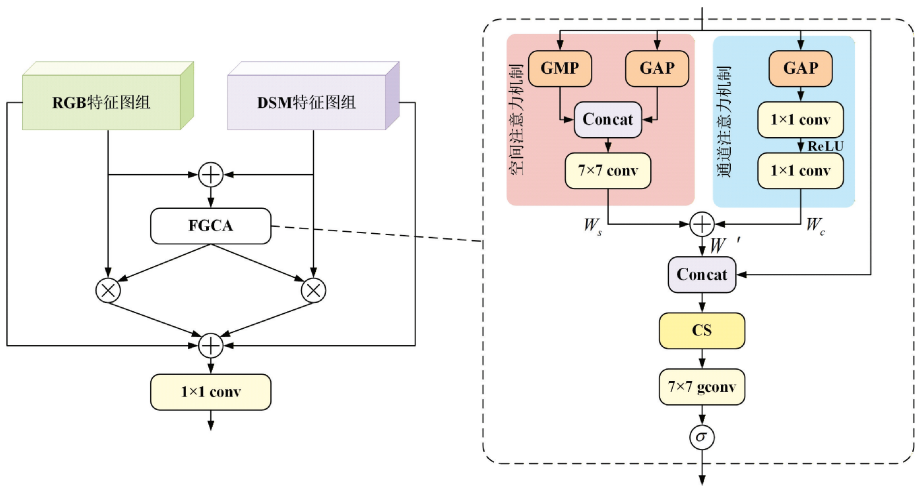

图 4 FGCA 模块结构

Fig. 4 FGCA module structure

为了降低模型参数量与复杂度，该模块首先采用卷积操作将通道数从 C 降至 C/r，r 为通道缩减比部分，随后通过二次卷积操作将所有通道数恢复至 C。设定 $r=16$，并将通道维度降为固定大小。根据广播机制，通过加法运算融合通道注意力权重 W_c 和空间注意力权重 W_s 得到空间重要性，如式(9)所示。

$$W' = W_c + W_s \tag{9}$$

为获得最优空间重要性，需依据输入特征对模型通道大小进行适当调整，给每个通道生成特殊专属的 SIM。通过通道混洗（channel shuffle，CS）操作可对 W' 和 X 的通道实现交错重排，如式(10)所示。

$$W = \sigma(gconv_{7\times7}(CS([X, W']))) \tag{10}$$

式中：σ 表示 sigmoid 激活函数；$CS(\cdot)$ 代表通道混洗操作计算。卷积核大小为 k 的卷积层采用分组卷积 gconv 计算，分组参数设定为 7。

FGCA 通过为每个通道分配独立的 SIM，引导模型关注各通道中的重要区域，从而增强特征中的关键信息表达，有效强化光学影像与 DSM 数据之间的联系。

3 实验结果

3.1 数据集

1）DroneOrthoSeg 数据集

DroneOrthoSeg 数据集是拍摄于城市郊外地区地貌的无人机采集图像数据集。为确保影像集分辨率达约 0.5 m，将飞行高度设定为 150 m。在无人机作业过程中，采用网格状飞行策略从不同视角获取影像。本文按 8∶2 的比例将 525 幅图像数据集拆分为训练子集和测试子集，将采集的倾斜影像数据导入 Smart3D 软件，可生成正射影像和 DSM 数据，如图 5 所示。

2）NWPU VHR-10 数据集

NWPU VHR-10 数据集[36-37]专为光学遥感图像而设

图 5　DroneOrthoSeg 的部分图像

Fig. 5　Some images of DroneOrthoSeg

计,并侧重于实例分割。包含飞机(AI)、篮球场(BC)、棒球场(BD)、桥梁(BR)、运动场(GTF)、港口(HB)、船舶(SH)、储罐(ST)、网球场(TC)和车辆(VC)10 个类别。该数据集包含 715 幅的光学遥感图像,其空间分辨率为 0.8 ~ 2 m。该数据集中还包含了 85 幅来自 Vaihingen 数据集[38]的彩色红外图像及 650 幅带注释的图像。在实验设计中,本文选择了 650 张带注释的图像进行训练,并将数据集划分为 80%用于训练,20%用于测试。NWPU VHR-10 数据集最初的建立是为了适配目标检测任务,作者后来在文献工作中补充了其在像素级层面的图像注释,使其可适用于实例分割任务模型的性能检测。

3)雾天 ShipInsseg 数据集

为了模拟无人机在实际检测环境中恶劣天气条件对模型检测部分产生的不良影响,并量化模型在极端条件下的数据监测的鲁棒性,本文还选择了雾天 ShipInsseg 数据集[39]作为一项评估模型在雨雾环境下的性能参考。该数据集在原标准 ShipInsseg 数据集的基础上引入了基于大气散射模型的合成雾叠加,可以模拟雾天环境中图像不同程度能见度降低。该数据集也保留了实例级标签,使本文可测试在清晰度降低时 MFFISNet 检测和分割的鲁棒性。对于该数据集,本文保留之前相同的训练配置,以严格测试本文模型与其他现有公开模型之间相比存在图像分布偏差情况下的零点稳健性。

3.2　实验环境配置

训练的关键超参数设置为 300 个 epochs、batch_size 为 4 张图像、4 个数据的加载线程,并在最后 10 个时期期间停用马赛克增强、激活自动混合精度(AMP),以减小不必要的计算消耗,提升模型的最终表现性能。实验环境配置如表 1 所示。

3.3　模型复杂度分析

为了能进一步有效评估本文所提出的 MFFISNet 模型在实际无人机线路巡检场景中的实用性优劣,全面的计算了模型的复杂度,如表 2 所示。表 2 涵盖参数量、每秒浮点运算次数(GFLOPs)以及标准分辨率输入下的推理速度(FPS)等指标。这些指标对嵌入式或机载无人机系统的边缘部署特别关键,这是因为提升模型性能的关键限制是提高模型计算效率并降低能耗。

表 1　实验环境配置

Table 1　Experimental environment setup

硬软件资源	选型
运行系统	UBUNTU 20.04
CPU	AMD Ryzen 7 5800H 16G
GPU	NVIDIA RTX 3090 24G × 2
编程语言	Python 3.8
算法框架	TORCH 1.13.1 + CUDA 11.7

MFFISNet 的参数量为 48.3×10^6,略高于部分对比模型,其增加主要归因于双路径架构与额外的 HWF-LM、DMSCA 与 FGCA 模块(表 2)。这些新功能在融合多模态特征时显著提升模型表达能力。

表 2　MFFISNet 与其他代表性模型的计算复杂度比较

Table 2　Comparison of computational complexity between MFFISNet and other representative models

模型	参数量/($\times10^6$)	计算量/GFLOPs	帧数/fps
U-Net	13.4	31.0	56
DeepLabv3+	43.5	86.7	24
MaskR-CNN	44.5	92.2	20
YOLACT	33.2	75.3	33
PointRend	49.0	101.3	19
QueryInst	54.8	124.5	18
Mask2Former	47.2	89.6	21
YOLOv8-seg	44.1	86.4	27
MFFISNet	48.3	93.7	22

MFFISNet 在准确率和计算效率之间取得了良好的平衡,因其设计上提供了出色的分割准确率和目标识别能力,同时在实际无人机部署场景中保持了计算可行性。

3.4　实验结果比较分析

将所提方法与当前最先进的实例分割模型进行比较,具体包括 U-net、DeepLabv3+、MaskR-CNN、PointRend、YOLACT、QueryInst、Mask2Former 和 YOLOv8-seg。在实验设计比较中,所有最先进的模型都统一使用 ResNet50 作为预训练主干。而 U-net 和 DeepLabv3+模型常被用于语义分割任务,在实际应用中,它们与实例分割模型存在功能差异,且针对场景理解设计的不同方面。本文也将上述数据集转换为语义分割数据集,并部署了相应的训练环境来训练这两个模型并实现相应的可视化处理,旨在更全面地展示这些提及的模型在分割任务中相较于传统语义分割模型存在的优势。然而,由于 实例分割任务与语义分割任务之间存在显著的功能差异,U-Net 和

DeepLabv3+均没有 mAP_{box}、mAP_{mask}、mAP^{50}_{mask}、mAP^{75}_{mask} 等衡量指标。本文也对 MFFISNet 的结果与其他在 DroneOrthoSeg、NWPU 和 FoggyShipInsseg 数据集上的最新方法的结果进行了定量比较,以 IoU 阈值计算为参考。具体而言,比较结果为 IoU 阈值为 50%的 520 个结果与 75%的 521 个结果(边界框,mAP75 个掩膜)以及总体平均的 mAP 大小。DroneOrthoSeg 数据集的对比实验中,MFFISNet 超越了目前所有其他方法,通过比较模型中的各种评估指标,MFFISNet 在边界框检测和掩码任务中均表现出显著优势,整体性能领先于其他方法,尤其是在实例分割任务展现了其强大的综合能力。实验数据如表 3 所示。

表 3 DroneOrthoSeg 数据集对比实验数据

Table 3 DroneOrthoSeg dataset comparative experiment data (%)

模型	mAP_{box}	mAP^{50}_{box}	mAP^{75}_{box}	mAP_{mask}	mAP^{50}_{mask}	mAP^{75}_{mask}
MaskR-CNN	35.6	63.8	30.1	32.0	60.2	30.6
YOLACT	18.4	46.9	11.6	16.5	41.5	12.3
PointRend	38.5	68.9	37.4	38.9	66.1	37.8
QueryInst	40.5	62.4	39.4	41.2	64.7	36.7
Mask2Former	41.8	69.2	40.1	42.1	68.8	38.2
YOLOv8-seg	39.13	63.91	38.81	38.63	66.13	36.23
MFFISNet	**42.63**	**70.74**	**40.85**	**42.69**	**70.30**	**38.41**

MFFISNet 在目标识别任务中实现指标为 42.63%的 mAP_{box},位居所有方法之首。与其他方法比较,尤其是与 Mask2Former(41.8%)和 YOLOv8-seg(39.13%)相比,MFFISNet 表现出明显优势。在较高的 IoU(75%)条件下,MFFISNet 仍保持了较高的准确率,展现了其强大的细粒度检测能力。在 Mask 分割任务中,MFFISNet 也保持领先地位,mAP_{mask} 为 42.69%,略高于排名第 2 的 Mask2Former(42.1%),并显著优于其他方法。例如,YOLOv8-seg 的 mAP_{mask} 为 38.63%,Mask R-CNN 为 32.0%,而 YOLACT 仅为 16.5%。MFFISNet 的表现性能也优于 DeepLabv3+(mIoU 为 35.8%)和 U-Net(mIoU 为 14.5%)等传统语义分割模型,在 mAP^{75}_{box} 指标上,展现了其在细节分割任务中的强大能力。在 mAP^{50}_{box} 和 mAP^{75}_{box} 的高 IoU 指标下,MFFISNet 分别保持了 70.74%和 40.85%的优异性能。高 IoU 指标意味着想要达到高性能需要更精确的分割边界,而 MFFISNet 可有效地解决这些挑战。虽然 Mask2Former 和 PointRend 在某些特定场景下表现出色,但 MFFISNet 在整体 mAP 和高 IoU 性能方面仍然更胜一筹。在需要高 IoU 的任务中,YOLOv8-seg 表现良好,但在 Mask 分割任务中的准确率明显低于 MFFISNet,这表明其在实例分割任务中存在局限性。根据实验结果和图 6 所示的视觉推理结果(不同模型可视化代码中的颜色结果有所不同),MFFISNet 在边界框检测和分割任务中均展现出显著优势,尤其是在需要高精度和高 IoU 的场景下。

如表 4 所示,在基于 NWPU VHR-10 数据集的对比实验中,MFFISNet 展现出优异的性能,在边界框识别和掩码分割任务中的多个评估指标上也超越了对比实验中的其他方法。

在掩膜任务中,MFFISNet 的 mAP_{mask} 为 72.59%,超过了 YOLOv8-seg(69.42%)和 Mask2Former(68.1%)。

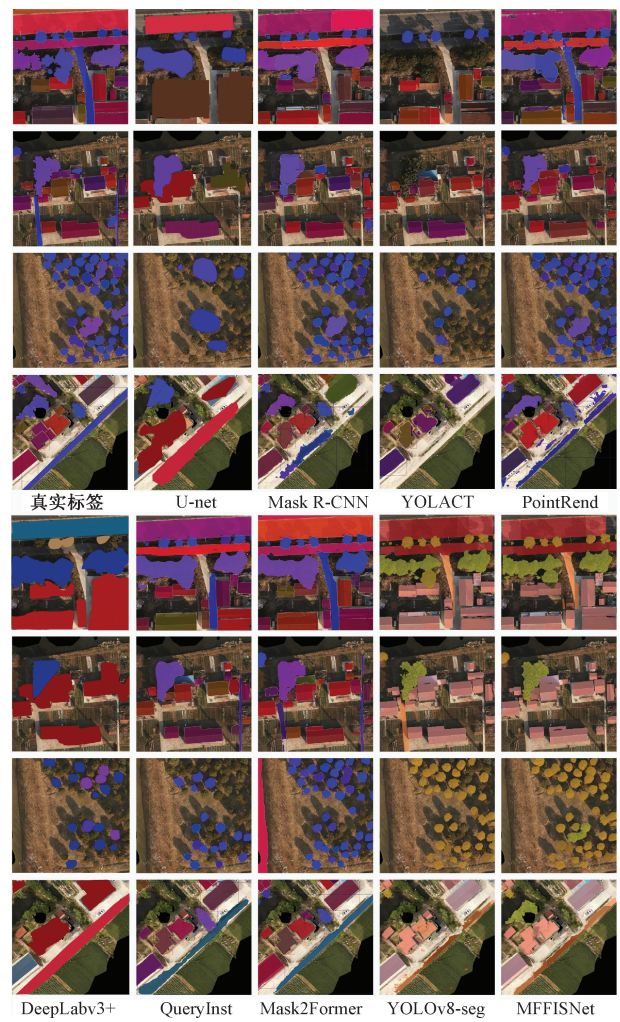

图 6 在 DroneOrthoSeg 数据集上使用各种实例分割算法进行目标识别的可视化结果

Fig. 6 Visualization results of object recognition using various instance segmentation algorithms on the DroneOrthoSeg dataset

表 4　NWPU VHR-10 数据集对比实验数据

Table 4　NWPU VHR-10 dataset comparative experiment data　　(%)

模型	mAP_{box}	mAP^{50}_{box}	mAP^{75}_{box}	mAP_{mask}	mAP^{50}_{mask}	mAP^{75}_{mask}
MaskR-CNN	63.4	90.3	74.0	59.9	89.6	63.4
YOLACT	46.9	85.9	48.6	41.6	76.6	41.2
PointRend	65.7	93.8	77.4	66.4	93.7	73.0
QueryInst	59.6	78.8	70.9	56.9	78.9	63.2
Mask2Former	68.8	90.5	79.0	68.1	92.6	74.5
YOLOv8-seg	75.85	94.46	84.89	69.42	94.17	76.95
MFFISNet	**77.86**	**96.29**	**86.64**	**72.59**	**95.63**	**83.26**

MFFISNet 的 mAP_{mask} 达 83.26%,远超其他方法,显示其在精细掩膜分割任务中的强大能力。相比之下,YOLOv8-seg 的掩膜精度为 76.95%,在高精度分割任务中仍有不足。在 mAP^{50}_{box}(96.29%)和 mAP^{75}_{box}(86.64%)的高 IoU 指标下,MFFISNet 的边界框和掩膜精度显著高于其他方法,有效地解决了高精度分割挑战,体现了其卓越的分割精度和鲁棒性。

与 YOLOv8-seg 相比,MFFISNet 不仅在边界框检测任务中表现出色,而且在掩模分割精度方面也显著优于 YOLOv8-seg。虽然 YOLOv8-seg 在边界框检测任务中表现良好,但在高精度掩模分割任务中的表现却远低于 MFFISNet。Mask2Former(mAP_{box} 为 68.8%,mAP_{mask} 为 68.1%)和 PointRend(mAP_{box} 为 65.7%,mAP_{mask} 为 66.4%)也表现出色;该模型的 mAP_{mask} 转换 mIoU 为 70.2%,远高于 U-Net 的 mIoU 为 32.1%,DeepLabv3+的 mIoU 为 51.8%。在高 IoU 条件下,准确率仍低于 MFFISNet。而 MFFISNet 在边界框检测和掩膜分割精度方面均占优。图 7 所示的推理可视化结果进一步验证了 MFFISNet 在边界框检测和掩膜分割任务上的综合优势使得其可在实例分割任务中达到最佳效果。

尽管 YOLOv8-seg 和 Mask2Former 在某些方面取得了突破,但 MFFISNet 在整体精度和处理高 IoU 任务方面仍然遥遥领先。由此可见,MFFISNet 不仅在实例分割领域表现出色,而且在需要高精度和高 IoU 的任务中也拥有显著优势,展现出其广泛的应用潜力。

基于 MFFISNe 在边界框检测与掩膜分割任务的出色表现,其在高精度和高 IoU 任务中展现出明显的优势。这些优势都足以证明其在高精度的复杂场景分割任务下,仍具有巨大的潜在价值。

比如在无人机实际室外线路巡检任务场景中,必须要求无人机系统时刻保持稳健运行,以具备应对不同潜在自然灾害干扰的能力。其中就包含应对雾、暴雨、低能见度等极端天气条件影响图像成像质量的关键任务。从数据层面上为能更有效直观地评估和体现本文模型在这种具有挑战性的室外环境中依旧具备良好的部署潜力,本文设计了本模型在海上场景下对船舶个体进行实例分割的模拟实验,以评估本文提出的模型在复杂环境下成像质量不高时的具备良好的鲁棒性和泛化能力。

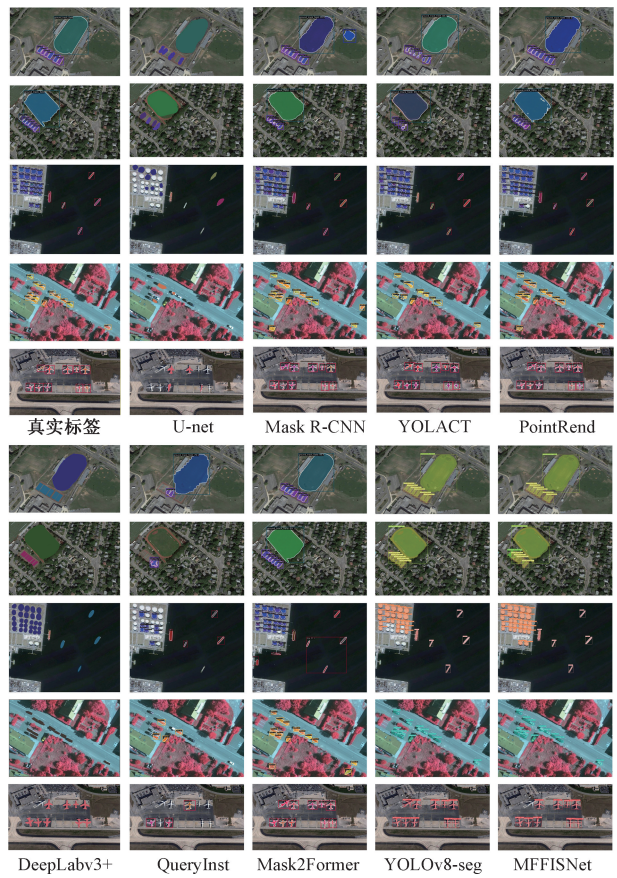

图 7　在 NWPU 数据集上使用各种实例分割算法进行目标识别的可视化结果

Fig. 7　Visualization results of object recognition using various instance segmentation algorithms on the NWPU dataset

从 FoggyShipInsseg 数据集上的对比实验可以看出,通过多项评估指标的计算分析可得,MFFISNet 在边界框检测和掩膜分割任务中都体现出较为明显的优势,与其他方法相比取得了领先的整体结果。综合实验结果如表 5 所示。

从 FoggyShipInsseg 数据集上的实验结果可以看出,

MFFISNet 在各项评测指标中均取得最优表现,展现出良好的实例分割性能。在边界框检测方面,MFFISNet 的 mAP_{box}、mAP^{50}_{box} 和 mAP^{75}_{box} 分别达到 63.86%、81.35% 和 70.24%,显著优于当前主流的其他方法。相比之下,YOLOv8-seg 的 mAP_{box} 为 60.85%,Mask2Form 仅为 54.2%,均落后于本模型。在掩模分割任务中,MFFISNet 的 mAP_{mask}、mAP^{50}_{mask}、mAP^{75}_{mask} 分别达到 59.69%、81.63% 和 65.27%,在分割精度和细粒度轮廓恢复方面展现出明显的优势。相比之下,YOLOv8-seg 的 mAP^{75}_{mask} 仅为 59.65%,Mask2Former 为 59.6%,均未能达到 MFFISNet 的高精度水平,在目标边界复杂或局部细节明显的场景下尤为明显,而 MFFISNet 的表现则更为稳定。此外,与 U-Net、DeepLabv3+等传统语义分割模型相比,MFFISNet 的优势更加明显。U-Net 的 mIoU 为 23.7%,DeepLabv3+ 的 mIoU 为 69.4%,均显著低于 MFFISNet 在实例级预测方面的表现。

表 5 FoggyShipInsseg 数据集对比实验数据

Table 5 FoggyShipInsseg dataset comparative experimental data (%)

模型	mAP_{box}	mAP^{50}_{box}	mAP^{75}_{box}	mAP_{mask}	mAP^{50}_{mask}	mAP^{75}_{mask}
MaskR-CNN	46.5	71.3	57.9	42.8	52.6	46.8
YOLACT	31.1	65.5	31.4	24.1	58.3	27.2
PointRend	43.5	72.6	63.7	45.1	72.4	39.5
QueryInst	45.4	67.3	51.2	35.9	71.9	57.2
Mask2Former	54.2	76.1	60.3	53.6	72.7	59.6
YOLOv8-seg	60.85	78.46	66.89	58.42	79.17	59.65
MFFISNet	**63.86**	**81.35**	**70.24**	**59.69**	**81.63**	**65.27**

由表 5 可知,传统方法相较于 Mask R-CNN、YOLACT 等模型在多个指标上存在较为明显的差距,而在高 IoU 条件下的准确率普遍相对较低。虽然 PointRend 和 QueryInstade 等方法在某些指标上具有竞争力,但在综合性能上仍然不如 MFFISNet。在 FoggyShipInsseg 数据集上,MFFISNet 在 mAP_{box} 和 mAP_{mask} 上分别达到 63.86% 与 59.69%,均明显优于对比方法,表明该模型在低能见度下的鲁棒性较强。这是因为 DMSCA 可抑制散射与雾导致的低频漂移,使雾天图像中全局对比度降低,而低频分量更容易被雾散射平滑,故 DMSCA 能捕捉并放大全频谱中对目标有判别力的频段,从而维持目标特征。HWF-LM 可保留边缘细节,小波分解在频域上能够将因雾而被模糊的边界成分分离并在高频通道中保留,使得边界恢复能力增强。FGCA 则强化了模态间的特征互补,即在雾天情况下,RGB 信噪比下降,DSM 的高度信息变得更重要,FGCA 允许 DSM 引导 RGB 在空间上给予更多权重,从而提升整体鲁棒性。

如图 8 所示,MFFISNet 可在反映掩膜细节水平的重要指标 mAP^{75}_{mask} 方面取得显著领先,验证了其在复杂环境中准确建模物体轮廓的能力。MFFISNet 不仅在 FoggyShipInsseg 数据集上实现了准确的边界框检测与掩膜分割任务的全面提升,还可在高 IoU 场景下展现更好的鲁棒性和泛化能力,为复杂天气环境下的无人机巡检线路场景下的目标实例分割提供了更为可靠的技术支持。

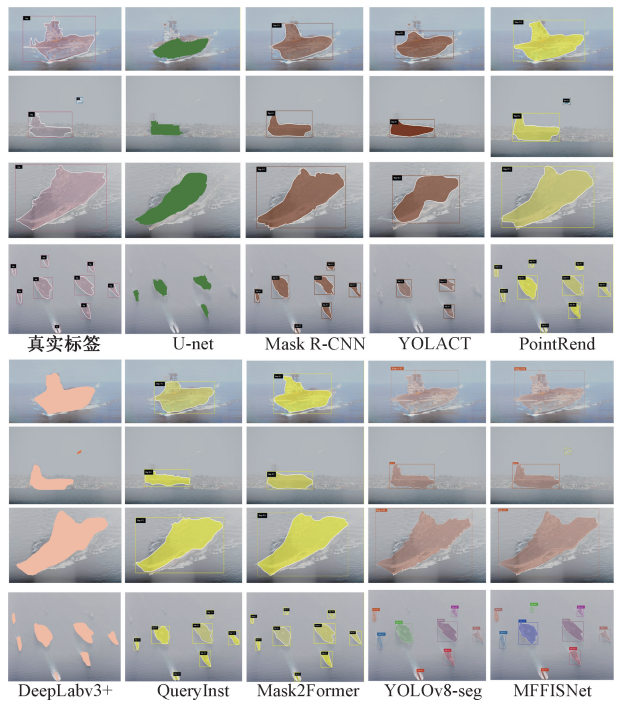

图 8 在 FoggyShipInsseg 数据集上使用各种实例分割算法进行目标识别的可视化结果

Fig. 8 Visualization results of object recognition using various instance segmentation algorithms on the FoggyShipInsseg dataset

3.5 消融实验

本文在 DroneOrthoSeg、NWPU VHR-10 和 FoggyShipInsseg 数据集上进行了一系列消融实验,以分析每个模块对模型性能的影响。设计了多个实验,逐步将多模态主干架构(multimodal backbone, MMB)、改进的

DSM 模态特征提取网络 MMB-DSM 和 FGCA 等引入基线模型中。基线模型采用以 YOLOv8 特征提取网络作为网络主干部分,FPN 为颈部部分,采用融合 YOLACT 边界框检测头和掩模检测头的头部设计,比较不同模块组合产生的实验结果,以全面评估每个模块在模型中对模型性能的贡献。实验结果如表 6 所示。

表 6　在 DroneOrthoSeg、NWPU VHR-10 和 FoggyShipInsseg 数据集上的消融实验结果

Table 6　Ablation experiment results on the DroneOrthoSeg, NWPU VHR-10, and FoggyShipInsseg dataset　(%)

数据集	MMB	MMB-DSM	FGCA	mAP_{box}	mAP^{50}_{box}	mAP^{75}_{box}	mAP_{mask}	mAP^{50}_{mask}	mAP^{75}_{mask}
DroneOrthoSeg				39.13	63.91	38.81	38.63	66.13	36.23
	√			40.32	65.46	38.86	38.88	66.19	36.59
	√	√		41.17	67.68	39.71	39.62	66.23	36.82
	√	√	√	**42.63**	**70.74**	**40.85**	**42.69**	**70.30**	**38.41**
NWPU VHR-10				75.85	94.46	84.89	69.42	94.17	76.95
	√			76.21	94.46	84.19	70.56	94.23	78.40
	√	√		77.60	95.93	86.53	72.39	95.62	81.08
	√	√	√	**77.86**	**96.29**	**86.64**	**72.59**	**95.63**	**83.26**
FoggyShipInsseg				60.12	77.81	65.46	56.94	78.52	60.87
	√			61.04	78.92	66.25	57.51	79.43	61.72
	√	√		62.37	79.65	68.02	58.34	80.32	63.14
	√	√	√	**63.86**	**81.35**	**70.24**	**59.69**	**81.63**	**65.27**

在对未增添任何模块予以改进时,基线模型呈现出来的性能指标分别是 39.13%以及 38.63%。这样的结果能够为后续相关模块的引入提供一定的技术参照。当引入了 MMB 之后,相应指标分别提升到了 40.32%与 38.88%。这表明了 MMB 模块具备整合多种模态信息的能力,并且明显强化了模型从不同模态当中提取特征的能力,因此本模型在目标检测方面的性能得以提升。在以 MMB 为基础的前提之下,本文又进一步引入了 MMB-DSM 模块设计,使相关指标分别达到了 41.17%和 39.62%。这一改进举措让 DSM 模块在提取多模态特征层面获得了有效提升,同时模型在细粒度分割任务上的完整程度也得到增强。当三者全部启用(MMB+MMB-DSM+FGCA)后,模型在 DroneOrthoSeg 上 mAP_{box} 与 mAP_{mask} 分别达到 42.63%与 42.69%,进一步强化了模型在高 IoU 阈值下的表现。与前两种模块不同,FGCA 并非仅提升单一模态的表达能力,而是通过生成通道独立的空间注意图,使 DSM 提供的几何线索与 RGB 的纹理特征实现非均匀融合,从机制上避免跨模态融合的平均化损失。其结果是,模型在精细实例边界、结构复杂区域以及颜色相似但高度不同的目标中表现最为明显。因此 FGCA 的提升幅度虽然出现在三阶段改进的最后,但其贡献不可替代。在 NWPU-VHR10 中 MMB-DSM 带来的提升高于 MMB,主要因为该数据集中小目标的比例更高,DSM 与频域增强机制能够提高小尺度区域信噪比与结构表达能力,使预测更加稳定一致。在以雾天为背景的 FoggyShipInsseg 数据集中,FGCA 与 DMSCA 的贡献尤为明显,其整体提升幅度大于清晰场景下的贡献。这是因为雾环境会削弱 RGB 的低频结构信息,使 DSM 与频谱信息成为判别核心,FGCA 则在跨模态融合时自适应提高 DSM 的权重,从而改善分割质量。

由表 6 可知,3 个模块的组合不仅带来逐步递进的指标提升,而且各自提升具有可解释的来源:MMB 扩大可表达信息空间,MMB-DSM 增强结构细节刻画能力,FGCA 提升跨模态融合决策质量。最终模型在 3 项模块共同作用下达到最优性能,验证了所提网络结构的合理性与必要性。

本文模型的 mAP_{box} 在 MMB-DSM 和 FGCA 模块的共同作用下提升到了 42.63%,且 mAP_{mask} 提升到了 42.69%。模型的 mAP_{box} 和 mAP_{mask} 在高 IoU 阈值的限制之下,精度较基准模型均有显著提升。实验表明,这些模块的逐步整合完善,显著地提升了模型精度指标,并最终达到了最佳理想结果。

4　结　论

本文提出了一种改进后的多模态特征融合实例分割网络 MFFISNet,其目的是在低空场景下的工程勘探作业里对于目标展开精确识别以及分割。近年来,在目标检测还有语义分割领域取得了一些比较明显的进展。大部分当前已有的主流模型,在对目标边界进行准确描绘以及对同类实例加以区分时依旧存在着欠缺,高精度要求下的工程场景应用与推广存在限制。这些研究展现出了多模态特征融合网络在当前遥感图像分析、目标提取、语义分割和显著性检测等领域的巨大潜力。通过合理设计

有效融合模块、引入多级特征融合及注意力机制等方式，可以利用不同模态数据的互补现有方法显著提升模型性能。本文面向低空遥感实例分割任务提出一种基于多模态融合的实例分割网络 MFFISNet。该网络引入 RGB 与 DSM 的双路径输入、HWF-LM 与 DMSCA 特征增强模块及 FGCA 跨模态融合机制。实验表明，模型在 3 个数据集上显著优于现有方法，尤其在小目标、复杂纹理和恶劣天气场景中提升明显。MFFISNet 同时具备高实时性与较低计算开销，适合在线巡检等工程应用。未来将进一步优化多模态融合策略并引入时序建模用于持续监测任务。MFFISNet 以多模态特征融合策略与精细实例分割能力为优势，可承担基于低空场景下地物勘探工程中的目标识别和测量任务。此模型在智慧城市基础设施监管、自然灾害监测与响应、生态普查等更宽广场景下仍具备着应用前景。

参考文献

[1] HAFIZ A M, BHAT G M. A survey on instance segmentation: State of the art[J]. International Journal of Multimedia Information Retrieval, 2020, 9(3): 171-189.

[2] GU W, BAI S, KONG L. A review on 2D instance segmentation based on deep neural networks[J]. Image and Vision Computing, 2022, 120: 104401.

[3] FU C, TANG X, YANG Y, et al. A survey of research progresses on instance segmentation based on deep learning[C]. International Conference on Big Data and Security. Singapore: Springer Nature Singapore, 2023: 138-151.

[4] ZHANG L. Automatic digital surface model (DSM) generation from linear array images [J]. ETH Zurich, 2005.

[5] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]. International Conference on Medical Image Computing and Computer-assisted Intervention. Cham: Springer International Publishing, 2015: 234-241.

[6] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [J]. arXiv preprint arXiv: 1802.02611, 2018.

[7] 李明, 魏利胜. 基于改进 Deeplabv3+的磁瓦表面缺陷分割[J]. 电子测量与仪器学报, 2025, 39(1): 50-56.
LI M, WEI L SH. Magnetic tile surface defect segmentation based on improved Deeplabv3+[J]. Journal of Electronic Measurement and Instrumentation, 2025, 39(1): 50-56.

[8] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2961-2969.

[9] BOLYA D, ZHOU C, XIAO F, et al. Yolact: Real-time instance segmentation [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9157-9166.

[10] JIANG P, ERGU D, LIU F, et al. A review of YOLO algorithm developments[J]. Procedia Computer Science, 2022, 199: 1066-1073.

[11] TERVEN J, CÓRDOVA-ESPARZA D M, ROMERO-GONZáLEZ J A. A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-nas [J]. Machine Learning and Knowledge Extraction, 2023, 5: 1680-1716.

[12] LEE Y H, KIM Y. Comparison of CNN and YOLO for object detection [J]. Journal of the Semiconductor & Display Technology, 2020, 19: 85-92.

[13] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.

[14] 蔡文彪, 李永锋, 吴怀诚, 等. 基于改进 Faster RCNN 模型的输电线缺陷检测方法[J]. 信息技术, 2023, 47(1): 148-153.
CAI W B, LI Y F, WU H CH, et al. Transmission line defect detection method based on improved Faster RCNN model [J]. Information Technology, 2023, 47(1): 148-153.

[15] KIRILLOV A, WU Y, HE K, et al. Pointrend: Image segmentation as rendering[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 9799-9808.

[16] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context [C]. European Conference on Computer Vision. Cham: Springer International Publishing, 2014: 740-755.

[17] CORDTS M, OMRAN M, RAMOS S, et al. The Cityscapes dataset for semantic urban scene understanding[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 3213-3223.

[18] FANG Y, YANG S, WANG X, et al. Instances as queries[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 6910-6919.

[19] LYU C, ZHANG W, HUANG H, et al. Rtmdet: An empirical study of designing real-time object detectors[J]. arXiv preprint arXiv:2212.07784, 2022.

[20] CHENG B, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal image segmentation [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 1290-1299.

[21] SUN X, TIAN Y, LU W, et al. From single-to multi-modal remote sensing imagery interpretation: A survey and taxonomy[J]. Science China Information Sciences, 2023, 66(4): 140301.

[22] 郭润泽, 孙备, 孙晓永, 等. 无人机弱光条件下多模态融合目标检测方法[J]. 仪器仪表学报, 2025, 46(1): 338-350.
GUO R Z, SUN B, SUN X Y, et al. Multi-modal fusion object detection method for UAV in low-light conditions[J]. Chinese Journal of Scientific Instrument, 2025, 46(1): 338-350.

[23] HOSSEINPOUR H, SAMADZADEGAN F, JAVAN F D. CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 184: 96-115.

[24] HU Q, WANG F, FANG J, et al. Semantic labeling of high-resolution images combining a self-cascaded multimodal fully convolution neural network with fully conditional random field [J]. Remote Sensing, 2024, 16(17): 3300.

[25] ZHANG P, DU P, LIN C, et al. A hybrid attention-aware fusion network (HAFNet) for building extraction from high-resolution imagery and LiDAR data [J]. Remote Sensing, 2020, 12(22): 3764.

[26] WEN H, YAN C, ZHOU X, et al. Dynamic selective network for RGB-D salient object detection [J]. IEEE Transactions on Image Processing, 2021, 30: 9179-9192.

[27] SOHAN M, SAI RAM T, REDDY R, et al. A review on yolov8 and its advancements [C]. Proceedings of the International Conference on Data Intelligence and Cognitive Informatics. Springer, 2024: 529-545.

[28] FARHADI A, REDMON J. Yolov3: An incremental improvement [C]. Computer Vision and Pattern Recognition, 2018, 1804: 1-6.

[29] 邵延华, 张铎, 楚红雨, 等. 基于深度学习的 YOLO 目标检测综述[J]. 电子与信息学报, 2022, 44(10): 3697-3708.
SHAO Y H, ZHANG D, CHU H Y, et al. A survey of YOLO object detection based on deep learning [J]. Journal of Electronics & Information Technology, 2022, 44(10): 3697-3708.

[30] 林哲, 潘慧琳, 陈丹. 融合改进 YOLO 和语义分割的遮挡目标抓取方法[J]. 电子测量与仪器学报, 2024, 38(12): 190-201.
LIN ZH, PAN H L, CHEN D. Occluded object grasping method combining improved YOLO and semantic segmentation[J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(12): 190-201.

[31] FUJIEDA S, TAKAYAMA K, HACHISUKA T. Wavelet convolutional neural networks[J]. arXiv preprint arXiv: 1805.08620, 2018.

[32] LIU P, ZHANG H, LIAN W, et al. Multi-level wavelet convolutional neural networks[J]. IEEE Access, 2019, 7: 74973-74985.

[33] WILLIAMS T, LI R. Wavelet pooling for convolutional neural networks [C]. Proceedings of the International Conference on Learning Representations, 2018.

[34] 肖海林, 孔祥婷, 王玉, 等. 基于改进奇异值分解和哈尔小波变换的图像水印算法[J]. 计算机应用, 2025, 45(3): 896-903.
XIAO H L, KONG X T, WANG Y, et al. Image watermarking algorithm based on improved singular value decomposition and Haar wavelet transform[J]. Journal of Computer Applications, 2025, 45(3): 896-903.

[35] AHMED N, NATARAJAN T, RAO K R. Discrete cosine transform[J]. IEEE Transactions on Computers, 1974, 100(1): 90-93.

[36] 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.

[37] SU H, WEI S, YAN M, et al. Object detection and instance segmentation in remote sensing imagery based on precise Mask R-CNN [C]. IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2019: 1454-1457.

[38] SU H, WEI S, LIU S, et al. HQ-ISNet: High-quality instance segmentation for remote sensing imagery [J]. Remote Sensing, 2020, 12(6): 989.

[39] SUN Y, SU L, LUO Y, et al. IRDCLNet: Instance segmentation of ship images based on interference reduction and dynamic contour learning in foggy scenes[J]. IEEE

Transactions on Circuits and Systems for Video Technology, 2022, 32(9): 6029-6043.

作者简介

张堃,2005 年于江苏大学获得学士学位,2010 年于浙江大学获得硕士学位,2016 年于上海大学获博士学位,现为南通大学教授,主要研究方向为医工结合人工智能、工业机器视觉、智能控制研究。

E-mail: zhangkun_nt@ 163. com

Zhang Kun received his B. Sc. degree from Jiangsu University in 2005, M. Sc. degree from Zhejiang University in 2010, and Ph. D. degree from Shanghai University in 2016. Now he is a professor at Nantong University. His main research interests include the integration of artificial intelligence in medicine and engineering, industrial machine vision, and intelligent control.

华亮(通信作者),2001 年于南通工学院获得学士学位,2007 年和 2014 年于浙江工业大学分别获得硕士学位和博士学位,现为南通大学教授,主要研究方向为机器人及控制、图像处理与模式识别。

E-mail: hualiang@ ntu. edu. cn

Hua Liang (Corresponding author) received the B. Sc. degree from Nantong Instituted Technology in 2001, received M. Sc. and Ph. D. degrees both from Zhejiang University of Technology in 2007 and 2014, respectively. Now he is a professor at Nantong University. His main research interests include robot control, machine vision, pattern recognition.