

UMM-Det: 面向异构多模态遥感影像的一体化目标检测框架

邹旻瑞^① 李宇轩^① 戴一冕^{*①②} 李翔^{①②} 程明明^{①②}

^①(南开大学计算机学院 天津 300350)

^②(南开国际先进研究院 深圳 518045)

摘要: 当前天基遥感目标检测任务面临着如何构建一个统一模型以有效处理合成孔径雷达(SAR)、可见光、红外等多模态异构数据的挑战。针对此, 该文提出一种异构多模态遥感影像一体化目标检测框架UMM-Det, 致力于通过单一架构实现对多源数据的高性能目标检测。该框架采用单一共享架构, 旨在实现对多源遥感数据的高效、统一检测。UMM-Det在基线模型SM3Det的基础上进行3点关键改进: 首先, 以具备动态采样与大感受野建模能力的InternImage替换原有ConvNeXt主干, 旨在提升对多尺度、低对比度目标的特征提取精度; 其次, 针对红外分支设计了基于时序信息的时空视觉提示模块, 通过精细化的帧差增强策略生成高对比度的运动特征, 以此作为先验知识辅助网络区分动态弱小目标; 最后, 针对红外序列中普遍存在的弱小目标正负样本极度不均衡问题, 引入概率性锚框分配策略(PAA)优化检测头, 显著提升了目标采样的精确性与检测性能。在SARDet-50K、DOTA与SatVideoIRSTD 3个公开数据集上的实验表明, UMM-Det在SAR与可见光检测任务中mAP@0.5:0.95分别提升2.40%和1.77%, 并且在红外序列弱小目标检测任务中较基线模型SM3Det检测率提升了2.54%。同时, 该模型在保证精度提升的前提下将参数量减少50%以上, 展现出精度、效率与轻量化的综合优势, 为新一代高性能天基遥感态统一检测框架的构建提供了有效路径。

关键词: 天基多模态统一检测框架; 多模态遥感检测; 红外序列感知; 弱小目标检测

中图分类号: TN953; TP751

文献标识码: A

文章编号: 1009-5896(2025)12-4704-10

DOI: 10.11999/JEIT250933

CSTR: 32379.14.JEIT250933

1 引言

随着对地观测能力的持续跃升, 天基多模态遥感检测模型正成为全球态势感知的核心手段^[1]。卫星是最早、最直接的观测入口^[2], 能够在第一时间获取并处理合成孔径雷达(Synthetic Aperture Radar, SAR)、可见光、红外等多模态数据, 为导弹来袭预警等战略任务提供前置支撑, 同时在应急救援与资源勘探中展现出广泛价值^[3-5]。然而, 现有方法多沿用“单模态—单网络”的分离式范式, 难以实现跨模态信息的高效共享。构建统一架构的天基多模态遥感检测框架, 既是提升检测实时性与准确性的必然选择, 也是推动遥感智能化发展的关键路径^[6]。

近年来, 多模态遥感大模型发展迅速, 利用卫星采集的大量SAR、可见光、红外等多模态数据在多种地球观测应用领域取得了显著进展^[7]。SkySense^[8]利用可见光、多光谱、SAR时序数据以及地理位置等信息, 在多类典型任务上展现出卓越的通用表征与泛化能力。不同于SkySense采用多个主干网, SkySense V2^[9]采用单一Transformer主干结合自适应模块与混合专家(Mixture of Experts, MoE)机制, 在多个任务上全面超越SkySense, 展现更强的泛化能力。为了追求更强的遥感图像解译能力, RingMoE^[10]结合层次化 MoE、物理感知自监督与动态专家剪枝等技术, 显著增强了模型对遥感影像的深层特征提取与语义理解能力, 其拥有147亿个参数, 是目前规模最大的多模态遥感大模型。为了实现任意空间分辨率和多光谱通道的高保真表征, FlexiMo^[11]通过分辨率感知与通道自适应模块实现了多任务下的高泛化性与鲁棒性。

此外, 还有不少工作关注视觉语言遥感模型^[12]以及针对时序数据^[13]或分割任务^[14]的研究, 但专门面向多模态遥感目标检测任务的模型仍然稀缺。SM3Det^[15]是一个专注于多模态遥感目标一体化检测的架构, 但其主要聚焦于SAR、可见光和红外单帧图像检测任务, 对于红外序列信息的利用较为有限。而在红外目标检测任务中, 单帧图像往往信噪

收稿日期: 2025-09-19; 改回日期: 2026-01-04; 网络出版: 2026-01-10

*通信作者: 戴一冕 yimian.dai@gmail.com

基金项目: 国家杰出青年科学基金(62225604), 国家自然科学基金(62301261, 62206134), 深圳市自然科学基金(JCYJ20240813114237048), 天津市自然科学基金(25JCQNJC01370), 南开大学超算中心

Foundation Items: The National Science Fund for Distinguished Young Scholar (62225604), The National Natural Science Foundation of China (62301261, 62206134), The General Program of Shenzhen Natural Science Foundation (JCYJ20240813114237048), The Natural Science Foundation of Tianjin(25JCQNJC01370), The Supercomputing Center of Nankai University (NKSC)

比较低,小目标易被背景干扰,导致检测性能受限。相比之下,红外序列能够提供丰富的时序上下文信息,有助于动态目标的连续跟踪与特征增强。因此,如何在天基多模态遥感检测模型中引入红外序列建模机制,充分利用时序特征提升小目标检测能力,是亟须解决的关键问题。

基于上述分析,本文工作在 SM3Det 架构基础上,提出了适合红外序列的异构多模态遥感影像一体化目标检测框架 UMM-Det(Unified Multi-Modal Detector),其在红外分支上引入了一种新颖的时空视觉提示机制,旨在将低信噪比的单帧检测任务转化为对高对比度运动特征的识别任务,同时保持可见光与 SAR 分支的原有优势,从而提升多模态联合检测能力。本文的具体贡献如下:

(1)在 SM3Det 结构中以 InternImage^[16] 替换 ConvNeXt^[17] 主干,利用其核心算子可变形卷积v3(Deformable Convolution v3, DCNv3)的动态采样机制提升对低对比度、多尺度遥感目标的特征建模能力,在微小目标检测与复杂背景抑制方面表现突出。

(2)设计了新颖的时空视觉提示模块,通过精细化的帧差增强策略生成一个强有力的运动特征提示,以此作为注意力先验引导主干网络高效定位与识别动态弱小目标。

(3)针对红外序列中普遍存在的弱小目标正负样本极度不均衡问题,引入了基于概率性锚框分配(Probabilistic Anchor Assignment, PAA)^[18]策略的检测头,通过动态、自适应的样本划分机制,显著提升了模型的收敛稳定性和对难检目标的召回能力。

本文的组织结构如下,第2节方法部分介绍 UMM-Det 框架的实现细节,第3节实验与结果分析介绍实验数据集、实验设置、消融实验、对比实验以及可视化分析等内容,第4节为结论。

2 方法

UMM-Det框架以SAR单帧图像、可见光单帧图像及红外图像序列为输入,旨在构建一个统一且高效的多模态遥感目标检测模型,其整体结构如图1所示。UMM-Det首先利用InternImage主干网络对所有输入模态的图像进行统一的深度特征提取。借助DCNv3算子的动态感受野机制,网络能够自适应地提取各模态的专属表征特征,包括可见光的丰富纹理细节、SAR的独特散射模式以及红外的热辐射特性,从而强化对目标上下文信息的建模。在推理阶段,模型采用任务驱动的动态路由机制。根据输入数据的模态,自动激活对应检测头(SAR、可见光或红外),实现单一权重文件对多种任务的灵活响应。

针对红外序列数据,本文设计了专属的时空视觉提示模块,通过一种精细化的帧差增强策略,生成替代原始红外序列输入的运动特征提示,以显著增强动态小目标的表征。最终,所有模态的高级语义表征被送入一个使用PAA策略增强的检测头,该检测头能够根据模型自身的学习状态自适应地进行正负样本采样,并结合交并比(Intersection over Union, IoU)预测分支,完成对目标的精准分类与边界框回归。

具体来讲,给定3种模态的输入集合

$$\mathcal{I} = \{\mathcal{I}_{\text{RGB}}, \mathcal{I}_{\text{SAR}}, \mathcal{I}_{\text{IR}}\}, \mathcal{I}_{\text{IR}} = \{\mathcal{I}_t \mid t = 1, 2, \dots, T\} \quad (1)$$

其中, $\mathcal{I}_{\text{RGB}}, \mathcal{I}_{\text{SAR}} \in \mathbb{R}^{3 \times H \times W}$ 分别表示可见光图像和SAR图像, \mathcal{I}_{IR} 为红外序列, $\mathcal{I}_t \in \mathbb{R}^{3 \times H \times W}$ 是 T 张连续的红外序列中的第 t 帧图像。为了充分挖掘红外序列所蕴含的时序上下文信息,本文先引入时空视觉提示模块对其时序线索进行专项增强,通过

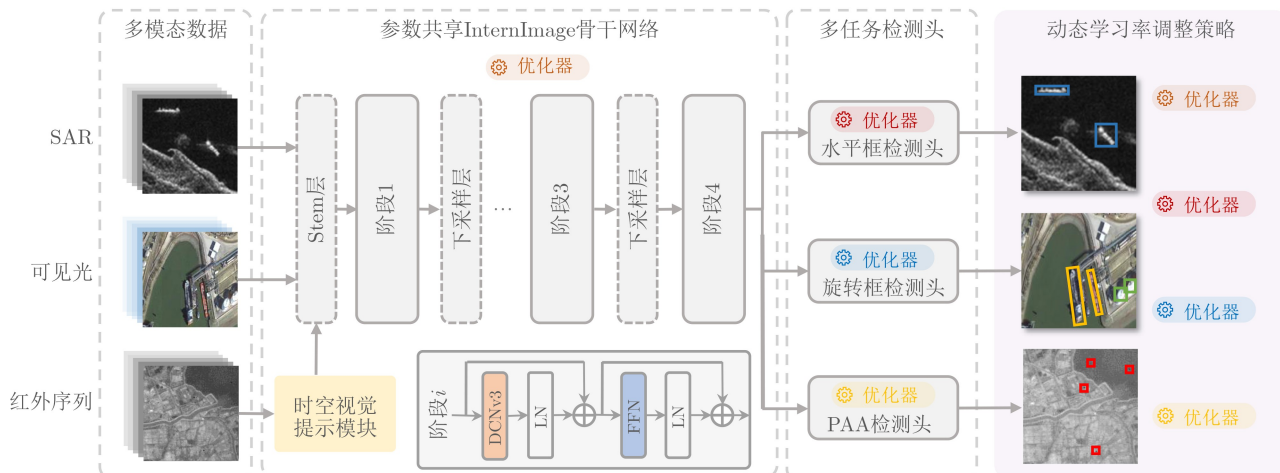


图1 UMM-Det 网络结构图(图中展示多个优化器节点代表不同检测任务的梯度回传路径)

相邻3帧的运动特征提取, 以实现动态弱小目标的连续跟踪与表征强化

$$\mathcal{I}_{\text{IRP}}^t = \Phi_{\text{prompt}}(\mathcal{I}_t, \mathcal{I}_{t-1}, \mathcal{I}_{t-2}), t \geq 3 \quad (2)$$

其中, $\Phi_{\text{prompt}}(\cdot)$ 是时空视觉提示模块, $\mathcal{I}_{\text{IRP}}^t$ 是增强后的红外序列提示, 其替换原始红外序列 \mathcal{I}_{IR} 。之后, 3个模态的数据 $\{\mathcal{I}_{\text{RGB}}, \mathcal{I}_{\text{SAR}}, \mathcal{I}_{\text{IRP}}\}$ 被输入到InternImage主干网络, 使各模态输入被映射为多尺度特征表示

$$\mathcal{F}_{\text{modal}} = \text{InternImage}(\mathcal{I}_{\text{modal}}), \text{modal} \in \{\text{RGB}, \text{SAR}, \text{IRP}\} \quad (3)$$

最后, 各级检测头基于各模态的高级语义特征进行目标检测, 即红外分支使用增强后特征 $\tilde{\mathcal{F}}_{\text{IR}}^t$, 可见光和SAR分支使用主干网络输出的特征 \mathcal{F}_{RGB} 和 \mathcal{F}_{SAR} , 最终检测结果记为(以红外时序分支为例, 其他分支与其一致, 但没有表示时序的上标 $(\cdot)^t$)

$$\mathbf{y}^t = \Psi(\tilde{\mathcal{F}}_{\text{IR}}^t), \quad (4)$$

其中, $\Psi(\cdot)$ 表示不同模态下的各级检测头, \mathbf{y}^t 包含类别预测集合、边界框预测集合与置信度分数预测集合。

2.1 时空视觉提示模块

为充分挖掘红外序列中蕴含的时序动态线索, 本文设计了时空视觉提示模块对其时序信息进行增强建模, 其示意图可见图2时空视觉提示模块示意图。该模块通过精细的帧差增强策略, 生成一个强有力的运动特征提示, 引导主干网络精准定位动态弱小目标。

模型从输入的视频序列(如40帧片段)中, 以滑动窗口的方式为当前待检测帧 \mathcal{I}_t 动态关联其前序帧 \mathcal{I}_{t-1} 和 \mathcal{I}_{t-2} , 构建局部时空上下文 $\{\mathcal{I}_{t-2}, \mathcal{I}_{t-1}, \mathcal{I}_t\}$ 。此外, 为了聚焦于整体亮度变化并简化计算, 本文首先对红外序列中的每一帧 $\mathcal{I}_k \in \mathbb{R}^{3 \times H \times W}$ 进行通道维度的均值池化操作(原始红外图像以RGB格式加载), 将其压缩为单通道灰度图 $\mathcal{I}_k' \in \mathbb{R}^{1 \times H \times W}$, 得到 $\{\mathcal{I}_{t-2}', \mathcal{I}_{t-1}', \mathcal{I}_t'\}$ 。如图2所示, 在获得单通道表示后, 通过多尺度帧差法来捕捉不同速度目标的运动信息, 即计算当前帧与前两帧的差分图 $D_{t,1}$ 和 $D_{t,2}$

$$D_{t,1} = |\mathcal{I}_t' - \mathcal{I}_{t-1}'|, D_{t,2} = |\mathcal{I}_t' - \mathcal{I}_{t-2}'| \quad (5)$$

随后, 当前帧的静态信息 \mathcal{I}_t' 与这两种尺度的动态信息 $D_{t,1}$ 和 $D_{t,2}$ 在通道维度上进行拼接, 从而构建出一个信息高度浓缩的三通道提示张量 $\mathcal{P}_t \in \mathbb{R}^{3 \times H \times W}$ 。通过将当前帧 \mathcal{I}_t 纳入拼接, 该模块确保了原始红外图像的热辐射强度信息得以完整保留, 从而避免了因仅关注运动特征而导致静止目标漏检的问题。最后, 该张量被送入一个轻量级的

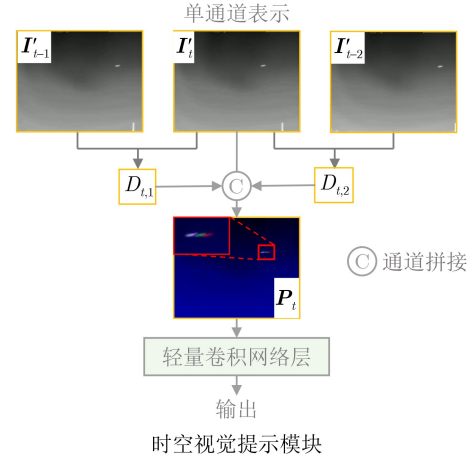


图2 时空视觉提示模块示意图

2D卷积网络层进行特征编码, 以高效融合静态空间特征与动态运动线索, 其计算过程可表达为

$$\begin{aligned} \mathcal{I}_{\text{IR-P}}^{c_{\text{out}}, i, j} &= \text{bias}(c_{\text{out}}) \\ &+ \sum_{c_{\text{in}}=1}^3 \sum_{l=1}^{K_h} \sum_{m=1}^{K_w} \mathcal{P}_t(c_{\text{in}}, i+l, j+m) \\ &\cdot W(c_{\text{out}}, c_{\text{in}}, l, m) \end{aligned} \quad (6)$$

其中, W 是卷积核权重, $c_{\text{in}}, c_{\text{out}}$ 分别表示输入通道数与输出通道数, K_h, K_w 则分别表示卷积核的长宽大小, bias 是卷积核偏置。 $\mathcal{I}_{\text{IRP}} \in \mathbb{R}^{C_{\text{out}} \times H' \times W'}$ 即为最终的时空视觉提示特征, 将替代原始红外数据 \mathcal{I}_{IR} , 与SAR图像和可见光图像数据一同输入主干网络进行特征抽取。

时空视觉提示模块生成的运动线索提示, 成功将检测任务从寻找低对比度目标转变为识别高对比度的运动特征, 显著提升了目标辨识度。该模块作为一个强大的注意力先验, 能够提示并引导骨干网络将特征提取能力聚焦于高概率的动态区域, 从而有效提升了检测率。

2.2 InternImage 主干网络

基于多模态遥感图像(SAR、可见光、红外)各自独特的成像机理, 针对红外图像低对比度、纹理细节缺失及噪声干扰严重等特点, 本文选用InternImage-Small作为特征提取骨干网络。其核心算子DCN v3突破了传统卷积固定采样模式的局限, 实现了对低信噪比红外图像中目标信息的高效捕获与表达, 为后续模块提供了高质量的特征基础。

DCNv3通过引入可学习的采样偏移和调制权重, 赋予了标准卷积动态调整其感受野形状和大小的能力。其核心运算过程可由以下公式精确描述

$$y(p_0) = \sum_{g=1}^G \sum_{k=1}^K w_g \cdot m_{gk} \cdot x_g(p_0 + p_k + \Delta p_{gk}) \quad (7)$$

其中, p_0, p_k 和 Δp_{gk} 分别表示分组的输入特征 x_g 的基准采样点、标准卷积核偏移位置以及学习到的可变偏移量, w_g 是组内卷积核权重, m_{gk} 是可学习的调制权重, 对每个采样点进行加权。 G 是分组数量, K 是每组卷积核的采样点数量, $y(p_0)$ 代表输出特征图上位置 p_0 的特征。

通过引入动态采样机制, 可变形卷积操作能够根据输入内容为每个采样点预测一个偏移量 Δp_{gk} , 从而打破了固定几何结构的限制, 实现了对长距离依赖关系的有效建模。此外, 通过一个自适应的空间聚合模块, 利用Softmax归一化的调制标量 m_{gk} 对采样特征进行加权, 使模型能够聚焦于信息丰富的区域。最后, 基于分组与权重共享策略, 将运算划分为 G 组以学习多样化的特征模式, 并在组内共享投影权重 w_g , 极大地提升了模型的参数效率。

2.3 概率性锚框分配检测头

为解决遥感图像目标检测模型在训练过程中正负样本分配不均衡的问题, 本文为红外序列分支采用了PAA策略构建检测头, 以实现目标的精准采样与检测。

2.3.1 概率性锚框分配原理

与传统的基于固定IoU阈值的硬性分配策略不同, PAA将锚框分配问题转化为一个概率建模问题。在该框架下, 锚框是否被划分为正样本, 不再仅由其与真实框的几何重叠度决定, 而是取决于一个能够综合反映模型当前分类与定位能力的动态分数。首先给每个锚框定义一个综合分数 S , 其指数形式被定义为分类损失 \mathcal{L}_{cls} 与定位损失 \mathcal{L}_{loc} 的加权和指数幂

$$S = e^{-(\mathcal{L}_{cls} + \mathcal{L}_{loc})}, \quad (8)$$

之后, PAA假设所有候选锚框的分数来自一个由正、负样本两个部分组成的高斯混合模型。其概率密度函数可表示为

$$P(s) = w_{pos} \mathcal{N}(s; \mu_{pos}, \sigma_{pos}) + w_{neg} \mathcal{N}(s; \mu_{neg}, \sigma_{neg}) \quad (9)$$

其中, w, μ, σ 分别代表各组分高斯分布的权重、均值和标准差。之后, PAA算法通过期望最大化算法估计模型参数后, 计算出每个锚框属于正样本分布的概率, 并据此进行动态、自适应的样本划分。

2.3.2 损失函数

UMM-Det红外序列分支的损失函数 \mathcal{L} 定义为在所有由PAA确定的正样本锚框集合 \mathcal{P} 和所有锚框集合 \mathcal{A} 上的损失之和

$$\mathcal{L} = \frac{1}{N_{pos}} \sum_{i \in \mathcal{P}} (\mathcal{L}_{cls}(p_i, g_i) + \lambda_1 \mathcal{L}_{reg}(b_i, g_i) + \lambda_2 \mathcal{L}_{IoUP}(iou_i, \widehat{iou}_i)) + \frac{1}{N_{all}} \sum_{j \in \mathcal{A}} \mathcal{L}_{cls}(p_j, g_j) \quad (10)$$

其中, p 是预测分类, g 是真实标注信息, b 是预测标注框; N_{pos} 是正样本数量, \mathcal{L}_{cls} 指分类分支的Focal Loss, \mathcal{L}_{reg} 是回归分支的GIoU Loss。 \mathcal{L}_{IoUP} 则是用于监督PAA额外引入的IoU预测分支的BCE Loss, 仅对正样本计算, 其定义为

$$\mathcal{L}_{IoUP}(iou_p, iou_g) = -[iou_g \ln(iou_p) + (1 - iou_g) \cdot \ln(1 - iou_p)] \quad (11)$$

其中, iou_p 是预测的IoU值, iou_g 是真实的IoU值。

3 实验与结果分析

为全面评估本文提出的UMM-Det框架在红外序列目标检测任务中的有效性与先进性, 本文分别在SAR、可见光、红外序列等3种数据集上进行了一系列详尽的实验, 包括关键模块的消融实验、与当前主流方法的性能对比分析以及定性可视化分析。

3.1 数据集

本文的实验评估基于3个公开基准数据集。针对SAR图像的目标检测任务, 采用了SARDet-50K^[15]数据集, 其作为SARDet-100K^[19]的子集, 涵盖了以水平边界框标注的飞机、船舶、汽车、桥梁、坦克与港口6个目标类别。在可见光遥感影像检测任务上, 则采用了大规模基准DOTA^[20], 该数据集包含15个类别的旋转边界框标注实例。其原始高分辨率图像经过800×800像素的滑窗切割(步长为400像素)处理, 生成了包含25 028张图像(337 728个实例)的训练集与17 041张图像(95 380个实例)的测试集。针对红外弱小目标序列检测任务, 本文选用了极具挑战性的SatVideoIRSTD^[21]数据集, 它整合了IRAir热红外卫星数据集^[22]、IRSatVideo-LEO半仿真数据集^[23]以及源自“武汉一号”卫星的真实在轨观测视频, 共计1 400段序列, 对算法的泛化能力与跨域适应性构成了严峻考验。对于SatVideoIRSTD数据集, 本文在训练与测试阶段统一从各视频序列中截取连续40帧作为样本输入。

3.2 实验环境及参数设置

所有实验均在配备4张NVIDIA RTX 4090 GPU的服务器上基于MMDetection框架实现, 为保证公平对比, 在主要结果和消融研究中, 除非有特别说明, 所有模型均采用AdamW优化器对模型进行训练, 其基础学习率设置为 1×10^{-4} , 权重衰减系数为0.05, betas参数设置为(0.9, 0.999)。为实现多模态数据的平衡训练, 本文采用混合批次采样策略。在每个训练迭代中, 同时从SAR、可见光和红外序列数据集中抽取样本组成一个批次。为简化训练过程, 本文对主干网络、颈部和各个检测头等所有模

块均采用统一的学习率,学习率更新策略仍采用SM3Det使用的动态学习率调整策略^[15]。

3.3 评价指标

在SAR和可见光图像目标检测任务上,本文采用业界公认的核心指标——平均精度均值(mean Average Precision, mAP)。该指标能够综合评估模型在不同置信度与IoU阈值下的整体性能,全面地反映其在精确率与召回率之间的权衡能力。实验中,本文主要报告了两种条件下的mAP值(即IoU阈值在0.5~0.95的平均值mAP@0.5:0.95与IoU阈值为0.5时的mAP@0.5),以进行详尽的性能比较。

针对红外小目标检测任务的特殊性,本文额外引入了检测率(Probability of detection, Pd)与虚警率(False alarm rate, Fa)。检测率指标旨在直接量化模型对真实目标的发现与捕获能力,以评估检测的完备性。而采用像素级计算的虚警率,则用于精细衡量模型对复杂背景噪声的抑制性能与控制误报的水平。在算法的实际运用中检测率和虚警率是极为关键的考量因素。

3.4 消融实验

为了清晰地验证本文所提各个核心组件的有效性,本文在SatVideoIRSTD数据集上进行了一系列模块消融实验,结果如表1不同模块的消融实验结果所示。

以基线模型为起点,本文逐步集成所提出的优化模块。首先,仅将骨干网络替换为InternImage,模型的Pd即提升了0.80个百分点,初步证明了其更强的特征提取能力。在此基础上,引入PAA策略的检测头使Pd进一步提升了0.51个百分点,验证了其在动态标签分配上的优势。最为关键的是,在最终集成本文核心的时空视觉提示模块后,模型性能实现了1.23个百分点的最大增幅,达到了79.67%的最佳Pd。值得注意的是,该时序模块在显著提升检测概率的同时,也使得Fa相较于前一配置有所降低,这有力地表明,利用时空上下文信息不仅能增强对真实目标的捕获,还能有效抑制瞬时背景噪声,是模型整体性能实现飞跃的核心。综上,这一系列递进的实验结果逐层验证了各

模块的贡献,并凸显了时空建模对于红外序列检测的决定性作用。

3.5 对比实验分析

为了进行全面的性能评估,本文不仅与核心基线SM3Det进行对比,还进一步将其与多种主流的骨干网络(如VAN^[24], LSKNet^[25]和PVT-v2^[26])和检测头(如RetinaNet^[27], Faster R-CNN^[28], Cascade R-CNN^[29], GFL^[30], RoI Transformer^[31]和S²ANet^[32])进行组合测试,以构建一系列强大的基线模型。所有实验结果如表2不同基线网络在3个模态数据集上的实验结果所示。

实验结果表明,UMM-Det在3个检测任务的性能指标上全面超越其他基线网络,这充分表明本文所提的UMM-Det模型在多模态遥感检测任务中具有显著优势。具体来讲,在可见光和SAR目标检测任务上,本文模型在SAR和可见光检测任务上均实现了全面的性能超越,其mAP@0.5:0.95指标相较基线模型SM3Det分别提升了2.4%和1.77%。尤为关键的是,这一替换在所有模态上均实现了超过50%的参数量削减,证明了新主干网络在提升不同遥感数据下的特征提取能力与实现模型轻量化部署潜力之间取得了出色的平衡。此外,针对红外序列弱小目标检测这一核心挑战,本文设计的时空视觉提示模块发挥了关键作用。实验证明,UMM-Det能够有效利用帧间上下文信息,使得检测率提升了2.54%,显著增强了在低信噪比场景下的目标捕获能力。尽管虚警率和计算量有所上升,如表1不同模块的消融实验结果数据所示,InternImage主干网络与PAA策略的引入虽然使虚警率分别上升至 5.34×10^{-4} 和 8.24×10^{-4} ,但这实质上反映了检测灵敏度提升所伴随的必然代价。鉴于红外弱小目标检测的首要任务是最大限度降低漏检率,这种性能权衡对于增强系统的整体态势感知能力是必要且具有重要价值的。与此同时,模型参数量的显著减少赋予了其在实际部署中的巨大优势,充分验证了UMM-Det兼具高有效性与轻量化的双重点。

3.6 结果可视化分析

为了更直观地展示UMM-Det模型的性能优势,本文选取检测性能出色的方法以及具有代表性的场景进行检测结果的可视化对比。

在SatVideoIRSTD红外序列数据集中,目标通常表现为信噪比低、尺寸微小的亮点,且易受复杂云层或地面背景干扰,如图3所示(黄色框表示虚警,绿色检测框表示检测结果,红色框为局部放大区域)。可以看出,基线模型SM3Det在部分场景中出现了漏检或将背景杂波误检为目标的情况。相比

表1 不同模块的消融实验结果

InternImage骨干	PAA检测头	时空视觉提示模块	Pd(%)	Fa
			77.13	3.24×10^{-4}
✓			77.93	5.34×10^{-4}
✓	✓		78.44	8.24×10^{-4}
✓	✓	✓	79.67	6.61×10^{-4}

表 2 不同基线网络在3个模态数据集上的实验结果

	SARDet-50K		DOTA		SatVideoIRSTD		计算量	参数量
	mAP@0.5:0.95	mAP@0.5	mAP@0.5:0.95	mAP@0.5	Pd	Fa		
RetinaNet ^[27]	53.04	83.99	-	-	66.55	1.19×10^{-4}	520.74G	206.69M
Faster RCNN ^[28]	54.56	85.62	-	-	45.05	7.99×10^{-5}	435.69G	173.55M
Cascade RCNN ^[29]	56.30	85.39	-	-	58.06	8.44×10^{-5}	463.44G	201.30M
GFL ^[30]	59.01	88.77	-	-	72.51	3.51×10^{-4}	733.85G	274.95M
RoI Transformer ^[31]	-	-	45.43	76.79	-	-	520.74G	206.69M
S ² ANet ^[32]	-	-	39.92	76.20	-	-	463.44G	201.30M
VAN-T ^[24]	49.28	80.85	43.60	74.73	70.52	3.05×10^{-4}	270.47G	45.32M
VAN-S ^[24]	57.98	88.36	45.50	76.66	74.84	4.94×10^{-4}	366.56G	64.87M
LSKNet-T ^[25]	49.95	81.76	43.56	75.44	70.81	3.51×10^{-4}	269.38G	45.03M
LSKNet-S ^[25]	58.41	88.48	44.80	76.69	74.51	6.13×10^{-4}	369.67G	65.37M
PVT-v2-T ^[26]	48.58	80.71	42.72	75.39	71.94	3.88×10^{-4}	236.92G	40.20M
PVT-v2-S ^[26]	54.53	85.48	44.37	77.53	75.19	6.83×10^{-4}	293.87G	51.45M
SM3Det ^[15]	60.64	89.94	46.47	77.88	77.13	3.24×10^{-4}	741.29G	164.29M
UMM-Det(本文方法)	63.04	91.55	48.24	80.91	79.67	6.61×10^{-4}	977.31G	76.64M

注: 粗体表示最优值。

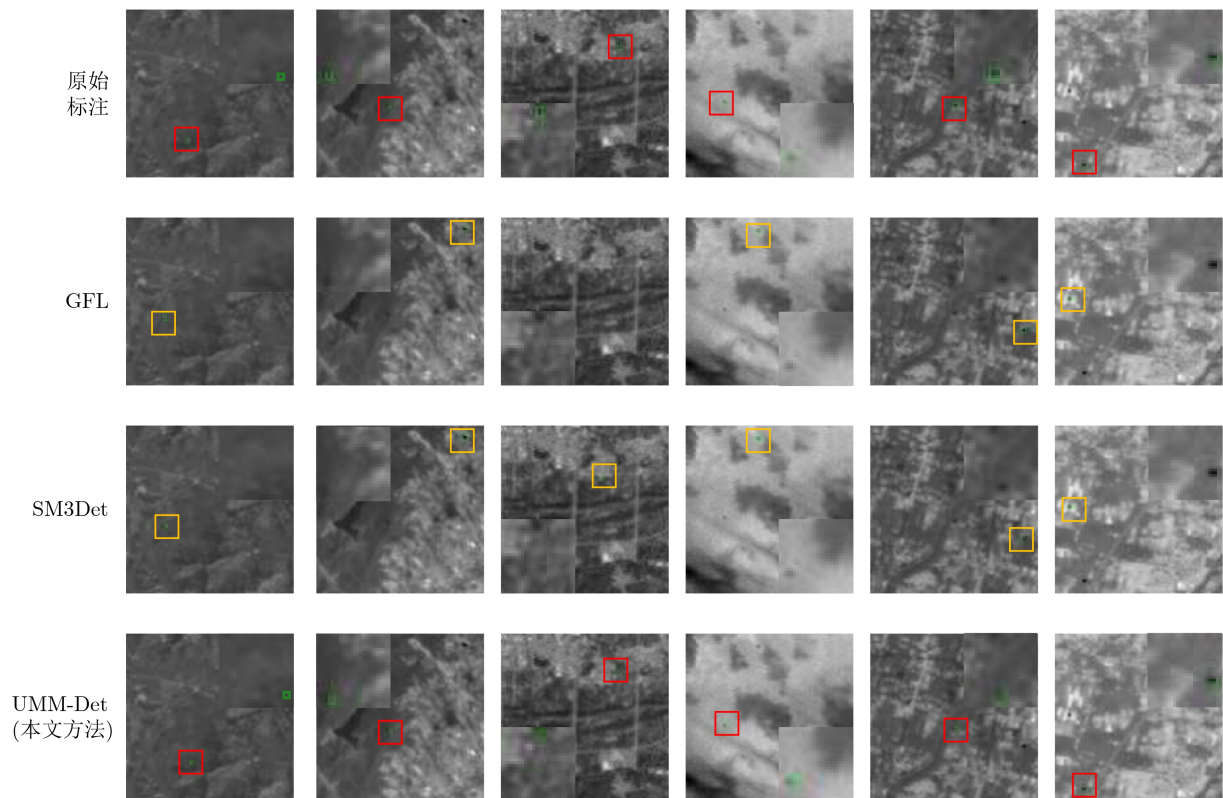


图 3 在红外序列小目标检测SatVideoIRSTD数据集上的可视化

之下, 本文提出的UMM-Det凭借其高效的时序信息利用能力, 能够在各种复杂背景下稳定、准确地检测出真实目标, 其检测结果与原始标注高度一致。这直观地证明了时序建模对于提升红外弱小目标检测鲁棒性的关键作用。

4 结论

针对当前天基多模态遥感检测模型仅支持单帧图像而忽视红外序列时序价值的局限性, 本文提出了适配红外序列的UMM-Det模型。此模型通过以动态特征提取能力更强的InternImage替换主干网

络,并为红外分支精心设计时空视觉提示模块,利用帧差增强策略生成运动特征以引导动态弱小目标聚焦,同时引入概率性锚框分配策略优化检测头,有效缓解了弱小目标训练中的正负样本失衡问题。实验表明,相较于基线SM3Det, UMM-Det在SatVideoIRSTD数据集上将红外序列弱小目标的检测率提升了2.54%,并在SAR与可见光任务上分别取得了2.40%和1.77%的mAP@0.5:0.95增益,同时将模型参数量压缩50%以上,实现了轻量化与高精度的统一。本研究验证了红外序列时序线索对弱小目标检测的不可替代性,为构建高精度、轻量化的天基遥感态势感知系统提供了可工程化落地的技术路径。

参考文献

- [1] 安成锦, 杨俊刚, 梁政宇, 等. 阵列相机图像邻近目标超分辨率方法[J]. 电子与信息学报, 2023, 45(11): 4050–4059. doi: [10.11999/JEIT230810](https://doi.org/10.11999/JEIT230810).
AN Chengjin, YANG Jungang, LIANG Zhengyu, *et al.* Closely spaced objects super-resolution method using array camera images[J]. *Journal of Electronics & Information Technology*, 2023, 45(11): 4050–4059. doi: [10.11999/JEIT230810](https://doi.org/10.11999/JEIT230810).
- [2] 杨俊刚, 刘婷, 刘永贤, 等. 基于非凸低秩塔克分解的红外弱小目标检测方法[J]. 红外与毫米波学报, 2025, 44(2): 311–325. doi: [10.11972/j.issn.1001-9014.2025.02.018](https://doi.org/10.11972/j.issn.1001-9014.2025.02.018).
YANG Jungang, LIU Ting, LIU Yongxian, *et al.* Infrared small target detection method based on nonconvex low-rank Tucker decomposition[J]. *Journal of Infrared and Millimeter Waves*, 2025, 44(2): 311–325. doi: [10.11972/j.issn.1001-9014.2025.02.018](https://doi.org/10.11972/j.issn.1001-9014.2025.02.018).
- [3] 林再平, 罗伊杭, 李博扬, 等. 基于梯度可感知通道注意力模块的红外弱小目标检测前去除噪网络[J]. 红外与毫米波学报, 2024, 43(2): 254–260. doi: [10.11972/j.issn.1001-9014.2024.02.015](https://doi.org/10.11972/j.issn.1001-9014.2024.02.015).
LIN Zaiping, LUO Yihang, LI Boyang, *et al.* Gradient-aware channel attention network for infrared small target image denoising before detection[J]. *Journal of Infrared and Millimeter Waves*, 2024, 43(2): 254–260. doi: [10.11972/j.issn.1001-9014.2024.02.015](https://doi.org/10.11972/j.issn.1001-9014.2024.02.015).
- [4] SHI Qian, HE Da, LIU Zhengyu, *et al.* Globe230k: A benchmark dense-pixel annotation dataset for global land cover mapping[J]. *Journal of Remote Sensing*, 2023, 3: 0078. doi: [10.34133/remotesensing.0078](https://doi.org/10.34133/remotesensing.0078).
- [5] TIAN Jiaqi, ZHU Xiaolin, SHEN Miaogen, *et al.* Effectiveness of spatiotemporal data fusion in fine-scale land surface phenology monitoring: A simulation study[J]. *Journal of Remote Sensing*, 2024, 4: 0118. doi: [10.34133/remotesensing.0118](https://doi.org/10.34133/remotesensing.0118).
- [6] LIU Shuaijun, LIU Jia, TAN Xiaoyue, *et al.* A hybrid spatiotemporal fusion method for high spatial resolution imagery: Fusion of gaofen-1 and sentinel-2 over agricultural landscapes[J]. *Journal of Remote Sensing*, 2024, 4: 0159. doi: [10.34133/remotesensing.0159](https://doi.org/10.34133/remotesensing.0159).
- [7] MEI Shaohui, LIAN Jiawei, WANG Xiaofei, *et al.* A comprehensive study on the robustness of deep learning-based image classification and object detection in remote sensing: Surveying and benchmarking[J]. *Journal of Remote Sensing*, 2024, 4: 0219. doi: [10.34133/remotesensing.0219](https://doi.org/10.34133/remotesensing.0219).
- [8] GUO Xin, LAO Jiangwei, DANG Bo, *et al.* SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2024: 27662–27673. doi: [10.1109/CVPR52733.2024.02613](https://doi.org/10.1109/CVPR52733.2024.02613).
- [9] ZHANG Yingying, RU Lixiang, Wu Kang, *et al.* SkySense V2: A unified foundation model for multi-modal remote sensing[C]. International Conference on Computer Vision, Honolulu, Hawaii, 2025: 9136–9146.
- [10] BI Hanbo, FENG Yingchao, TONG Boyuan, *et al.* RingMoE: Mixture-of-modality-experts multi-modal foundation models for universal remote sensing image interpretation[J]. arXiv: 2504.03166, 2025. doi: [10.48550/arXiv.2504.03166](https://doi.org/10.48550/arXiv.2504.03166).
- [11] LI Xuyang, LI Chenyu, GHAMISI P, *et al.* FlexiMo: A flexible remote sensing foundation model[J]. arXiv: 2503.23844, 2025. doi: [10.48550/arXiv.2503.23844](https://doi.org/10.48550/arXiv.2503.23844).
- [12] YAO Kelu, XU Nuo, YANG Rong, *et al.* Falcon: A remote sensing vision-language foundation model (Technical Report)[J]. arXiv: 2503.11070, 2025. doi: [10.48550/arXiv.2503.11070](https://doi.org/10.48550/arXiv.2503.11070).
- [13] QIN Xiaolei, WANG Di, ZHANG Jing, *et al.* TiMo: Spatiotemporal foundation model for satellite image time series[J]. arXiv: 2505.08723, 2025. doi: [10.48550/arXiv.2505.08723](https://doi.org/10.48550/arXiv.2505.08723).
- [14] YAO Liang, LIU Fan, CHEN Delong, *et al.* RemoteSAM: Towards segment anything for earth observation[C]. The 33rd ACM International Conference on Multimedia, Dublin, Ireland, 2025: 3027–3036. doi: [10.1145/3746027.3754950](https://doi.org/10.1145/3746027.3754950).
- [15] LI Yuxuan, LI Xiang, LI Yunheng, *et al.* SM3Det: A unified model for multi-modal remote sensing object detection[J]. arXiv: 2412.20665, 2024. doi: [10.48550/arXiv.2412.20665](https://doi.org/10.48550/arXiv.2412.20665).
- [16] WANG Wenhai, DAI Jifeng, CHEN Zhe, *et al.* InternImage: Exploring large-scale vision foundation models with deformable convolutions[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, 2023: 14408–14419. doi: [10.1109/CVPR52729.2023.01385](https://doi.org/10.1109/CVPR52729.2023.01385).
- [17] LIU Zhuang, MAO Hanzi, WU Chaoyuan, *et al.* A ConvNet for the 2020s[C]. 2022 IEEE/CVF Conference on Computer

- Vision and Pattern Recognition (CVPR), New Orleans, USA, 2022: 11966–11976. doi: [10.1109/CVPR52688.2022.01167](https://doi.org/10.1109/CVPR52688.2022.01167).
- [18] KIM K and LEE H S. Probabilistic anchor assignment with IoU prediction for object detection[C]. 16th European Conference on Computer Vision – ECCV 2020, Glasgow, UK, 2020: 355–371. doi: [10.1007/978-3-030-58595-2_22](https://doi.org/10.1007/978-3-030-58595-2_22).
- [19] LI Yuxuan, LI Xiang, LI Weijie, *et al.* SARDet-100K: Towards open-source benchmark and toolkit for large-scale SAR object detection[C]. The 38th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2024: 4079.
- [20] XIA Guisong, BAI Xiang, DING Jian, *et al.* DOTA: A large-scale dataset for object detection in aerial images[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 3974–3983. doi: [10.1109/CVPR.2018.00418](https://doi.org/10.1109/CVPR.2018.00418).
- [21] LI Ruoqing, AN Wei, YING Xinyi, *et al.* Probing deep into temporal profile makes the infrared small target detector much better[J]. arXiv: 2506.12766, 2025. doi: [10.48550/arXiv.2506.12766](https://doi.org/10.48550/arXiv.2506.12766).
- [22] 李朝旭, 徐清宇, 安玮, 等. 红外图像暗弱目标轻量级检测网络[J]. 红外与毫米波学报, 2025, 44(2): 299–310. doi: [10.11972/j.issn.1001-9014.2025.02.017](https://doi.org/10.11972/j.issn.1001-9014.2025.02.017).
- LI Zhaoxu, XU Qingyu, AN Wei, *et al.* A lightweight dark object detection network for infrared images[J]. *Journal of Infrared and Millimeter Waves*, 2025, 44(2): 299–310. doi: [10.11972/j.issn.1001-9014.2025.02.017](https://doi.org/10.11972/j.issn.1001-9014.2025.02.017).
- [23] YING Xinyi, LIU Li, LIN Zaipin, *et al.* Infrared small target detection in satellite videos: A new dataset and a novel recurrent feature refinement framework[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2025, 63: 5002818. doi: [10.1109/TGRS.2025.3542368](https://doi.org/10.1109/TGRS.2025.3542368).
- [24] GUO Menghao, LU Chengze, LIU Zhengning, *et al.* Visual attention network[J]. *Computational Visual Media*, 2023, 9(4): 733–752. doi: [10.1007/s41095-023-0364-2](https://doi.org/10.1007/s41095-023-0364-2).
- [25] LI Yuxuan, HOU Qibin, ZHENG Zhaohui, *et al.* Large selective kernel network for remote sensing object detection[C]. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023: 16748–16759. doi: [10.1109/ICCV51070.2023.01540](https://doi.org/10.1109/ICCV51070.2023.01540).
- [26] WANG Wenhai, XIE Enze, LI Xiang, *et al.* PVT v2: Improved baselines with pyramid vision transformer[J]. *Computational Visual Media*, 2022, 8(3): 415–424. doi: [10.1007/s41095-022-0274-8](https://doi.org/10.1007/s41095-022-0274-8).
- [27] LIN T Y, GOYAL P, GIRSHICK R, *et al.* Focal loss for dense object detection[C]. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017: 2999–3007. doi: [10.1109/ICCV.2017.324](https://doi.org/10.1109/ICCV.2017.324).
- [28] REN Shaoqing, HE Kaiming, 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*, 2017, 39(6): 1137–1149. doi: [10.1109/TPAMI.2016.2577031](https://doi.org/10.1109/TPAMI.2016.2577031).
- [29] CAI Zhaowei, VASCONCELOS N. Cascade R-CNN: Delving into high quality object detection[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 6154–6162. doi: [10.1109/CVPR.2018.00644](https://doi.org/10.1109/CVPR.2018.00644).
- [30] LI Xiang, WANG Wenhai, WU Lijun, *et al.* Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection[C]. The 34th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2020: 1763.
- [31] DING Jian, XUE Nan, LONG Yang, *et al.* Learning RoI transformer for oriented object detection in aerial images[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, 2019: 2844–2853. doi: [10.1109/CVPR.2019.00296](https://doi.org/10.1109/CVPR.2019.00296).
- [32] LIU Yujie, SUN Xiaorui, SHAO Wenbin, *et al.* S²ANet: Combining local spectral and spatial point grouping for point cloud processing[J]. *Virtual Reality & Intelligent Hardware*, 2024, 6(4): 267–279. doi: [10.1016/j.vrih.2023.06.005](https://doi.org/10.1016/j.vrih.2023.06.005).
- 邹旻瑞: 男, 博士生, 研究方向为遥感目标检测等。
李宇轩: 男, 博士生, 研究方向为遥感目标检测等。
戴一冕: 男, 副教授, 研究方向为计算机视觉、遥感目标检测等。
李翔: 男, 副教授, 研究方向为计算机视觉、图像识别与检测等。
程明明: 男, 教授, 研究方向为人工智能、计算机视觉等。

责任编辑: 廖海贝

UMM-Det: A Unified Object Detection Framework for Heterogeneous Multimodal Remote Sensing Imagery

ZOU Minrui^① LI Yuxuan^① DAI Yimian^{①②} LI Xiang^{①②} CHENG Mingming^{①②}

^①(College of Computer Science, Nankai University, Tianjin 300350, China)

^②(Nankai International Advanced Research Institute, Shenzhen 518045, China)

Abstract:

Objective With the increasing demand for space-based situational awareness, object detection across multiple modalities has become a fundamental yet challenging task. Existing large-scale multimodal detection models for space-based remote sensing mainly operate on single-frame images from visible light, Synthetic Aperture Radar (SAR), and infrared modalities. Although these models achieve acceptable performance in conventional detection tasks, they largely neglect the critical role of infrared video sequences in improving weak and small target detection. Temporal information in sequential infrared data provides discriminative cues for separating dynamic targets from complex clutter, which cannot be captured by single-frame detectors. To address this limitation, this study proposes UMM-Det, a unified detection model designed for infrared sequences. The proposed model extends existing space-based multimodal frameworks to sequential data and demonstrates that exploiting temporal dynamics is essential for next-generation high-precision space-based sensing systems.

Methods UMM-Det is developed based on the unified multimodal detection framework SM3Det and introduces three key innovations. First, the ConvNeXt backbone is replaced with InternImage, a state-of-the-art architecture with dynamic sampling and large receptive field modeling. This replacement improves feature extraction robustness under multi-scale variations and low-contrast conditions that are typical of weak and small targets. Second, a spatiotemporal visual prompting module is designed for the infrared branch. This module generates high-contrast motion features using a refined frame-difference enhancement strategy. The resulting temporal priors guide the backbone to focus on dynamic target regions, thereby reducing interference from static background clutter. Third, to address the imbalance between positive and negative samples during training, Probabilistic Anchor Assignment (PAA) is incorporated into the infrared detection head. This strategy improves anchor selection reliability and enhances small target detection under highly skewed data distributions. The overall pipeline is shown in Fig. 1, and the structure of the spatiotemporal visual prompting module is illustrated in Fig. 2.

Results and Discussions Extensive experiments are conducted on three public benchmarks: SatVideoIRSTD for infrared sequence detection, SARDet-50K for SAR target detection, and DOTA for visible light remote sensing detection. The results in Table 2 show that UMM-Det consistently outperforms the baseline SM3Det across all modalities while significantly improving efficiency. For infrared sequence small target detection, UMM-Det improves detection accuracy by 2.54% compared with SM3Det, confirming the effectiveness of temporal priors. In SAR target detection, the model achieves a 2.40% improvement in mAP@0.5:0.95. In visible light detection, an improvement of 1.77% is observed. These results demonstrate the strong generalization capability of the proposed framework across heterogeneous modalities. In addition, UMM-Det reduces the number of parameters by more than 50% relative to SM3Det, which supports efficient and lightweight deployment in space-based systems. Qualitative results in Fig. 3 show that UMM-Det detects low-contrast and dynamic weak targets that are missed by the baseline model. The analysis highlights three main findings. First, the spatiotemporal visual prompting strategy effectively converts frame-to-frame variations into salient motion-aware cues, which are critical for distinguishing small dynamic targets from clutter in complex infrared scenes. Second, the use of InternImage substantially strengthens multi-scale representation capability, improving

robustness to variations in target size and contrast. Third, PAA alleviates training imbalance, leading to more stable optimization and higher detection reliability. Together, these components produce a synergistic effect, resulting in superior performance on both sequential infrared data and static SAR and visible light imagery.

Conclusions This study proposes UMM-Det, a space-based multimodal detection model that explicitly integrates infrared sequence information into a unified detection framework. By adopting InternImage for feature extraction, a spatiotemporal visual prompting module for motion-aware enhancement, and PAA for balanced training, UMM-Det achieves notable improvements in detection accuracy while reducing computational cost by more than 50%. Experimental results on SatVideoIRSTD, SARDet-50K, and DOTA demonstrate state-of-the-art performance across infrared, SAR, and visible light modalities, with accuracy gains of 2.54%, 2.40%, and 1.77%, respectively. The proposed framework provides a practical solution for future high-performance space-based situational awareness systems, where accuracy, efficiency, and lightweight design are all required. Future work may extend this framework to multi-satellite cooperative sensing and real-time onboard deployment.

Key words: Space-based multimodal unified detection framework; Multimodal remote sensing detection; Infrared sequence perception; Small target detection