自动化学报
ACTA AUTOMATICA SINICA

面向复杂环境的频域导向稀疏融合多光谱目标检测算法

肖海东 任志刚 蔡声泽 许超 吴宗泽

A Frequency-guided Sparse Fusion Algorithm for Multispectral Object Detection in Complex Environments

XIAO Hai-Dong, REN Zhi-Gang, CAI Sheng-Ze, XU Chao, WU Zong-Ze

在线阅读 View online: https://doi.org/10.16383/j.aas.c250634

面向复杂环境的频域导向稀疏融合多光谱目标检测算法

肖海东[1] 任志刚[1] 蔡声泽[2] 许 超[2] 吴宗泽[3]

摘 要 随着多光谱感知与智能视觉技术的发展, 如何在复杂环境中实现稳定而精确的目标检测已成为自动化视觉检测领域的重要研究方向. 针对传统单模态可见光目标检测在夜间、大雾及低照度等复杂环境中性能下降的问题, 提出一种基于频域特征细化导向的多光谱稀疏融合目标检测方法. 该方法利用共享权重的双分支编码器分别提取可见光与红外光特征, 并通过组稀疏自注意力模块实现跨模态长距离特征筛选, 以抑制冗余信息、增强显著特征表达. 同时, 设计频域自适应加权模块, 在频域空间中进行多光谱特征解耦与自适应融合, 实现不同光谱模态间的高效语义交互与动态权重分配. 该方法可在端到端框架下实现跨模态特征的高精度对齐与融合, 有效提升模型的检测精度与鲁棒性. 在 M^{3}FD 和 FLIR 数据集上取得 83.5% 和 81.6% 的 mAP_{50} 结果, 在 KAIST 数据集上取得 76.2% 的 AP_{50} 结果, 显著优于现有多光谱目标检测算法, 验证了所提方法在复杂场景下的优越性能和泛化能力.

关键词 目标检测; 跨光谱; 特征筛选; 跨模态特征对齐

引用格式 肖海东, 任志刚, 蔡声泽, 许超, 吴宗泽. 面向复杂环境的频域导向稀疏融合多光谱目标检测算法. 自动化学报, 2026, **52**(7): 1401−1412

DOI 10.16383/j.aas.c250634 **CSTR** 32138.14.j.aas.c250634

A Frequency-guided Sparse Fusion Algorithm for Multispectral Object Detection in Complex Environments

XIAO Hai-Dong[1] REN Zhi-Gang[1] CAI Sheng-Ze[2] XU Chao[2] WU Zong-Ze[3]

Abstract With the development of multispectral perception and intelligent vision technologies, achieving reliable and accurate object detection in complex environments has become a critical research focus in automatic vision detection. To address the performance degradation problem of traditional unimodal visible-light object detection in complex environments such as darkness, dense fog, and low illumination, this paper proposes a multispectral sparse fusion object detection method guided by frequency-domain feature refinement. The proposed method employs a dual-branch encoder with shared weights to extract features from visible and infrared modalities. A group sparse self-attention module is designed to perform cross-modal long-range feature selection, suppress redundant information, and enhance salient feature representation. Furthermore, a frequency-domain adaptive weighting module is designed to decouple and adaptively fuse multispectral features in the frequency-domain space, achieving efficient semantic interaction and dynamic weighting allocation across different spectral modalities. Under an end-to-end architecture, this method realizes high-precision cross-modal feature alignment and fusion, significantly improving detection accuracy and robustness. The proposed method achieved mAP_{50} results of 83.5% and 81.6% on the M^{3}FD and FLIR datasets, and AP_{50} results of 76.2% on the KAIST dataset. These results significantly outperform those of existing multispectral object detection algorithms, and validate the superior performance and generalization capability of the proposed method in complex scenes.

Keywords object detection; cross-spectral; feature selection; cross-modal feature alignment

Citation Xiao Hai-Dong, Ren Zhi-Gang, Cai Sheng-Ze, Xu Chao, Wu Zong-Ze. A frequency-guided sparse fusion algorithm for multispectral object detection in complex environments. *Acta Automatica Sinica*, 2026, **52**(7): 1401−1412

近年来, 多光谱成像在夜间监控、自动驾驶及无人机遥感等领域展现出重要应用价值[1]. 随着多光谱成像与智能视觉技术的快速发展, 如何在复杂环境中实现稳定、精确的目标检测, 已成为自动化

收稿日期 2025-11-14 录用日期 2026-03-19
Manuscript received November 14, 2025; accepted March 19, 2026
广东省基础与应用基础研究基金 (2024A1515011768), 广东省基础与应用基础研究重大专项项目 (2023B0303000009), 中国烟草总公司重点研发项目 (110202402018) 资助
Supported by Guangdong Basic and Applied Basic Research Foundation (2024A1515011768), Guangdong Major Project of Basic and Applied Basic Research (2023B0303000009), and Key R&D Project of China National Tobacco Corporation (110202402018)
本文责任编委 罗彪
Recommended by Associate Editor LUO Biao
1. 广东工业大学自动化学院 广州 510006 2. 浙江大学控制科学与工程学院 杭州 310027 3. 深圳大学机电与控制工程学院 深圳 518052
1. School of Automation, Guangdong University of Technology, Guangzhou 510006 2. College of Control Science and Engineering, Zhejiang University, Hangzhou 310027 3. College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen 518052

视觉检测领域的重要研究方向. 在监控识别、无人驾驶和无人机遥感等任务中, 目标检测算法通常依赖于可见光图像来获取目标的外观与纹理信息. 然而, 由于可见光成像机制受环境光照限制, 在夜间、大雾、雨雪等复杂场景下, 其检测性能显著下降, 常出现漏检或误检现象, 严重影响系统的可靠性与实用性[2−3]. 为弥补可见光图像在特定场景下信息不足的问题, 研究者引入能够反映目标热辐射特性的红外成像技术. 红外光图像具有对光照变化不敏感的优势, 可提供清晰的目标轮廓信息, 从而在低光照环境下有效增强目标检测性能. 可见光与红外光图像在光谱域上具有显著互补性, 如何高效融合两种模态的信息以获得稳健的检测性能, 成为近年来的研究热点[4−5].

如图 1 所示, 红外光图像可以在特殊场景下为可见光图像提供更多的信息参考. 然而, 目前的多光谱跨模态目标检测任务仍然存在诸多难题, 可见光图像含有丰富的色彩和纹理信息, 红外光图像则侧重于热轮廓信息, 二者的光谱频域差异大[6−7]. 因此简单的融合方法难以高效地对齐跨模态信息. 侧重于热轮廓信息的红外光图像往往存在大量空白画面信息, 引入过多的冗余信息不仅会降低模型的泛化能力, 还会干扰模型对关键信息的编码表征能力, 造成检测性能下降[8]. 同时, 不同的成像方式对目标有不同的敏感度, 因此在不同场景下, 不同光谱的特征图信息对模型表征的贡献度也不同. 在多模态目标检测任务中, 往往不容易确定可见光和红外光信息在不同场景下的主导地位. 针对这些问题, 需要构建一种统一的范式来对齐可见光和红外光图像的互补信息, 通过高效的特征筛选实现跨模态信息聚焦, 同时需要合理地动态分配不同模态的权重来提高模型鲁棒性.

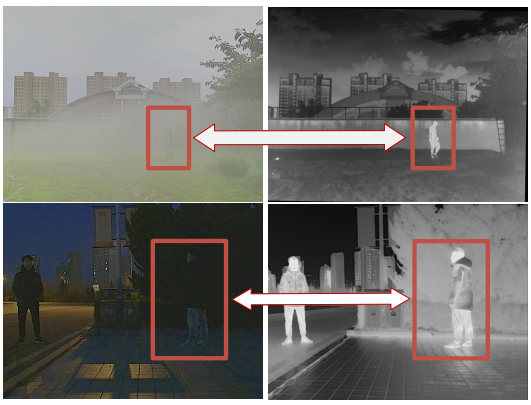


图 1 大雾场景下的可见光 (左上) 与红外光 (右上) 及黑夜场景下的可见光 (左下) 与红外光 (右下) 示例图

Fig. 1 Example images of visible light (top-left) and infrared light (top-right) under dense fog scenes, and visible light (bottom-left) and infrared light (bottom-right) under night-time scenes

针对上述痛点, 学界已开展从像素级到特征级融合的广泛研究, 但现有方法仍难以兼顾特征稀疏性与跨模态语义的一致性. 基于此, 本文提出一种基于频域特征细化导向的多光谱稀疏融合目标检测方法. 该方法通过共享权重的双分支编码结构, 实现可见光与红外光特征的协同提取; 设计组稀疏自注意力模块 (group sparse self-attention module, GSAM) 用于长距离特征选择与冗余抑制; 提出频域自适应加权模块 (frequency-domain adaptive weighting module, FDAW), 在频域空间实现跨模态特征的动态加权与细粒度融合, 从而有效提升模型的检测精度与鲁棒性. 本文的主要贡献概述如下:

1) 提出基于组稀疏自注意力的多光谱长距离特征筛选机制, 利用多组稀疏自注意力实现远程特征依赖的动态建模;

2) 设计频域导向的自适应融合策略, 有效实现跨模态特征的细粒度动态对齐, 强化跨模态语义交互;

3) 构建高效端到端多光谱检测框架, 在检测精度与模型复杂度之间取得有效平衡, 并在参数量可控的条件下实现显著性能提升.

本文剩余部分结构如下: 第 1 节介绍可见光与红外光图像目标检测领域的相关研究工作; 第 2 节阐述构建多光谱目标检测模型所采用的方法论; 第 3 节详细说明所进行的相关定性与定量实验; 第 4 节则对本研究进行总结并对未来工作进行规划与展望.

1 相关工作

1.1 单光谱目标检测

目标检测作为计算机视觉的基础任务之一, 在智能监控、自动驾驶及遥感监测等应用中发挥着关键作用. 早期目标检测方法多采用两阶段结构, 包括候选区域生成与目标分类, 例如 R-CNN 系列模型通过选择性搜索获取候选框并使用 SVM 分类器实现检测[9−10]. 然而, 该类方法计算复杂、推理速度慢, 难以满足实时应用需求. 随着深度学习的发展, 端到端的单阶段检测器成为主流. 代表性方法如 SSD (single shot multibox detector) 与 YOLO (you only look once) 系列[11−12], 通过一次前向传播实现目标定位与分类, 极大提高了实时检测性能. 近年来, Transformer 结构被引入计算机视觉领域, 如 ViT (vision Transformer)[13] 与 DETR (detection

Transformer) 模型[14], 通过全局自注意力机制显著提升了特征表达与上下文建模能力, 为多模态检测任务提供了新的理论基础.

1.2　多光谱目标检测

受限于光学成像机制, 传统单模态检测算法在大雾、低照度等极端场景下性能退化严重. 为此, 能够充分利用光谱互补特性的多模态探测方法得到了广泛研究[15]. 现有的多光谱目标检测策略主要可归纳为两类: 先融合后检测的两阶段策略与端到端的特征层融合架构.

先融合后检测的两阶段策略通常在数据层面进行像素级融合, 即先利用自监督或生成式网络将可见光与红外光图像合成单模态图像, 再输入标准检测器进行识别. 代表性工作如文献 [16] 提出的无监督融合网络以及文献 [17] 采用的双层优化方案. 此外, 文献 [18] 通过边缘引导注意力机制使融合过程聚焦共同结构. 尽管该类方法能直接利用现成的检测器, 但由于可见光与红外光存在巨大的频域差异, 单一的融合图像往往难以全面表征双模态的关键特征, 且臃肿的架构易造成原始信息的稀释.

相比之下, 另一种主流策略是端到端的特征层融合架构, 该策略利用双分支编码器在语义层面实现信息交互. 文献 [19] 通过跨模态 Transformer 实现了融合与检测的一体化设计; 文献 [20] 提出了迭代交叉注意力方法以捕获局部与全局特征; 文献 [21] 则在 RT-DETR[22] 基础上引入了模态竞争查询策略. 然而此类模型受限于其自注意力机制的二次复杂度计算, 往往伴随高计算开销的全局架构. 随着模型的迭代优化, 这种高度耦合的端到端架构在实时性与泛用性上展现出显著优势.

总体而言, 现有两阶段策略存在较大的信息损耗, 而端到端架构虽能保持跨模态特性, 但受限于跨频域对齐效率, 仍难以有效兼顾特征稀疏性与语义互补性. 鉴于此, 本文提出频域细化导向与组稀疏融合的双模块架构, 旨在通过高效的信息提炼实现更精准的多光谱目标检测.

2　多光谱稀疏融合目标检测方法

2.1　检测网络架构与优化目标

本文提出的基于频域特征细化导向的多光谱稀疏融合目标检测网络 (frequency-domain guided sparse fusion network, FGSFNet) 整体框架如图 2 所示, 采用双分支共享主干结构. 设可见光输入为 $\boldsymbol{X}_{vi}$, 红外光输入为 $\boldsymbol{X}_{ir}$. 两路输入分别通过共享权重的双分支主干进行多尺度特征提取, 得到不同阶段的特征图. 为了在特征层实现高效的跨模态信息交互与冗余抑制, 本文在主干的若干阶段引入两类核心模块:

1) 组稀疏自注意力模块, 用于在语义层面进行长距离的稀疏编码与特征筛选, 从而抑制冗余背景信息并增强对显著目标区域的响应;

2) 频域自适应加权模块, 用于在频域对多光谱融合特征进行解耦、加权与自适应融合, 以实现不同频段间的细粒度信息重分配.

模块间通过残差与跳跃连接保持信息流的稳定, 最终将多尺度融合特征送入标准检测头得到目

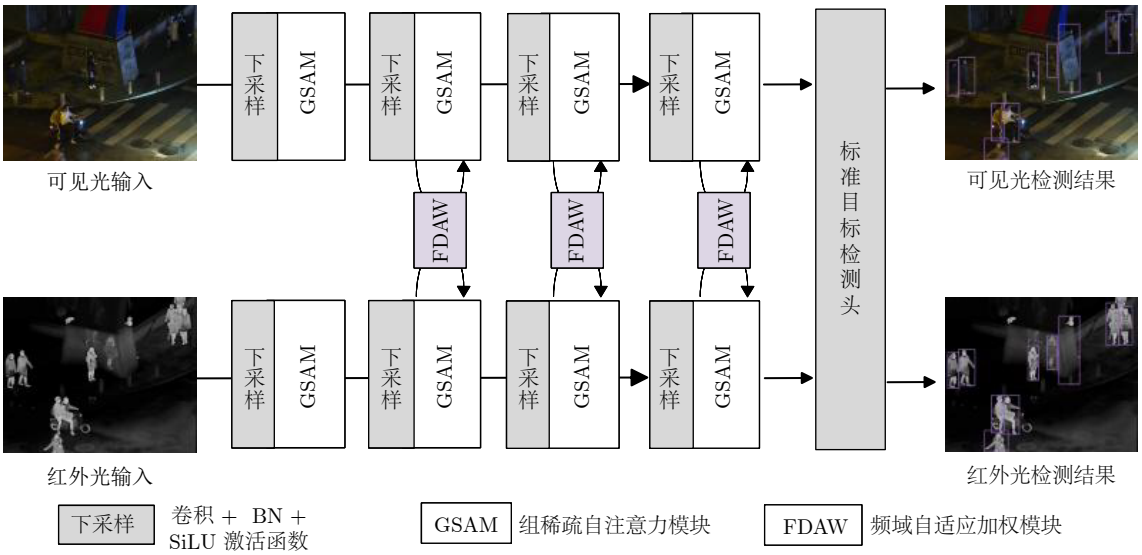


图 2　网络结构图

Fig. 2　Network structure diagram

标检测结果. 具体地, 本文设计了以自注意力机制为核心的堆叠结构作为特征提取模块, 通过四个阶段的降采样将输入尺寸为 $H \times W \times C$ 的原始图像编码为 $\frac{H}{8} \times \frac{W}{8} \times C$ 的特征图, 接着合并可见光和红外光特征图的编码信息并使用标准检测头进行预测输出. 与主流架构不同的是, 在下采样阶段使用了一个 3 × 3 的卷积块以及批归一化 (batch normalization, BN) 和 SiLU 激活函数来调整输入尺寸以提供长距离特征细化. 编码块部分采用针对多光谱场景的窗口分组稀疏自注意力模块和多层感知机模块的标准化设计, 并添加层归一化以及残差连接以保证训练的稳定性, 在每个阶段通过堆叠 N_i $(i = 1, 2, 3, 4)$ 次组稀疏自注意力模块来获取长距离依赖. 同时, 为了促进不同光谱模态之间的信息交流, 构建了以可见光和红外光特征频域特性为导向的频域自适应加权模块. 该模块通过对一组不同光谱的输入特征根据其频域特性进行加权, 在双编码器的不同阶段构建多模态之间的有效频域信息交互, 以得到更高效的跨光谱全局语义特征.

为了指导网络进行端到端的学习, 本文检测头采用标准的检测损失分解, 网络训练采用如下加权总损失函数:

$$Loss = \lambda_{box} L_{box} + \lambda_{obj} L_{obj} + \lambda_{cls} L_{cls} \tag{1}$$

其中, L_{box} 表示边界框回归损失; L_{obj} 表示目标置信度损失; L_{cls} 表示分类损失; 各分量的权重 λ_{box}、λ_{obj}、λ_{cls} 可根据数据集与任务需求进行调节以平衡定位与分类的训练目标. 最后通过目标识别架构的检测头获取可见光和红外光的输出, 得到最终的检测结果.

2.2 组稀疏自注意力模块 (GSAM)

得益于其强大的长距离建模能力, 标准的 ViT 视觉编码器在各类计算机视觉任务中展现了优异的效果. 在多光谱的场景下, 考虑到可见光图像富含色彩和纹理信息而红外光图像侧重于热轮廓信息, 我们希望后者能为前者提供轮廓信息指引, 然而这样不可避免地会引入大量冗余的空白区域信息. 因此, 为了解决标准自注意力在高分辨率特征图上的计算负担和多模态冗余问题, 本文在自注意力机制的基础上, 针对上述问题设计了组稀疏自注意力模块, 模块结构如图 3 所示. 输入特征图首先被划分成 $n \times n$ 个窗口, 接着对其进行扁平化处理至特征维度, 划分窗口的数量会随着特征图分辨率下降而减少. 接着用一个 1×1 卷积和 3×3 深度可分离卷积进一步扩展输入的特征维度并从中拆分获得线性变换后的查询矩阵 $\boldsymbol{Q} \in \mathbf{R}^{L \times d}$、键矩阵 $\boldsymbol{K} \in \mathbf{R}^{L \times d}$ 以及值矩阵 $\boldsymbol{V} \in \mathbf{R}^{L \times d}$, 其中 L 为特征序列长度, d 为嵌入维度. 计算得到的原始注意力矩阵 $\boldsymbol{A} \in \mathbf{R}^{L \times L}$ 用于捕获长距离特征依赖:

$$I_i = \text{TopKIndex}(\boldsymbol{A}_{i\cdot}, k), \quad i = 1, 2, \cdots, L$$

其中, $\text{TopKIndex}(\cdot, k)$ 算子用于提取矩阵 $\boldsymbol{A}$ 每一行中响应强度前 k 高的索引, 并基于位置索引构建与注意力矩阵同维度的二值掩码矩阵 $\boldsymbol{M} \in \mathbf{R}^{L \times L}$:

$$\boldsymbol{M}_{ij} = \begin{cases} 1, & j \in I_i \\ 0, & j \notin I_i \end{cases}$$

其中, 掩码矩阵 $\boldsymbol{M}$ 的行索引 i 与列索引 j 分别对应 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 的序列位置. 接着利用掩码矩阵对注意力矩阵进行过滤, 低于稀疏率阈值的位置填充为 $-\infty$, 以便在 Softmax 之后被抑制.

$$\boldsymbol{A}^k_{ij} = \begin{cases} \boldsymbol{A}_{ij}, & \boldsymbol{M}_{ij} > 0 \\ -\infty, & \boldsymbol{M}_{ij} = 0 \end{cases}$$

最终的组稀疏注意力计算为:

$$\mathcal{A}_{\text{sparse}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \left(\sum_{g=1}^{s} \text{Softmax}(\boldsymbol{A}^{k_g}) \right) \boldsymbol{V}$$

其中, s 表示稀疏率组的数量; k_g 表示第 g 组对应的稀疏率阈值. 该模块通过分组与 TopK 筛选显著

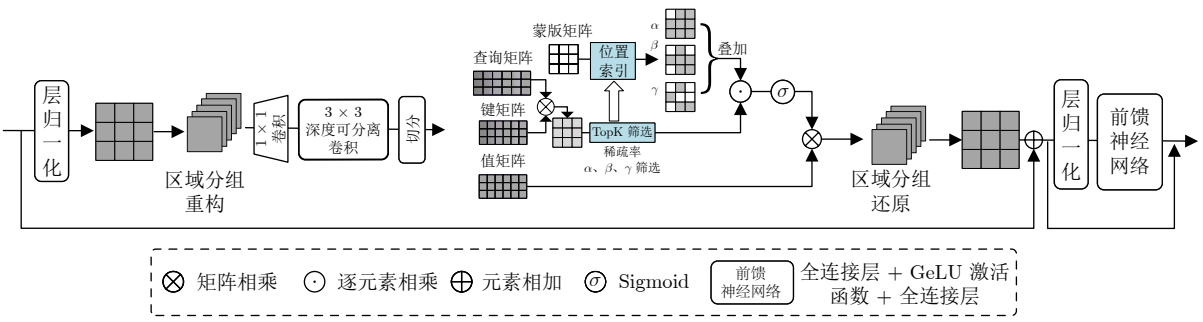


图 3 组稀疏自注意力模块结构图

Fig. 3 Group sparse self-attention module structure diagram

降低无关位置的注意力响应, 从而达到稀疏化冗余特征、突出关键信息的目的. 此外, 这种多组稀疏率叠加与残差连接的协同设计, 确保了模型在目标密集的复杂场景下仍能通过残差路径保留关键特征, 避免了高稀疏率筛选可能带来的信息截断, 增强了算法在不同目标密度下的鲁棒性. 模块后接前馈神经网络、层归一化和残差连接以保证训练稳定性.

2.3　频域自适应加权模块 (FDAW)

多光谱模态在频域上具有不同特性: 可见光传感器 (波长通常分布在 0.38 ~ 0.78 μm) 主要捕获环境光源在物体表面的细微纹理、颜色及几何结构, 使得可见光图像在空间域内富含精细的纹理特征, 在频域空间则表现为显著的高频分量. 红外传感器 (波长通常在 8 ~ 14 μm) 主要反映目标的热轮廓信息, 在频域中表现为明显的低频分布特性, 且具有极强的抗光照干扰与穿透烟雾能力[23−24]. 基于此, 本文在双分支编码器的若干阶段引入了自适应加权模块, 其理论依据在于利用快速傅里叶变换 (fast Fourier transform, FFT) 将空间特征投影至频域, 从而实现对可见光高频纹理与红外低频轮廓的显式解耦. 这种基于物理机理的频域导向策略能够更精准地实现跨模态特征的对齐与语义交互, 其整体架构如图 4 所示.

首先, 将来自两种模态的特征在通道维度拼接并通过 1×1 卷积调整通道尺寸, 得到初步的融合特征 $\boldsymbol{X}_{\text{fuse}}\in\mathbf{R}^{C\times H\times W}$. 随后, 对 $\boldsymbol{X}_{\text{fuse}}$ 进行分块延展 (patch unfolding) 处理, 得到局部块特征 $\boldsymbol{X}_{\text{patch}}\in\mathbf{R}^{C\times h\times w\times p\times p}$, 其中 p 为块大小, h、w 为分块后的空间分辨率. 接着对每个区块进行快速傅里叶变换 (FFT), 得到频域表示 $\widetilde{\boldsymbol{X}}_{\text{fuse}}\in\mathbf{C}^{C\times h\times w\times p\times p'}$, 其中 $p'=\lfloor p/2\rfloor+1$, $\lfloor\ \rfloor$ 表示向下取整. 该过程如式 (2) ~ (3) 所示:

$$\boldsymbol{X}_{\text{patch}}=\mathcal{F}^{\text{PU}}\left(\mathcal{F}^{C}_{1\times1}\left(\text{Concat}(\boldsymbol{X}_{vi},\ \boldsymbol{X}_{ir})\right)\right)\tag{2}$$

$$\widetilde{\boldsymbol{X}}_{\text{fuse}}[c,\ h,\ w,\ k,\ l]=\sum_{m=0}^{p-1}\sum_{n=0}^{p-1}\boldsymbol{X}_{\text{patch}}[c,\ h,\ w,\ m,\ n]\cdot \mathrm{e}^{-\mathrm{j}2\pi\left(\frac{km}{p}+\frac{ln}{p}\right)}\tag{3}$$

其中, $\mathcal{F}^{\text{PU}}$ 表示分块延展操作; $\mathcal{F}^{C}_{1\times1}$ 表示 1×1 卷积操作; Concat(·) 表示通道维度的特征拼接操作; m 和 n 分别表示局部特征块 (patch) 在空间域的水平与垂直坐标索引; "·"表示标量乘法; j 为虚数单位; $k\in\{0,\ 1,\ \cdots,\ p-1\}$、$l\in\{0,\ 1,\ \cdots,\ \lfloor p/2\rfloor\}$ 和 $c\in\{0,\ 1,\ \cdots,\ C-1\}$ 分别表示频域的水平、垂直频率索引和通道索引. 接着对频域表示取实部并在

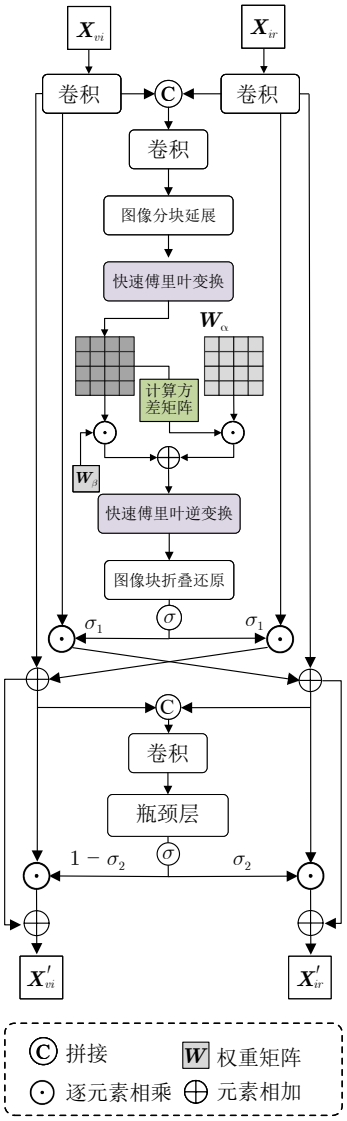


图 4　频域自适应加权模块结构图

Fig. 4　Frequency-domain adaptive weighting module structure diagram

h、w 维度计算方差以提取频域统计向量 $\boldsymbol{v}$:

$$\boldsymbol{v}=\text{Var}(\text{Re}(\widetilde{\boldsymbol{X}}_{\text{fuse}}))$$

其中, Re(·) 表示取复数的实部操作; Var(·) 表示沿空间维度 $(h,\ w)$ 计算样本方差的操作. 频域方差统计量 $\boldsymbol{v}$ 在信号处理层面反映了各频率分量的结构复杂度与能量分布强度. 由于目标边缘与细粒度纹理在频域内具有更显著的波动特性 (高方差), 而冗余背景通常表现为平滑分布 (低方差), 通过与学习权值 $\boldsymbol{W}_{\alpha}$、$\boldsymbol{W}_{\beta}$ 结合, 模型能够自适应地量化不同模态在各频带下的贡献度, 引导网络聚焦于具有更高信息增益的跨模态特征, 实现重要性驱动的动态融合. 具体的加权计算如下式所示:

$$\widetilde{\boldsymbol{X}}_{w}=\boldsymbol{W}_{\alpha}\odot\boldsymbol{v}+\boldsymbol{W}_{\beta}\odot\widetilde{\boldsymbol{X}}_{\text{fuse}}$$

其中, $\odot$ 表示逐元素相乘. 接着对加权后的频域特征执行快速傅里叶逆变换 (inverse fast Fourier transform, IFFT) 并将其还原为空间块表示, 如式 (4) ~ (5) 所示:

$$\boldsymbol{X}'_w[c,\ h,\ w,\ m,\ n] = \frac{1}{p^2}\sum_{k=0}^{p-1}\sum_{l=0}^{\lfloor p/2\rfloor}\widetilde{\boldsymbol{X}}_{\text{fuse}}[c,\ h,\ w,\ k,\ l]\cdot \mathrm{e}^{\mathrm{j}2\pi\left(\frac{km}{p}+\frac{ln}{p}\right)} \tag{4}$$

$$\widetilde{\boldsymbol{X}} = \mathcal{F}^{\text{PF}}(\boldsymbol{X}'_w) \tag{5}$$

其中, 式 (4) 表示 IFFT; $\mathcal{F}^{\text{PF}}$ 表示分块折叠 (patch folding) 操作, 用于将空间局部块还原为完整的空间特征图表示. 最后, 对逆变换结果进行 Sigmoid 映射得到融合权重 σ_1, 并与原始模态特征进行加权叠加, 具体过程如式 (6) ~ (8) 所示:

$$\sigma_1 = \text{Sigmoid}(\widetilde{\boldsymbol{X}}) \tag{6}$$

$$\widetilde{\boldsymbol{X}}_{vi} = \sigma_1 \odot \boldsymbol{X}_{vi} + \boldsymbol{X}_{ir} \tag{7}$$

$$\widetilde{\boldsymbol{X}}_{ir} = \sigma_1 \odot \boldsymbol{X}_{ir} + \boldsymbol{X}_{vi} \tag{8}$$

随后, 在通道维度拼接加权后的多光谱特征, 经过 1×1 卷积和瓶颈层编码得到第二阶段的自适应权重 σ_2, 并通过式 (9) ~ (11) 对特征进行最终加权与残差融合:

$$\sigma_2 = \text{Sigmoid}(\mathcal{F}^{\text{B}}(\mathcal{F}^{\text{C}}_{1\times 1}(\text{Concat}(\widetilde{\boldsymbol{X}}_{vi},\ \widetilde{\boldsymbol{X}}_{ir})))) \tag{9}$$

$$\boldsymbol{X}'_{vi} = \sigma_2 \odot \widetilde{\boldsymbol{X}}_{vi} + \widetilde{\boldsymbol{X}}_{vi} \tag{10}$$

$$\boldsymbol{X}'_{ir} = (1-\sigma_2) \odot \widetilde{\boldsymbol{X}}_{ir} + \widetilde{\boldsymbol{X}}_{ir} \tag{11}$$

其中, $\mathcal{F}^{\text{B}}$ 表示瓶颈层编码操作. 在计算时间复杂度方面, FDAW 的计算主要集中在频域变换和加权操作上. 由于 FFT 和 IFFT 操作的时间复杂度为 $\text{O}(p^2\log p)$, 而加权操作为 $\text{O}(c\times h\times w\times p^2)$, 因此整体复杂度约为 $\text{O}(c\times h\times w\times p^2 + p^2\log p)$. 尽管存在一定的计算开销, 但相较于主干的卷积层, FDAW 的计算量主要取决于空间分辨率而非通道数的平方, 这使得它在处理多模态特征时具有显著的效率优势, 能够在不同频段和不同模态之间实现细粒度的信息重分配, 从而提升跨模态特征对齐的准确性和鲁棒性.

3 实验设置与结果分析

3.1 实验设置

为验证本文提出的多光谱稀疏融合目标检测网络 (FGSFNet) 的有效性与泛化性能, 本文在 M³FD、KAIST 和 FLIR 三个公开多光谱目标检测数据集上进行了实验. 实验平台配置为: NVIDIA GeForce RTX 3090 GPU. 模型基于 PyTorch 实现, 损失函数采用式 (1) 定义的加权组合. 在模型构建方面, 组稀疏自注意力模块 (GSAM) 的堆叠次数 N_i 设为 {3, 9, 9, 3}; 窗口分组数 n 设置为 {8, 4, 2, 1}; 分块数 p 设置为 {4, 2, 1}. 优化器采用 SGD, 权重衰减为 5×10^{-4}, 动量为 0.9, 初始学习率为 0.01. 在每个数据集上, 训练与测试集按照 80% 与 20% 划分, 批量大小为 8. 在 M³FD[17] 数据集上训练 250 个轮次, 在 KAIST 和 FLIR 数据集上训练 100 个轮次, 该差异化设置的原因在于 KAIST[25] 与 FLIR[26] 数据集目标相对集中, 实验观察显示其在 100 轮左右即可达到性能饱和, 而 M³FD 数据集需要更长的训练过程以实现全类别的特征收敛. 值得注意的是, 对于"先融合后检测"类模型, 本文首先利用原始融合模型生成融合数据集, 再统一使用标准 YOLOv5 模型进行训练与检测, 以保证实验公平性.

3.2 数据集

本文在三个公开基准数据集上进行实验, 包括 M³FD、KAIST 和 FLIR. 上述数据均涵盖了从乡村道路、森林等稀疏场景到繁华城市街道、拥挤人群等极具挑战性的密集目标场景. M³FD 数据集是一个包含 4 200 对高分辨率多光谱图像的数据集, 其中包含了行人 (person)、轿车 (car)、巴士 (bus)、摩托车 (motorcycle)、灯 (lamp) 和卡车 (truck) 六个目标; 该数据集包含了城市、森林、浓烟等不同场景, 其多样性有助于提高算法的泛化能力. KAIST 数据集是一个涵盖了校园、居民区、街道等场景在不同时间段以及不同天气场景的多模态行人数据集, 本文选用了其中 10 346 对图像进行实验. FLIR 数据集是一个涵盖城市街道、乡村道路等场景的可见光、红外光图像数据集, 主要包含行人、汽车、自行车等常见的道路目标; 本文使用了一个公开可用的对齐版本, 其中包含 5 039 对图像.

3.3 评价指标

本文采用精度 (Precision)、召回率 (Recall)、平均精度均值 (mean average precision, mAP)、参数量 (Params) 以及每秒传输帧数 (frames per second, FPS) 作为评价指标.

1) 精度 (Precision) 表示预测为正样本的集合中真正对应目标的比例:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{12}$$

其中, TP (true positive) 表示正确检测到的正样本数量; FP (false positive) 表示误报为正样本的负样本数量.

2) 召回率 (Recall) 表示所有真实目标中被正确检测到的比例:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

其中, FN (false negative) 表示未被检测到的真实目标数量.

3) mAP_{50} 指在交并比 (intersection over union, IoU) 阈值为 0.50 时, 所有类别的平均精度的均值:

$$\text{mAP}_{50} = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} \text{AP}_i^{\text{IoU}=0.50}$$

其中, AP (average precision) 表示单个类别的平均精度; N_{cls} 表示类别数量.

4) $\text{mAP}_{50:95}$ 是在 IoU 阈值以 0.05 为步长从 0.50 到 0.95 的一系列阈值下, 所有类别的平均精度的均值:

$$\text{mAP}_{50:95} = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} \frac{1}{10} \sum_{j=0}^{9} \text{AP}_i^{\text{IoU}=0.50+0.05j}$$

5) Params 指在模型中所有可学习参数的总量, 是衡量模型空间复杂度和存储资源消耗的核心指标. 模型总参数量由网络中所有参数化层的参数累加得到.

6) FPS 用于衡量模型的推理速度, 是评估算法实时性能的核心指标. 它表示模型在单位时间内能够处理的图像数量, 计算公式如下:

$$\text{FPS} = \frac{1}{T_{\text{avg}}} \tag{14}$$

其中, T_{avg} 表示模型处理单张图像的平均推理时间. FPS 数值越高, 代表模型的推理效率越高, 实时处理能力越强.

3.4　与其他算法的性能比较

本文与代表两类主流多光谱目标检测范式的几种最先进模型进行了比较分析. 首先是先融合后检测的架构, 其中包括 U2Fusion[16]、RFN[27]、Tardal[17]、MFEIF[18] 和 HALDeR[28]. 上述模型均为融合模型, 首先将多模态图像对融合为一张图像后再经由标准 YOLO 检测器对融合后的图像进行单模态训练以及推理得到检测结果. 另一种架构则是端到端的多光谱目标检测模型, 包括 DEYOLO[29]、ICAFusion[20]、CFT[19]、DAMSDet[21] 以及 SuperYOLO[30]. 这类模型通过端到端的架构直接对可见光和红外光图像进行训练推理得到检测结果, 本文提出的模型也属于该架构.

3.4.1　架构类型比较

本文所对比的两种主流架构的实验结果如表 1 和表 2 所示, 红色和蓝色数据分别表示最优和次优的结果. 从结果可以看出, 端到端的多光谱目标检

表 1　不同方法在 M³FD 数据集上的实验评价结果比较

Table 1　Comparative experimental evaluation results of different methods on the M³FD dataset

架构	方法	mAP_{50} (%)						精度 (%)	召回率 (%)	mAP_{50} (%)	$\text{mAP}_{50:95}$ (%)	Params (M)	FPS
		巴士	轿车	灯	摩托车	行人	卡车						
单光谱模态	可见光	90.9	88.5	79.6	75.2	67.7	85.3	89.4	73.4	81.2	50.5	1.77	175.4
	红外光	90.7	86.5	57.8	69.5	79.0	83.9	85.1	71.5	77.9	48.2	1.77	175.4
融合检测架构	U2Fusion[16]	80.8	86.3	73.8	56.4	73.2	78.2	84.1	70.1	74.8	45.4	—	—
	RFN[27]	79.2	85.7	69.9	53.1	71.4	76.4	88.5	65.1	72.6	44.9	—	—
	Tardal[17]	77.7	85.0	67.1	56.7	73.1	74.4	84.4	65.7	72.3	43.6	—	—
	MFEIF[18]	79.7	84.9	68.2	56.9	72.9	73.5	85.2	65.9	72.7	44.6	—	—
	HALDeR[28]	79.8	86.0	70.8	52.7	71.2	75.4	86.7	66.8	72.7	44.4	—	—
端到端架构	Twin-YOLOv5-n	89.1	87.8	77.1	66.2	77.0	83.8	85.0	75.1	80.2	47.8	2.84	192.3
	Twin-YOLOv8-n	89.8	88.6	78.2	67.5	78.3	85.1	85.2	73.0	81.2	49.6	4.28	128.2
	DEYOLO[29]	89.3	87.5	77.7	67.8	77.4	85.2	84.0	75.9	80.8	49.3	4.57	188.5
	ICAFusion[20]	90.4	87.6	78.5	65.2	77.2	86.0	83.6	76.1	80.8	48.5	5.94	144.9
	CFT[19]	90.4	87.9	77.0	70.2	76.6	85.1	87.7	74.3	81.2	48.1	11.20	154.0
	DAMSDet[21]	83.1	92.7	71.1	73.5	73.6	78.6	78.7	83.1	78.8	51.0	78.90	88.5
	SuperYOLO[30]	89.1	87.6	74.7	68.9	76.6	83.2	88.2	70.1	80.0	48.0	1.80	212.7
	FGSFNet	92.1	89.0	80.5	74.3	79.0	86.1	86.7	77.3	83.5	51.2	7.44	208.3

表 2 不同方法在 KAIST 和 FLIR 数据集上的实验评价结果比较 (%)

Table 2 Comparative experimental evaluation results of different methods on the KAIST and FLIR datasets (%)

架构	方法	KAIST (单类别: 行人)				FLIR (多类别)						
		精度	召回率	AP_{50}	$AP_{50:95}$	mAP_{50}			精度	召回率	mAP_{50}	$mAP_{50:95}$
						自行车	汽车	行人				
单光谱模态	可见光	71.6	54.4	62.0	24.6	68.9	83.5	70.0	80.2	67.6	74.1	34.8
	红外光	75.8	65.4	73.5	32.3	75.8	87.5	81.2	84.5	72.0	81.5	40.6
融合检测架构	U2Fusion[16]	65.4	45.9	52.2	21.1	63.1	83.0	71.6	79.4	65.9	72.6	33.8
	RFN[27]	59.6	43.7	48.9	18.9	63.1	82.0	71.4	78.7	65.2	72.1	33.4
	Tardal[17]	62.4	45.0	50.6	20.6	60.1	81.5	68.5	77.1	63.3	70.0	32.3
	MFEIF[18]	62.1	43.8	50.0	20.6	62.9	80.9	69.7	76.0	66.2	71.2	32.8
	HALDeR[28]	63.3	45.7	51.6	21.1	62.9	82.1	70.1	79.6	64.9	71.7	32.9
端到端架构	Twin-YOLOv5-n	76.7	68.7	74.8	32.8	74.9	87.2	82.0	84.0	74.6	81.4	39.4
	Twin-YOLOv8-n	78.2	68.0	74.2	32.9	73.8	87.5	81.2	83.0	72.6	80.8	39.7
	DEYOLO[29]	74.9	68.2	74.4	31.9	74.5	86.9	82.0	84.7	72.7	81.1	40.0
	ICAFusion[20]	75.5	69.1	75.7	33.2	74.8	87.4	81.6	82.2	75.9	81.3	39.6
	CFT[19]	77.6	67.2	75.7	32.9	72.9	87.4	82.1	82.9	73.4	80.8	39.2
	SuperYOLO[30]	74.3	65.5	71.8	31.5	70.6	86.3	78.3	82.1	69.1	78.4	37.7
	FGSFNet	77.6	71.1	76.2	33.8	75.0	87.7	82.1	83.4	75.0	81.6	40.2

测架构的表现远远优于先融合后检测的架构. 在先融合后检测的架构中, 单一的融合图像往往难以表征丰富的多模态信息. 臃肿的二段式架构由于包含了检测器的二次训练, 其最终检测精度也很大程度上会取决于所选用的检测器的性能, 因此大多数先融合后检测模型在相同配置下的检测精度也比较相近. 而端到端多光谱检测模型能够高效地利用可见光和红外光图像的跨模态信息, 在小尺寸模型框架下对比单模态基线 (可见光或红外光), 普遍都有着较大的精度提升.

3.4.2 定量结果分析

表 1 和表 2 分别列出各个算法在 M^{3}FD、KAIST 和 FLIR 数据集上的实验结果, 其中 Twin-YOLO 为不含融合模块和额外采样模块的双分支 YOLO 基线模型. 在 M^{3}FD 数据集上的定量实验结果如表 1 所示, 本文提出的算法在 mAP_{50} 和 $mAP_{50:95}$ 方面取得了最优结果, 分别达到了 83.5% 和 51.2%. 在 mAP_{50} 指标上, 本文算法以更小的参数量, 在精度上超过次优方法 2.3%. 在 $mAP_{50:95}$ 方面, 虽然仅以 0.2% 的微弱优势超过次优方法 (DAMSDet), 但参数量远小于后者. 值得注意的是, 在推理效率方面, 本文模型展现出显著的实时检测优势, 推理速度达到了 208.3. 相比于其他主流端到端架构, 这一数据充分证明了 FGSFNet 在高精度与实时性之间取得了良好的平衡, 能够较好地适配复杂环境下的实时视觉检测任务. 在 KAIST 数据集上的定量实验结果如表 2 左侧所示, 本文提出的网络在精度和召回率方面分别取得了次优和最优结果, 在 mAP_{50} 和 $mAP_{50:95}$ 指标上均取得了最优性能, 分别达到 76.2% 和 33.8%; 相较于其他端到端多光谱目标检测模型, 分别提升了 0.5% ~ 4.4% 和 0.6% ~ 2.3%. 在 FLIR 数据集上的定量实验结果如表 2 右侧所示, 本文模型在 mAP_{50} 和 $mAP_{50:95}$ 方面分别取得了最优和次优结果, 达到了 81.6% 和 40.2%, 均超越了对比模型; 相较于可见光单模态基线, 分别提升了 7.5% 和 5.4%. 实验结果充分体现了本文所提模型的优越性能.

3.4.3 定性结果分析

本文所提出模型的检测结果示意图如图 5 所示, 其中 VI 和 IR 分别表示可见光和红外光图像. 实心框表示模型的检测结果, 红色虚线框表示遗漏检测 (假阴性). 从本文选取的较为代表性的示例图中可以看出, 先融合后检测模型在处理复杂的场景目标时, 受限于其单一模态的图像信息, 往往会出现漏检 (第 1 行). 部分端到端检测模型在面对不明显的小目标时往往也会表现不佳 (第 2、3 行). 而本文模型得益于其多窗口分组筛选特征的模式, 在面对不明显的小目标时也能高效地整合多模态间的不同频域的互补信息, 即使在具有挑战性的场景下也能展现出强大的检测性能.

3.5 额外实验研究

3.5.1 组稀疏自注意力模块稀疏率参数研究

在构建组稀疏自注意力模块时, 本文采用了由

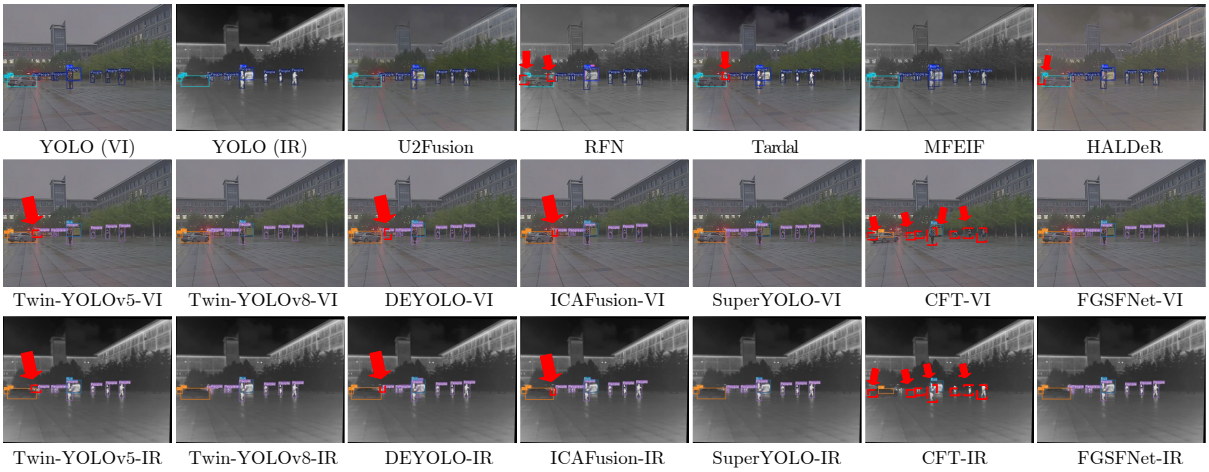


图 5　不同方法在 M³FD 数据集上的检测结果图

Fig. 5　Detection results diagram of different methods on the M³FD dataset

多个不同稀疏率的选择模块组合叠加设计. 为了探究不同稀疏参数组合对模型的影响, 选择了几种比较有代表性的组合, 并在 M³FD 数据集上进行了一系列实验, 结果如表 3 所示, 其中红色和蓝色数据分别表示最优和次优的结果. 当稀疏参数只有一组时, 模型的表现不佳 (实验 I). 随着参数组合数增加, 模型的检测精度也随之上升 (实验 II ～ 实验 IV), 当组合为 {50%, 67%, 80%}(实验 III) 时达到最佳效果, 实现了 83.5% 和 51.2% 的 mAP_{50} 和 $mAP_{50:95}$. 当参数组合数达到 4 时 (实验 V), 模型的性能表现不佳. 因此, 本文选择实验 III 中的参数组合作为组稀疏自注意力模块的稀疏率参数组.

表 3　稀疏率组合选取

Table 3　Sparsity rate combination selection

实验	稀疏率参数组合				实验结果 (%)	
	50%	67%	75%	80%	mAP_{50}	$mAP_{50:95}$
I	0	0	0	1	78.4	46.2
II	0	1	0	1	82.3	49.7
III	1	1	0	1	83.5	51.2
IV	1	1	1	0	83.5	50.2
V	1	1	1	1	81.4	49.1

3.5.2　融合阶段配置策略研究

为探究本文所提出的双分支编码结构中的频域自适应加权模块的最佳设置策略, 本文在 M³FD 数据集上根据频域自适应加权模块的设置位置评估三种不同的融合策略, 如图 6 所示. 早期融合: 在进入主干编码器之前进行可见光和红外光的特征输入融合; 中期融合: 在双分支编码器之间不同阶段插入频域自适应加权模块, 以融合不同尺度的跨模态特征信息; 晚期融合: 来自可见光和红外光分支的特征仅在编码器末端, 即进入检测头的位置进行融合. 在上述实验中, 其他网络组件和参数配置保持一致.

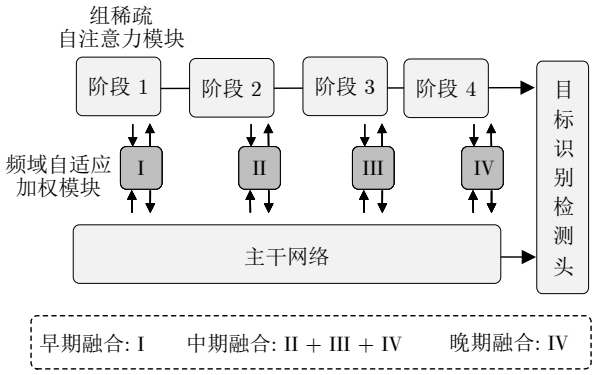


图 6　特征融合阶段示意图

Fig. 6　Schematic diagram of feature fusion stage

三种不同融合策略的定量结果如表 4 所示, 其中红色和蓝色数据分别表示最优和次优的结果. 中期融合策略在 mAP_{50} 和 $mAP_{50:95}$ 方面分别表现出最好的检测结果. 在参数量相近的情况下, 相较于早期融合策略和晚期融合策略在 mAP_{50} 方面分别提升了 3.4% 和 1.0%. 相较于没有融合模块的由组

表 4　不同融合阶段性能研究

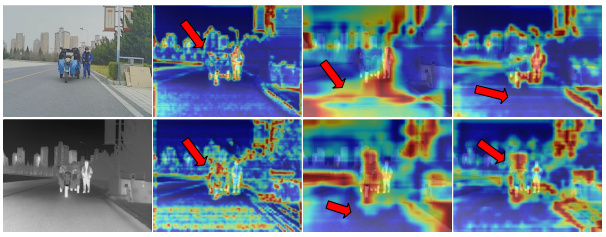
Table 4　Performance study at different fusion stages

融合策略	精度 (%)	召回率 (%)	mAP_{50} (%)	$mAP_{50:95}$ (%)	Params (M)
基线	85.0	75.1	81.2	47.7	7.12
早期融合	86.3	71.8	80.1	48.1	7.12
中期融合	86.7	77.3	83.5	51.2	7.44
晚期融合	86.8	75.2	82.5	49.1	7.39

稀疏自注意力模块构成的双编码器模型, 中期融合的配置仅增加了 0.32 M 的参数量. 这表明在多语义层面逐步融合, 能够更加有效地捕获多光谱特征图的跨模态信息, 从而提高模型的检测能力.

3.5.3 可视化研究

为了进一步研究所设计模块的效果, 在本文所提出模型的每个频域自适应加权模块前后的位置生成了注意力热图. 其中频域自适应加权模块之前的位置代表了来自不同光谱的特征编码块 (组稀疏自注意力模块) 的编码结果, 频域自适应加权模块之后的位置代表了其对多模态特征图进行融合表征的结果. 如图 7 所示, 列 (a) 是可见光和红外光原始输入图像, 列 (b)、(c)、(d) 上下两张图分别表示三个在模型中由浅入深位置的加权模块前后位置的注意力热图. 从横向比较来看, 随着特征图经过编码块多阶段的编码表征, 模型对图中大面积的冗余无关区域的注意力逐渐变得稀疏; 从纵向比较来看, 多光谱特征图经过融合编码, 模型更加聚焦于画面中的关键目标, 注意力变得更集中. 可以看出本文提出的组稀疏自注意力模块和频域自适应加权模块对模型注意力的正向引导, 促使模型对冗余区域注意力稀疏化, 对感兴趣区域注意力集中化, 从而提高了模型的检测能力.



(a) (b) (c) (d)

图 7 不同阶段注意力热力图

Fig. 7 Attention heatmaps across different stages

3.5.4 消融实验

为了量化分析每一个组件对模型整体检测性能的贡献, 本文在 M^{3}FD 数据集上进行了一系列消融研究, 结果如表 5 所示, 其中红色和蓝色数据分别表示最优和次优的结果. 本文选用了没有添加任何额外编码模块和融合模块的双分支 YOLOv5 编码器作为基线 (实验 I).

当单独采用组稀疏自注意力模块作为网络的编码器时 (实验 II), 模型在 mAP$_{50}$ 方面取得了 81.2% 的检测结果, 相较于基线提升了 1.0%. 模型的参数量虽然小有提升, 但相较于如 DAMSDet、CFT 等模型, 该提升属于可接受范围. 当在基线设置上单独添加频域自适应加权模块作为融合模块时 (实验 III), 在仅提升 0.36 M 参数量的情况下, 模型的检测精度有较大的提升, 在 mAP$_{50}$ 方面达到了 82.8%. 最后, 结合组稀疏自注意力模块和频域自适应加权模块 (实验 IV), 构成了完整模型, 取得了 83.5% 的 mAP$_{50}$, 相较于基线实现了 3.3% 的最大提升. 组稀疏自注意力模块在长距离语义编码过程中对特征信息进行区域分组筛选, 促使模型对冗余特征注意力稀疏化. 频域自适应加权模块作为多模态信息交互的桥梁, 其高效的频域感知能力促使模型更好地捕获关键的多光谱信息. 这样的消融实验结果表明了两大结构的协同效应, 验证了本文模型的优越性.

表 5 消融实验

Table 5 Ablation experiments

实验	组稀疏自注意力模块	频域自适应加权模块	mAP$_{50}$ (%)	mAP$_{50:95}$ (%)	Params (M)
I	×	×	80.2	47.8	2.84
II	√	×	81.2	47.7	7.12
III	×	√	82.8	49.5	3.20
IV	√	√	83.5	51.2	7.44

4 结束语

本文针对传统多光谱目标检测在复杂环境下的性能退化问题, 提出了 FGSFNet. 该方法通过引入 GSAM 实现长距离特征依赖的稀疏建模, 有效抑制冗余信息并增强显著目标的特征表达, 同时通过 FDAW 完成多光谱特征的细粒度融合与动态权重调整, 显著提升了模态互补性与跨域语义一致性. 在端到端的高效率框架下, FGSFNet 在保持较低计算复杂度的同时实现了可见光与红外光特征的高效融合和精准对齐. 实验结果显示, 在 M^{3}FD、KAIST 和 FLIR 三个公开数据集上, FGSFNet在 mAP$_{50}$ 和 mAP$_{50:95}$ 指标上均显著优于现有主流方法, 验证了所提模型在复杂环境下的有效性、鲁棒性及良好泛化能力.

尽管 FGSFNet 在检测精度和实时性方面取得了显著提升, 但仍存在进一步优化的空间. 未来可从模型轻量化和部署优化入手, 将重点探索借助大核卷积或深度可分离卷积等方法将频域自适应加权逻辑通过数学等效变换, 转换为 NPU 友好型空间域算子. 此外, 研究针对特定边缘端架构的算子融合与指令集优化技术, 以确保 FGSFNet 在多样化嵌入式平台上的高性能运行. 同时, 将静态图像检测拓展至时序或视频级检测, 结合跨帧特征关联与动态信息建模, 有望进一步提升多光谱目标跟踪与

连续检测能力. 此外, 本文框架可扩展至更多传感模态, 如毫米波雷达或激光雷达, 探索自适应模态选择与动态权重分配机制, 以增强多源数据下的泛化性能. 在智能交通、无人安防及工业检测等实际工程场景中, FGSFNet 具备潜在应用价值, 可结合场景特征进行结构优化, 实现高可靠的实时目标检测和预警.

参考文献

1 Mao J G, Shi S S, Wang X G, Li H S. 3D object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 2023, **131**(8): 1909−1963

2 Shi X L, Song A J. Defog YOLO for road object detection in foggy weather. *The Computer Journal*, 2024, **67**(11): 3115−3127

3 Wang J, Yang P, Liu Y S, Shang D, Hui X, Song J H, et al. Research on improved YOLOv5 for low-light environment object detection. *Electronics*, 2023, **12**(14): Article No. 3089

4 Yan Meng-Kai, Qian Jian-Jun, Yang Jian. Weakly aligned cross-spectral face detection. *Acta Automatica Sinica*, 2023, **49**(1): 135−147
(闫梦凯, 钱建军, 杨健. 弱对齐的跨光谱人脸检测. 自动化学报, 2023, **49**(1): 135−147)

5 Zhao Xing-Ke, Li Ming-Lei, Zhang Gong, Li Ning, Li Jia-Song. Object detection method based on saliency map fusion for UAV-borne thermal images. *Acta Automatica Sinica*, 2021, **47**(9): 2120−2131
(赵兴科, 李明磊, 张弓, 黎宁, 李家松. 基于显著图融合的无人机载热红外图像目标检测方法. 自动化学报, 2021, **47**(9): 2120−2131)

6 Hu S M, Zhao F, Lu H Z, Deng Y J, Du J M, Shen X L. Improving YOLOv7-tiny for infrared and visible light image object detection on drones. *Remote Sensing*, 2023, **15**(13): Article No. 3214

7 Wang P, Wu J S, Fang A Q, Zhu Z X, Wang C W. Multi-spectral image fusion for moving object detection. *Infrared Physics & Technology*, 2024, **141**: Article No. 105489

8 He X, Tang C, Zou X, Zhang W. Multispectral object detection via cross-modal conflict-aware learning. In: Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM, 2023. 1465−1474

9 Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 580−587

10 Zhang Peng, Lei Wei-Min, Zhao Xin-Lei, Dong Li-Jia, Lin Zhao-Nan, Jing Qing-Yang. A survey on multi-target multi-camera tracking methods. *Chinese Journal of Computers*, 2024, **47**(2): 287−309
(张鹏, 雷为民, 赵新蕾, 董力嘉, 林兆楠, 景庆阳. 跨摄像头多目标跟踪方法综述. 计算机学报, 2024, **47**(2): 287−309)

11 Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 779−788

12 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv preprint arXiv: 1804.02767, 2018.

13 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 6000−6010

14 Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2020.

15 Chen Y T, Shi J H, Ye Z L, Mertz C, Ramanan D, Kong S. Multimodal object detection via probabilistic ensembling. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 139−158

16 Xu H, Ma J Y, Jiang J J, Guo X J, Ling H B. U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(1): 502−518

17 Liu J Y, Fan X, Huang Z B, Wu G Y, Liu R S, Zhong W, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 5792−5801

18 Liu J Y, Fan X, Jiang J, Liu R S, Luo Z X. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, **32**(1): 105−119

19 Fang Q Y, Han D P, Wang Z K. Cross-modality fusion Transformer for multispectral object detection. arXiv preprint arXiv: 2111.00273, 2021.

20 Shen J F, Chen Y F, Liu Y, Zuo X, Fan H, Yang W K. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 2024, **145**: Article No. 109913

21 Guo J J, Gao C Q, Liu F C, Meng D Y, Gao X B. DAMSDet: Dynamic adaptive multispectral detection Transformer with competitive query selection and adaptive feature fusion. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2024. 464−481

22 Zhu X Z, Su W J, Lu L W, Li B, Wang X G, Dai J F. Deformable DETR: Deformable Transformers for end-to-end object detection. In: Proceedings of the 9th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2021.

23 Xiao G B, Tang Z M, Guo H L, Yu J, Shen H T. FAFusion: Learning for infrared and visible image fusion via frequency awareness. *IEEE Transactions on Instrumentation and Measurement*, 2024, **73**: Article No. 5015011

24 Ma J Y, Ma Y, Li C. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 2019, **45**: 153−178

25 Hwang S, Park J, Kim N, Choi Y, Kweon I S. Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 1037−1045

26 Zhang H, Fromont E, Lefevre S, Avignon B. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). Abu Dhabi, United Arab Emirates: IEEE, 2020. 276−280

27 Li H, Wu X J, Kittler J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 2021, **73**: 72−86

28 Liu J Y, Shang J J, Liu R S, Fan X. Halder: Hierarchical attention-guided learning with detail-refinement for multi-exposure image fusion. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Shenzhen, China: IEEE, 2021. 1−6

29 Chen Y S, Wang B R, Guo X Y, Zhu W B, He J S, Liu X B, et al. DEYOLO: Dual-feature-enhancement YOLO for cross-modality object detection. In: Proceedings of the 27th International Conference on Pattern Recognition. Kolkata, India: Springer, 2024. 236−252

30 Zhang J Q, Lei J, Xie W Y, Fang Z M, Li Y S, Du Q. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, **61**: Article No. 5605415

肖海东 广东工业大学自动化学院硕士研究生. 主要研究方向为目标检测与计算机视觉.
E-mail: 2112304427@mail2.gdut.edu.cn
(**XIAO Hai-Dong** Master student at the School of Automation, Guangdong University of Technology. His research interests include object detection and computer vision.)

任志刚 广东工业大学自动化学院教授. 主要研究方向为控制科学与工程, 模式识别. 本文通信作者.
E-mail: renzhigang@gdut.edu.cn
(**REN Zhi-Gang** Professor at the School of Automation, Guangdong University of Technology. His research interests include control science and engineering, and pattern recognition. Corresponding author of this paper.)

蔡声泽 浙江大学控制科学与工程学院教授. 主要研究方向为智能控制与机器学习.
E-mail: shengze_cai@zju.edu.cn
(**CAI Sheng-Ze** Professor at the College of Control Science and Engineering, Zhejiang University. His research interests include intelligent control and machine learning.)

许 超 浙江大学控制科学与工程学院教授, 湖州研究院院长. 主要研究方向为深度学习的控制论框架, 机器人世界模型.
E-mail: cxu@zju.edu.cn
(**XU Chao** Professor at the College of Control Science and Engineering, dean at Huzhou Institute, Zhejiang University. His research interests include cybernetic architectures for deep learning and robot world models.)

吴宗泽 深圳大学机电与控制工程学院教授. 主要研究方向为智能制造, 工业互联网, 大数据分析与深度学习.
E-mail: zzwu@gdut.edu.cn
(**WU Zong-Ze** Professor at the College of Mechatronics and Control Engineering, Shenzhen University. His research interests include intelligent manufacturing, industrial Internet, big data analysis, and deep learning.)