

可见光和红外特征自适应融合的多模态目标检测方法

喻智睿, 尹展鹏, 王俊宇, 周亮, 叶沅鑫

西南交通大学 地球科学与工程学院, 成都 611756

摘要: 针对可见光目标检测在被遮挡、弱光等复杂环境下, 特征丰富度易受影响, 导致目标检测准确性降低的问题, 本文引入红外模态影像来弥补可见光影像的不足, 提出红外与可见光特征自适应融合的多模态目标检测方法。该方法采用YOLOv8目标检测框架作为基础网络提取多尺度特征信息; 在此基础上, 基于可见光影像拥有更丰富的纹理特征而红外影像自身的边缘轮廓较纹理更明显的特点, 构建多模态混合注意力模块, 进行跨模态的信息权重交流和重组以实现不同光照条件下的优势特征; 然后, 利用可见光模态特征丰富度与环境光照强度的关系, 设计了以环境光照强度为指标的可见光—红外权重动态分配模块, 并将权重作为参考, 纳入多模态特征融合模块进行自适应融合, 实现基于多模态特征融合的目标检测。最后, 通过在公开的街景数据集M³FD及航拍车辆数据集DroneVehicle上进行实验。结果表明, 相较于现有的单模态与多模态目标检测算法, 本文提出方法能够获得更高的检测精度。

关键词: 目标检测, 多模态, 卷积神经网络, 特征融合, 注意力机制, 可见光影像, 红外影像, 深度学习

中图分类号: P237/P2

引用格式: 喻智睿, 尹展鹏, 王俊宇, 周亮, 叶沅鑫. 2025. 可见光和红外特征自适应融合的多模态目标检测方法. 遥感学报, 29(10): 3006-3019

Yu Z R, Yin Z P, Wang J Y, Zhou L and Ye Y X. 2025. Multimodal object detection method using adaptive fusion of infrared and visible features. National Remote Sensing Bulletin, 29(10): 3006-3019 [DOI: 10.11834/jrs.20254358]

1 引言

随着人工智能的发展和硬件水平的提升, 基于深度学习的目标检测技术成为了遥感领域的焦点 (Dong 等, 2021; 罗会兰和陈鸿坤, 2020)。然而, 目前大多数目标检测方法主要基于可见光图像 (张顺 等, 2019), 其利用目标反射的环境光源成像, 虽然其能够充分展现目标的细节与纹理特征, 但在目标模糊、遮挡、光照过强或过暗等情况下, 仅靠可见光传感器难以充分获取目标信息, 从而影响目标检测效果 (Zuo 等, 2022)。相比之下, 红外传感器具有抗环境干扰能力强、对光照不敏感、能够反映目标温度等优点, 能够弥补可见光成像的不足, 因此将红外特征融入基于可见光的目标检测, 研究多传感器信息协同作用的目

标检测算法, 不仅能提升目标检测准确性和鲁棒性, 还能更好地适用于实际场景 (Liu 等, 2018)。然而, 红外影像自身也存在分辨率较低、目标细节不清晰等缺点, 在环境光丰富的条件下检测精度不如可见光影像 (Zhang 等, 2023c; 贾明东 等, 2023)。因此, 如何在根据可见光与红外模态的互补特性进行自适应融合, 是当前多模态目标检测任务研究的热点和难点 (Osco 等, 2021)。随着深度学习技术的发展及计算机性能的不断突破, 许多研究者开始将深度学习引入多模态目标检测任务中, 并取得了较好的效果 (Yang 等, 2022)。Fang 等 (2022) 首次将Transformer应用到多光谱目标检测领域, 有效地整合了不同模态之间的互补特征, 但其引入的Transformer带来了巨大的参数量和计算量; Liu 等 (2022a) 将生成对抗网络

收稿日期: 2024-09-04; 预印本: 2025-04-23

基金项目: 国家重点研发计划 (编号: 2024YFC3015404); 国家自然科学基金 (编号: 42271446); 四川省中央引导地方科技发展专项 (编号: 2024ZYD0084)

第一作者简介: 喻智睿, 研究方向为多模态图像处理与分析。E-mail: 2669521041@qq.com

通信作者简介: 叶沅鑫, 研究方向为多模态遥感图像处理与分析。E-mail: yeyuanxin110@163.com

GAN (Generative Adversarial Network) 引入多模态目标检测任务, 通过学习可见光和红外影像之间的差异与共同点保留目标的轮廓和纹理信息, 但 GAN 在生成过程中可能出现一部分目标的失真, 且在环境光强度低时表现较差; Zhang 等 (2023a) 考虑到遥感场景下小目标多的特点, 提出的卷积神经网络采用了超分辨率的思想, 以高分辨率遥感图像进行训练, 推理过程只使用原始低分辨率图像, 但其采用在骨干网络之前进行多模态融合的方式使目标的高层级特征不能被有效提取; Li 等 (2024) 采用多尺度跨模态同质增强和置信感知特征融合以提取多模态特征, 结合多尺度跨模态同质增强、通道注意力机制和置信方法实现多模态特征的融合和目标检测; Zhao 等 (2024) 在骨干网络前设置基于频率域的冗余频谱去除模块, 并在骨干网络后采用动态特征选择模块促进不同尺度特征的互补融合, 以实现可见光和红外影像特征的由粗到精融合。然而, 这些方法往往没有考虑不同场景下可见光与红外影像的特征优势, 忽略了各模态之间的信息特征差异所导致的特征模态不平衡问题, 而是直接在维度上将两者进行融合, 不能够充分发挥可见光与红外影像在不同环境下的互补优势 (孙颖等, 2023)。

因此, 本研究针对上述红外—可见光目标检测方法存在的不足, 并结合复杂环境下可见光与红外影像的特征丰富度不同的特点, 提出了一种基于可见光与红外多模态图像自适应融合的目标检测算法, 并与先进的目标检测算法进行比较, 通过在公开的街景数据集 M³FD 及航拍车辆数据集 DroneVehicle 上进行实验, 以期当前可见光和红外多模态目标检测方法提供一种新思路。

2 MAF-YOLO 目标检测算法

2.1 网络模型

本文所提出方法的整体结构如图 1 所示, 该网络在 YOLOv8 网络 (Terven 和 Cordova-Esparza, 2024) 的基础上提出了一种多模态自适应融合 MAF-YOLO (Multimodal Adaptive Fusion-YOLO) 的目标检测网络, 实现了多模态图像特征的自适应提取与融合, 解决了复杂环境下可见光图像目标检测精度低下的问题。MAF-YOLO 网络由骨干网络、颈部与检测头组成:

(1) 骨干网络部分主要由双模态特征提取网络、跨模态注意力和特征自适应融合 3 部分组成。MAF-YOLO 的双模态特征提取网络以 YOLOv8 网络为基础, 通过多个 CBS 模块 (Conv+BN+SiLU) 及 C2f 模块 (CSP Bottleneck with 2 Conv) 提取多模态影像的多尺度特征, 并采用快速空间金字塔池化 SPPF (Spatial Pyramid Pooling Fast) 将不同尺度的特征图拼接在一起, 提高对不同尺寸目标的检测能力 (Li 等, 2023)。在此基础上, 本文构建跨模态混合注意力模块 (CHAM) 以有针对性地提取可见光与红外影像共同及差异的特征, 对双模态的空间和通道特征进行重组, 并设计特征自适应融合结构 (CAFM 和 IAM) 结合自主学习与环境光强度监督, 对可见光与红外模态的互补特征进行自主选择融合, 以减少模态特征不平衡问题的影响, 实现高性能的跨模态特征自适应特征提取与融合。

(2) 颈部网络主要由 CBS、C2f 模块以及一系列的上采样和拼接操作组成, 将来自不同阶段的特征图进行融合, 增强特征表示能力, 以实现尺度特征融合 (Hussain, 2023)。

(3) 检测头分别采用分类和边界回归损失, 避免了重复计算并减少了检测时可能出现的干扰, 进而提高了分类和定位的精度, 实现对不同目标的准确检测。

2.2 跨模态混合注意力

对于多模态目标检测, 不同模态的图像通常具有不同的特征 (Liu 等, 2022b)。如白天的可见光图像具有丰富的细节和纹理特征, 但夜间的可见光图像质量较差 (Tang 等, 2023)。另一方面, 白天强烈的阳光会对红外图像的成像造成干扰, 但夜间拍摄的红外图像通常能呈现出清晰的目标轮廓和形状特征 (Simonyan 和 Zisserman, 2015)。此外, 在现实情况下, 模态之间仍然存在信息间的差异。例如, 可见光图像中的目标在光照不足和被遮挡的情况下可能会丢失一些特征, 从而干扰分类和定位, 目标在单个图像中可能因光照不均而呈现出与实际形状不同的轮廓等 (Zhang 等, 2020)。

为充分对影像特征进行分析, 使得网络更加聚焦于信息丰富的部分, 注意力机制被广泛应用, 其中空间和通道注意力分别从特征的 3 个维度对其进行全面的分析, 因此其思想也被广泛地应用于多种不同的注意力模块中, 如卷积注意力模块、坐

标注意力等 (Shen等, 2024)。但是常用的单模态注意力机制无法根据不同模态的特征进行提取和分

析, 使其难以被直接应用于多模态目标检测任务中。

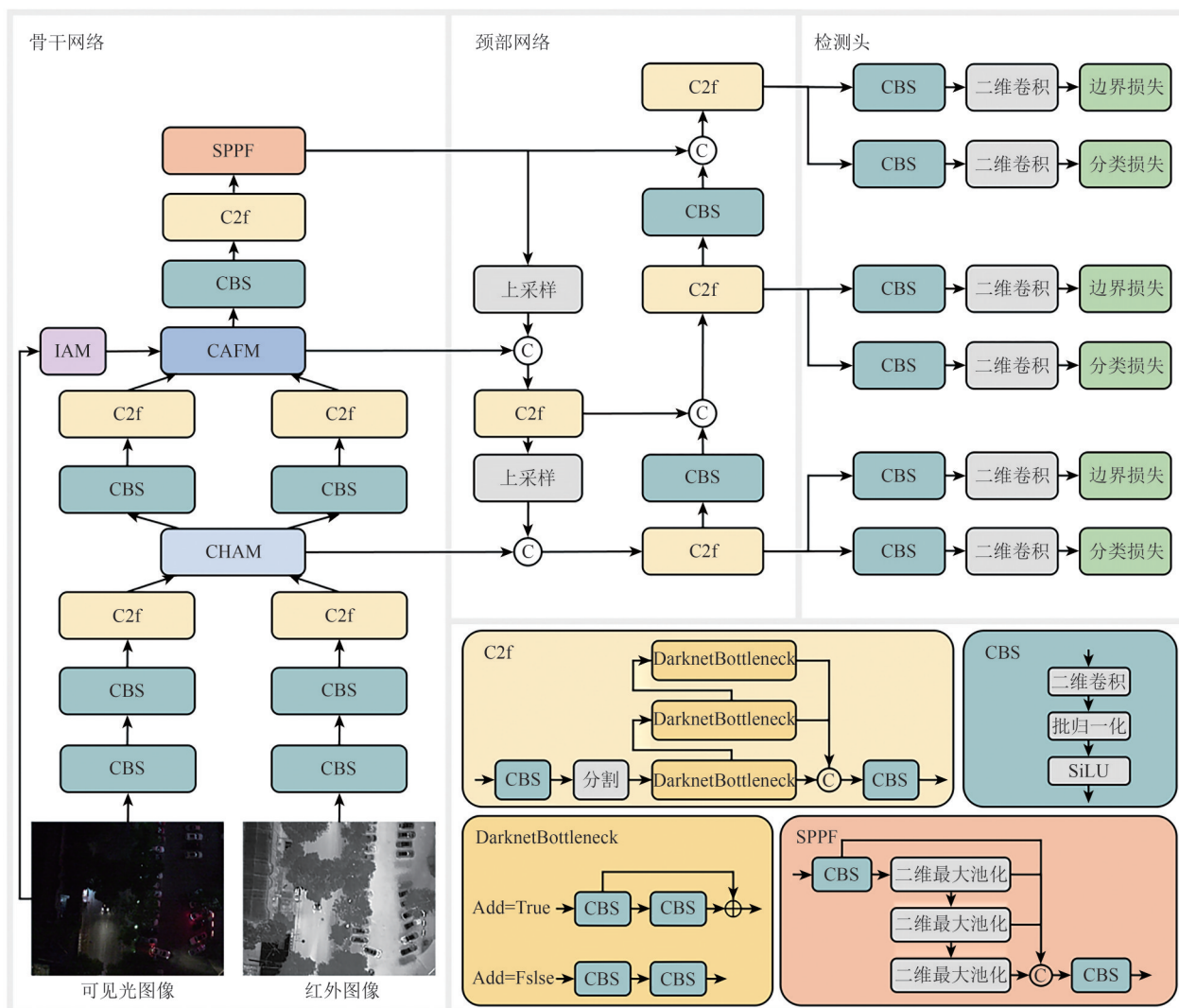


图1 MAF-YOLO网络结构

Fig. 1 The architecture of MAF-YOLO network

为了解决上述挑战, 本文提出了一种跨模态混合注意力模块CHAM (Cross-modal Hybrid Attention Module), 通过自适应优化可见光和红外特征在通道和空间上的权重, 加强多模态特征中信息更丰富的部分, 同时抑制信息较少的部分, 从而最大程度地利用了可见光和红外模态的特征。CHAM的组成如图2所示。该模块由两个线性相连的子模块组成, 分别为通道注意力子模块和空间注意力子模块, 这使得可见光和红外特征提取网络都可以分别在通道和空间轴上学习, 同时通过对两个模态的信息进行交互, 提取共同特征并弥补不足 (Woo等, 2018)。

2.2.1 通道注意力子模块

CHAM的通道注意力子模块的结构如图3所示。可见该模块利用特征的通道间关系生成了一个通道注意力图, 为了有效地计算通道注意力, 本算法压缩了输入特征映射的空间维度。对于空间信息的聚合, 目前普遍采用的有最大池化和平均池化。最大池化保留图像中最显著的特征, 适于提取边缘轮廓等细节部分; 而平均池化计算特征均值, 保留更多的整体信息。由于红外图像更多地体现了对象的边缘轮廓特征, 而可见光图像中对象的纹理和色彩能更好地被展示, 为了充分提取两种模态各自的特征, 本文提出方法使用平均池化提取可见光特征, 而使用最大池化提取红外特征。

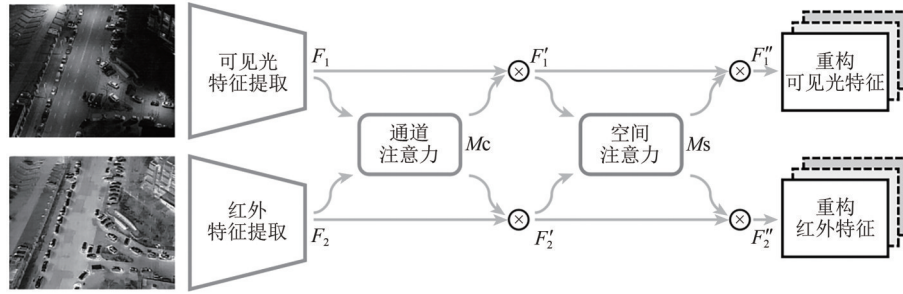


图2 跨模态混合注意力模块

Fig. 2 Cross-modal Hybrid Attention Module(CHAM)

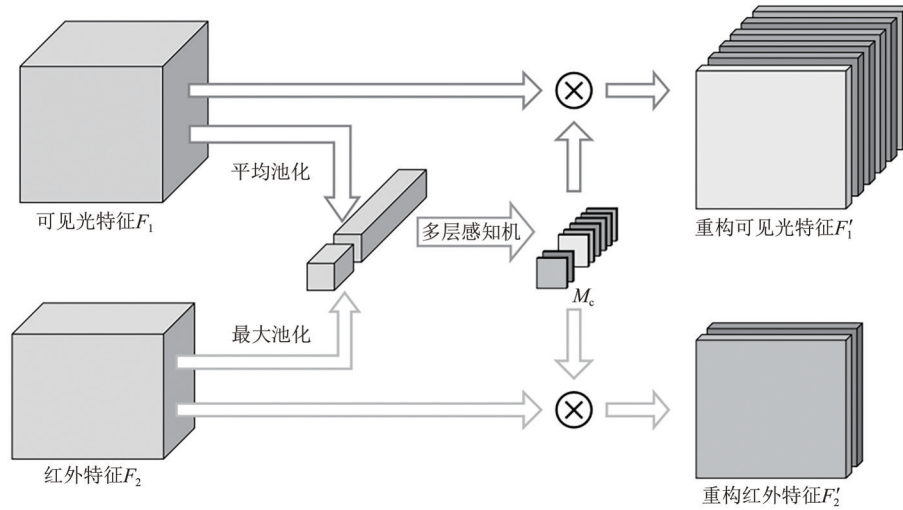


图3 通道注意力子模块

Fig. 3 Channel attention sub-model of CHAM

对于输入的可见光特征 $F_1 \in \mathbb{R}^{3 \times C \times W \times H}$ 和红外特征 $F_2 \in \mathbb{R}^{C \times W \times H}$, 本模块首先通过平均池化和最大池化对两个模态特征的空间特征进行聚合, 生成两个不同空间特征描述符。然后, 这两个描述符被拼接并由多层感知机 MLP (Multi-Layer Perceptron) 生成通道注意力图 $M_c \in \mathbb{R}^{4 \times 1 \times 1}$ 。即通道注意力计算为

$$M_c = \sigma \left(\text{MLP} \left(\text{AvgPool}(F_1) \right) - \text{MLP} \left(\text{MaxPool}(F_2) \right) \right) \quad (1)$$

式中, σ 为 Sigmoid 激活函数, F_1 和 F_2 分别为可见光和红外特征。最终生成的通道注意力参数与原始的可见光特征与红外特征进行逐通道相乘, 得到经过重构的可见光特征 $F'_1 \in \mathbb{R}^{3 \times C \times W \times H}$ 和红外特征 $F'_2 \in \mathbb{R}^{C \times W \times H}$ 。

2.2.2 空间注意力子模块

与通道注意力不同, 空间注意力关注特征丰富度在空间上的分布。如图4所示, 为了计算空间注意力, 本模块首先沿通道轴对可见光和红外特征分别进行平均池化和最大池化, 然后将其连接生成一个高度抽象的特征描述符, 以突出信息丰富的区域。在连接的特征描述符上, 本模块采用

7×7 的卷积层来生成一个空间注意力图 $M_s \in \mathbb{R}^{2 \times H \times W}$, 它包含了需要强调或抑制的位置信息。

具体来说, 空间注意力子模块对可见光和红外模态的重构特征 F'_1 、 F'_2 分别采用平均池化和最大池化聚合两者的通道信息, 得到两个 $\mathbb{R}^{1 \times H \times W}$ 的特征图, 然后将这两个特征图基于通道进行拼接, 通过卷积操作将其降维为 $\mathbb{R}^{1 \times H \times W}$ 。再经过 Sigmoid 函数激活生成空间注意力特征 M_s 。最后将该特征和原始的输入特征到经过重构的可见光特征 $F'_1 \in \mathbb{R}^{3 \times C \times W \times H}$ 和红外特征 $F'_2 \in \mathbb{R}^{C \times W \times H}$ 。做矩阵点乘, 得到最终生成的注意力特征 F''_1 和 F''_2 。

2.3 基于光照感知的跨模态自适应融合

由于可见光影像中的特征丰富度往往与环境光照强度呈现正相关, 许多方法都根据环境光照强度分配可见光与红外影像的融合权重, 但现有方法往往采用在检测网络中添加一个用于区分昼夜的二分类子网络来实现 (Tang等, 2022; Zhang等, 2023c), 这不仅过于简化了光照强度的受影响因素,

将线性的光照强度变化简单地划分为白天和

黑夜, 而且要求训练数据新增昼夜标签进行监督学习, 大大增加了数据集制作与网络训练成本。对此, 本文提出了一种简单有效的光照强度评估

模块, 通过评估可见光图像的亮度和饱和度以衡量可见光分支特征的融合权重, 将融合权重纳入跨模态自适应融合模块指导特征权重的分配。

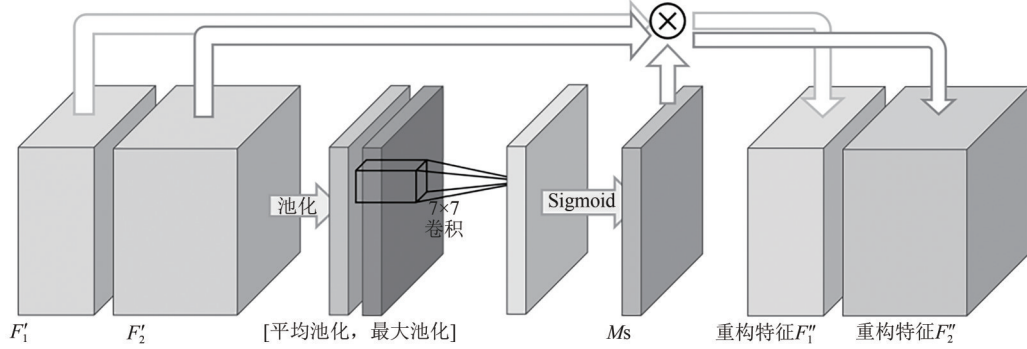


图4 空间注意力子模块

Fig. 4 Spatial Attention of CHAM

由于红外图像的特征丰富程度多受热源数量与强度影响, 难以反映环境照明条件, 因此本文仅采用可见光图像为输入, 设计了一个简单有效的光照感知模块IAM (Illumination Aware Modul), 用于获取可见光影像的饱和度与亮度信息, 并根据其综合计算光照强度指数 ω_{illu} , 用于在后续环节指导融合的进行。IAM模块的流程如图5所示。

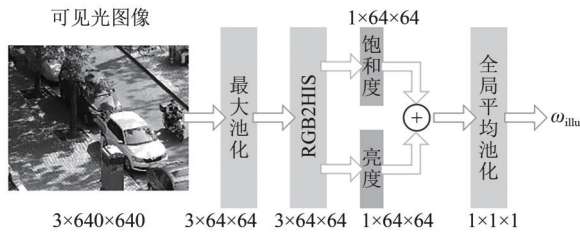


图5 光照感知模块

Fig. 5 Illumination Aware Module(IAM)

首先, 为了降低计算复杂度, 原始尺寸为640×640的可见光图像首先通过最大池化被降采样为64×64, 并将RGB信息转换为HSI信息以得到输入影像的饱和度 (Saturation) 与亮度 (Intensity)。由于饱和度信息能够较好地反映可见光图像的色彩丰富程度, 而亮度信息能够较好地反映环境光照强度, 这两个参数被共同用于计算可见光的光照强度指导后续融合环节可见光和红外两个模态各自的权重。具体来说, 光照感知模块最终获取环境光照强度权重 ω_{illu} , 并将该值归一化至 $[0, 1]$ 之间, 后续输入特征融合模块, 与可见光和红外的特征丰富度权重共同得到综合权重, 对两个模态的特征融合权重进行分配。

为充分结合可见光与红外模态的特征, 以确定优势模态信息并进行自适应融合, 本文提出了一种跨模态自适应融合模块CAFM (Cross-modal Adaptive Fusion Module), 该模块根据特征质量的动态比较自适应地选择红外特征与可见光特征的权重进行融合。如图6所示, CAFM模块中分别存在3条并行的分支, 分别为可见光特征提取模块、红外特征提取模块和跨模态权重分配模块。在可见光特征注意力模块和红外特征提取模块中, 坐标注意力 (Hou 等, 2021) 通过沿水平和垂直空间方向聚合特征, 将通道注意力分解为两个一维特征编码过程, 以在保持精确位置信息的同时捕捉长程依赖关系。这使得两个模态各自的特征信息得到凸显, 并通过将处理后的结果与原始特征逐层相加, 经过一个标准的卷积模块后采用残差结构与原始特征相加, 保证了信息的流动并加速网络的训练与收敛。在跨模态权重分配模块中, 首先将输入的可见光模态特征 F_{RGB} 与红外模态特征 F_{IR} 进行连接得到特征 F , 将特征经过多层感知机得到可见光特征的融合权重 ω_1 和红外特征的融合权重 ω_2 。同时, 引入此前IAW模块提取的可见光模态光照权重 ω_{illu} , 与可见光特征融合权重 ω_1 相加, 而 $1-\omega_{\text{illu}}$ 被视为红外模态的光照强度权重, 与红外特征融合权重 ω_2 相加。

令处理后的可见光特征为 F'_{RGB} , 处理后的红外特征为 F'_{IR} , 融合后的特征为 F_{fusion} , 则自适应融合公式为

$$F_{\text{fusion}} = ((\omega_1 + \omega_{\text{illu}}) F'_{\text{RGB}}; (\omega_2 + (1 - \omega_{\text{illu}})) F'_{\text{IR}}) \quad (2)$$

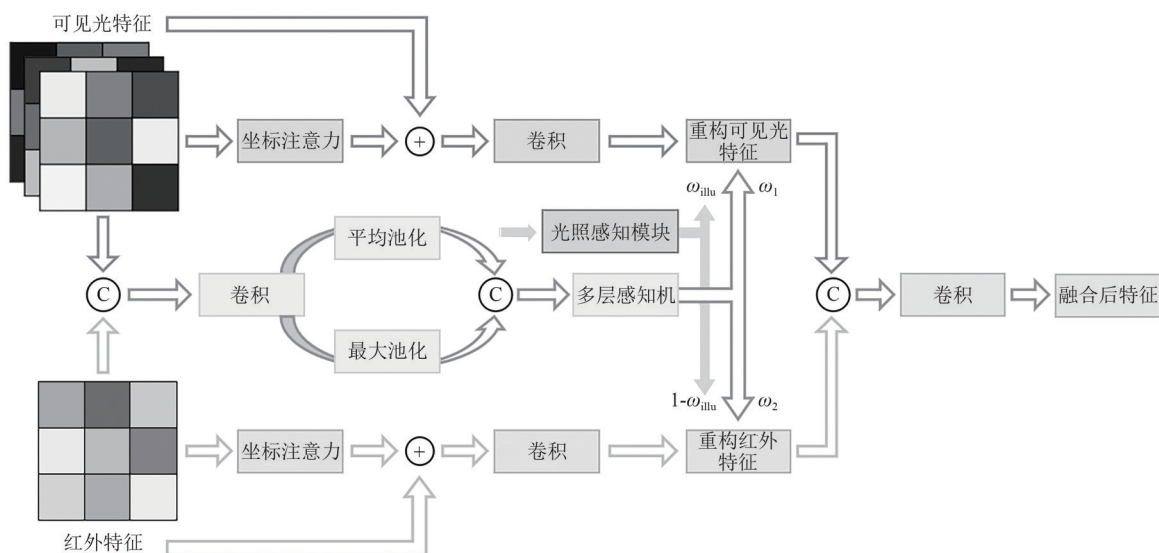


图6 跨模态自适应融合模块

Fig. 6 Cross-modal Adaptive Fusion Module(CAFM)

2.4 损失函数

将上述构建的 CHAM, CAFM 和 IAM 嵌入 MAF-YOLO 的骨干网络中, 进行多模态特征提取和融合, 再通过颈部网络进行多尺度特征融合, 并将其输入到检测头网络进行目标检测。为提高检测的准确性并加快模型的收敛速度, MAF-YOLO 的检测头网络由目标分类和边界框回归两条独立的分支组成。因此本算法使用融合损失函数记为 L_{total} , 对于分类任务, 本算法使用焦点损失 L_F , 而对于预测边界框的回归, 使用完备交并比损失 L_{CioU} 和分布焦点损失 L_{DFL} 。完整的损失函数可以被表述为

$$L_{total} = \lambda_1 L_F + \lambda_2 L_{CioU} + \lambda_3 L_{DFL} \quad (3)$$

式中, λ_1 、 λ_2 与 λ_3 为用于平衡不同损失之间的系数。

在本文中, L_F 通过调整易分类样本和难分类样本的权重, 减少多数类样本对损失的影响, 进而解决类别不平衡问题的影响; L_{CioU} 在交并比损失 L_{IoU} 的基础上弥补了两者不相交时 L_{IoU} 无法被计算阻碍反向传播的不足; L_{DFL} 为空间分布的识别提供了一个方向, 通过增强训练特征的有效性来保留详细的空间信息, 进而提高整体性能。

3 实验设计与结果分析

3.1 数据集

本文共选用两个公开的红外—可见光多模态数据集对本文所提出算法进行评估, 分别为 M³FD

数据集和 Drone Vehicle 数据集, 在实验中数据集被随机按照 7 : 2 : 1 的比例随机抽选并划分为训练集、验证集和测试集。

(1) M³FD 数据集 (Liu 等, 2022a): 该数据集采用自制的可见光—红外同步成像系统获取可见光和热红外影像, 其数据集包含 4200 对图像, 且图像对经过了几何校正以实现严格的时空对齐。M³FD 数据集中的图像对包含 4 种典型时间场景, 分别为白天、阴天、夜晚和具有挑战性的场景 (如烟雾和景观遮挡)。在 M³FD 数据集中标注了 6 种类型的对象, 分别为人、轿车、公交车、汽车、摩托车、卡车和灯 (通常为红绿灯)。

(2) DroneVehicle 数据集 (Sun 等, 2022): 该数据集采用搭载 Zenmuse XT 2 相机的大疆 M200 无人机拍摄, 数据集中包括 28439 对像素尺寸为 640×512 的可见光—红外影像对, 覆盖了不同的物体和场景 (如城市道路, 农村地区, 居民区, 停车场和高速公路等)。数据集中涵盖 5 类目标, 分别是轿车, 卡车, 公交车, 面包车和货车。在 DroneVehicle 数据集的采集过程中设置了不同的高度、角度和光照条件来模拟现实中数据的分布, 使得数据集中的目标特征更加丰富, 这使得该数据集更具挑战性, 以帮助研究人员应对现实中的各种复杂条件。

3.2 实验设置及评价指标

本文实验硬件 CPU 型号为 Intel (R) Core (TM) i7-12700KF, GPU 为 NVIDIA GeForce RTX 3090 显

卡，显存大小为 24 GB。对于软件环境，本实验的网络模型采用 Python 3.9.18 和 CUDA 11.7。本实验设置输入图像尺寸为 640×640，输入可见光图像为三波段，红外图像为单波段，采用 SGD 优化器，设置 Batch Size 为 16，共进行 300 轮训练，初始学习率为 0.01，权值衰减为 0.0005。

为更好地评价本文提出检测方法的表现，实验采用平均精度 AP（Average Precision）和平均精度均值 mAP（mean Average Precision）作为评价指标。AP 结合精确率和召回率之间的关系，以更好地评估模型的性能，更高的 AP 值意味着模型在不同阈值下能保持较高的精确率和召回率，说明模型在分类任务中表现更好，综合能力更强。为了综合考量模型对多种目标类型的检测性能，mAP 对所有目标检测结果类别的 AP 值取平均，以评估模型在所有类别上的整体性能（Padilla 等，2021）。

3.3 对比实验及分析

为了充分展示本文所提出 MAF-YOLO 模型的有效性，多个经典的目标检测模型和当前先进的检测方法被用于与 MAF-YOLO 进行对比实验。其中，选用的经典目标检测模型包括 Faster R-CNN（Ren 等，2015），YOLOv8（Sohan 等，2024）及 YOLOv10（Wang 等，2024），由于它们都是针对单一模态影像的目标检测模型，本文分别在可见光和红外模态上进行了实验；当前先进的模型包括首次将 Transformer 应用到多光谱目标检测

领域的 CFT（Feng 等，2022）；使用生成对抗网络生成融合影像的 TarDAL（Liu 等，2022a）；基于 YOLOv5 实现超分辨率重构，提升多模态目标检测精度的 SuperYOLO（Zhang 等，2023c）；采用多尺度跨模态同质增强和置信感知特征融合的 MCHE-CF（Li 等，2024）；以及采用冗余频谱去除及动态特征选择实现多模态特征实现由粗到精融合的 RsDet（Zhao 等，2024）。对比实验在 M³FD 街景数据集和 DroneVehicle 航拍车辆数据集上进行，以测试本文方法在不同光照环境和场景下的稳健性。

3.3.1 基于 M³FD 数据集的对比实验

为了验证本文所提方法的有效性，本文以 M³FD 数据集中的 6 种目标类型（人、轿车、巴士、摩托、灯、卡车）的 AP 和 mAP 为指标，对多种目标检测模型的结果进行评估。表 1 展示了各个目标检测模型的对比实验结果。根据实验结果可知，由于该数据集涵盖较多低光照场景下的目标，采用红外模态的单模态目标检测模型相较于可见光模态下的目标检测模型有着更加高的准确性，而由于多模态融合的目标检测模型能够提取不同模态之间的互补特征，其准确性相对单模态的目标检测方法更高。其中，MAF-YOLO 在 RGB+IR 模态下的 mAP 为 84.2%，在所有对比方法之中取得了最高的分数，表明该方法在检测目标时取得了相对最好的综合性能。

表 1 基于 M³FD 数据集的对比实验
Table 1 Comparison of different methods based on M³FD dataset

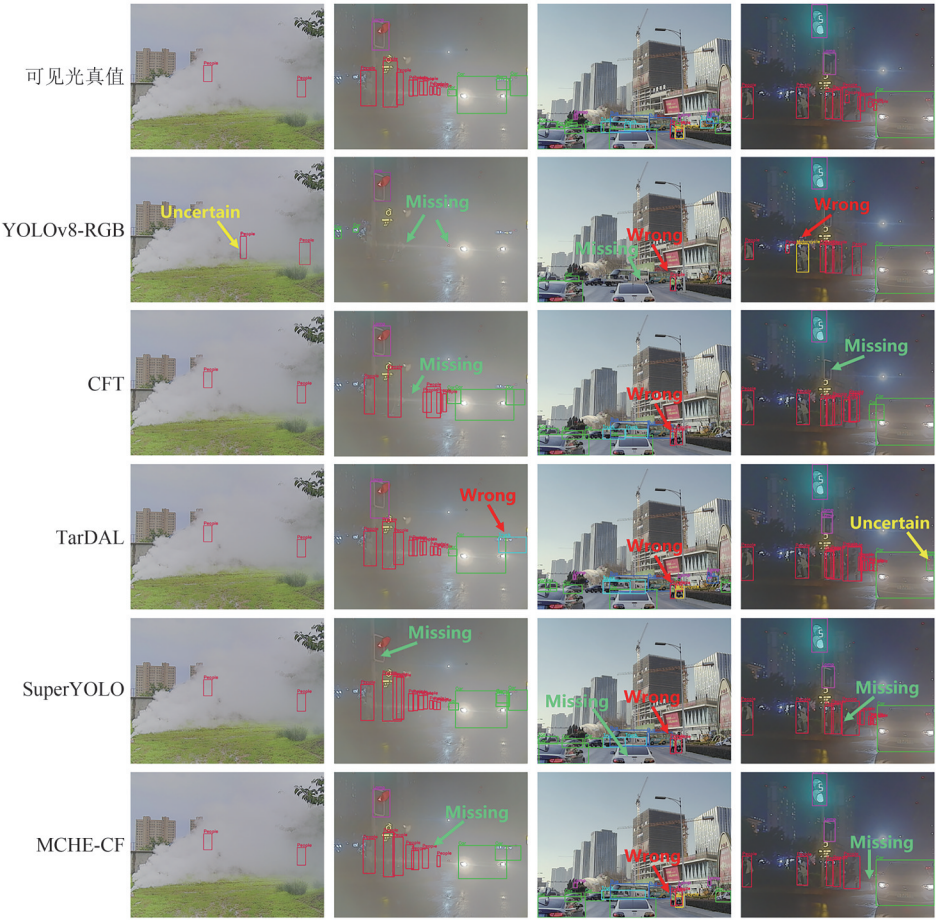
/%								
模型	数据模态	AP _{person}	AP _{car}	AP _{bus}	AP _{motorcycle}	AP _{lamp}	AP _{lorry}	mAP
Faster R-CNN	RGB	48.4	75.3	81.8	55.2	54.0	63.7	63.1
	IR	61.3	84.0	83.8	62.2	69.7	59.8	70.1
YOLOv8	RGB	50.4	82.9	83.4	48.1	58.3	68.6	65.3
	IR	66.7	82.9	85.1	66.6	72.7	72.3	73.0
YOLOv10	RGB	56.4	77.6	80.3	47.0	64.1	70.6	66.0
	IR	61.1	79.3	83.7	70.4	73.8	69.0	72.9
CFT	RGB+IR	82.8	76.3	90.9	82.2	80.6	81.0	82.3
TarDAL	RGB+IR	75.8	93.3	86.3	76.7	84.3	74.4	81.3
SuperYOLO	RGB+IR	80.4	88.6	87.5	81.1	83.4	78.9	83.3
MCHE-CF	RGB+IR	81.6	82.6	75.4	77.1	76.4	82.8	79.3
RsDet	RGB+IR	78.3	84.9	79.7	80.1	82.9	75.0	80.2
本文方法	RGB+IR	83.1	92.8	91.1	77.5	84.9	75.6	84.2

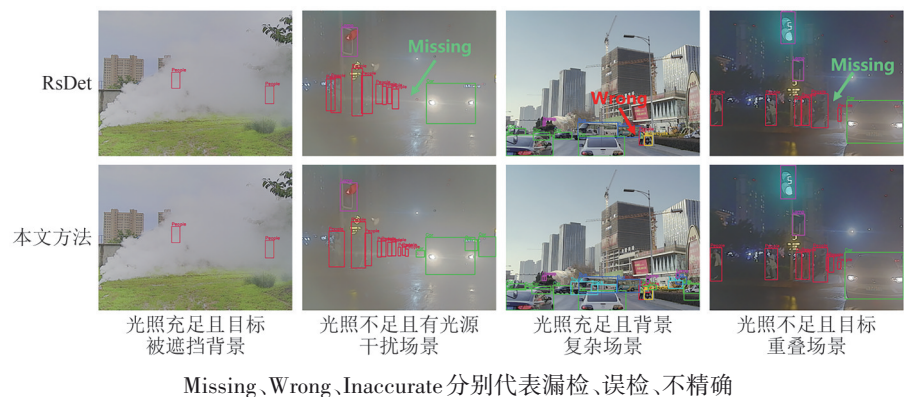
注：表中加粗黑体数值为同一列最佳结果。AP 代表平均精度，其下标代表被检测类别。mAP 代表所有类别的平均精度。

在具体目标类别上，MAF-YOLO在多个类别上均表现出较高的检测精度，其中在人、巴士和灯的类别上的检测精度尤为突出，分别达到了83.1%、91.1%和84.9%。但也可以注意到MAF-YOLO在巴士检测上效果较好，但对其他类型车辆的检测表现稍逊，这可能是由于其采用了更多的红外模态信息，而巴士的目标较大，其红外特征更为明显。对于其他类型的车辆，尤其是小型车辆，可能由于特征相对较小或形状多样，导致检测效果稍逊。总的来说，通过采用可见光加红外的多模态数据进行目标检测，MAF-YOLO方法显著提高了检测精度，尤其是在利用红外信息的情况下，可以更好地应对光照变化、低光照等环境条件下的目标检测任务。

为了更加直观地展示本文所提出的MAF-YOLO与其他目标检测模型之间的检测效果，本方法及部分对比方法在M³FD数据集上具有代表性场景下的目标检测结果见图7。由图7第1列可见在环境光充足但可见光目标被遮挡的情况下，仅采用可见光信息进行目标检测的YOLOv8-RGB模型在面对遮挡时难以进行正确检测；YOLOv8-IR模型利

用穿透性强的红外信息，能够在不被干扰的情况下进行目标检测任务。与此同时，由于多模态目标检测模型能够结合可见光与红外信息，其都可以实现在可见光信息被缺失的情况下准确地完成目标检测任务。图7第2列展示了环境光强度较弱且具有光源干扰的情况下，各个模型的目标检测结果。在该条件下，由于可见光信息受到干扰，仅采用可见光信息的目标检测模型受到较大影响，而使用红外信息和对可见光—红外多模态信息进行融合检测的模型能够较好地完成检测。图7第3列展示了环境光充足且待检测目标与背景复杂的条件下，不同目标检测模型的表现。对于存在多个目标重叠且环境光强度较大导致的红外信息背景影响，可见光信息能够更多地包含目标的特征，在此场景下，MAF-YOLO通过自适应融合提取可见光信息，在多个尺度上进行目标检测，得到更好的结果。图7第4列为环境光强度较弱且目标存在重叠的场景，在此情况下，MAF-YOLO不仅能正确地识别车辆及信号灯等单独的目标，对大量重叠的行人目标也仍保持着较高的检测准确性。





Missing、Wrong、Inaccurate 分别代表漏检、误检、不精确
图7 M³FD数据集上的目标检测可视化结果
Fig. 7 Visualization of detection result based on the M³FD dataset

3.3.2 基于 DroneVehicle 数据集的对比实验

DroneVehicle 数据集下各目标检测网络在不同类型目标下的 AP 和 mAP 如表 2 所示。观察表 2 可以发现，由于 DroneVehicle 数据集存在较多的黑暗场景，可见光影像往往难以展现丰富的纹理和色彩信息，而红外影像则能清晰呈现目标的轮廓特征，有助于实现高效的目标检测。因此，采用红外影像的单模态目标检测模型在这种情况下表现更优。与单模态目标检测相比，多模态检测模型能够充分利用不同模态之间的互补特征，因此多模态模型的性能普遍更出色。

表 2 基于 DroneVehicle 数据集的对比实验
Table 2 Comparison of different methods based on DroneVehicle dataset

		/%						
模型	数据模态	AP _{car}	AP _{lorry}	AP _{bus}	AP _{van}	AP _{truck}	mAP	
Faster R-CNN	RGB	88.2	64.2	91.8	60.2	67.5	74.4	
	IR	88.5	77.2	90.1	71.5	68.4	79.2	
YOLOv8	RGB	91.5	71.6	91.1	63.7	61.5	75.9	
	IR	90.6	72.3	90.0	70.0	67.1	78.0	
YOLOv10	RGB	87.6	73.1	92.1	64.2	64.7	76.3	
	IR	90.2	75.7	87.3	67.5	68.4	77.8	
CFT	RGB+IR	95.3	80.4	94.6	63.8	69.4	80.7	
TarDAL	RGB+IR	94.5	77.9	95.8	72.8	70.5	82.3	
SuperYOLO	RGB+IR	93.1	82.9	93.6	70.4	75.8	83.2	
MCHE-CF	RGB+IR	93.4	82.6	92.7	65.9	71.4	81.2	
RsDet	RGB+IR	92.7	80.1	93.9	74.6	67.4	81.7	
本文方法	RGB+IR	94.9	83.2	96.0	73.1	73.5	84.1	

注：表中加粗黑体为同一列最佳结果。AP 代表平均精度，其下标代表被检测类别。mAP 代表所有类别的平均精度。

在进行多模态目标检测模型的横向比较时，可以发现各模型在不同类别的目标检测中表现不同。CFT 模型在轿车类别的检测中取得了较高的平均精度（AP），但在其他类别的检测中表现较差，

尤其在面包车类别的检测中 AP 甚至低于部分单模态目标检测模型。TarDAL 模型在巴士类别的检测中表现较好，同时在轿车类别的检测精度也较高。SuperYOLO 模型在卡车和货车类别的检测中取得了良好的结果，RsDet 则在面包车上的检测结果中取得了最好的结果。而本文提出的 MAF-YOLO 模型不仅在轿车、卡车和巴士类别的检测效果最佳，而且在其他类别的检测结果中也表现出色，获得了最高的 mAP，呈现出了最好的检测性能。综上所述，本实验证明 MAF-YOLO 模型在黑暗场景下的多模态目标检测中具有较高的检测精度和鲁棒性。

为进一步更直观的展示 MAF-YOLO 与其他模型的目标检测效果，DroneVehicle 数据集中几组典型的场景被选取，详见图 8。可见在弱光条件下，各目标检测模型的准确性均产生了不同程度的下降。其中，CFT、MCHE-CF 及 RsDet 模型出现了较多的漏检及部分错检；TarDAL 由于使用生成对抗网络，难以从可见光影像中提取共同特征，使得其漏检现象较为严重；而 SuperYOLO 和本节提出的 MAF-YOLO 在弱光条件下的检测效果较好。

图 9 为环境光强度极低且存在多种目标的场景下各网络的检测结果。可见，在这种挑战性的条件下，各多模态目标检测模型普遍面临着漏检和错检的问题。其中：FT、TarDAL、SuperYOLO、MCHE-CF 及 RsDet 的性能受到了更为显著的影响，出现了较多的漏检和错检现象；相比之下，本文提出方法 MAF-YOLO 在这种极端环境下表现更为鲁棒，虽然仍然存在一定数量的漏检和错检，但相较于其他模型展现出更强的适应能力和鲁棒性。综上，相较于其他模型，本文提出方法 MAF-YOLO 在这种复杂场景下显示出最少的错检和漏检率，证明了其在多种环境条件下的优越性。

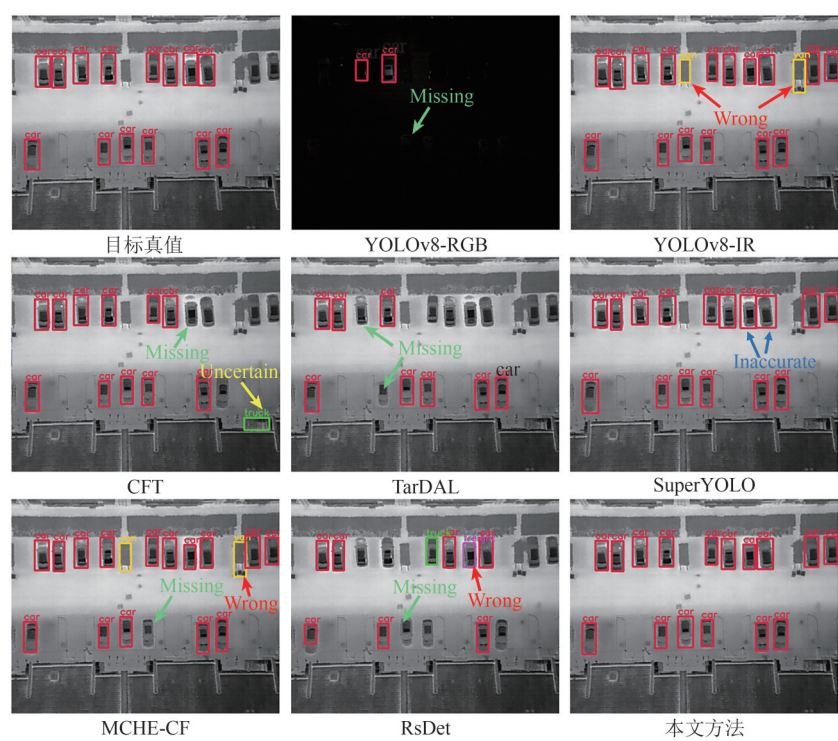


图8 DroneVehicle数据集弱光条件下的目标检测结果

Fig. 8 Object detection results under low light condition based on DroneVehicle dataset

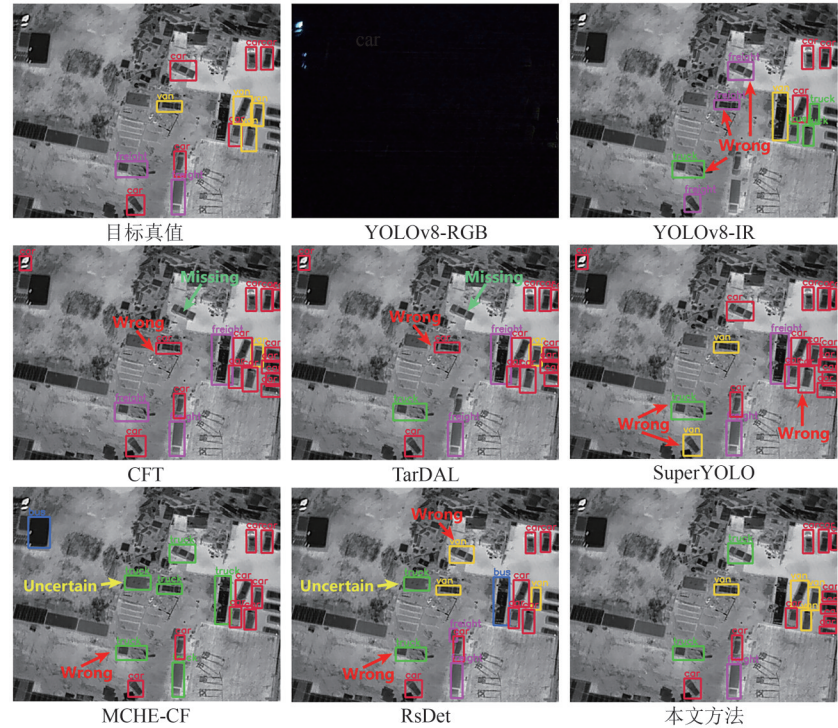


图9 DroneVehicle数据集弱光复杂条件下的目标检测结果

Fig. 9 Object detection results under low light complex condition based on DroneVehicle dataset

3.4 消融实验

为了说明所构建的跨模态混合注意力模块 (CHAM)、跨模态自适应融合模块 (CAFM) 和光照感知模块 (IAM) 模块的有效性, 本文基于

DroneVehicle 数据集对 MAF-YOLO 模型进行了消融实验。实验采用的基准模型为修改为双流的 YOLOv8l 网络, 其通过直接的拼接操作对双流信息进行融合。本节实验在此基准模型的基础上

依次添加这3个模块了以各类别的AP、mAP和GLOPs来评估改进的有效性。消融实验的结果如表3所示。通过消融实验的结果可以看出，随着逐步引入跨模态混合注意力模块、跨模态自适应融合模块和光照感知模块，目标检测的精度有所提升，而模型的复杂度也相应有些许增加。

第1个模型为基准模型，该模型在卡车、面包车和货车等类别的AP较低，主要因为这3种目标之间的相似性较高，直接拼接叠加的融合方式难以保留不同模态的互补特征，尤其在光照不足的条件下容易对三者产生误检。第2个模型在引入跨模态混合注意力模块后，mAP从基准模型的78.9%提升到了81.8%，说明跨模态混合注意力模块的引入对提升目标检测精度起到了积极作用，尤其在面包车这一类上有了5.7%的提升，这表明该模块能够有效地提取不同模态之间的优势特征，从而提高模型对不同类别的分类精度。第3个模型引入了不包含光照感知的跨模态自适应融合模块，通

过融合多模态数据的特征信息显著提高了模型在各类别上的分类性能。第4个模型引入了跨模态混合注意力和不包含光照感知的跨模态自适应融合模块，使得目标检测结果的mAP均有所提高，并在公交车这一类别取得了最高的AP。最后一个模型通过添加跨模态混合注意力模块、光照感知模块和跨模态自适应融合模块，在多类目标的AP及mAP均取得了最好的结果，说明本文提出的改进能够有效地增强模型对于复杂环境的适应能力，通过融合跨模态信息，使得模型取得更好的检测效果。

由表3可见，MAF-YOLO通过添加跨模态混合注意力、自适应融合与光照感知模块后，模型在各类别上的分类精度和模型整体的性能表现均得到了提升，其模型复杂度虽有所增加但其幅度较小，表明本文所提出的创新有助于提升可见光—红外的跨模态特征融合性能，并能够实现更高准确性的多模态目标检测。

表3 本文所提方法基于DroneVehicle数据集的消融实验
Table 3 Ablation experiments on the DroneVehicle dataset

CHAM	CAFM	IAM	AP _{car} /%	AP _{lorry} /%	AP _{bus} /%	AP _{van} /%	AP _{truck} /%	mAP/%	GLOPs
×	×	×	91.0	76.5	91	67.3	68.7	78.9	170.8
✓	×	×	92.0	79.6	94.1	73.0	70.4	81.8	178.2
×	✓	×	91.4	78	93.9	70.4	68.5	80.4	180.2
✓	✓	×	92.2	79.5	94.5	72.8	71.5	82.1	188.7
✓	✓	✓	92.7	80.4	94.3	75.8	71.5	82.9	190.5

注：表中“√”代表加入模块；“×”代表去除模块；加粗黑体表示同列最佳结果。

4 结 论

本文针对可见光目标检测在遮挡、弱光等复杂环境下特征提取能力不足、检测精度下降的核心问题，提出一种基于可见光与红外特征自适应融合的多模态目标检测方法（MAF-YOLO），通过创新模块设计与网络优化，有效实现了多模态优势特征的协同利用，主要研究结论如下：

（1）本文设计的跨模态混合注意力模块（CHAM），通过通道与空间注意力的分层交互，针对性提取可见光纹理细节与红外边缘轮廓的互补特征，联合进行通道与空间上的注意力分析，实现跨模态信息权重的交流重组，提高了不同模态间互补特征的感知能力。

（2）本文构建了以环境光照强度为指标的可

见光—红外特征自适应融合模块，首先设计了光照感知模块（IAM），根据可见光图像的光照强度评估其包含目标的特征丰富度，将其纳入跨模态自适应融合模块CAFM（Cross-modal Adaptive Fusion Module）指导自适应融合的进行，解决了常规的融合方法难以根据多模态数据特征实现动态处理的问题。

（3）基于YOLOv8构建的MAF-YOLO网络，通过骨干网络的双模态特征提取、颈部的多尺度融合与检测头的独立分类回归分支，在公开的街景数据集M³FD及航拍车辆数据集DroneVehicle上展现出优异的检测精度与鲁棒性。

尽管本文方法取得了较好效果，仍存在可优化空间：（1）针对多模态传感器成像的时空对准误差，后续将结合相机标定与特征点匹配技术，实

现模态间的亚像素级配准;(2)针对航拍场景中
小目标占比高的特点,将通过引入超分辨率重建
子网络与多尺度检测头优化,进一步提升小目标
检测精度;(3)将探索MAF-YOLO在YOLOv10、
YOLOv11等新一代检测框架上的迁移应用,以实
现精度与推理速度的协同提升,推动方法向实际
工程场景落地。

参考文献(References)

- Dong S, Wang P and Abbas K. 2021. A survey on deep learning and its applications. *Computer Science Review*, 40: 100379 [DOI: 10.1016/j.cosrev.2021.100379]
- Fang Q Y, Han D P and Wang Z K. 2022. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv: 2111.00273* [DOI: 10.48550/arXiv.2111.00273]
- Hou Q B, Zhou D Q and Feng J S. 2021. Coordinate attention for efficient mobile network design//*Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville: IEEE: 13708-13717 [DOI: 10.1109/CVPR46437.2021.01350]
- Hussain M. 2023. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. *Machines*, 11(7): 677 [DOI: 10.3390/machines11070677]
- Jia M D, Li G, He S W, Li H W, Kang Y and Wang C Y. 2023. Review of two-band infrared target detection. *Flight Control and Detection*, 6(2): 60-69 (贾明东, 李根, 贺仕伟, 李红伟, 亢颖, 汪陈跃. 2023. 双波段红外目标检测综述. *飞控与探测*, 6(2): 60-69)
- Li R M, Xiang J J, Sun F X, Yuan Y, Yuan L W and Gou S P. 2024. Multiscale cross-modal homogeneity enhancement and confidence-aware fusion for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 26: 852-863 [DOI: 10.1109/TMM.2023.3272471]
- Li Y T, Fan Q S, Huang H S, Han Z G and Gu Q. 2023. A modified YOLOv8 detection network for UAV aerial image recognition. *Drones*, 7(5): 304 [DOI: 10.3390/drones7050304]
- Liu J Y, Fan X, Huang Z B, Wu G Y, Liu R S, Zhong W and Luo Z X. 2022a. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection//*Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans: IEEE: 5792-5801 [DOI: 10.1109/CVPR52688.2022.00571]
- Liu R, Wang D J, Jia P and Sun H. 2018. An omnidirectional morphological method for aerial point target detection based on infrared dual-band model. *Remote Sensing*, 10(7): 1054 [DOI: 10.3390/rs10071054]
- Liu W Y, Ren G F, Yu R S, Guo S, Zhu J K and Zhang L. 2022b. Image-adaptive YOLO for object detection in adverse weather conditions//*Proceedings of the 36th AAAI Conference on Artificial Intelligence*. Virtually: AAAI: 1792-1800 [DOI: 10.1609/aaai.v36i2.20072]
- Luo H L and Chen H K. 2020. Survey of object detection based on deep learning. *Acta Electronica Sinica*, 48(6): 1230-1239 (罗会兰, 陈鸿坤. 2020. 基于深度学习的目标检测研究综述. *电子学报*, 48(6): 1230-1239) [DOI: 10.3969/j.issn.0372-2112.2020.06.026]
- Osco L P, Junior J M, Ramos A P M, De Castro Jorge L A, Fatholahi S N, De Andrade Silva J, Matsubara E T, Pistori H, Gonçalves W N and Li J. 2021. A review on deep learning in UAV remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 102: 102456 [DOI: 10.1016/j.jag.2021.102456]
- Padilla R, Passos W L, Dias T L B, Netto S L and Da Silva E A B. 2021. A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics*, 10(3): 279 [DOI: 10.3390/electronics10030279]
- Ren S Q, He K M, Girshick R and Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks//*Proceedings of the 29th International Conference on Neural Information Processing Systems*. Montreal: MIT Press, 91-99
- Shen J F, Chen Y F, Liu Y, Zuo X, Fan H and Yang W K. 2024. ICAFusion: iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145: 109913 [DOI: 10.1016/j.patcog.2023.109913]
- Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv: 1409.1556* [DOI: 10.48550/arXiv.1409.1556]
- Sohan M, Sai Ram T and Rami Reddy C V. 2024. A review on YOLOv8 and its advancements//*International Conference on Data Intelligence and Cognitive Informatics*. Tirunelveli: Springer: 529-545 [DOI: 10.1007/978-981-99-7962-2_39]
- Sun Y, Hou Z Q, Yang C, Ma S G and Fan J L. 2023. Object detection algorithm based on dual-modal fusion network. *Acta Photonica Sinica*, 52(1): 0110002 (孙颖, 侯志强, 杨晨, 马素刚, 范九伦. 2023. 基于双模态融合网络的目标检测算法. *光子学报*, 52(1): 0110002) [DOI: 10.3788/gzxb20235201.0110002]
- Sun Y M, Cao B, Zhu P F and Hu Q H. 2022. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10): 6700-6713 [DOI: 10.1109/TCSVT.2022.3168279]
- Tang G Y, Ni J J, Zhao Y H, Gu Y and Cao W D. 2023. A survey of object detection for UAVs based on deep learning. *Remote Sensing*, 16(1): 149 [DOI: 10.3390/rs16010149]
- Tang L F, Yuan J T, Zhang H, Jiang X Y and Ma J Y. 2022. PIAFusion: a progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83-84: 79-92 [DOI: 10.1016/j.inffus.2022.03.007]
- Terven J and Cordova-Esparza D. 2024. A comprehensive review of YOLO: from YOLOv1 and beyond. *arXiv preprint arXiv: 2304.*

- 00501 [DOI: 10.48550/arXiv.2304.00501]
- Wang A, Chen H, Liu L H, Chen K, Lin Z J, Han J G and Ding G G. 2024. YOLOv10: real-time end-to-end object detection. arXiv preprint arXiv:2405.14458 [DOI: 10.48550/arXiv.2405.14458]
- Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: convolutional block attention module//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]
- Yang X X, Qian Y Q, Zhu H J, Wang C X and Yang M. 2022. BAANet: learning bi-directional adaptive attention gates for multispectral pedestrian detection//2022 International Conference on Robotics and Automation (ICRA). Philadelphia: IEEE: 2920-2926 [DOI: 10.1109/ICRA46639.2022.9811999]
- Zhang J Q, Lei J, Xie W Y, Fang Z M, Li Y S and Du Q. 2023a. Super-YOLO: super resolution assisted object detection in multimodal remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 61: 5605415 [DOI: 10.1109/TGRS.2023.3258666]
- Zhang J Q, Lei J, Xie W Y, Li Y S, Yang G and Jia X P. 2023b. Guided hybrid quantization for object detection in remote sensing imagery via one-to-one self-teaching. IEEE Transactions on Geoscience and Remote Sensing, 61: 5614815 [DOI: 10.1109/TGRS.2023.3293147]
- Zhang S, Gong Y H and Wang J J. 2019. The development of deep convolution neural network and its applications on computer vision. Chinese Journal of Computers, 42(3): 453-482 (张顺, 龚怡宏, 王进军. 2019. 深度卷积神经网络的发展及其在计算机视觉领域的应用. 计算机学报, 42(3): 453-482) [DOI: 10.11897/SP.J.1016.2019.00453]
- Zhang Y, Yu H, He Y J, Wang X Y and Yang W. 2023c. Illumination-guided RGBT object detection with inter- and intra-modality fusion. IEEE Transactions on Instrumentation and Measurement, 72: 2508013 [DOI: 10.1109/TIM.2023.3251414]
- Zhang Y T, Yin Z S, Nie L Z and Huang S. 2020. Attention based multi-layer fusion of multispectral images for pedestrian detection. IEEE Access, 8: 165071-165084 [DOI: 10.1109/ACCESS.2020.3022623]
- Zhao T Y, Yuan M X, Jiang F, Wang N and Wei X X. 2024. Removal then selection: a coarse-to-fine fusion perspective for RGB-infrared object detection. arXiv preprint arXiv:2401.10731 [DOI: 10.48550/arXiv.2401.10731]
- Zuo C, Qian J M, Feng S J, Yin W, Li Y X, Fan P F, Han J, Qian K M and Chen Q. 2022. Deep learning in optical metrology: a review. Light: Science and Applications, 11(1): 39 [DOI: 10.1038/s41377-022-00714-x]

Multimodal object detection method using adaptive fusion of infrared and visible features

YU Zhirui, YIN Zhanpeng, WANG Junyu, ZHOU Liang, YE Yuanxin

Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu 611756, China

Abstract: Target detection methods are mainly based on visible light images. Although they can fully display the details and texture features of the target, when the target is blurred or blocked, or the light is excessively strong or dark, the target information is difficult to fully obtain by relying solely on visible light sensors, thus affecting the target detection effect. By contrast, infrared sensors have the advantages of strong resistance to environmental interference, insensitivity to light, and ability to reflect target temperature, which can compensate for the shortcomings of visible light imaging. Therefore, infrared features are integrated into target detection based on visible light.

This study proposes a multimodal target detection method that adaptively fuses infrared and visible light features. This method uses the YOLOv8 target detection framework as the basic network to extract multiscale feature information. On this basis, a new multimodal hybrid attention module (cross-modal hybrid attention module) is constructed. This module extracts the complementary features of visible light and infrared images and jointly performs channel and spatial attention analyses to achieve the communication and reorganization of cross-modal information weights, thereby improving the perception of complementary features between different modalities. In addition, a visible light-infrared feature adaptive fusion module with ambient light intensity as an indicator is constructed. An illumination awareness module is first designed to evaluate the feature richness of the target contained in the visible light image in accordance with the light intensity of the image. It is then incorporated into a cross-modal adaptive fusion module to guide the adaptive fusion process, solving the problem that conventional fusion methods hardly achieve dynamic processing based on multimodal data features. Thus, target detection based on multimodal feature fusion is realized.

Several classic target detection models and current advanced detection methods are used to conduct comparative experiments with the MAF-YOLO model proposed in this paper to fully demonstrate its effectiveness. The classic target detection models include Faster R-CNN and YOLOv8. Because they are both target detection models for single-modality images, this study conducts experiments on visible light and infrared modalities. Current advanced models include CFT, which is the first application of Transformer to the field of multispectral target detection; TarDAL, which uses generative adversarial networks to generate fused images; and SuperYOLO, which achieves super-

resolution reconstruction based on YOLOv5 and improves the accuracy of multimodal target detection. Comparative experiments are conducted on the M³FD street view dataset and the DroneVehicle aerial vehicle dataset to test the robustness of these methods in different lighting environments and scenes. Experimental results show that the proposed method achieves higher detection accuracy compared with current state-of-the-art single-modal and multimodal object detection methods in both datasets.

On the basis of the YOLOv8 model, this study proposes a target detection network based on visible light – infrared multimodal feature fusion (MAF-YOLO). A new cross-modal hybrid attention mechanism is designed to fully utilize the complementary features of visible light and infrared information. Moreover, an illumination perception module and an adaptive fusion module are constructed to perform midterm fusion of dual-stream information, extracting the complementary features of the two modalities for target detection. Comparative experiments of the proposed method with various existing target detection models on the DroneVehicle and M³FD datasets show that MAF-YOLO can achieve good detection performance and robustness in complex environments. Experimental findings prove that the proposed method can effectively solve the problem of insufficient visible light modal target features in complex environments and realize accurate target detection with the fusion of infrared and visible light multimodal features.

Key words: target detection, multimodal, convolutional neural network, feature fusion, attention mechanism, visible light imaging, infrared imaging, deep learning

Supported by National Key Research and Development Program (No. 2024YFC3015404); National Natural Science Foundation of China (No. 42271446); Sichuan Province Central Government Guidance for Local Science and Technology Development Special Project (No. 2024ZYD0084)