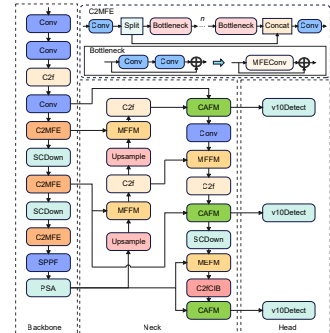


基于 YOLO 的自适应多尺度红外目标检测网络

汪佳旭^{1,2}, 杨俊^{2*}, 许聪源²

¹浙江理工大学信息科学与工程学院(网络安全学院), 浙江 杭州 310018;

²嘉兴大学人工智能学院, 浙江 嘉兴 314001



摘要: 现有基于深度学习的红外目标检测算法 (infrared target detection algorithm, ITDA) 对红外图像中特征的提取和处理能力不足, 导致漏检、误检率较高。多数算法不能兼顾不同尺度的红外目标, 红外小目标检测依赖高灵敏度特征提取以捕捉细微特征, 当检测目标尺度增大时, 因需要更强的全局理解能力, 易出现局部纹理过拟合, 导致性能降低, 使算法在跨场景部署时准确性降低。为解决上述问题, 本文提出了一种基于 YOLO 的自适应多尺度红外目标检测的网络 (adaptive multi-scale infrared target detection network based on YOLO, AFITDYOLO), 该网络通过多层特征提取模块及多层特征融合模块增强多尺度红外目标检测能力。提出多尺度特征融合模块 MFFM, 通过增强特征金字塔中不同层特征的相关性, 提升了多尺度特征融合的表达力; 提出多核特征提取卷积 MFECConv, 通过异构卷积组增大感受野, 更好地与不同尺度目标的空间分布保持一致; 提出交叉注意力融合模块 CAFM, 通过网络中不同层输出特征图的对比交互, 增强重要特征信息, 提升特征表达能力。为了评估 AFITDYOLO 的性能, 在无人机飞鸟数据集 SIRST-UAVB 和道路行人车辆数据集 CTIR 上进行实验, mAP50 分别达到了 88.9% 和 90.7%, 对比 YOLOv10n 分别提高了 5.6% 和 6.5%, 与目前主流方法对比, 本文方法检测精度最佳。本文提出的红外目标检测算法, 在多尺度红外目标检测中, 展现了优异的准确性和适应性, 可跨场景部署。

关键词: 红外目标检测; 特征融合; 轻量级网络结构; 注意力机制; YOLO

The English title and abstract appear at the end of the paper.

DOI: 10.12086/oe.2026.250292 | CSTR: 32245.14.oe.2026.250292 | 中图分类号: TP391 | 文献标识码: A

汪佳旭, 杨俊, 许聪源. 基于 YOLO 的自适应多尺度红外目标检测网络 [J]. 光电工程, 2026, 53(4): 250292

Wang J X, Yang J, Xu C Y. YOLO-based adaptive multi-scale infrared target detection network[J]. Opto-Electron Eng, 2026, 53(4): 250292

1 引言

红外目标检测凭借其在低光照、恶劣天气等复杂环境下独特的感知能力, 在安防监控、自动驾驶、军事侦察等众多关键领域中发挥着举足轻重的作用。红外图像具有对比度低、细节模糊、噪声干扰明显等特性^[1], 导致

其检测时易出现目标边缘不清晰、纹理特征缺失等情况。

早期红外目标检测采用传统模型驱动方法^[2-5], 需要基于先验知识并手动调整参数, 难以有效提取特征, 检测精度和实时性均面临瓶颈。随着深度卷积神经网络 (convolutional neural network, CNN) 在视觉任务中的成功应用, 基于深度学习的红外目标检测算法^[6-8]也由此兴

收稿日期: 2025-09-27; 修回日期: 2026-01-28; 录用日期: 2026-01-16; 网络出版日期: 2026-04-25

基金项目: 国家自然科学基金项目 (62302197); 浙江省高等教育“十四五”研究生教学改革项目 (JGGC2024034); 嘉兴市科技计划基金项目 (2024AD10045, 2024AY40010)

*通信作者: 杨俊, juneryoung@zjxu.edu.cn

版权所有©2026 中国科学院光电技术研究所

起, 此类算法具有高准确性和轻量化的优势, 有助于提高红外目标检测精度, 从而提高运营效率和成本效益。现有基于深度学习的红外目标检测算法通过新颖的网络架构提升了红外检测的性能, 如 Yang 等人根据红外小目标的成像特点提出了风车形卷积 PConv^[9] 并引入检测网络, 增强了底层特征提取, 提升了红外小目标检测精度; Zhao 等人提出了一种基于深度注意力机制的红外行人检测方法^[10], 通过构建四尺度特征金字塔网络并结合注意力模块, 显著提高了近距离行人目标的检测性能。我们在分析红外目标检测任务中得出, 红外小目标检测算法依赖高灵敏度特征提取以捕捉细微特征, 但在检测目标尺度增大时, 易出现局部纹理过拟合和误检, 导致性能降低。在实际应用中, 检测环境复杂多变, 目标尺度不一, 多尺度检测能力是确保算法在复杂真实环境中保持高可靠性、适应性的关键。但目前多数算法均针对单一尺度目标进行改进, 这些算法难以同时满足小目标的高精度定位与大目标的语义理解需求, 本文针对该问题提出一种基于 YOLO 的自适应多尺度红外目标检测网络 AFITDYOLO, 目的是同时提升检测网络的适用性和性能, 该网络接收不同尺度红外目标图像并使用多层特征提取模块及多层特征融合模块增强多尺度红外目标检测能力。本文的主要贡献如下:

1) 提出一种基于 YOLO 的自适应多尺度红外目标检测网络 AFITDYOLO, 通过多层特征提取及融合增强多尺度红外目标检测能力。

2) 提出多尺度特征融合模块 (multi scale feature fusion module, MFPM), 通过增强特征金字塔中不同层特征之间的相关性, 提升多尺度特征融合的表达能力。

3) 构建多核特征提取卷积 (multi-kernel feature extraction convolution, MFEConv), 通过异构卷积组, 增大感受野, 增强模型特征提取能力,

4) 设计交叉注意力融合模块 (cross-attention fusion module, CAFM), 通过检测网络中不同层输出特征图的对比交互, 利用特征图之间的互补信息, 抑制图像红外噪声, 进一步增强特征表达能力。

2 相关工作

2.1 传统目标检测算法

传统目标检测方法依赖手工特征提取与模式分类这一核心流程。这些方法通常包含预处理、分割、特征提

取和分类等步骤, 核心思想是利用图像的特性, 如目标与背景的差异、目标形状等进行检测。Dalal 等人基于手工特征提出了 HOG+SVM^[11], 该方法使用面向梯度 HOG 描述符的直方图, 并将检测窗口与重叠的 HOG 描述符组合, 最后使用基于 SVM 的窗口分类器对组合的特征向量进行分类, 实现目标行人检测。He 等人提出了一种基于 SVM 和红外热成像技术的高压绝缘子污染等级检测新方法^[12], 使用基于梯度信息的自适应平滑滤波去除图像噪声, 分割并提取目标区域, 从绝缘子表面提取温度、最大与最小温度对比度、表面温度标准差以及前 10% 亮度像素与总像素比例四类特征, 采用多类 SVM 进行污染等级检测。Felzenszwalb 等人在 HOG 基础上提出基于部件的模型 (discriminatively trained part based models, DPM)^[13], 通过将目标对象分解为多个部件, 然后对每个部件进行独立的训练, 最后这些部件组合形成完整的目标检测模型。

传统的目标检测方法, 提高了目标的检测精度, 但泛化能力差, 很难部署至红外目标检测, 同时多步骤串联导致实时性受限。

2.2 基于深度学习的红外目标检测算法

早期由 Sermanet 等人提出的 Overfeat 算法^[14]首次使用 CNN 代替手工特征来提取特征, 显著提升了特征表达能力。近年来, 许多基于深度学习的方法从不同角度分析和改进了红外目标检测。

特征金字塔 (Feature pyramid networks, FPN) 的构建是红外目标检测任务中的关键步骤, 并且是现代检测器的组成部分, 形成了解决目标多尺度问题的基础。对于较小的红外目标, 特征图通常只包含来自几个甚至一个像素的有效信息; 对于较大的红外目标, 特征图则是由多个组成部分和多层次特征构成的复杂实体。因此, 特征融合方法的研究对于准确表示红外目标的特征信息显得尤为重要。FPN 构建了自上而下的路径, 结合了各个级别的特征以实现多尺度特征融合。PANet^[15] 在 FPN 的基础上引入了自下而上的路径, 有利于高分辨率信息与更强语义特征的融合。随后, NAS-FPN^[16] 和 BiFPN^[17] 被提出来增强多尺度特征的融合。与许多特征融合方法不同, AFPN^[18] 探索聚合特征上的节点操作, 利用注意力机制来指导特征融合。虽然这种方法增强了检测性能, 但它显著增加了计算复杂度。在本文中, 我们研究多尺度特征融合, 提出多尺度特征融合模块 MFPM, 以低于 FPN 的计算复杂度实现了模型优越的检测性能。

在基于 CNN 的红外目标检测中, Dai 等人提出了 OSCAR 单级联级优化网络^[19], Yang 等人引入了 EFLNet^[7] 以解决目标与背景不平衡问题。ACM 模块^[20] 将底层细节嵌入高级特征中, 而 DNANet^[21] 则通过融合机制挖掘上下文信息。这些基于 CNN 的网络专注于构建特征提取网络, 但忽视了通过卷积模块增强 ITDA 性能的潜力。ISNet 模型^[22], 它结合了可变形卷积, 提高了检测性能, 但增加了训练时间和网络参数。因此, 本文通过构建轻量化卷积 MFECov, 增强模型提取底层特征的能力。

红外图像包含丰富的多层次信息, 不同信息对检测任务的贡献程度各异。为了更聚焦于图像中的重要特征与结构信息, Woo 等人提出了 CBAM^[23], 实现红外目标准确提取。此后, 各种注意力机制^[24-27] 被证明是提升红外目标检测性能的有效手段。但这些方法未充分挖掘图像的互补信息, 本文设计交叉注意力融合模块 CAFM, 通过检测网络中不同层输出特征图之间的对比交互, 突出重要特征信息, 抑制图像红外噪声, 实现模型特征表达能力的增强。

3 所提方法

3.1 AFITDYOLO

本文提出一种基于 YOLO 的自适应多尺度红外目标检测网络 AFITDYOLO, 由主干 (backbone)、颈部 (neck) 和检测头 (head) 三个部分组成, 整体结构如图 1 所示。

AFITDYOLO 模型结合了 C2MFE 网络架构, 实现高效的特征提取, 该架构使用 MFECov 模块替换 C2f 中 Bottleneck^[28-30] 的标准卷积层, 用于提升 C2f 模块的特征提取能力, 同时实现更高效、更灵活的卷积运算。为了优化网络结构, 增强模型多尺度特征融合环节, 提出 MFFM 模块, 用于合并不同层次的特征图, 该模块通过改进特征金字塔中不同层次特征的相关性, 提升了多尺度特征融合的表达能力。为了进一步提升网络检测性能, 设计交叉注意力融合模块 CAFM 提高检测精确度, 在检测头的 P3 (高分辨率)、P4 (中尺度)、P5 (低分辨率) 层输出前分别部署 CAFM, 针对性处理不同尺度的

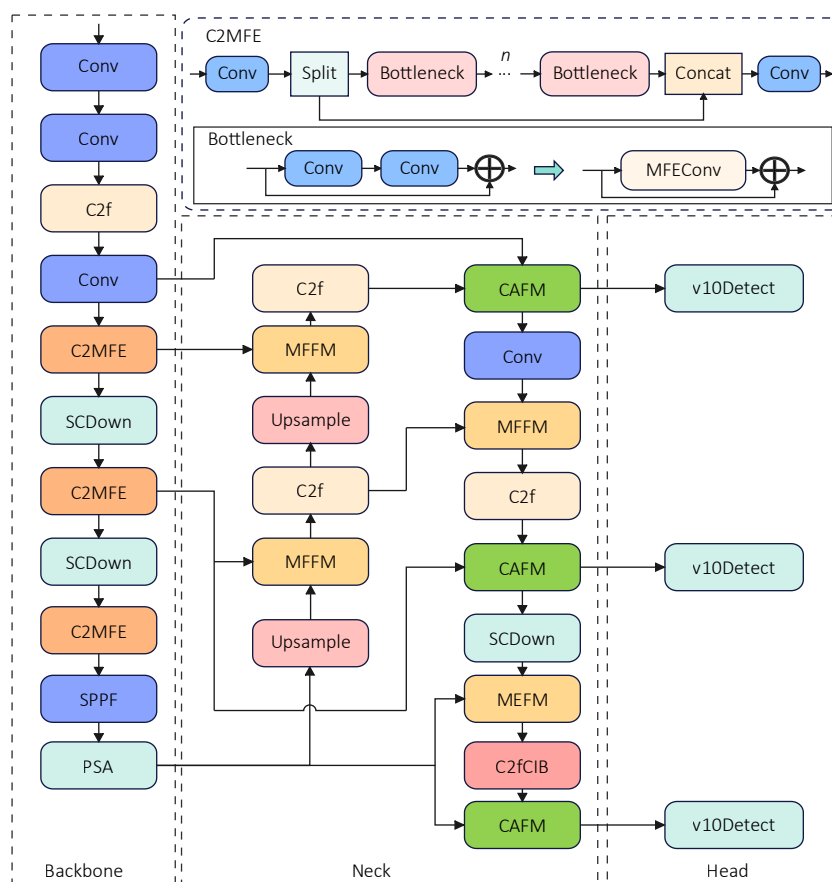


图 1 AFITDYOLO 结构

Fig. 1 The structures of AFITDYOLO

红外目标, 有效提升了红外目标检测的精度与鲁棒性。

3.2 MFFM

在红外目标检测任务中, 通过多尺度特征融合机制^[29]整合浅层空间细节与深层语义信息能够有效处理红外图像低分辨率、低对比度、目标尺度极端变化三大问题。但传统多尺度特征融合机制存在缺陷, 仅通过堆叠和通道融合对低分辨率特征进行上采样并与相邻层融合, 没有考虑它们的相关性, 从而限制了各层相关特征的利用, 针对该问题, 本文设计了多尺度有效融合模块 MFFM, 模块结构如图 2 所示。首先对输入特征进行预处理与全局特征融合, 使用 1×1 卷积层来获取模块的浅层特征并保持通道数的一致, 将浅层特征进行元素级相加生成全局特征 X_G , 使用三支并行处理全局特征。

通道注意力机制 CBAM 通过引入注意力模块学习每个通道的权重以动态调整每个通道的重要性。第一个分支引入通道注意力机制 CBAM 对 X_G 进行处理, 得到通道注意力特征 Y_1 。该分支确保模块能有效聚焦在重要的通道特征上, 同时保留原始特征细节, 防止丢失关键

目标信息。

第二个分支中通过 Sigmoid 激活函数^[31]处理 X_G 生成筛选系数 S , 并与全局特征相乘得到筛选后的特征 X_S 。将 X_S 沿通道维度均分为 4 组, 得到 $X_s^i, i \in [1, 2, 3, 4]$ 。每组特征经由卷积模块 Conv2 处理, 产生一个用于捕捉通道间特征相关性的注意力掩码。该掩码被应用于特征细化过程, 随后各组细化特征 X_{proc}^4 拼接融合^[32](Cat), 以形成具有聚合性和高度相关性的相邻特征 $X_{proc}(Y_2)$, 计算方法见式 (1)。

$$\begin{aligned} X_{proc}^i &= X_s^i \cdot \text{Softmax}\left\{BN\left[Conv_{1 \times 1}\left(X_s^i\right)\right]\right\} \\ Y_2 &= X_{proc} = \text{Cat}\left(X_{proc}^1, \dots, X_{proc}^4\right) \end{aligned} \quad (1)$$

通过分组处理, 模型能更灵活地捕捉红外图像的多尺度特征。

第三个分支将来自不同深度层级特征图中包含的丰富信息和较弱信息分离出来, 并进行独立处理。具体而言, 如图 2 中所示, 输入特征 X_{i-1} , X_i 由 1×1 卷积处理后, 进行批量归一化 (BN)^[33]并由 Sigmoid 函数激活, 分别生成每个特征对应的空间信息权重 ω_1, ω_2 , 表

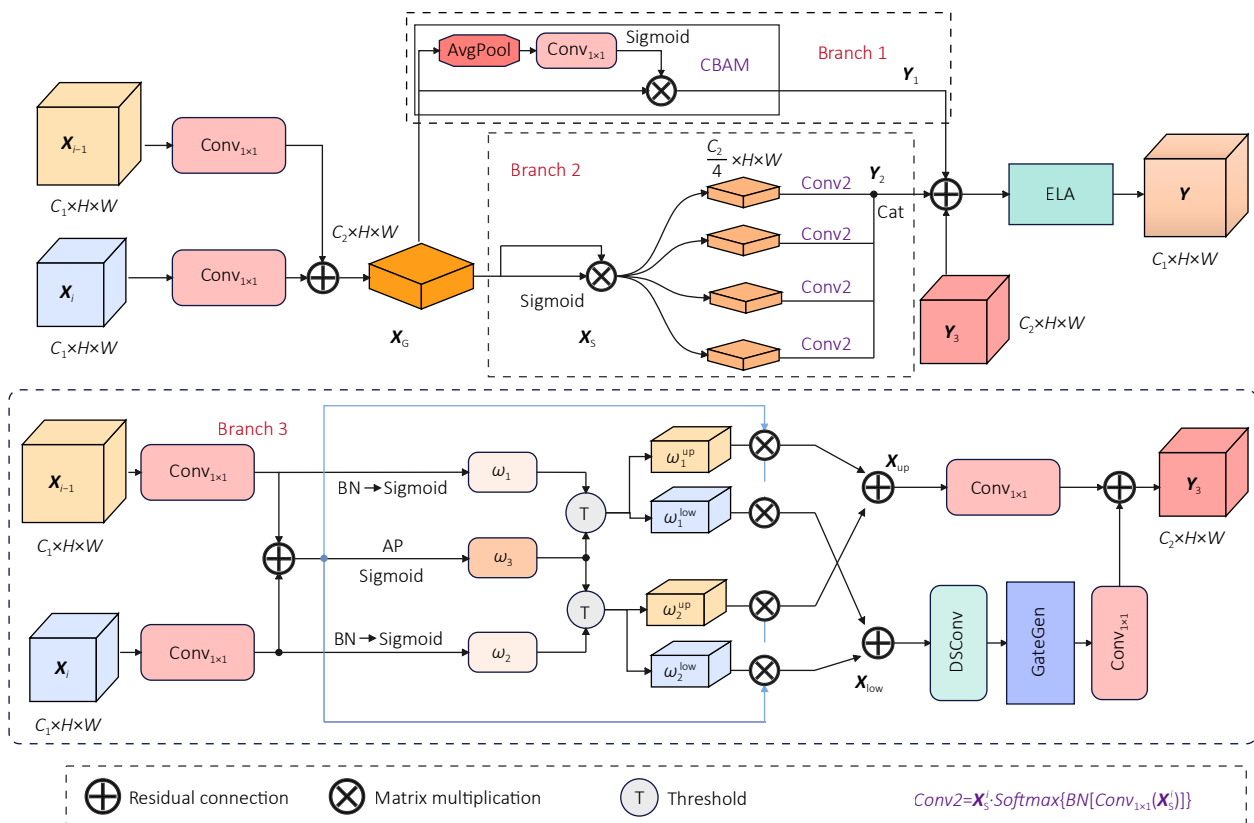


图 2 MFFM 架构

Fig. 2 The structures of MFFM

明不同阶段特征的重要性。对 X_G 进行自适应平均池化操作 AP (average pooling)^[34] 并通过 Sigmoid 激活函数得到特征权重阈值 ω_3 , 利用阈值函数 Threshold^[35], 将来自不同阶段特征的权重信息 ω_1, ω_2 与特征权重阈值 ω_3 进行比较, 获得捕捉空间信息强度的注意力图, 计算方法如下:

$$\begin{aligned} (\omega_1^{\text{low}}, \omega_1^{\text{up}}) &= \text{Threshold}(\omega_1, \omega_3), \\ (\omega_2^{\text{low}}, \omega_2^{\text{up}}) &= \text{Threshold}(\omega_2, \omega_3). \end{aligned} \quad (2)$$

通过注意力图将来自网络中不同层的强特征分离并聚合成强特征集 X_{up} , 不同层的弱特征分离并聚合成弱特征集 X_{low} , 计算方法见式 (3), 其中 $\otimes$ 表示逐元素乘法:

$$\begin{aligned} X_{\text{up}} &= (\omega_1^{\text{up}} + \omega_2^{\text{up}}) \otimes X_G, \\ X_{\text{low}} &= (\omega_1^{\text{low}} + \omega_2^{\text{low}}) \otimes X_G. \end{aligned} \quad (3)$$

然后, 将强、弱特征集分别进行独立处理, 采用深度可分离卷积 (DSConv)^[36] 和门控生成器 GateGen 处理弱特征集 X_{low} , 生成具有更丰富语义信息的特征; 应用 1×1 卷积保持强特征集与处理后的弱特征集的通道一致。聚合得到此分支的最终输出 Y_3 。通过具有更丰富语义信息的特征与展示更多详细信息的强特征集进行聚合, 使第三分支输出特征既包含详细信息, 也包含跨通道交换信息。

将上述三个分支输出进行残差连接, 并在输出前引入高效局部注意力 (ELA)^[37], 通过动态地聚焦目标关键区域进而抑制无关噪声。MFFM 增强 FPN 中不同层输出特征的相关性, 使深层语义特征与浅层空间细节特征更加协调, 提升了多尺度特征融合的表达力。

3.3 MFConv

基于卷积神经网络的红外目标检测方法多数采用标准卷积, 从而忽略了红外目标分布的空间特征^[8]。红外场景下往往存在多个待检测目标, 且尺度不同、分布不均, 标准卷积无法全面有效捕捉。本文针对上述问题设计了多核特征提取卷积 MFConv, 模块架构如图 3 所示。

由通道移位模块 (shift modules)^[38] 对输入特征进行预处理, 对输入特征图进行不同方向的移位并将移位后的特征图进行叠加, 通过该处理增强输入特征的多样性。在 MFConv 中, 对预处理后的特征图 X_2 按通道维度均分为 6 组, 通过异构卷积组处理, 卷积组由水平左右方向 $\text{Conv}_{1 \times k}$ 、垂直上下方向 $\text{Conv}_{k \times 1}$ 及左右对角线方向 $\text{Conv}_{k \times k}$ 6 个卷积核构成。为了提高训练稳定性和速度,

本文在每次卷积后应用 BN 和 Sigmoid 线性单元 (SiLU)^[39], 并行卷积计算见式 (4)。

$$\begin{aligned} X_H^{1,2} &= X_H^{2,2} = \text{SiLU}[\text{BN}(X_H^1 * \text{Conv}_{1 \times k}^{C_B})], \\ X_V^{1,2} &= X_V^{2,2} = \text{SiLU}[\text{BN}(X_V^1 * \text{Conv}_{k \times 1}^{C_B})], \\ X_S^{1,2} &= X_S^{2,2} = \text{SiLU}[\text{BN}(X_S^1 * \text{Conv}_{k \times k}^{C_B})], \end{aligned} \quad (4)$$

式中: $*$ 表示卷积算子; 通道数 $C_B = C/6$; k 表示卷积核的长度, 其取值越大模块的参数量越大, 模块中 k 的值取 3; 下角标 H、V、S 分别表示水平方向、垂直方向、对角线方向; $X_H^{1,2}, X_H^{2,2}$ 分别表示由水平左右方向卷积处理后的特征, 6 分支并行卷积输出拼接融合 (Cat):

$$X_3 = \text{Cat}(X_H^{1,2}, X_V^{1,2}, X_S^{1,2}, X_H^{2,2}, X_V^{2,2}, X_S^{2,2}). \quad (5)$$

通过异构卷积组增大感受野, 捕获特征图中多方向分布特征, 同时有效应对红外目标尺度变化问题。

设计联合注意力机制模块 (fusion attention mechanism, FAM) 引入 MFConv, 进一步增强多尺度特征提取能力。具体而言, 使用 1×1 卷积处理特征 X_3 压缩至单通道, 生成 FAM 输入特征 X_{FAM} 。应用 3×3 卷积与 Sigmoid 函数处理 X_{FAM} 生成热辐射权重 W 。引入 Softmax 函数^[40] 处理权重 W 与输入特征 X_{FAM} 获取通道注意力 A_C , 计算方法见式 (6)。

$$\begin{aligned} A_C &= \text{Softmax}\{\lambda \cdot \text{AvgPool}(X_3 \odot W) \\ &\quad + (1 - \lambda) \text{MaxPool}[X_3 \odot (1 - W)]\}, \end{aligned} \quad (6)$$

式中: AvgPool 与 MaxPool 分别为热加权平均池化, 反向加权最大池化; $\lambda \in [0, 1]$ 为可学习热融合系数, 默认为 0.5; $\odot$ 表示加权乘法。通道注意力聚焦主通道, 利于大目标的语义理解。空间注意力 A_S 计算方法见式 (7), 生成过程中, 权重 W 值大的目标区域在计算 $K^T V$ 时获得更高权重, 权重 W 值小的目标区域则降低其空间关联性, 从而强化高温区域细节, 利于小目标的定位。其中 $\sqrt{C}$ 为温度缩放因子, 用于稳定训练, 上角标 T 表示矩阵转置:

$$\begin{aligned} A_S &= \text{Sigmoid}\left(\frac{K^T V}{\sqrt{C}} \odot W\right), \\ K = V &= \text{Conv}_{1 \times 1}(X_{\text{FAM}}). \end{aligned} \quad (7)$$

MFConv 末端使用残差连接, 其中 γ 为可学习缩放参数, 初始值为 1, 输出计算如下:

$$Y = X_3 + \gamma[(A_S \odot A_C \odot W) \odot X_2]. \quad (8)$$

MFConv 中的异构卷积组, 可增强特征图中不同方向分布特征的提取; 利用联合注意力机制, 可以在抑制噪声的同时, 对目标区域进行特征强化。

C2f 是 YOLO 网络中的高效特征提取模块, 通过拼

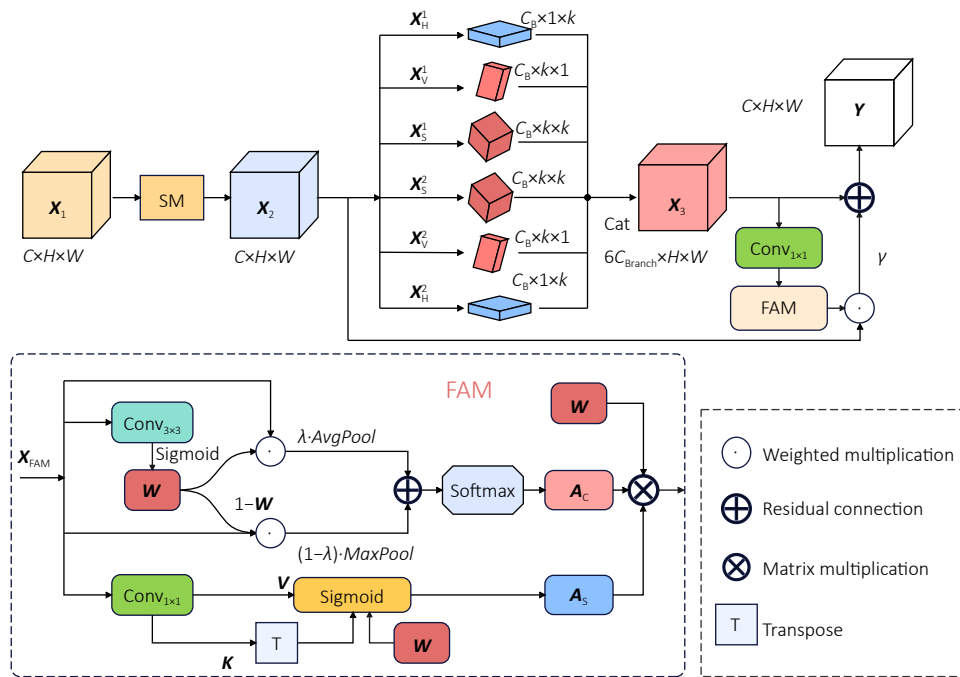


图 3 MFEConv 架构

Fig. 3 The structures of MFEConv

接不同 Bottleneck 模块的输出和原始特征图实现特征提取, Bottleneck 的标准卷积层由一个 3×3 卷积和 1×1 卷积叠加组成。本文提出 MFEConv 替换 Bottleneck 中的标准卷积层, 构建改进模块 C2MFE, 在提升效果的同时, 实现轻量化改造, 提高检测效率, 克服 C2f 在红外场景下面临感受野不足、噪声敏感、细节丢失等问题。

3.4 CAFM

传统目标检测模型对所有区域或通道赋予相同权重, 导致背景噪声干扰或关键特征被稀释, 利用注意力机制可以使模型聚焦于目标区域和重要通道。然而, 在红外目标检测中, 现有多数注意力机制如 SE^[41]、CBAM 不能有效捕获红外目标必需的局部上下文, 加之采用固定权重相加或拼接, 无法根据目标尺寸动态调整融合策略, 导致出现小目标漏检、大目标模糊。针对上述问题, 本文设计了交叉注意力融合模块 CAFM, 由交叉交互注意力 CA (cross-interaction attention) 和动态简易门控模块 DG (dynamic gating) 构成, 如图 4 所示。

CA 建立左右分支特征 (浅层特征与深层特征) 的跨层级语义关联, 增强多尺度特征交互能力。如图中 CA 部分所示, 左右分支输入特征通过投影层分别生成查询矩阵 V_R 和 V_L , 计算方法见式 (9), 查询矩阵 V_R 进行转置

后用于下一步运算。通过矩阵乘法计算后引入 Softmax 函数处理, 生成跨分支的注意力权重矩阵 **Attention**, 计算方法见式 (10), 计算过程中引入 $\sqrt{C}$ 用于优化分布, 防止点积过大导致梯度不稳。

$$V_L = DSCConv(X_L), V_R = DSCConv(X_R), \quad (9)$$

$$\mathbf{Attention} = \text{Softmax}\left(\frac{V_R^T V_L}{\sqrt{C}}\right) \in \mathbb{R}^{H \times W \times W}. \quad (10)$$

利用注意力权重矩阵加权, 实现有效的左右视图交互, 增强左右输入特征并应用 1×1 卷积层保持左右增强特征通道数一致。投影层采用深度可分离卷积进行轻量化改造。CA 通过左右分支特征互补捕获全局关联语义, 实现图像的跨视图信息增强, 特征增强计算如下:

$$\begin{aligned} F_{L \rightarrow R} &= \text{Conv}_{1 \times 1}(\mathbf{Attention} \otimes V_R), \\ F_{R \rightarrow L} &= \text{Conv}_{1 \times 1}(\mathbf{Attention}^T \otimes V_L), \end{aligned} \quad (11)$$

DG 融合局部细节与全局语义线索, 动态生成空间-通道自适应的权重矩阵, 实现像素级特征优选, 解决小目标细节丢失和噪声干扰问题, 原始特征 X_L 、 X_R 与交叉注意力增强特征 $F_{R \rightarrow L}$ 、 $F_{L \rightarrow R}$ 沿通道维度拼接:

$$X_{Concat} = \text{Concat}(X_L, X_R, F_{R \rightarrow L}, F_{L \rightarrow R}). \quad (12)$$

使用内核长度为 5 的卷积来捕获局部上下文, 并引入 Sigmoid 函数生成注意力权重 α , 让模型自身获得了

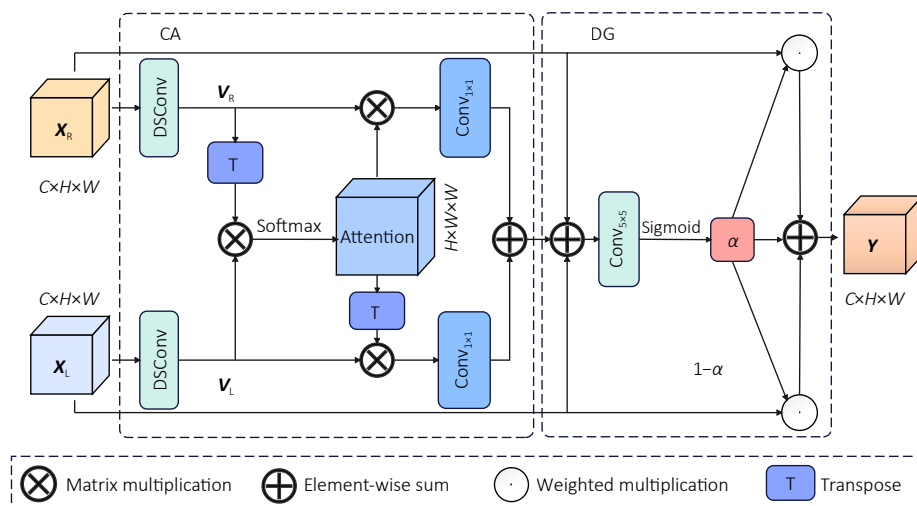


图 4 CAFM 架构

Fig. 4 The structures of CAFM

根据输入内容动态调整感受野的能力。将该注意力权重归一化至 $[0, 1]$, α 动态调节左右分支贡献比例生成抗噪声门控输出:

$$Y = \alpha \odot X_R + (1 - \alpha) \odot X_L. \quad (13)$$

CAFM 通过轻量化的跨视图交互机制, 对左右视图互补信息进行有效融合, 实现多尺度特征高效提取。AFITDYOLO 颈部网络用于红外目标多尺度特征融合与提取, 在颈部网络输出前, 使用交叉注意力融合模块 CAFM, 优化检测头的输入特征质量。该模块左右输入分别对应浅层输出特征图和深层输出特征图, 通过动态多尺度特征增强与噪声感知抑制, 显著提升了网络的红外目标检测的精度与鲁棒性。

4 实验

本节首先描述数据集、评估指标和训练细节, 然后展示不同模块对模型的性能影响, 并进行目标检测可视化, 与现有主流模块模型对比, 最后验证提出模型的泛化能力。

4.1 实验配置

数据集与评价指标。为了验证本文方法对不同尺度的红外目标检测能力, 使用两个不同复杂背景下的红外数据集, 单帧无人机飞鸟数据集 SIRST-UAVB^[8] 与红外监控摄像头实拍下的道路行人车辆数据集 CTIR^[42]。CTIR 包含城市道路和高速公路场景下, 在夜间及多种

恶劣天气条件中采集的 11938 张红外热图像, 灰度 8~14, 其中目标包含 34078 辆汽车 (car)、31035 个行人 (pedestrian)、16524 个骑自行车人 (cyclist)、2404 辆公共汽车 (bus) 和 1886 辆卡车 (truck), 每张红外图像中目标种类较多, 且同时包含大小尺度目标。SIRST-UAVB, 由 3000 张针对无人机和鸟类的红外图像组成, 这些图像包括 1742 个鸟类 (bird) 和 2955 个无人机 (UAV), 是一年多来收集的不同季节、天气条件和复杂背景的图像。该数据集中目标方向、尺度不同且受遮挡等挑战, 微小尺度目标占比较高。数据集中部分红外图像如图 5 所示。本文使用精度 P 和召回率 R 来评估模型检测的准确性, mAP_{50} 及 mAP_{95} 来评估检测目标位置准确性, 参数量 Params 作为轻量化指标, 浮点运算量 FLOPs 衡量网络计算复杂度, 每秒处理帧率 FPS 衡量检测速度。如式 (14) 所示, mAP_{50} 是指在 IoU 阈值为 0.5 时计算得到的平均精度均值, mAP_{95} 是一项更为综合的评估指标, 它在 IoU 从 0.5 到 0.95 的范围内, 以 0.05 为步长, 取多个不同阈值下的平均精度结果。TP 表示真正例, FP 为假正例, FN 为假反例, n 为缺陷的类别数量。

$$\begin{aligned} P &= \frac{TP}{TP + FP}, \\ R &= \frac{TP}{TP + FN}, \\ AP &= \int_0^1 P(R) dR, \\ mAP &= \frac{1}{n} \sum_{i=1}^n AP_i. \end{aligned} \quad (14)$$

训练细节。为了验证本文提出的方法可以显著提高

模型检测能力, 在包含多尺度目标的 CTIR 数据集进行训练, 同时为了进一步验证方法的自适应性, 在背景更为复杂、目标尺度更小的 SIRST-UAVB 数据集上也进行训练。两个数据集均按 6:2:2 的比例划分为训练集、测试集和验证集。本文实验均使用 RTX 3090 GPU 上的 PyTorch 框架, 在训练阶段, 采用 SGD 作为优化器, 初始学习率为 0.01, 耐心值设置为 80, 其余超参数均为默认。输入图像尺寸为 640 pixel×640 pixel、batchsize 大小为 40、epochs 为 500。

4.2 消融实验

为了评估每个组件对 AFITDYOLO 的单独影响, 使用 CTIR 及 SIRST-UAVB 两个红外数据集, 对模型进行消融实验。为保证消融实验的精度和准确性, 保持各组实验的环境和超参数一致, 实验结果如表 1 所示, 其中最优性能数据以粗体和下划线突出显示。本文选择

YOLOv10n 为基线。由表 1 中实验结果可知, 第一组实验基于 MFFM 模块改进了颈部网络, 增强了网络的多尺度特征融合的表达能力, 各项指标均有提升, CTIR 数据集中, mAP50、mAP95 涨幅更明显, 均达 4.1%。使用 MFECConv 模块改造主干网络中的 C2f 模块, 对模型轻量化改造, MFECConv 模块实现了检测精度的提升和参数量的降低。使得主干网络输出特征更具灵活性和适应性, 能够在复杂场景下保持较高性能, 相比于基线 *P*、*R*、mAP50、mAP95 均有着明显的增长, Params、FLOPs 和 FPS 均为最佳, 提高了模型检测效率。模型颈部网络输出前使用 CAFM, 通过浅层特征与深层特征的对比较交互处理, 使模型在略微增加计算复杂度和模型参数数量的情况下, 显著提高了检测的准确性。通过任务对齐机制, 以及分类和回归分支的动态选择机制, 提高了特征利用率和检测精度, *P* 与 mAP50 提升均大于 3%, mAP95 涨幅也达 2.7%。第 5 组在第 2 组基础上继续集

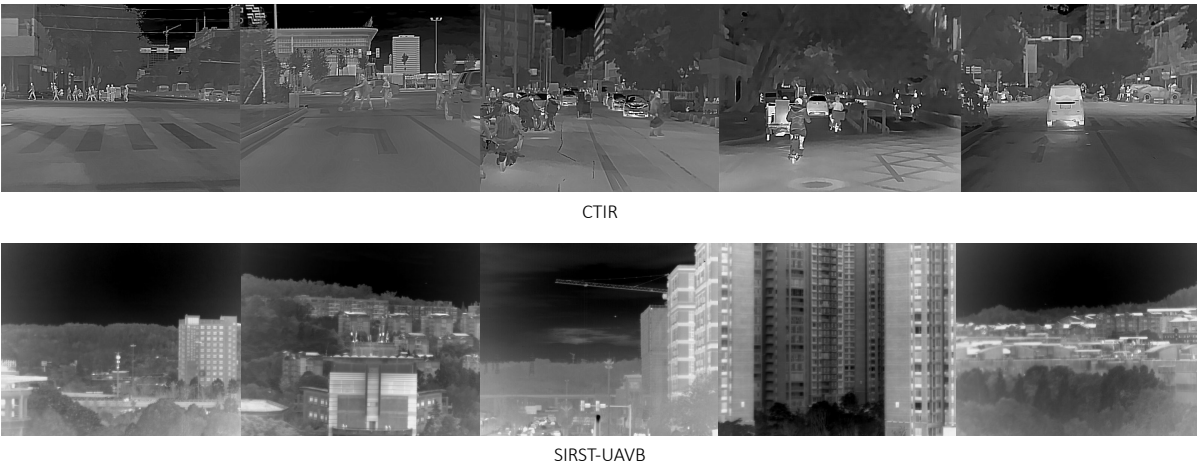


图 5 数据集

Fig. 5 Infrared data sample illustration

表 1 各个模块消融结果

Table 1 The ablation results for each module

Module+			Params/M	FPS/FLOPs	CTIR/%				SIRST-UAVB/%			
MFFM	MFECConv	CAFM			<i>P</i>	<i>R</i>	mAP50	mAP95	<i>P</i>	<i>R</i>	mAP50	mAP95
			2.71	119/8.4	79.2	75.2	84.2	59.2	91.2	74.6	83.3	55.1
✓			2.69	122/8.4	83.8	77.9	88.3	63.3	91.8	75.9	84.7	56.5
	✓		2.32	127/7.5	82.5	77.4	87.7	62.7	92.5	76.7	85.5	57.2
		✓	2.79	119/8.8	82.6	78.8	87.3	61.3	93.3	76.9	86.9	57.8
✓	✓		2.32	124/8.0	84.3	79.2	88.7	63.8	93.2	77.8	87.6	58.1
✓	✓	✓	2.56	122/8.2	84.7	80.4	90.7	64.2	93.7	77.8	88.9	58.8

成 MFConv 模块, 观察到使用双模块的实验结果中, 各项指标中均优于单模块。第 6 组在第 5 组基础上集成 CAFM 模块, 即为本文方法, 以更少的参数量 Params 和浮点运算量 FLOPs, 实现了更高的推理速度 FPS, 效率优势明显。在 CTIR 数据集上, 其余各项评估指标均达到最高。随后, 在更具挑战的 SIRST-UAVB 数据集上依旧表现稳健, 进一步证明了本文方法高效、精准且适应性强。

4.3 可视化实验

AFITDYOLO 模型主干、颈部和检测头网络的是基于 YOLOv10 模型上的改进, AFITDYOLO 与 YOLOv10 结构相似。因此选择轻量化检测模型 YOLOv10n 作为可视化实验的对照组, 实验结果如图 6 所示, 图中每个数据集的第一行为 YOLOv10n 可视化结果, 第二行为 AFITDYOLO 可视化结果。可以看出, 对于 CTIR 数据集, 本文所提方法在背景复杂且目标较多的情况下可以以更准确检测到不同尺度目标, 准确检测到各种遮挡的车辆与行人目标, 减少了漏检和误检的发生概率。在

SIRST-UAVB 数据集中, 对于红外远距离的检测目标置信度更高, 因此微小目标的定位也更精准。

AFITDYOLO 很大程度上减少了漏检和误检的概率, 该模型在多种复杂环境下的大小尺度目标的红外检测中, 取得了良好的检测结果, 可应用于多种环境中的检测任务, 实现跨场景部署。

4.4 与其他方法比较

4.4.1 MFConv 卷积对比实验

本文将 MFConv 卷积模块与现主流卷积模块进行对比实验, 结果如表 2 所示。本文选择以下卷积模块作为对照组: 可变核卷积 AKConv^[39]、深度可分离卷积 DSConv、大型选择性核卷积 LSKConv^[43] 及风车型卷积 PConv, 为保证实验的有效性和可行性, 其他模块均采用与本文 MFConv 相同的模块集成方法, 每个卷积模块替换 C2f 中 Bottleneck 的 Conv 层。对于 SIRST-UAVB 数据集, 可以看出 MFConv 在 $k=3$ 时优于 PConv 之外的所有卷积模块, 而在目标尺度更大的 CTIR 数据集中, 其表现则优于包括 PConv 的所有卷积

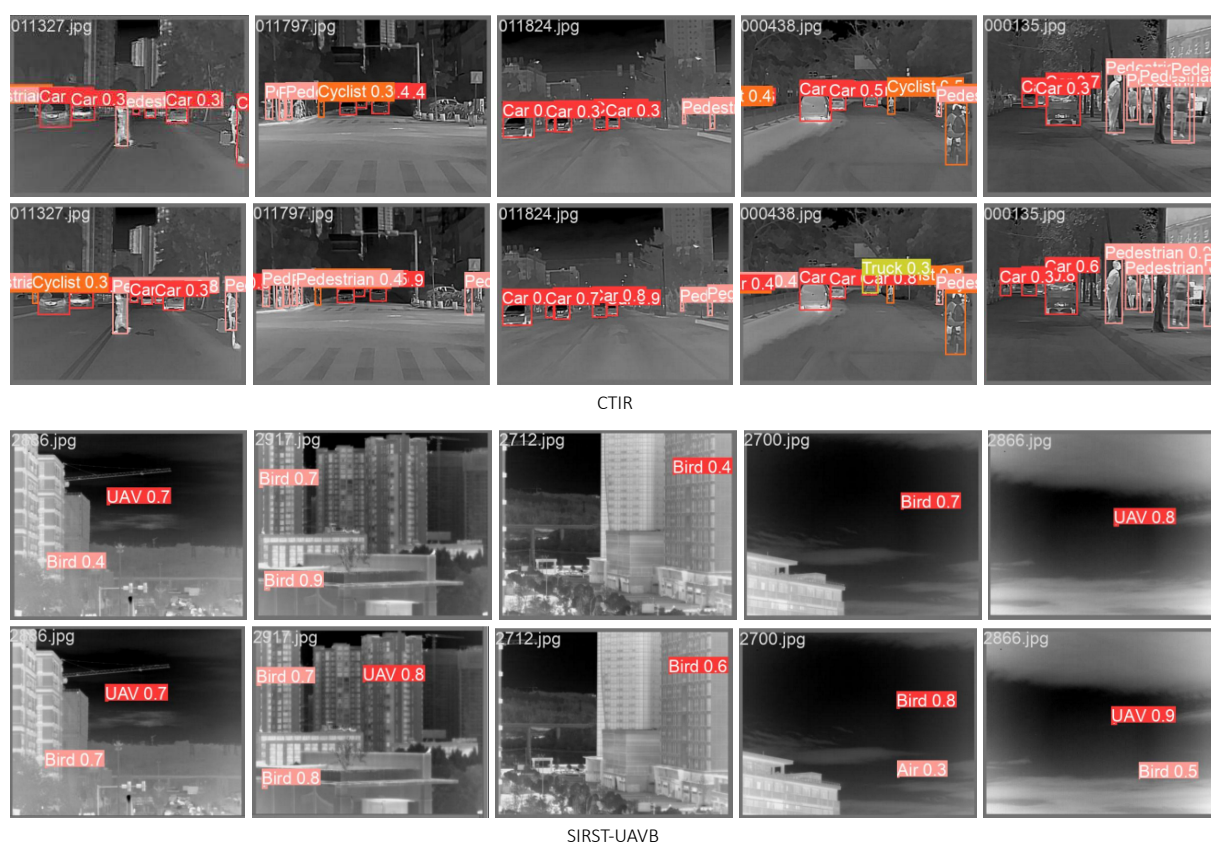


图 6 对比可视化

Fig. 6 Compare visualization results

模块。MFEConv 在 $k = 7$ 时性能达到最优，但 Params 和 FLOPs 为最高值，同时 FPS 下降明显。

通过堆叠奇数卷积核或增加卷积核长度能扩展感受野^[41]。由表 2 中可以看出增加 MFEConv 的卷积核的长度并没有产生可观的性能增益，同时卷积核长度的增加，参数量和浮点计算量会相应增加，降低检测速度。对于 MFEConv 异构卷积组，采用 6 个方向的卷积核进行构建相比于 3 个方向的卷积核参数量略微增加，但检测精度大幅提高。MFEConv 通过堆叠奇数卷积核来扩展感受野，并且构建了由 6 个短卷积核构成的卷积层，实现该模块感受野全覆盖。由表 2 可以看出，设定卷积核数量为 6，卷积核长度 $k = 3$ 的情况下，在损失较少性能的同时实现最佳的模块轻量化效果，并且考虑到整体网络参数量，本文设定卷积核长为 3，使整体网络的综合性能最佳。

4.4.2 与主流模型对比

为了进一步验证模型的有效性，按照相同的实验环境，在本文的数据集上，将本文所提算法与其他主流深度学习检测算法进行对比。实验结果由表 3 所示，可以看出，AFITDYOLO 检测精度优于实验所提单模态目标检测模型 (1~6 行)，其中 FR-CNN^[44] 为经典的二阶检测算法，mAP50 达到了 84.0%，但其参数量大，检测速度低。在 YOLO 系列算法中，YOLOv11n^[45]、YOLOv12n^[46] 通过轻量化改造，减少了模型的参数量，但损失了一定的检测性能。D-FINE-N^[47]、DEIM^[48] 是基于 Transformer 架构的两个目标检测模型，在不增加额外推理和训练成本的情况下，实现良好的检测性能，但准确性和检测速度均低于本文方法。

综上，从实验结果可知，本文提出的模型参数量低于目前主流模型，轻量化效果优秀，同时实现了良好检

表 2 MFEConv 与其他卷积的比较
Table 2 Comparison of MFEConv with other convolutions

Model	Params/M	FPS/FLOPs	CTIR/%				SIRST-UAVB/%			
			<i>P</i>	<i>R</i>	mAP50	mAP95	<i>P</i>	<i>R</i>	mAP50	mAP95
Conv	2.71	119/8.4	79.2	75.2	84.2	59.2	91.2	74.6	83.3	55.1
AKConv	2.69	126/7.8	80.2	74.7	83.6	58.8	91.1	74.3	81.4	54.9
DSConv	2.32	133/7.4	77.9	75.4	83.9	59.0	89.7	74.8	82.6	54.7
LSKConv	2.72	117/8.4	79.7	76.7	85.2	60.2	90.3	75.6	83.9	55.6
PConv	2.32	129/7.6	82.9	77.9	86.5	61.3	92.6	76.8	85.7	57.6
MFEConv(3) _{k=5}	2.38	121/7.7	81.2	77.5	86.9	62.1	92.3	76.7	84.6	56.3
MFEConv(3) _{k=3}	2.32	127/7.5	81.2	77.3	86.3	61.6	92.2	76.7	84.3	56.1
MFEConv(6) _{k=7}	2.5	112/8.6	83.2	77.7	87.7	62.9	92.8	77.3	86.1	57.6
MFEConv(6) _{k=5}	2.41	118/8.0	82.9	77.3	87.8	63.1	92.6	76.6	85.6	57.2
MFEConv(6) _{k=3}	2.32	127/7.5	83.1	77.4	87.7	62.7	92.5	76.7	85.5	57.2

表 3 与主流模型的比较
Table 3 Comparison results with mainstream models

Model	Params/M	FPS/FLOPs	CTIR/%				SIRST-UAVB/%			
			<i>P</i>	<i>R</i>	mAP50	mAP95	<i>P</i>	<i>R</i>	mAP50	mAP95
FR-CNN	41.38	56/18.3	80.3	76.2	83.9	58.8	87.2	74.8	83.5	55.1
YOLOv10s	7.21	95/24.4	82.7	76.5	85.1	59.2	91.3	75.2	83.7	55.3
YOLOv10m	15.43	83/64	82.9	77	85.3	59.4	91.1	75.1	84.1	55.6
YOLOv10n	2.71	119/8.4	79.2	75.2	84.2	59.2	91.2	74.6	83.3	55.1
YOLOv11n	2.63	133/6.6	82.9	78.9	87.7	62.7	91	76.9	84.9	57.4
YOLOv12n	2.61	141/6.0	77.3	75.5	81.3	58.2	86.3	71.5	80.4	54.1
D-FINE-N	3.97	76/7.0	80.2	77.8	86.6	62.1	91.7	78.1	86.4	57.5
DEIM-N	5.35	81/7.03	82.5	78.2	86.5	62.3	91.3	78.3	86.3	57.2
AFITDYOLO	2.56	122/8.2	84.7	80.4	90.7	64.2	93.7	77.7	88.9	58.8

测精度, 能够满足实时检测需求。

4.5 泛化性实验

本文选取高空无人机红外热成像数据集 HIT-UAV^[41]和经典红外小目标数据集IRSTD-1k^[49]进行泛化性实验。HIT-UAV 数据集包含不同场景下无人机拍摄的 2898 张红外图像。该数据集的图像包括人员、自行车、汽车、其他车辆等目标。IRSTD-1k 数据集由 1000 张红外图像组成, 图像尺寸为 512 pixel×512 pixel, 包含不同环境下的无人机、生物、船只和车辆四个目标。选择 YOLOv10n、YOLOv12n 与 DEIM-N 作为实验对照组。实验结果如表 4 所示, HIT-UAV 数据集上, AFITDYOLO 与其他算法相比, *P* 值略差于 DEIM-N, 其余指标均达到最高值。在 ISTD-1k 数据集上, 各项指标均最佳。可以看出, 本文方法在泛化实验所用的红外数据集上也实现了检测精度的提高, 验证了本文所提方法的泛化性, 进一步说明本文方法可以跨场景部署。

表 4 泛化性实验结果/%
Table 4 Generalization experiment results/%

Dataset	Model	<i>P</i>	<i>R</i>	mAP50	mAP95
HIT-UAV	YOLOv10n	91.1	69.2	77.8	49.2
	YOLOv12n	90.3	68.5	76.5	48.9
	DEIM-N	92.1	73.2	82.5	51.3
	AFITDYOLO	91.7	74.7	83.7	52.3
IRSTD-1k	YOLOv10n	89.0	83.7	87.4	56.3
	YOLOv12n	89.2	82.9	86.8	55.7
	DEIM-N	90.3	83.3	89.8	58.3
	AFITDYOLO	90.7	83.9	90.9	58.8

5 结 论

本文提出的基于 YOLO 的自适应多尺度红外目标检测网络 AFITDYOLO, 在较低参数量情况下提高了不同背景下, 不同尺度红外目标的检测精度。本文提出多尺度特征融合模块 MFFM, 通过改进 FPN 中不同层特征之间的相关性, 增强了模型多尺度特征融合的表达能。本文还设计了轻量化卷积模块 MFECov, 利用红外图像的目标分布特性, 以最小的参数实现高效、更大的感受野。接着引入交叉注意力融合模块 CAFM, 通过不同层输出特征图的对比交互, 突出重要特性信息, 滤除无关背景信息和抑制噪声, 进一步增强模型特征表达能力。

实验结果表明, 本文方法优于目前主流的方法, 展示了良好的检测精度、轻量化程度和泛化能力, 并且具有跨场景部署能力。

作者贡献

汪佳旭: 研究资料的收集、分析, 实验的设计与完成, 论文写作; 杨俊, 许聪源: 指导论文选题、论文修改。

利益冲突

所有作者声明无利益冲突

参考文献

1. Liu R M, Lu Y H, Gong C L, et al. Infrared point target detection with improved template matching[J]. *Infrared Phys Technol*, 2012, **55**(4): 380–387.

2. Rivest J F, Fortin R. Detection of dim targets in digital infrared imagery by morphological image processing[J]. *Opt Eng*, 1996, **35**(7): 1886–1893.

3. Dai Y M, Wu Y Q, Zhou F, et al. Attentional local contrast networks for infrared small target detection[J]. *IEEE Trans Geosci Remote Sensing*, 2021, **59**(11): 9813–9824.

4. Du P, Hamdulla A. Infrared small target detection using homogeneity-weighted local contrast measure[J]. *IEEE Geosci Remote Sensing Lett*, 2020, **17**(3): 514–518.

5. Liu Y J, Liu X Y, Hao X Y, et al. Single-frame infrared small target detection by high local variance, low-rank and sparse decomposition[J]. *IEEE Trans Geosci Remote Sensing*, 2023, **61**: 5614317.

6. Li R H, Shen Y. YOLOSR-IST: a deep learning method for small target detection in infrared remote sensing images based on super-resolution and YOLO[J]. *Signal Proc*, 2023, **208**: 108962.

7. Yang B, Zhang X Y, Zhang J, et al. EFLNet: enhancing feature learning network for infrared small target detection[J]. *IEEE Trans Geosci Remote Sensing*, 2024, **62**: 5906511.

8. GUPTA A, GUPTA U. Real time target detection for infrared images[C]//2020 Fourth International Conference on Inventive Systems and Control (ICISC), 2020: 570–574. <https://doi.org/10.1109/ICISC47916.2020.9171208>.

9. Yang J N, Liu S L, Wu J J, et al. Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection[C]//Proceedings of the 39th AAAI Conference on Artificial Intelligence, 2025: 9202–9210. <https://doi.org/10.1609/aaai.v39i9.32996>.

10. Zhao B, Wang C P, Fu Q, et al. Multi-scale infrared pedestrian detection based on deep attention mechanism[J]. *Acta Opt Sin*, 2020, **40**(5): 0504001. 赵斌, 王春平, 付强, 等. 基于深度注意力机制的多尺度红外行人检测[J]. *光学学报*, 2020, **40**(5): 0504001.

11. Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005: 886–893. <https://doi.org/10.1109/CVPR.2005.177>.

12. He H Y, Yao J G, Jiang Z L, et al. Infrared thermal image detecting of high voltage insulator contamination grades based on support vector machine[J]. *Autom Electr Power Syst*, 2005, **29**(24): 70–74,82. 何洪英, 姚建刚, 蒋正龙, 等. 基于支持向量机的高压绝缘子污秽等级红外热像检测[J]. *电力系统自动化*, 2005, **29**(24): 70–74,82.

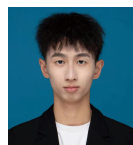
13. Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. *IEEE Trans Pattern Anal Mach Intell*, 2010, **32**(9): 1627–1645.

14. Mathieu M, LeCun Y, Fergus R, et al. OverFeat: integrated recognition, localization and detection using convolutional networks[C]//*International Conference on Learning Representations*, 2014.
15. Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation[C]//*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 8759–8768.
<https://doi.org/10.1109/CVPR.2018.00913>.
16. Ghiasi G, Lin T Y, Le Q V. NAS-FPN: learning scalable feature pyramid architecture for object detection[C]//*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019: 7029–7038.
<https://doi.org/10.1109/CVPR.2019.00720>.
17. Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection[C]//*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020: 10778–10787.
<https://doi.org/10.1109/CVPR42600.2020.01079>.
18. Min K, Lee G H, Lee S W. Attentional feature pyramid network for small object detection[J]. *Neural Netw*, 2022, **155**: 439–450.
19. Dai Y M, Li X, Zhou F, et al. One-stage cascade refinement networks for infrared small target detection[J]. *IEEE Trans Geosci Remote Sensing*, 2023, **61**: 5000917.
20. Dai Y M, Wu Y Q, Zhou F, et al. Asymmetric contextual modulation for infrared small target detection[C]. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021: 949–958.
<https://doi.org/10.1109/WACV48630.2021.00099>.
21. Ren D D, Li J B, Han M, et al. DNANet: dense nested attention network for single image dehazing[C]//*ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021: 2035–2039.
<https://doi.org/10.1109/ICASSP39728.2021.9414179>.
22. Fan W Q, Xu X M, Cai B L, et al. ISNet: individual standardization network for speech emotion recognition[J]. *IEEE/ACM Trans Audio Sp Lang Process*, 2022, **30**: 1803–1814.
23. Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]//*15th European Conference on Computer Vision*, 2018: 3–19.
https://doi.org/10.1007/978-3-030-01234-2_1.
24. Dai T, Cai J R, Zhang Y B, et al. Second-order attention network for single image super-resolution[C]//*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019: 11057–11066.
<https://doi.org/10.1109/CVPR.2019.01132>.
25. Dai T, Zha H, Jiang Y, et al. Image super-resolution via residual block attention networks[C]//*2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019: 3879–3886.
<https://doi.org/10.1109/ICCVW.2019.00481>.
26. Niu B, Wen W L, Ren W Q, et al. Single image super-resolution via a holistic attention network[C]//*16th European Conference on Computer Vision – ECCV 2020*, 2020: 191–207.
https://doi.org/10.1007/978-3-030-58610-2_12.
27. Wang L, Shen J, Tang E, et al. Multi-scale attention network for image super-resolution[J]. *J Vis Commun Image Repres*, 2021, **80**: 103300.
28. Varghese R, M S. YOLOv8: a novel object detection algorithm with enhanced performance and robustness[C]//*2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 2024: 1–6.
<https://doi.org/10.1109/ADICS58448.2024.10533619>.
29. Terven J, Córdova-Esparza D M, Romero-González J A. A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS[J]. *Mach Learn Knowl Extr*, 2023, **5**(4): 1680–1716.
30. Chen H, Chen K, Ding G G, et al. YOLOv10: real-time end-to-end object detection[C]//*Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024: 107984–108011.
<https://doi.org/10.52202/079017-3429>.
31. Zhao F, Li S J, Zhang J J, et al. Convolution transformer fusion splicing network for hyperspectral image classification[J]. *IEEE Geosci Remote Sensing Lett*, 2023, **20**: 5501005.
32. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]//*Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010: 249–256.
33. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//*Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, 2015: 448–456.
34. Bieder F, Sandkühler R, Cattin P C. Comparison of methods generalizing max- and average-pooling[Z]. arXiv: 2103.01746, 2021.
<https://doi.org/10.48550/arXiv.2103.01746>.
35. Kusupati A, Ramanujan V, Somani R, et al. Soft threshold weight reparameterization for learnable sparsity[C]//*Proceedings of the 37th International Conference on Machine Learning*, 2020: 5544–5555.
36. Chollet F. Xception: deep learning with depthwise separable convolutions[C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017: 1800–1807.
<https://doi.org/10.1109/CVPR.2017.195>.
37. Xu W, Wan Y. ELA: efficient local attention for deep convolutional neural networks[Z]. arXiv: 2403.01123, 2024.
<https://doi.org/10.48550/arXiv.2403.01123>.
38. Fløystad G. Shift modules, strongly stable ideals, and their dualities[J]. *Trans Am Math Soc Ser B*, 2023, **10**(21): 670–714.
39. Zhang X, Song Y Z, Song T T, et al. AKConv: convolutional kernel with arbitrary sampled shapes and arbitrary number of parameters[Z]. arXiv: 2311.11587, 2023.
<https://doi.org/10.48550/arXiv.2311.11587>.
40. Luo Y, Wong Y, Kankanhalli M, et al. G-softmax: improving intraclass compactness and interclass separability of features[J]. *IEEE Trans Neural Netw Learn Syst*, 2020, **31**(2): 685–699.
41. Zhao X F, Zhang W W, Zhang H, et al. ITD-YOLOv8: an infrared target detection model based on YOLOv8 for unmanned aerial vehicles[J]. *Drones*, 2024, **8**(4): 161.
42. Li Y X, Zou W B, Wei Q M, et al. Multi-level feature fusion network for lightweight stereo image super-resolution[C]. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024: 6489–6498.
<https://doi.org/10.1109/CVPRW63382.2024.00649>.
43. Wang X, He N, Hong C, et al. Improved YOLOX-X Based UAV aerial photography object detection algorithm[J]. *Image Vision Comput*, 2023, **135**: 104697.
44. Maity M, Banerjee S, Chaudhuri S S. Faster R-CNN and YOLO based vehicle detection: a survey[C]//*2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021: 1442–1447.
<https://doi.org/10.1109/ICCMC51019.2021.9418274>.
45. Sani A R, Zolfagharian A, Kouzani A Z. Automated defects detection in extrusion 3D printing using YOLO models[J]. *J Intell Manuf*, 2024.
<https://doi.org/10.1007/s10845-024-02543-8>.
46. Alif M A R, Hussain M. YOLOv12: a breakdown of the key architectural features[Z]. arXiv: 2502.14740, 2025.
<https://doi.org/10.48550/arXiv.2502.14740>.
47. Peng Y S, Li H B, Wu P X, et al. D-FINE: redefine regression task in DETRs as fine-grained distribution refinement[Z]. arXiv: 2410.13842, 2024.
<https://doi.org/10.48550/arXiv.2410.13842>.
48. Huang S H, Lu Z C, Cun X, et al. DEIM: DETR with improved matching for fast convergence[C]//*2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025: 15162–15171.
<https://doi.org/10.1109/CVPR52734.2025.01412>.

49. Zhang M J, Zhang R, Yang Y X, et al. ISNet: shape matters for infrared small target detection[C]//2022 IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR), 2022: 867–876.
<https://doi.org/10.1109/CVPR52688.2022.00095>.

作者简介



汪佳旭 (2001-), 男, 浙江理工大学信息科学与工程学院 (网络安全学院) 硕士研究生, 研究方向为机器视觉、深度学习理论及应用。

E-mail: 2936598421@qq.com



许聪源 (1990-), 男, 工学博士, 嘉兴大学人工智能学院讲师, 研究方向为机器视觉、网络空间安全。

E-mail: cyxu@zjxu.edu.cn



【通信作者】杨俊 (1978-), 男, 工学博士, 硕士研究生导师, 嘉兴大学人工智能学院副教授, 研究方向为智能多媒体信息分析与处理, 机器学习, 计算机视觉。

E-mail: juneryoung@zjxu.edu.cn

YOLO-based adaptive multi-scale infrared target detection network

Wang Jiaxu^{1,2}, Yang Jun^{2*}, Xu Congyuan²

¹School of Information Science and Engineering(School of Cyber Science and Technology), Zhejiang Sci-Tech University, Hangzhou, Zhejiang 310018, China;

²College of Artificial Intelligence, Jiaxing University, Jiaxing, Zhejiang 314001, China

Abstract

Objective Infrared target detection plays an indispensable role in numerous critical domains, such as security surveillance, autonomous driving, and military reconnaissance, owing to its unique perceptual capability under complex environments (e.g., low-light conditions and severe weather). However, infrared images inherently suffer from low contrast, blurred details, and significant noise interference, which often lead to ambiguous target edges, missing texture features, and other challenges during the detection process. Existing deep learning-based infrared target detection algorithms (ITDA) exhibit inadequate performance in feature extraction and processing for infrared images, resulting in relatively high rates of missed detection and false detection. Moreover, our systematic analysis of infrared target detection tasks reveals that algorithms tailored for small infrared targets rely heavily on high-sensitivity feature extraction to capture subtle characteristics. Nevertheless, as the scale of detected targets increases, these algorithms tend to encounter overfitting to local textures and elevated false detection rates, thereby degrading overall performance. In practical applications, detection environments are dynamically changing with targets of varying scales; thus, multi-scale detection capability is critical to ensuring algorithms maintain high reliability and adaptability in complex real-world scenarios. Unfortunately, most state-of-the-art algorithms are optimized for single-scale targets, making it challenging to simultaneously satisfy the requirements of high-precision localization for small targets and effective semantic understanding for large targets.

Methods To address the above issues, this paper proposes an adaptive multi-scale infrared target detection network based on YOLO (AFITDYOLO). This network is designed to receive infrared target images of different scales and employs a multi-layer feature extraction module and a multi-layer feature fusion module to enhance its multi-scale infrared target detection capability. Firstly, a multi-scale feature fusion module (MFFM) is proposed. This module enhances the correlation between features of different layers in the feature pyramid network (FPN), coordinates deep semantic features with shallow spatial detail features more effectively, and thereby improves the representational ability of multi-scale feature fusion. Secondly, a multi-kernel feature extraction convolution (MFEConv) is constructed. By

Foundation item: Supported by: National Natural Science Foundation of China (62302197); Zhejiang Province Higher Education "14th Five-Year Plan" Postgraduate Teaching Reform Project (JGGC2024034); Jiaxing City Science and Technology Project (2024AD10045, 2024AY40010)

*Correspondence: Yang J, E-mail: juneryoung@zjxu.edu.cn

utilizing heterogeneous convolution groups, MFEConv expands the receptive field and strengthens the model's feature extraction capability. Additionally, a cross-attention fusion module (CAFM) is designed. Through the comparative interaction of feature maps output by different layers in the detection network, CAFM leverages the complementary information among these feature maps to suppress infrared noise in images and further enhance feature representation capability.

Results and Discussions To validate the effectiveness of the proposed method in improving detection performance, extensive training and evaluation are conducted on the CTIR dataset, which comprises road pedestrians and vehicles with multi-scale infrared targets. To further verify the adaptability of the method, additional experiments are performed on the SIRST-UAVB dataset—a single-frame UAV bird dataset characterized by more complex backgrounds and smaller target scales. Experimental results on these two datasets demonstrate that AFITDYOLO achieves mean average precision at 50% intersection over union (mAP50) of 88.9% and 90.7%, respectively, representing significant improvements of 5.6% and 6.5% compared with YOLOv10n. In terms of lightweight optimization, the proposed method achieves higher inference speed (measured in frames per second, FPS) while utilizing fewer model parameters (params) and floating-point operations (FLOPs). When compared with current mainstream methods, AFITDYOLO exhibits the highest detection accuracy, the lowest parameter count and FLOPs, and the fastest inference speed, demonstrating distinct advantages. Additionally, to evaluate the generalization ability of the proposed method, cross-dataset experiments are carried out on the HIT-UAV dataset (a high-altitude UAV infrared thermal imaging dataset) and theIRSTD-1k dataset (a classic infrared small target dataset). Experimental results indicate that while the precision (P) value of AFITDYOLO is slightly inferior to that of DEIM-N, it outperforms all other mainstream methods in all remaining evaluation metrics. These findings confirm that the proposed method achieves improved detection accuracy on the infrared datasets used in the generalization experiments, validating its strong generalization capability and further demonstrating its feasibility for cross-scenario deployment. Overall, the proposed method simultaneously achieves enhanced detection accuracy and lightweight optimization of the detection model, fully meeting the requirements of real-time detection applications.

Conclusions The AFITDYOLO network proposed in this paper, which is an adaptive multi-scale infrared target detection network based on YOLO, enhances the detection accuracy of infrared targets of different scales under various backgrounds with a relatively small number of parameters. The proposed MFFM enhances the model's representational ability in multi-scale feature fusion by improving the correlation between features of different layers in FPN. Additionally, the lightweight convolution module MFEConv is designed to achieve an efficient and larger receptive field with minimal parameters by leveraging the target distribution characteristics of infrared images. Furthermore, the CAFM is introduced to highlight important feature information, filter out irrelevant background information, and suppress noise through the comparative interaction of feature maps output by different layers, thereby further boosting the model's feature representation capability. Experimental results demonstrate that the proposed method outperforms current mainstream algorithms, exhibiting excellent detection accuracy, lightweight performance, and generalization ability, along with cross-scenario deployment capabilities.

Keywords: infrared target detection; feature fusion; lightweight network architecture; attention mechanism; YOLO

Wang J X, Yang J, Xu C Y. YOLO-based adaptive multi-scale infrared target detection network[J]. *Opto-Electron Eng*, 2026, 53(4): 250292; DOI: [10.12086/oe.2026.250292](https://doi.org/10.12086/oe.2026.250292)

