OEE | 光电工程 Opto-Electronic Engineering

Article

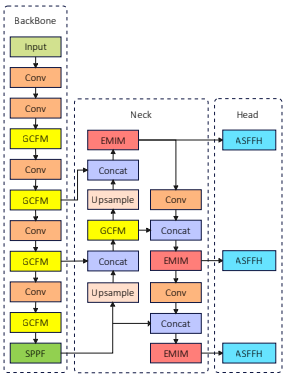
2026 年 5 月, 第 53 卷, 第 5 期

YOLO-MITD：多尺度红外目标检测算法

梁学才[1], 杜海娟[1*], 李锦龙[1], 周羿阳[1], 杨　赟[1], 石　博[1], 杨　俊[1*], 许聪源[1], 汪佳旭[1,2]

[1]嘉兴大学人工智能学院，浙江 嘉兴，314001；

[2]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



摘要：针对无人机红外目标检测中存在计算资源受限、背景干扰强及目标尺度多变等挑战，本文提出了一种基于 YOLO 的多尺度红外目标检测算法 YOLO-MITD。首先，设计 Ghost 增强型跨级融合模块 (ghost-enhanced cross-stage fusion module, GCFM)，通过 GhostConv 轻量化卷积重构特征融合路径，显著减少模型参数量并增强多尺度特征提取能力；其次，构建高效的多尺度交互模块 (efficient multi-scale interaction module, EMIM)，通过跨维度交互强化关键特征通道与空间关联性，提升复杂背景下目标显著性表达；然后，设计自适应空间特征融合检测头 (adaptive spatial feature fusion head, ASFFH)，通过动态权重分配实现多尺度特征融合并抑制特征冲突，有效提升目标检测的尺度一致性；最后，构造自适应阈值焦点损失函数 (adaptive threshold focal loss, ATFL)，通过动态调整难例挖掘阈值优化训练过程，缓解类别不平衡问题。在 HIT-UAV 红外数据集上实验结果表明，所提算法以少量计算开销，mAP50 达到 86.7%，较基准 YOLOv8n 提升 6.1%。在 CTIR 与 SIRST 数据集上的跨场景泛化实验进一步验证了算法的鲁棒性，mAP50 分别达到 85.4% 与 67.2%，显著优于主流模型。

关键词：红外目标检测；轻量卷积；注意力机制；多尺度融合；自适应损失

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

DOI: 10.12086/oee.2026.250322 | CSTR: 32245.14.oee.2026.250322 | 中图分类号：O434.3 | 文献标识码：A

梁学才，杜海娟，李锦龙，等. YOLO-MITD：多尺度红外目标检测算法 [J]. 光电工程，2026，**53**(5): 250322

Liang X C, Du H J, Li J L, et al. YOLO-MITD: a multi-scale infrared target detection algorithm[J]. *Opto-Electron Eng*, 2026, **53**(5): 250322

1　引　言

随着微小型无人机在军事侦察、交通巡查、灾害监测及边境安防等领域的广泛应用，基于红外成像的对地目标检测技术已成为研究热点[1]。与可见光成像相比，红外成像具备全天时、全天候工作能力及较强的抗干扰性能，能够在夜间、雾霾等低能见度条件下稳定提供目标热辐射信息[2]。然而，无人机平台受飞行高度与传感器限制，所获取的红外图像普遍存在空间分辨率低、信噪比差、目标尺度跨度大等问题。特别是在远距离拍摄条件下，目标往往呈现为多尺度分布特征，从数像素的小目标到覆盖数十像素的中大目标共存于同一场景。传统检测方法在复杂背景下易出现误检率高、漏检严重的现象[3]。因此，如何在资源受限的无人机平台上以及复杂环境下实现高精度的红外目标检测，仍是当前的技术难题。

收稿日期：2025-10-28；修回日期：2026-01-25；录用日期：2026-01-26；网络出版日期：2026-05-25

*通信作者：杜海娟，444567707@qq.com；杨俊，juneryoung@zjxu.edu.cn

版权所有©2026 中国科学院光电技术研究所

近年来，基于深度学习的目标检测算法显著推动了红外目标检测的发展。YOLO (You only look once)[4] 系列算法凭借其端到端、单阶段推理架构，在检测速度与精度间取得了较好平衡。YOLOv3[5] 首次引入多尺度预测机制，增强了对小目标的感知能力；YOLOv5[6] 通过优化特征金字塔结构与自适应锚框机制，提升了多尺度目标的检测鲁棒性；YOLOv8n[7] 进一步采用解耦检测头与 Anchor-Free 设计，通过将分类和回归任务分离处理，提升了检测精度与训练稳定性，同时在保持较高精度的同时显著降低计算开销，更适合嵌入式平台部署。然而，面对无人机红外图像中目标尺度变化大、背景热噪声复杂、边缘纹理模糊等挑战，上述通用模型在多尺度特征提取能力、轻量化结构设计、训练过程对难分类样本敏感程度等方面仍存在一些短板，难以在精度与模型复杂度之间达到理想平衡。

为应对上述问题，研究者们从结构优化、注意力机制、特征融合等多个维度对 YOLO 模型进行了改进。Ding 等[8] 提出的 CEDP-YOLO 通过复杂的上下文增强金字塔和动态感知头，有效提升了无人机场景下的检测精度。然而，该方法导致模型参数量增加了一倍，不利于轻量化部署。Cao 等[9] 提出了 YOLO-TSL，通过引入三重注意力机制 (triplet attention) 与 Slim-Neck 结构，在降低参数量的同时提升了检测速度。然而，其注意力机制仅在主干网络末端作用，未嵌入多尺度融合路径，导致浅层小目标特征在传递过程中被稀释。Wang 等[10] 提出了 LFIR-YOLO，通过引入可扩张残差模块 (dilation-wise residual, DWR) 与通道-空间注意力机制 (content-guided attention, CGA)，在红外道路场景中提升了车辆与行人检测精度。然而，该方法仅针对小目标增强感受野，未对多尺度特征融合结构优化，导致在中大目标上表现下降，且模型参数量较大，难以部署于资源受限的无人机平台。Luo 等[11] 提出了 IR-YOLO，专门设计小目标检测头并优化回归损失函数，缓解了目标遮挡导致的漏检问题。但其特征融合仍沿用传统 FPN 结构，缺乏跨尺度动态权重调整机制，在多尺度目标共存场景下 (如人车混合、远近目标并存) 易出现语义冲突，检测一致性差。刘奕阳等[12] 提出基于 YOLOv8n 的改进方法，通过增加小目标检测层并引入 BiFormer 注意力机制，虽然在 HIT-UAV 数据集上取得了一定提升，但模型复杂度显著增加，不利于轻量化部署。

现有方法在保持高精度与模型复杂度之间尚未达到理想平衡，尤其在应对红外图像中目标尺度多样、背景干扰强的场景时，缺乏一种既能有效提取多尺度特征，又具备高效推理能力的轻量化检测架构。针对上述局限，本文提出一种基于改进 YOLOv8n 的多尺度红外目标检测算法 YOLO-MITD，旨在通过结构重构与损失优化，在少量增加计算负担的前提下，提升模型对多尺度红外目标的检测能力与跨场景泛化性能。本文主要贡献包括以下几个部分。

1) 在主干网络中嵌入 GhostConv 轻量化卷积模块，通过线性变换替代部分标准卷积，构建 GCFM 模块，在减少参数量和提高运行效率的同时维持特征表达能力。

2) 在特征融合阶段引入高效多尺度注意力机制 (efficient multi-scale attention, EMA)，通过跨维度交互增强空间与通道特征关联，构建集成 EMA 的 EMIM 模块，提升复杂背景下目标的显著性特征表达。

3) 设计基于自适应空间特征融合 (adaptive spatial feature fusion, ASFF) 的改进检测头 ASFFH，对来自不同层级的特征进行自适应加权融合，有效增强多尺度特征的表征能力，缓解小目标因下采样导致的信息稀释。

4) 构造自适应目标聚焦损失 (ATFL)，替代分类损失，通过动态调整损失权重，提升难样本关注度与收敛稳定性。

2 YOLO-MITD 检测算法

改进后的 YOLOv8n 网络由图 1 所示，称为基于 YOLO 的多尺度红外目标检测算法 (YOLO-based algorithm for multi-scale infrared target detection, YOLO-MITD)，该模型通过设计 GCFM 模块、EMIM 模块及 ASFFH 模块，并采用构造的 ATFL 损失函数，为资源受限的无人机平台提供了一种高精度、低参数量和较好泛化能力的红外目标检测算法。

2.1 GCFM 模块

卷积神经网络 (Convolutional neural network，CNN) 在目标检测任务中取得了显著成效，然而其深层特征图中存在大量信息冗余，导致模型参数量与计算复杂度居高不下，限制了在资源受限平台上的部署能力[13]。为平衡模型性能与计算效率，本文引入 GhostConv 轻量化卷积模块[14]，其核心思想是通过特征重用的策略，维持特征表达能力，同时降低参数量与计算复杂度。

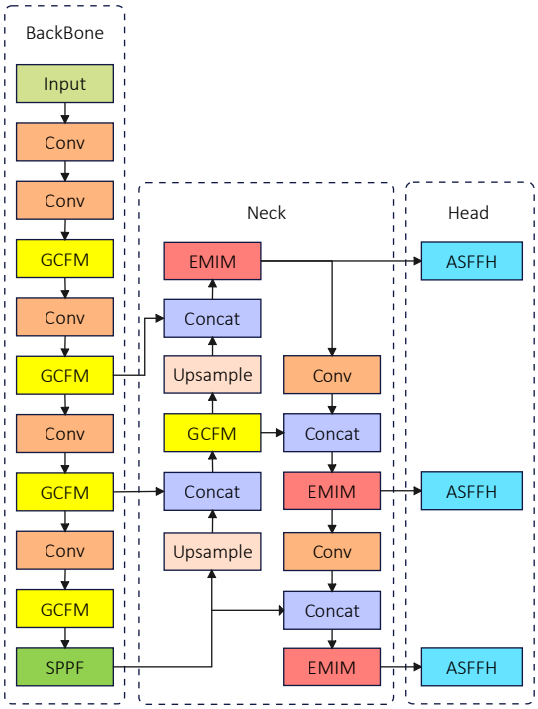


图 1 YOLO-MITD 网络结构

Fig. 1 YOLO-MITD network structure

GhostConv 模块采用两阶段特征生成机制，其结构如图 2 所示。给定尺寸为$C\times H\times W$的特征图 $\boldsymbol{X}$ 作为输入，首先通过标准卷积操作提取本质特征

$$\boldsymbol{Y}' = \boldsymbol{X} * \boldsymbol{f}' , \tag{1}$$

式中：$\boldsymbol{f}'$为卷积核权重；*表示卷积操作；输出特征图 $\boldsymbol{Y}'$包含m个通道，满足$m = n/s$，n为最终输出通道数，s为冗余倍数 (用于控制特征丰富度与计算成本的平衡)。

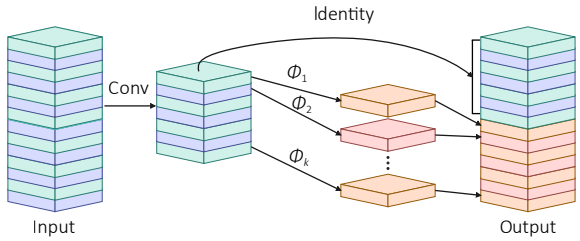


图 2 GhostConv 模块结构

Fig. 2 GhostConv module structure

然后，对$\boldsymbol{Y}'$中每张特征图$\boldsymbol{y}_i'$应用深度可分离卷积进行廉价变换。该变换使用3×3深度卷积核实现，其操作定义为

$$y_{ij} = \Phi_{ij}(\boldsymbol{y}_i') . \tag{2}$$

通过$s-1$次廉价变换，为每个本质特征图生成$s-1$张幻影特征图。

最后，将本质特征与冗余特征沿通道维拼接，得到完整输出

$$\boldsymbol{Y} = [y_{11}, y_{12}, \cdots, y_{ms}] , \tag{3}$$

其通道数$n = m\times s$，与标准卷积保持一致，但参数量显著降低。

此外，对于一个输入通道数为 c、输出通道数为 n 的卷积层而言，标准卷积的理论计算量为

$$F_{\text{std}} = n\times h'\times w'\times c\times k\times k , \tag{4}$$

式中：h'和w'为输出特征图尺寸；$k\times k$为卷积核大小，$k = 3$。相比之下，GhostConv 模块的计算量由主卷积和线性变换两部分构成：

$$\begin{aligned} F_{\text{ghost}} = &\frac{n}{s}\times h'\times w'\times c\times 1\times 1 + \frac{n}{s}\times h' \\ &\times w'\times (s-1)\times 3\times 3 \end{aligned} , \tag{5}$$

式中：第一项为主卷积 (1×1) 的计算量；第二项为廉价变换 (3×3深度卷积) 的计算量。因此，GhostConv 相对于标准卷积的理论加速比为

$$r_s = \frac{F_{\text{std}}}{F_{\text{ghost}}} \approx \frac{s\times c}{s + c - 1} , \tag{6}$$

式中：$s \ll c$，其加速比r_s可近似等于 s。通过引入 GhostConv 模块，理论上可以实现约 s 倍的模型加速效果。

本文将 YOLOv8 原有 C2f 模块内所有 Bottleneck 的 3×3标准卷积替换为 GhostConv 操作，构建 GCFM 模块，其结构如图 3 所示，在保持多尺度特征融合能力的基础上，实现轻量化。

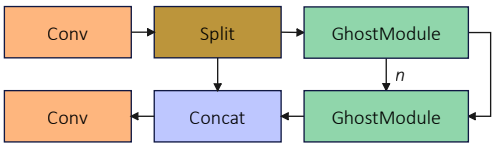


图 3 GCFM 模块结构

Fig. 3 GCFM module structure

2.2 EMIM 模块

红外图像依赖热辐射成像，其空间分辨率较低，导致小目标边缘与纹理在一次下采样后即出现显著模糊，传统加性或拼接式融合方法难以有效恢复此类细节[15]。尽管引入注意力机制可抑制背景干扰，但 SE 模块[16]

仅进行通道权重校准，对空间信息的建模相对有限；CBAM 模块[17] 虽采用通道与空间权重的串行计算，二者仍相互独立；Triplet Attention[18] 则需三次张量旋转操作，计算复杂度相对较高。为此，本文引入高效多尺度注意力机制 EMA[19]，构建 EMIM 模块，其结构如图 4 所示。

联合重标定，其具体实现机制如下：

给定尺寸为$C\times H\times W$的特征图 $\boldsymbol{X}$ 作为输入，首先沿通道维度将其均匀划分为 G 组，每组特征独立并行地输入两条支路进行处理。

1) 1×1分支：通过水平和垂直方向的一维全局平均池化分别捕获长程空间依赖关系，再结合1×1卷积与 GroupNorm 操作生成方向敏感的空间特征图，实现了通道间的跨空间信息交互。

2) 3×3分支：采用填充为 1 的3×3卷积保持空间分辨率，提取局部跨通道交互信息，进一步增强了空间与通道特征的关联性。

两分支输出分别重塑为维度$(b\times G, C/G, H, W)$，随后各自进行二维全局平均池化，并通过 Softmax 函数获得通道-空间联合注意力权重，其中 b 表示批次大小 (batch size)。该权重与原输入特征图通过逐元素乘法 (哈达玛积) 实现像素级重校准。该过程无需类似 SE 模块的降维与升维操作，计算复杂度控制在$O(C\times H\times W)$量级。

分组数 G 的取值是通过动态调整机制确定的,其表达式为

$$G=\min(F,C), \tag{7}$$

式中：F 是预设因子 (默认为 32)；C 是输入特征图的通道数。当$C\leqslant F$时，$G=C$，确保通道间的独立语义信息得以保留；当$C>F$时，$G=F$，从而限制计算复杂度并保持高效的特征表达能力。该动态调整策略在保证跨通道交互能力的同时，有效控制了计算开销。

本文将 EMA 模块嵌入 C2f 结构的每个 Bottleneck 模块的末端，构成 Bottleneck-EMA 子模块 (其结构如图 5 所示)，进而构建 EMIM 模块 (整体结构如图 6 所示)。每个 Bottleneck 模块保留原有的残差连接结构。其主路径依次经过标准1×1卷积、3×3卷积及1×3卷积处理后，与捷径连接的特征进行拼接融合。融合后的特征进一步输入至 EMA 子网络进行注意力加权，生成最终输出。该设计依托 C2f 结构固有的多分支特性与残差连接，有效促进了梯度流动与特征复用。同时，通过引入

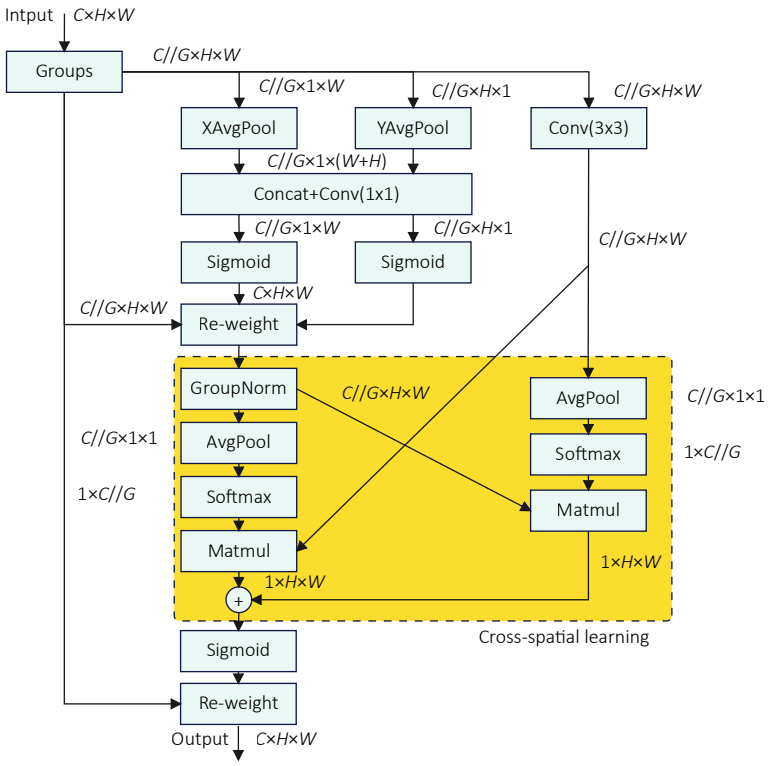


图 4 EMA 注意力机制结构

Fig. 4 EMA attention mechanism structure

EMA 模块实现了不需降维的跨维度特征筛选机制，在不增加计算开销的前提下，显著增强了模型对红外小目标的感知能力。

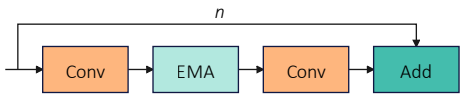


图 5 Bottleneck-EMA 模块结构

Fig. 5 Bottleneck-EMA module structure

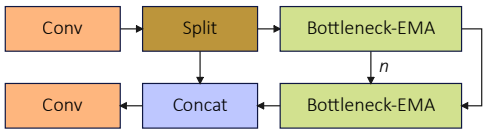


图 6 EMIM 模块结构

Fig. 6 EMIM module structure

2.3 ASFFH 模块

YOLOv8 原有的检测头存在特征金字塔不一致性的核心缺陷。各尺度特征图独立执行预测任务，导致同一目标在某一层级被视为正样本，而在其他层级则被判定为负样本。这种矛盾的梯度信号干扰模型训练过程，制约模型性能，尤其损害中小目标检测精度。

为此，本文提出一种基于自适应空间特征融合(ASFF)机制[20]的目标检测头 ASFFH。ASFFH 模块结构如图 7 所示，其中 Level-1、Level-2 和 Level-3 代表特征金字塔中具有不同空间分辨率的层级特征。ASFF-1、ASFF-2 和 ASFF-3 则表示应用 ASFF 机制进行特征融合后的各层级输出。该模块通过自适应融合来自不同金字塔层级的特征图，实现多尺度特征融合，从而增强模型对多尺度目标的检测能力。其采用的特征冲突抑制方法，有效解决了特征金字塔中跨尺度特征间的不一致性问题，如梯度冲突、语义冲突，提升了训练稳定性与检测精度。该模块显著提升了中小目标的检测性能，同时保持了较高的推理速度。

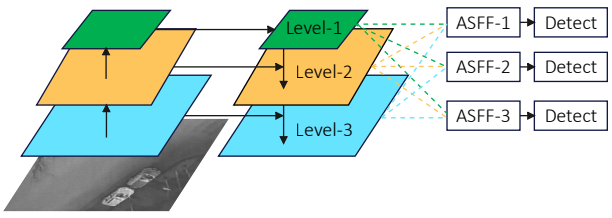


图 7 ASFFH 模块结构

Fig. 7 ASFFH module structure

ASFFH 的工作原理可分为以下三个步骤。

1) 特征重缩放

对于当前目标层级的特征图，将其他层级的特征图通过空间分辨率对齐操作统一至与该层级相同的尺度。具体来说，首先使用1×1卷积将各个输入特征图的通道数调整为统一维度，随后根据分辨率差异分别进行处理。对于低分辨率特征图，采用双线性插值方法进行上采样；对于高分辨率特征图，则使用步长为 2 的卷积操作进行下采样。这一策略在实现尺度统一的同时，最大程度保持了特征的结构完整性与语义一致性。

2) 自适应融合

对每个空间位置(i,j)，融合后的特征y_{ij}^{l}由以下公式计算：

$$y_{ij}^{l}=\alpha_{ij}^{l}\cdot x_{ij}^{1\to l}+\beta_{ij}^{l}\cdot x_{ij}^{2\to l}+\gamma_{ij}^{l}\cdot x_{ij}^{3\to l}, \tag{8}$$

式中：$x_{ij}^{k\to l}$表示第 k 层特征重缩放后在第 l 层位置(i,j)处的值；$\alpha_{ij}^{l},\beta_{ij}^{l},\gamma_{ij}^{l}$是通过 softmax 归一化的空间权重，由网络学习得到：

$$\alpha_{ij}^{l}=\frac{e^{\lambda_{\alpha_{ij}}^{l}}}{e^{\lambda_{\alpha_{ij}}^{l}}+e^{\lambda_{\beta_{ij}}^{l}}+e^{\lambda_{\gamma_{ij}}^{l}}}, \tag{9}$$

式中：权重图λ_{α}^{l}、λ_{β}^{l}、λ_{γ}^{l}由各层级特征经1×1卷积预测得到，具备空间敏感性。

3) 一致性保证

通过上述自适应权重机制，网络能够自动抑制来自冲突区域的梯度，如背景中被误检为正样本的区域，从而提升训练稳定性与特征表示的一致性。该机制尤其有利于中小目标在不同尺度下的语义对齐，增强模型对多尺度目标的检测鲁棒性。

2.4 损失函数优化

在红外目标检测任务中，通常存在类别不平衡问题，某些类别实例数量远少于其他类别。此外，由于背景占图像的大部分，训练过程中模型往往过多关注背景特征而忽视目标特征。因此，即使在网络结构层面引入了先进的注意力机制或多尺度融合策略以增强特征表达，若损失函数仍采用静态加权方式，模型在优化过程中依然会倾向于拟合占主导地位的简单背景样本，导致对关键但稀疏的目标区域学习不足[21]。传统 Focal Loss 虽然通过调制因子$(1-p_t)^{\gamma}$缓解了易分类样本的权重，但其固定参数γ无法适应不同样本的难易程度差异[22]。这使得网络在训练早期即陷入局部最优，大量梯度被简单样本稀释，稀有类别特征得不到充分学习。为此，本文构造自

适应阈值焦点损失 (ATFL)，其动态调整损失权重的机制能够更好地解决类别不平衡问题，同时提升训练的稳定性和收敛速度。

ATFL 的核心思想是通过分段加权机制动态调整损失函数的曲率，以适应不同置信度区间的样本分布。具体来说，ATFL 将样本分为高置信度 (易分类) 和低置信度 (难分类) 两类，并对两部分分别施加不同的调制策略：对于高置信度样本 ($p_t > 0.5$)，使用指数衰减机制将损失权重压缩至较低水平；对于低置信度样本 ($p_t \leqslant 0.5$)，则借助爆炸性增长机制将损失权重放大，从而加速梯度回传。其整体公式如下：

$$L_{\mathrm{ATFL}} = \begin{cases} -(1-p_t)^{-\ln(\hat{p}_c)}\log(p_t), & p_t > 0.5 \\ -(\lambda - p_t)^{-\ln(p_t)}\log(p_t), & p_t \leqslant 0.5 \end{cases}, \tag{10}$$

式中：p_t是样本属于真实类别的预测概率；λ是用于控制低置信度区间调制强度的参数；通过调整λ，可以进一步平衡难分类样本与易分类样本的损失权重，本文实验中将其值设置为 1.5。$\hat{p}_c$是当前批次预测置信度的滑动平均值，它平滑了批次间波动，同时能够及时反映模型在最新参数空间下的平均置信度，其计算公式为

$$\hat{p}_c = 0.05\frac{1}{k-1}\sum_{i=0}^{k-1}\overline{p}_i + 0.95p_k, \tag{11}$$

式中：k是当前训练轮次，$\overline{p}_i$是第i个训练轮次的平均预测置信度，p_k是当前训练轮次的平均预测置信度。

当整体预测置信度较高 ($\hat{p}_c$趋近于 1) 时，$-\ln(\hat{p}_c)$逐渐趋近于 0，从而避免对易分类样本的过度抑制；当整体预测置信度较低 ($\hat{p}_c$趋近于 0) 时$-\ln(\hat{p}_c)$趋近于正无穷，提高所有样本权重，防止训练停滞。

3 实　验

本节首先阐述数据集构成、评估指标及训练细节，然后通过消融实验探究各改进模块对模型性能的作用；随后，与当前主流方法开展对比实验，以验证本文所提出模型的优越性；此外，还针对损失函数开展对比实验，以验证本文所提出损失函数的有效性；接着，对目标检测结果进行可视化分析，最终通过不同数据集验证模型的泛化能力。

3.1 数据集

为了评估本文检测方法在红外目标检测的有效性，本文选用了 HIT-UAV[23] 数据集进行验证。HIT-UAV 是目前首个面向高海拔场景、公开发布的无人机红外热成像目标检测数据集，采集场景覆盖学校、停车场、道路、操场等多类真实环境，并同时包含白天与夜晚两种光照条件。

HIT-UAV 数据集共含 2898 张无人机红外航拍图像，分辨率均为 640 pixel×512 pixel。本文按 7:2:1 的比例将其划分为训练集、验证集与测试集，因此训练集、验证集和测试集分别有 2029 张、579 张和 290 张图片。划分过程采用随机抽样，确保各类场景与天气条件在三部分中均匀分布，以充分验证模型的泛化性能。

此外，数据集在目标尺寸分布上呈现出显著的多样性特征。具体而言，小尺寸目标占比约为 70%；中等尺寸目标占比约为 28%；而大尺寸目标则相对较少，仅占数据集总量的约 2%。这种涵盖了从小型到大型目标的尺寸范围，能够有效验证模型的多尺度适应与学习能力。

3.2 评价指标

本文通过 5 个指标来评估性能，即模型的精确率 (precision, P)、召回率 (recall, R)、平均精度 (mAP50, mAP50-95)、浮点计算量 (FLOPs) 和参数量 (Params)。评价指标计算公式如下：

$$P = \frac{TP}{TP+FP}\times 100\%, \tag{12}$$

$$R = \frac{TP}{TP+FN}\times 100\%, \tag{13}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n}AP_i, \tag{14}$$

式中：TP (true positive) 是指被正确识别的正类样本；FP (false positive) 是指被错误识别为正类的负类样本；FN (false negative) 是指被错误预测为负类的正类样本[24]；精确率 P 表示预测为正类的样本中实际为正类的比例；召回率 R 表示所有真实正样本中正确检测的正样本占的比例。

mAP50 是 IoU 阈值为 0.5 时所有目标类别的平均检测精度；而 mAP50-95 则表示在 IoU 阈值从 0.5 到 0.95、步长为 0.05 的情况下，各 mAP 的平均值[25]。FLOPs 和 Params 共同体现了模型的复杂度。FLOPs 数值越低，推理与检测所需的计算资源就越少。Params 越小，网络对硬件设备配置的要求就越低，也更易于在低端设备上实现部署和移植。

3.3 实验环境

本文实验使用 Ubuntu 22.04.3 操作系统，CPU 为 Intel Xeon Platinum 8558，GPU 为 NVIDIA GeForce RTX 4090D。编程语言采用 Python 3.11.9。深度学习框架选用 PyTorch 2.7.0。

本文所有实验都使用相同配置，参数设置也一致。输入图像分辨率统一设定为 640 pixel×640 pixel，该尺寸适配 HIT-UAV 数据集的图像特征。批尺寸设置为 32，可以充分利用服务器硬件资源，同时避免因批尺寸过大导致的过拟合风险或过小引起的训练波动。优化器采用 SGD，经实验验证，其相较于 Adam 等优化器在本文数据集上能获得更高的检测精度。动量因子和初始学习率使用默认值，分别为 0.937 和 0.01。而训练总轮数则设置为 400 轮，此时模型收敛效果最佳。

3.4 消融实验

为验证本文模块有效性，同时最大程度排除模型基础架构差异带来的干扰，本文以所提模型的基础架构 YOLOv8n 为基准，在 HIT-UAV 数据集上进行消融实验，如表 1 所示。

第二行表示在 YOLOv8n 基础上引入 GCFM 模块，相比于 YOLOv8n，在参数量减少 0.485 M、FLOPs 下降 6.5 G 的前提下，mAP50 提高 1.8%，mAP50-95 提高了 1.9%，召回率上升 2.8%，体现出 GCFM 在压缩模型规模的同时，仍能有效增强多尺度特征表达能力，兼顾轻量与性能。

第三行表示在 YOLOv8n 基础上引入 EMIM 模块，相比于 YOLOv8n，mAP50 提高了 4.2%，mAP50-95 提高了 3.4%，精确率和召回率分别提高了 2.3% 与 5.4%，显示出 EMIM 对检测精度的显著性提升。

第四行表示在 YOLOv8n 基础上引入 ASFFH 模块，相比于 YOLOv8n，mAP50 提高了 2.6%，mAP50-95 提高了 2.5%，精确率提高了 3.5%，但 FLOPs 增加 2.2 G，参数量增加 1.373 M，可以看出 ASFFH 模块能有效提升检测精度但会增加模型复杂度。

第五行表示在将 YOLOv8n 的损失函数替换为 ATFL 损失函数，相比于 YOLOv8n，该配置在计算量与参数量均未变化的前提下，mAP50 和 mAP50-95 均提高了 3.0%，精确率和召回率分别提高了 4.2% 和 4.0%，说明 ATFL 通过动态调整训练过程中的权重，有效缓解了类别不平衡问题。

第六行表示在 YOLOv8n 基础上融合 GCFM+EMIM 模块，相比于 YOLOv8n，mAP50 和 mAP50-95 分别提高了 3.5% 和 2.6%，精确率和召回率也分别提高了 3.3% 和 4.8%。虽然比起单独增加 EMIM 模块，mAP50 和 mAP50-95 略有降低，但是精确率大幅提高，体现出轻量化卷积与注意力机制在压缩模型体积的同时，仍可协同提升检测性能。

第七行表示在 YOLOv8n+GCFM+EMIM 基础上进一步引入 ASFFH 检测头，形成完整模型结构。相比于 YOLOv8n，虽然计算量增强 0.6 G，参数量增加 0.888 M，但 mAP50 提高了 4.4%，mAP50-95 提高了 3.2%，精确率提高了 6.9%，显示出多尺度融合策略中检测头对模型精度提升具有关键作用。

第八行表示在 YOLOv8n+GCFM+EMIM 基础上进一步引入了 ATFL 损失函数，比最终模型少了 ASFFH 检测头。虽然与 YOLOv8n+GCFM+EMIM 模型相比，计算量和参数量未额外增加，但是性能提升也十分有限，仅精确率提升较高，达到 2.0%。可以看出，在没有 ASFFH 检测头的情况下，精度提升有限，验证了其重

表 1 消融实验结果

Table 1 Ablation experiment results

Model	*P*/%	*R*/%	mAP50/%	mAP50-95/%	FLOPs/G	Params/M
YOLOv8n	84.1	75.0	80.6	53.0	8.1	3.007
+GCFM	80.4	77.8	82.4	54.9	**6.5**	**2.522**
+EMIM	83.9	81.4	84.8	56.4	8.1	3.007
+ASFFH	87.6	77.6	83.2	55.5	10.3	4.380
+ATFL	88.3	79.4	83.6	56.0	8.1	3.007
+GCFM+EMIM	87.4	79.8	84.1	55.6	**6.5**	**2.522**
+GCFM+EMIM+ASFFH	**91.0**	75.5	85.0	56.2	8.7	3.895
+GCFM+EMIM+ATFL	89.4	79.1	84.2	55.7	**6.5**	**2.522**
+GCFM+EMIM+ASFFH+ATFL	89.7	**83.3**	**86.7**	**57.3**	8.7	3.895

要性。

最终模型在融合四个模块后，精确率、召回率、mAP50 和 mAP50-95 分别达到 89.7%、83.3%、86.7% 和 57.3%，比起 YOLOv8n 模型分别提高了 5.6%、8.3%、6.1% 和 4.3%。验证了各模块在功能上的互补性与协同性，整体架构在保持轻量化的同时实现了检测性能的全面提升。其中加粗数据为各列最优结果。

3.5 对比实验

为了验证本文提出的模型的优越性，本文在 HIT-UAV 数据集上与 RT-DETR-l[26]、YOLO-SRMX[27] 和 YOLO 系列的 YOLOv8n、YOLOv8s、YOLOv10n、YOLOv10s、YOLOv11n、YOLOv12n 模型进行了对比实验，实验的结果如表 2 所示。

由表 2 可见，本文提出的方法相较于当下的 RT-DETR-l、YOLO-SRMX，mAP50 分别提升了 6.6% 和 3.9%，精度都有明显提升，在模型轻量化方面，与 YOLOv8n、YOLOv10n、YOLOv11n、YOLOv12n 等主流 YOLO 轻量模型相比，虽然模型参数量和 FLOPs 均有略微扩大，但是做到在提升了高精度的同时保持模型体量和参数大小。以上实验结果说明本文提出的算法在有优越的检测性能的同时仍保持模型轻量级。

3.6 损失函数对比实验

为进一步验证本文提出的 ATFL 在红外目标检测任务中的有效性，本文设计了损失函数对比实验。我们选取了二元交叉熵损失 (binary cross entropy, BCE)、Focal Loss[28]、Varifocal Loss[29] 以及不同超参数 λ 配置下的 ATFL 进行对比，旨在分析不同损失函数在处理类别不平衡与难易样本分布问题上的性能差异，实验结果如表 3 所示。

其中，BCE 是 YOLOv8 中默认使用的分类损失函数，未对类别的不平衡进行处理；Focal Loss 通过固定调制因子 γ 降低易分类样本的权重，增强对难分类样本的关注；Varifocal Loss 在 Focal Loss 基础上引入非对称加权策略，进一步提升对正样本的敏感度。

由表 3 可见，Focal Loss 与 Varifocal Loss 相较 BCE 在 mAP50 与召回率等指标上均有提升，表明二者对类

表 2 对比实验结果

Table 2 Comparative experiment results

Model	*P*/%	*R*/%	mAP50 /%	mAP50-95/%	FLOPs/G	Params/M
RT-DETR-l	82.2	71.3	75.6	47.1	102.9	31.272
YOLO-SRMX	82.0	78.0	82.8	50.0	10.0	11.400
YOLOv8n	84.1	75.0	80.6	53.0	8.1	3.007
YOLOv8s	85.9	78.0	83.5	56.5	28.4	11.127
YOLOv10n	89.6	76.0	83.0	55.8	8.2	2.696
YOLOv10s	86.1	76.6	83.3	54.8	24.5	8.039
YOLOv11n	87.4	74.8	82.5	54.3	6.3	2.583
YOLOv12n	84.1	76.6	81.7	54.1	**6.0**	**2.520**
Ours	**89.7**	**83.3**	**86.7**	**57.3**	8.7	3.895

表 3 损失函数对比实验结果

Table 3 Comparative experiment results of loss functions

Loss	λ	*P*/%	*R*/%	mAP50 /%	mAP50-95/%
BCE	-	84.1	75.0	80.6	53.0
Focal loss	-	85.5	79.3	82.3	53.8
Varifocal loss	-	83.9	76.2	82.4	54.4
ATFL	1.2	85.7	74.8	80.9	54.0
ATFL	1.5	**88.3**	**79.4**	**83.6**	**56.0**
ATFL	1.7	87.3	76.5	83.4	54.8
ATFL	2.0	85.0	76.9	83.3	54.4

别不平衡问题具有一定缓解作用，但性能提升有限，且无法动态适应训练过程中的样本分布变化。

对于 ATFL，当 λ=1.2 时模型对难样本关注不足，λ=1.7 和 λ=2.0 时则可能过度抑制易样本，导致训练不稳定。在 λ=1.5 时则在各项指标上均达到最优，证明本文所选 λ 值的合理性，也体现出 ATFL 在动态平衡难易样本上的优势与其强大的适应能力。

3.7 可视化实验

本文选择 YOLOv8n 与 YOLOv12n 作为可视化实验的对照组。实验结果如图 8 所示，第 1 行是原图，第 2 行是 YOLOv8n 模型预测结果，第 3 行是 YOLOv12n 模型预测结果，第 4 行是本文模型预测结果。从第 1、3、4 张图片可以看出，YOLOv8n 模型对于小目标，复杂场景和检测目标轮廓残缺的情况下，存在一定的漏检现象，YOLOv12n 也存在漏检和误检，尤其在第 3 张图片中出现明显误检。从第 2 和 5 张图片可看出，对于训练样本较少的类别，YOLOv8n 模型和 YOLOv12n 模型容易误检，精度欠佳。而本文方法减少了漏检和误检，同时在背景复杂的场景下，能更加准确地检测到目标，定位也更为精准，就算检测目标轮廓残缺，仍能检测出来。

通过上面的对比实验，验证了本文所提出算法的可行性，可应用于多种环境中的检测任务，实现跨场景部署。

3.8 泛化性实验

为进一步验证本文提出的 YOLO-MITD 算法在不同场景下的泛化能力与稳定性，本文在 CTIR 数据集[30] 与 SIRST-UAVB 数据集[31] 上进行了泛化性实验。其中，CTIR 数据集主要针对复杂热红外场景，包含多种复杂背景以及多尺度目标；而 SIRST-UAVB 数据集则专注于单帧红外图像中的小目标检测任务，目标尺度极小、

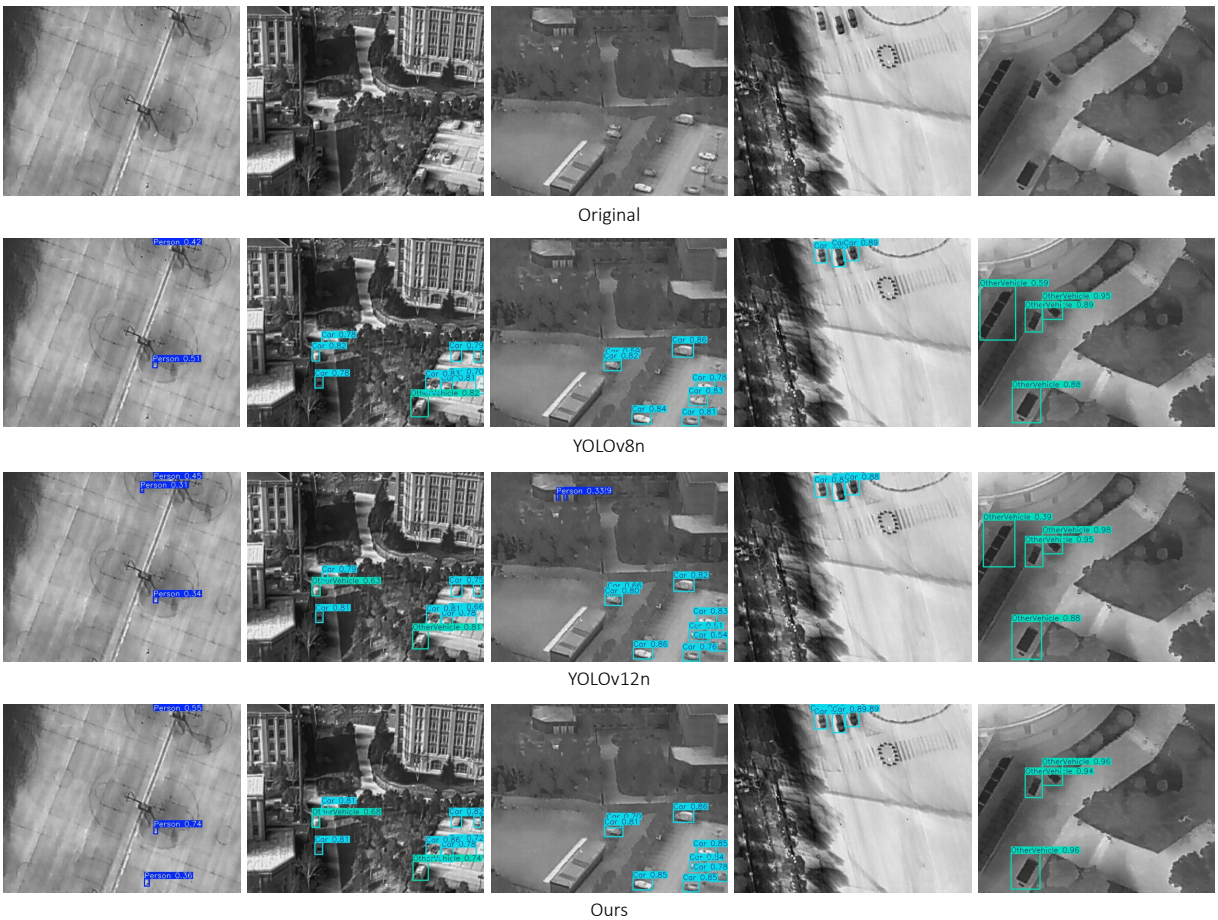


图 8 可视化对比

Fig. 8 Visualization comparison

信噪比低。这两个数据集在目标尺度、背景干扰强度和图像质量等方面具有显著差异，能够全面且深入地考察算法在多样化实际环境中的适应性、鲁棒性以及泛化性能。

实验结果如表 4 与表 5 所示。在 CTIR 数据集上，YOLO-MITD 的 mAP50 达到 85.4%，较基线模型提升 2.5%，mAP50-95 也提升了 3.0%，召回率与精确率也均有提升；在 SIRST 数据集上，YOLO-MITD 的召回率提升 6.7%，mAP50 和 mAP50-95 分别提升 6.5% 和 2.5%，表明模型对小目标具有更强的感知能力与更低的漏检率。

表 4 CTIR 数据集实验结果/%
Table 4 CTIR dataset experimental results/%

Model	*P*	*R*	mAP50	mAP50-95
YOLOv8n	77.9	77.0	82.9	57.0
Ours	**81.9**	**77.2**	**85.4**	**60.0**

表 5 SIRST 数据集实验结果/%
Table 5 SIRSTdataset experimental results/%

Model	*P*	*R*	mAP50	mAP50-95
YOLOv8n	85.0	48.8	60.7	31.7
Ours	**85.1**	**55.5**	**67.2**	**34.2**

实验结果表明，YOLO-MITD 在保持较低计算开销的前提下，在不同类型的红外数据集上均表现出良好的检测一致性与鲁棒性，具备较强的跨场景泛化能力。

4 结 论

针对无人机红外图像中小目标检测存在的计算资源受限、背景复杂以及目标尺度多样化等挑战，本文提出了一种基于 YOLOv8n 的改进算法，并设计了 GCFM 模块、EMIM 模块、ASFFH 模块与 ATFL 损失函数，在保持模型较低复杂度的同时显著提升检测精度。

所提算法在 HIT-UAV 数据集上，mAP50 较 YOLOv8n 提升 6.1%，召回率与精确率分别达到 83.3% 和 89.7%，同时模型计算量仅增加 0.6 G，参数量控制在 3.895 M，在精度与模型复杂度之间达到了较好的平衡。与各主流模型相比，本文方法在保持较低计算开销的前提下，检测性能显著优于对比模型，尤其在多尺度目标与复杂背景环境下表现突出，验证了其在实际无人机平台部署的可行性。

本文模型在极端低照度或目标高度遮挡场景下仍存在一定漏检风险，且当前模型尚未进行硬件部署与加速优化。下一步工作将推进算法在嵌入式平台上的部署实践，进一步优化模型结构与推理流程，提升实时性能与能效比，以适应更严苛的工程应用需求。

作者贡献

梁学才，李锦龙，周羿阳，杨赟，石博：研究资料的收集、分析，实验的设计与完成，论文写作；杜海娟，杨俊，许聪源，汪佳旭：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. Xu Q, Han N, Dong W A, et al. Infrared Target detection techniques for unmanned aerial vehicles: state-of-the-art and prospects[J]. *Radio Eng*, 2025, **55**(8): 1764−1774.
徐倩, 韩楠, 董文安, 等. 无人机红外目标检测技术: 现状与展望[J]. 无线电工程, 2025, **55**(8): 1764−1774.
2. Sun W N, Liu Y H, Wang F, et al. A study on flame detection method combining visible light and thermal infrared multimodal images[J]. *Fire Technol*, 2025, **61**(4): 2167−2188.
3. Qiao Q Y, Cheng H X. Infrared UAV target detection algorithm based on improved YOLOv8[J]. *Laser Infrared*, 2024, **54**(12): 1941−1947.
乔庆元, 程换新. 基于改进 YOLOv8 的红外无人机目标检测算法[J]. 激光与红外, 2024, **54**(12): 1941−1947.
4. Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016: 779–788.
https://doi.org/10.1109/CVPR.2016.91.
5. Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. arXiv: 1804.02767, 2018.
https://doi.org/10.48550/arXiv.1804.02767.
6. Jocher G. Ultralytics YOLOv5: version 7.0 [EB/OL]. (2022-11-22)[2026-01-30]
https://github.com/ultralytics/YOLOv5.
7. Yaseen M. What is YOLOv8: an in-depth exploration of the internal features of the next-generation object detector[Z]. arXiv: 2408.15857, 2024. https://doi.org/10.48550/arXiv.2408.15857.
8. Ding Z H, Zhang Z B, Yuan M, et al. CEDP-YOLO: UAV object detection based on context enhancement and dynamic perception[C]//*7th Chinese Conference on Pattern Recognition and Computer Vision*, Urumqi, China, 2025: 350–363.
https://doi.org/10.1007/978-981-97-8502-5_25.
9. Cao L, Wang Q, Luo Y H, et al. YOLO-TSL: a lightweight target detection algorithm for UAV infrared images based on Triplet attention and Slim-neck[J]. *Infrared Phys Technol*, 2024, **141**:

105487.
10. Wang Q, Liu F Y, Cao Y, et al. LFIR-YOLO: lightweight model for infrared vehicle and pedestrian detection[J]. *Sensors*, 2024, **24**(20): 6609.
11. Luo X, Zhu H, Zhang Z L. IR-YOLO: real-time infrared vehicle and pedestrian detection[J]. *Comput Mater Continua*, 2024, **78**(2): 2667−2687.
12. Liu Y Y, Wei Y. Infrared small target detection algorithm for UAV perspective based onYOLOv8n[J]. *Comput Technol Dev*, 2026, **36**(2): 71−77
刘奕阳, 魏延. 基于 YOLOv8n 的无人机视角红外小目标检测算法[J]. 计算机技术与发展, 2026, **36**(2): 71−77
13. Li C, Li J X. Orthogonal features extraction method and its application in convolution neural network[J]. *J Shanghai Jiaotong Univ*, 2021, **55**(10): 1320−1329.
李辰, 李建勋. 卷积神经网络的正交性特征提取方法及其应用[J]. 上海交通大学学报, 2021, **55**(10): 1320−1329.
14. Han K, Wang Y H, Tian Q, et al. GhostNet: more features from cheap operations[C]//*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020: 1577–1586. https://doi.org/10.1109/CVPR42600.2020.00165.
15. Yang H Y, Wang J, Bo Y M, et al. ISTD-DETR: a deep learning algorithm based on DETR and Super-resolution for infrared small target detection[J]. *Neurocomputing*, 2025, **621**: 129289.
16. Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. *IEEE Trans Pattern Anal Mach Intell*, 2020, **42**(8): 2011−2023.
17. Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]//*15th European Conference on Computer Vision*, Munich, Germany, 2018: 3–19.
https://doi.org/10.1007/978-3-030-01234-2_1.
18. Misra D, Nalamada T, Arasanipalai A U, et al. Rotate to attend: convolutional triplet attention module[C]//*2021 IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, HI, USA, 2021: 3138–3147. https://doi.org/10.1109/WACV48630.2021.00318.
19. Ouyang D L, He S, Zhan J Z, et al. Efficient multi-scale attention module with cross-spatial learning[C]//*2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, 2023: 1–5.
https://doi.org/10.1109/ICASSP49357.2023.10096516.
20. Liu S T, Huang D, Wang Y H. Learning spatial fusion for single-shot object detection[EB/OL]. arXiv: 1911.09516, 2019.
https://doi.org/10.48550/arXiv.1911.09516.
21. Zhang Z B, Lv G H, Zhao G X, et al. BS-YOLOV5S: insulator defect detection with attention mechanism and multi-scale fusion[C]//*2023 IEEE International Conference on Image Processing*, Kuala Lumpur, Malaysia, 2023: 2365–2369.
https://doi.org/10.1109/ICIP49359.2023.10222163.
22. Talamantes-Roman A, Ramirez-Alonso G, Gaxiola F, et al. A TransUNet model with an adaptive fuzzy focal loss for medical image segmentation[J]. *Soft Comput*, 2024, **28**(20): 12359−12375.
23. Suo J S, Wang T Y, Zhang X Z, et al. HIT-UAV: a high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection[J]. *Sci Data*, 2023, **10**(1): 227.
24. Ding K Y, Ding Z H, Zhang Z B, et al. SCD-YOLO: a novel object detection method for efficient road crack detection[J]. *Multimedia Syst*, 2024, **30**(6): 351.
25. DOI: 10. 12086/oee. 2026.250292。red target detection network[J]. *Opto-Electron Eng*, 2026, **53**(4): 250292.
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.
26. Zhao Y A, Lv W Y, Xu S L, et al. DETRs beat YOLOs on real-time object detection[C]//*2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2024: 16965–16974. https://doi.org/10.1109/CVPR52733.2024.01605.
27. Weng S M, Wang H, Wang J S, et al. YOLO-SRMX: a lightweight model for real-time object detection on unmanned aerial vehicles[J]. *Remote Sens*, 2025, **17**(13): 2313.
28. Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[J]. *IEEE Trans Pattern Anal Mach Intell*, 2020, **42**(2): 318−327.
29. Zhang H Y, Wang Y, Dayoub F, et al. VarifocalNet: an IoU-aware dense object detector[C]//*2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2020: 8510–8519. https://doi.org/10.1109/CVPR46437.2021.00841.
30. 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*, Seattle, WA, USA, 2024: 6489–6498.
https://doi.org/10.1109/CVPRW63382.2024.00649.
31. Yang J N, Liu S L, Wu J J, et al. Pinwheel-shaped convolution and scasle-based dynamic loss for infrared small target detection[C]// *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, Philadelphia, PA, USA, 2025: 9202–9210.
https://doi.org/10.1609/aaai.v39i9.32996.

作者简介



梁学才 (2005-)，男，本科生，研究方向为机器视觉、深度学习。
E-mail：2802723940@qq.com



李锦龙 (2005-)，男，本科生，研究方向为机器学习、深度学习理论及应用。
E-mail：2721154740@qq.com



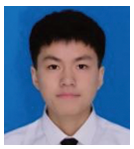
【通信作者】杜海娟 (1974-)，女，工学硕士，嘉兴大学人工智能学院讲师，研究方向为计算机视觉、多模态数据融合。
E-mail：444567707@qq.com



周羿阳 (2005-)，男，本科生，研究方向为计算机视觉、深度学习应用。
E-mail：aoaoo2025@163.com



杨赟 (2006-)，女，本科生，研究方向为机器视觉、深度学习应用。
E-mail：2409335092@qq.com



石博 (2004-)，男，本科生，研究方向为机器视觉、深度学习应用。
E-mail：2010427055@qq.com



【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



许聪源 (1990-)，男，工学博士，硕士研究生导师，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



汪佳旭 (2001-)，男，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

YOLO-MITD: a multi-scale infrared target detection algorithm

Liang Xuecai[1], Du Haijuan[1*], Li Jinlong[1], Zhou Yiyang[1], Yang Yun[1], Shi Bo[1], Yang Jun[1*], Xu Congyuan[1], Wang Jiaxu[1,2]

[1] College of Artificial Intelligence, Jiaxing University, Jiaxing, Zhejiang 314001, China;
[2] School of Information Science and Engineering(School of Cyber Science and Technology), Zhejiang Sci-Tech University, Hangzhou, Zhejiang 310018, China

Abstract

Objective Unmanned aerial vehicle (UAV) infrared target detection faces significant challenges including limited computational resources, strong background interference, and multi-scale target variations. Infrared imaging, compared with visible light imaging, provides all-day and all-weather working capabilities with strong anti-interference performance. However, due to flight altitude and sensor limitations, UAV-acquired infrared images commonly suffer from low spatial resolution, poor signal-to-noise ratio, and large target scale spans. In long-distance shooting conditions, targets often exhibit multi-scale distributions, with small targets of several pixels and medium-to-large targets covering tens of pixels coexisting in the same scene. Traditional detection methods tend to produce high false positive rates and severe missed detection in complex backgrounds. To address these issues, this paper proposes a multi-scale infrared target detection algorithm based on YOLO, named YOLO-MITD. The algorithm aims to achieve high detection accuracy while maintaining low computational complexity, making it suitable for deployment on resource-constrained UAV platforms.

Methods The proposed YOLO-MITD algorithm introduces four core modules to enhance detection performance. First, the ghost-enhanced cross-stage fusion module (GCFM) is designed by replacing standard convolutions in the original C2f module with GhostConv lightweight convolutions. GhostConv adopts a two-stage feature generation mechanism that uses standard convolution to extract essential features and then applies cheap transformations through depthwise separable convolution to generate phantom features, which significantly reduces model parameters while enhancing multi-scale feature extraction capabilities. The theoretical acceleration ratio approximately equals the redundancy factor s. Second, the efficient multi-scale interaction module (EMIM) incorporates the efficient multi-scale attention (EMA) mechanism into each bottleneck of the C2f structure. EMA adopts a "feature grouping plus cross-spatial learning" strategy to achieve one-time joint recalibration of channel and spatial information without dimensionality reduction in the channel dimension. The feature grouping number G is dynamically adjusted based on the channel count C, ensuring efficient feature representation while controlling computational complexity. Third, the adaptive spatial feature fusion head (ASFFH) is developed based on the adaptive spatial feature fusion (ASFF) mechanism, which consists of three steps: feature rescaling through interpolation or strided convolution to align

*Correspondence: Du H J, E-mail: 444567707@qq.com; Yang J, E-mail: juneryoung@zjxu.edu.cn

spatial resolutions, adaptive fusion through softmax-normalized spatial weights learned by the network, and consistency guarantee to automatically suppress gradients from conflicting regions. This module dynamically allocates fusion weights for multi-scale features and suppresses feature conflicts, effectively improving scale consistency in target detection. Fourth, the adaptive threshold focal loss (ATFL) function is constructed with a piecewise weighting mechanism that dynamically adjusts the loss curvature. For high-confidence samples ($p_t > 0.5$), exponential decay mechanism compresses loss weights to low levels; for low-confidence samples ($p_t \leqslant 0.5$), explosive growth mechanism amplifies loss weights to accelerate gradient backpropagation. The threshold parameter lambda is adaptively adjusted based on the exponential moving average of prediction confidence across training batches, thereby alleviating class imbalance issues during training.

Results and Discussions Extensive experiments were conducted on the HIT-UAV dataset, which is the first publicly available UAV infrared thermal imaging target detection dataset for high-altitude scenes. The dataset contains 2898 UAV infrared aerial images covering multiple real-world scenarios including schools, parking lots, roads, and playgrounds, under both daytime and nighttime conditions. Notably, the target size distribution exhibits significant diversity, with small targets accounting for approximately 70%, medium targets about 28%, and large targets only 2% of the dataset. The proposed YOLO-MITD algorithm achieved an mAP50 of 86.7%, representing a 6.1% improvement over the baseline YOLOv8n. The precision and recall reached 89.7% and 83.3%, respectively, with increases of 5.6% and 8.3% compared to YOLOv8n. Despite these significant performance gains, the computational overhead increased by only 0.6 GFLOPs and the parameter count was controlled at 3.895 M, achieving an excellent balance between accuracy and complexity. Ablation experiments demonstrated the effectiveness of each module: GCFM reduced parameters by 0.485 M and FLOPs by 6.5 G while improving mAP50 by 1.8 percentage points; EMIM improved mAP50 by 4.2 percentage points with the most significant accuracy enhancement; ASFFH enhanced mAP50 by 2.6% but increased FLOPs by 2.2 G; ATFL improved mAP50 by 3.0% without increasing model complexity. The complete model combining all four modules achieved optimal performance, verifying the complementary and synergistic nature of the modules. Comparative experiments against mainstream models including RT-DETR-l, YOLO-SRMX, YOLOv8n, YOLOv8s, YOLOv10n, YOLOv10s, YOLOv11n, and YOLOv12n showed that YOLO-MITD achieves superior detection performance while maintaining lightweight characteristics. Specifically, compared with RT-DETR-l and YOLO-SRMX, mAP50 improved by 6.6% and 3.9%, respectively. Cross-scene generalization experiments on the CTIR and SIRST datasets further validated the algorithm's robustness, with mAP50 reaching 85.4% and 67.2%, respectively, significantly outperforming existing methods. Visualization results demonstrated that the proposed method reduces missed and false detections, particularly in complex backgrounds and for targets with incomplete contours.

Conclusions This paper presents an improved YOLOv8n-based algorithm for multi-scale infrared target detection in UAV applications. The GCFM module achieves lightweight feature extraction through GhostConv by replacing standard convolutions with cheap operations, maintaining feature expressiveness while reducing computational burden. The EMIM module enhances feature representation via the EMA attention mechanism, which captures both channel and spatial dependencies through cross-dimensional interaction without dimensionality reduction. The ASFFH module improves multi-scale fusion through adaptive weight allocation, effectively resolving feature pyramid inconsistencies and suppressing gradient conflicts. The ATFL loss function optimizes the training process through dynamic threshold adjustment, applying different modulation strategies to samples based on confidence levels and alleviating class imbalance issues. The experimental results demonstrate that the proposed algorithm strikes an excellent balance between detection accuracy and model complexity, with strong cross-scene generalization capability. The method achieves state-of-the-art performance on multiple infrared datasets while maintaining low computational requirements, making it well-suited for real-time infrared target detection tasks on resource-constrained UAV platforms. Future work will focus on hardware deployment optimization, model quantization and pruning techniques, and further improvements to real-time performance and energy efficiency to adapt to more demanding engineering application requirements.

Keywords: infrared target detection; lightweight convolution; attention mechanism; multi-scale fusion; adaptive loss

Liang X C, Du H J, Li J L, et al. YOLO-MITD: a multi-scale infrared target detection algorithm[J]. *Opto-Electron Eng*, 2026, **53**(5): 250322; DOI: 10.12086/oee.2026.250322

Opto-Electronic Journals Group
www.oejournal.org

扫描二维码，获取 PDF 全文