DOI:10.19651/j.cnki.emt.2520308

IF-DETR:改进 RT-DETR 的航拍图像检测算法*

杨 洋 魏为民 郑书情 马凯楠 张劲阳
(上海电力大学人工智能学部 上海 201306)

摘 要:针对无人机航拍图像中存在的复杂光照干扰,小目标分辨率低、容易漏检误检以及目标结构模糊等问题,提出了一种基于改进 RT-DETR 的无人机航拍图像目标检测方法 IF-DETR。首先,在骨干网络中设计了跨阶段光照门控单元,通过在特征层面建模光照与反射分量,增强了模型在复杂光照环境下对目标的提取能力。其次,在颈部网络构建了小目标多尺度特征融合结构,有效提升了对小目标的特征提取与辨识能力。此外,提出了一种高斯特征增强模块,利用高斯先验与空间注意力机制增强与目标相关的结构信息,同时抑制背景噪声。最后,构建 Focaler-Powerful-IoU 损失函数,实现了更稳定、更精准的目标定位。实验结果表明,与 RT-DETR 模型相比,改进后的 IF-DETR 算法在 Visdrone2019 数据集上,平均精度均值(mAP)mAP_{50}、$mAP_{50:95}$、召回率和准确率分别提升了 3.0%、2.3%、3.9%、2.0%,能够有效改善无人机航拍图像检测的漏检、误检问题并且提高了检测性能,具有广泛的应用前景。

关键词:RT-DETR;无人机图像;目标检测;光照感知;结构信息增强
中图分类号:TN911.73;TP391.41 **文献标识码**:A **国家标准学科分类代码**:520.6040

IF-DETR: Improved RT-DETR-based aerial image detection algorithm

Yang Yang Wei Weimin Zheng Shuqing Ma Kainan Zhang Jinyang
(Aritifical Intelligence Department, Shanghai University of Electricity Power, Shanghai 201306, China)

Abstract: Aiming at the problems in UAV aerial images, such as complex illumination interference, low resolution of small targets with high susceptibility to missing detection and false detection and blurred target structures, a target detection method for UAV aerial images based on improved RT-DETR is proposed, denoted as IF-DETR. Firstly, a cross-stage illumination gating unit is designed in the backbone network, which enhances the model's ability to extract targets under complex illumination environments by modeling illumination and reflection components at the feature level. Secondly, a multi-scale feature fusion structure for small targets is constructed in the neck network, effectively improving the feature extraction and recognition capabilities for small targets. In addition, a Gaussian feature enhancement module is proposed, which utilizes Gaussian prior and spatial attention mechanism to enhance target-related structural information while suppressing background noise. Finally, a Focaler-Powerful-IoU loss function is constructed to achieve more stable and accurate target localization. Experimental results show that compared with the RT-DETR model, the improved IF-DETR algorithm on the Visdrone2019 dataset achieves improvements of 3.0%, 2.3%, 3.9% and 2.0% in mAP_{50}, $mAP_{50:95}$, recall and precision respectively. It can effectively alleviate the problems of missing detection and false detection in UAV aerial image detection, improve detection performance, and has broad application prospects.

Keywords: RT-DETR;UAV imagery;object detection;illumination aware;structural information enhancement

0 引 言

在智能化、自动化不断深入的时代背景下,无人机技术在智能交通管控、生态环境保护与建设、灾害应对与治理和军事指挥作战等领域得到了广泛应用[1]。然而,无人机航拍图像的目标检测仍然面临着许多挑战。首先,图像质量受限。无人机航拍过程中易受到复杂光照干扰,目标本就微弱的特征会进一步弱化甚至丢失,影响检测性能。其次,背景干扰与目标遮挡并存。无人机飞行的多数场景中存在树木、建筑物和电线等物体对目标物体造成部分或完全遮

收稿日期:2025-11-29
*基金项目:国家自然科学基金(62576203)项目资助

挡,致使目标信息缺失。再者,尺度和视角变化。由于无人机的旋转会导致视角偏移,导致目标结构模糊、定位不够精准,容易造成漏检和误检[2]。针对上述问题,该领域学者进行了新颖广泛的研究。郭润泽等[3]提出了一种耦合光照条件和对比度的多尺度差分注意力融合检测方法,通过可见光与红外多模态特征的深度融合,显著提升了无人机弱光环境下的目标检测性能。孙晓永等[4]提出了一种面向多无人机协同的多模态目标检测方法,采用可见光与红外图像融合及主动感知策略,克服单无人机视场受限、目标易遮挡等问题,从而提高了无人机目标检测的范围和精度。

近年来,基于深度学习的目标检测方法已经广泛应用于无人机航拍任务,且主要分为两大类:一类是基于传统深度卷积神经网络(convolutional neutral network, CNN)的目标检测方法;另一类是基于 Transformer 架构的目标检测方法。基于 CNN 架构的目标检测模型可以分为两类:单阶段检测算法和双阶段检测算法。以 RegionCNN(RCNN)[5]为代表的两阶段算法将检测问题分为两个部分:感兴趣区域和选择部分。以 YOLO[6]为代表的单阶段检测算法则不需要生成感兴趣区域,能够快速获取目标的类别和边界框。相较于 CNN 架构,Transformer 架构在全局感知、细节捕获和序列建模方面具有较大优势,利用全局感知的机制在不同层次和尺度上对整张图像进行细致分析,有利于在复杂场景下准确定位和识别目标。DETR[7]是一种创新的目标检测方法,它将目标检测视为集合预测问题,利用 Transformer 架构简化了检测流程并且摒弃了传统方法中的锚框和非极大值抑制后处理步骤,但 DETR 也存在训练收敛慢和查询优化难的问题。针对这些问题,衍生出多种 DETR 变体:文献[8]通过引入浅层主干和无参数注意力机制有效降低计算量从而加快收敛;SDH-DETR[9]通过引入动态上采样和小波下采样加速收敛;Deformable-DETR[10]通过引入可变形注意力加速收敛的同时降低计算量;Conditional-DETR[11]通过引入条件空间查询解决查询与特征对齐的模糊性并加速收敛;Anchor-DETR[12]使用 2D 锚点坐标作为查询简化优化过程;DAB-DETR[13]使用 4D 锚框作为查询,并逐层更新优化预测框;Co-DETR[14]通过在训练中并行加入多个辅助头部极大地增强了对 Encoder 的密集监督;DINO[15]通过结合之前工作的思想并加以改进,取得了更好的效果。在此基础上,百度实现了一个在速度和准确性方面取得了显著进步的端到端目标检测算法 RT-DETR[16]。

目前,已有多项面对无人机航拍场景的最新研究选择基于 RT-DETR 模型进行改进。如 UAV-DETR[17]通过引入多尺度特征融合与频率增强、频率感知下采样和语义对齐与校准等模块,在特征提取、下采样和跨尺度融合过程中同时利用空间域与频率域信息,更好地保留高频细节与上下文语义,从而显著提升对小目标和遮挡目标的表征能力,但由于频域变换和多分支结构带来的额外计算,使计算量大幅增加、帧率下降严重;SO-DETR[18]通过空间域与频率域融合的混合编码器增强细节与高分辨率特征表达;利用扩展 IoU 驱动的动态查询选择优化查询初始化、加快收敛;结合知识蒸馏和轻量骨干减少参数与计算量、提升推理效率,但模型对大目标检测精度下降、训练流程复杂。

1　RT-DETR 模型简介

RT-DETR 是一种实时的端到端目标检测算法,解决了传统 DETR 推理速度慢的痛点。通过采用 Anchor-free 检测头与 Transformer 解码器直接输出检测结果。这样的设计省去了耗时的非极大值抑制后处理步骤,同时避免了多阶段训练的繁琐流程和人工组件,成为兼顾了速度与性能的主流检测框架之一,其网络结构如图 1 所示。

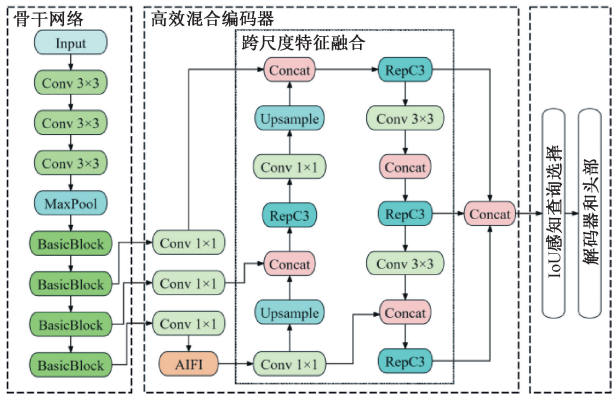


图 1　RT-DETR-R18 模型结构

Fig. 1　RT-DETR-R18 module structure

RT-DETR 整体主要由骨干网络、高效混合编码器和解码器 3 大模块组成。骨干网络通过卷积层、池化层与 BasicBlock 组合,对输入图像完成初步的特征提取。随后高效混合编码器先通过尺度内特征交互技术(attention-based intrascale feature interaction, AIFI)对深层次特征图进行处理,并将处理后的多尺度特征传入跨尺度特征融合(cross-scale feature fusion, CCFF)实现特征整合。最后,从编码器输出的特征序列中筛选出固定数量的特征作为初始查询,通过解码器迭代优化后,输出最终的检测框和类别置信度分数。

2　IF-DETR 模型

本文以 RT-DETR 为基线提出 IF-DETR,从光照鲁棒性、小目标表征、复杂背景辨识和定位精度 4 个方面进行改进:通过跨阶段光照门控单元(cross-stage illumination gating unit, CS-IGU)模拟光照与反射分量,增强对复杂光照变化的鲁棒性;构建小目标跨尺度特征融合结构(small object cross-scale feature fusion, SO-CCFF),有效保留并强化微小目标细节;设计高斯特征增强模块(Gaussian feature enhancement module, GFEM)分离并强化高频结

构信息，抑制背景噪声，提高复杂背景下的可辨识度；同时设计 Focaler-Powerful-IoU 损失函数，聚焦困难样本并引入几何约束，实现更快收敛和更精准的边界框定位。具体的网络结构如图 2 所示。

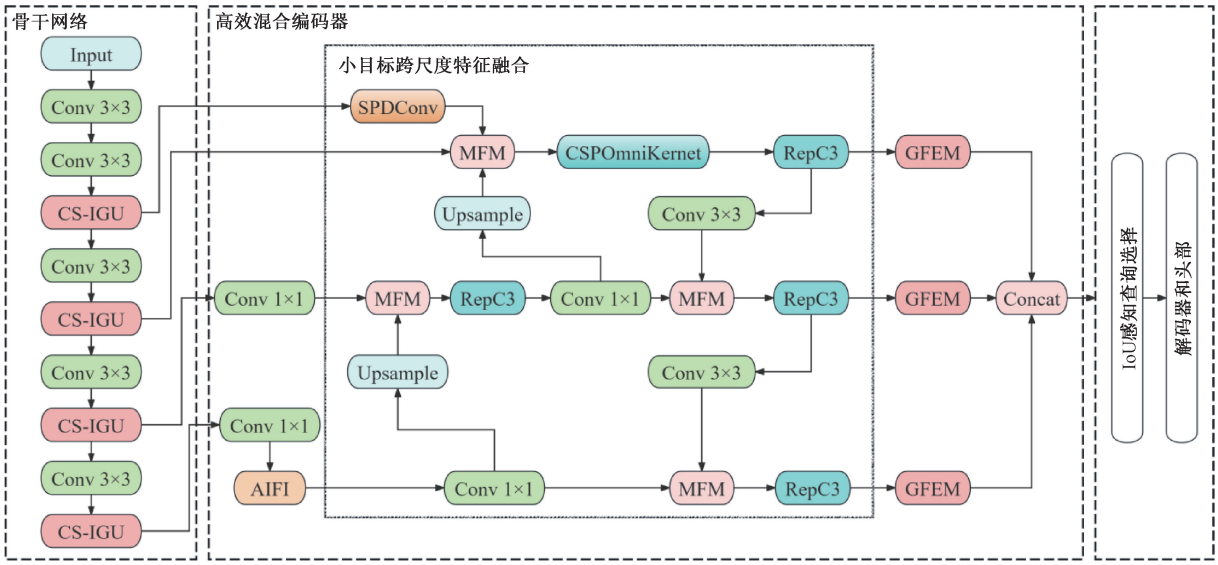
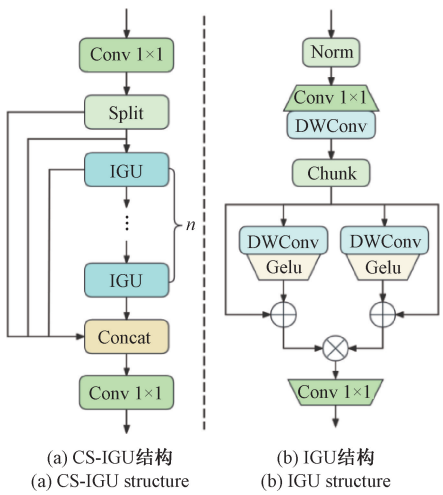


图 2　IF-DETR 模型结构

Fig. 2　IF-DETR module structure

2.1　跨阶段光照门控单元

在无人机航拍各种复杂场景中，目标检测任务面临着复杂多变的光照环境所带来的严峻挑战。RT-DETR 的主干网络在设计上缺乏针对光照变化的调节机制，导致在复杂光照条件下，图像中暗部或亮部区域的目标特征信息容易丢失，从而严重影响检测性能。为解决这一问题，本文对 RT-DETR 的主干网络进行了重构，提出了 CS-IGU。受跨阶段局部网络[19]设计思想的启发，CS-IGU 采用了高效的特征复用与融合策略，在提升网络对复杂光照适应性的同时，保持了结构的轻量化与高效性，其网络结构如图 3(a)所示。



(a) CS-IGU结构
(a) CS-IGU structure

(b) IGU结构
(b) IGU structure

图 3　网络结构

Fig. 3　Network structures

将上层网络传入的特征图表示为张量 $\boldsymbol{X}$。首先通过 1×1 卷积对输入特征 X 进行通道扩展，随后，处理后的特征图被分割为并行的主干分支 X_a 和残差分支 X_b。残差分支 X_b 被直接保留，用于后续的跨阶段特征融合。而主干分支 X_a 则被送入 n 个 CS-IGU 的核心模块光照门控单元(illumination gating unit，IGU)中。经过 n 次迭代后，将原始的残差分支 X_b 与主干分支最终输出以及所有中间过程的输出在通道维度上进行拼接。最后通过一个 1×1 卷积对拼接后的高维特征进行信息融合与通道调整，得到 CS-IGU 的最终输出。

IGU 的结构如图 3(b)所示，其设计来源经典 Retinex 理论，该理论将图像 I 建模为物体表面反射率(物体的固有属性)R 与环境光照 L(光照条件)的乘积，如式(1)所示。

$$I = L \times R \tag{1}$$

IGU 通过在特征空间中模拟这种乘法交互，来构建一个具有特定归纳偏置的乘法门控单元。对于输入 IGU 的特征张量 $\boldsymbol{X}$，其内部运算流程为输入特征图 X 首先进行归一化处理($\mathrm{Norm}_{\mathrm{BN}}$)稳定特征分布，随后通过一个 1×1 卷积进行通道扩展以及一个深度可分离卷积(DWConv)进行初步的空间特征提取，如式(2)所示。

$$X_{\mathrm{proj}} = \mathrm{DWConv}(\mathrm{Conv}_{1\times1}(\mathrm{Norm}_{\mathrm{BN}}(X))) \tag{2}$$

式中：X_{proj} 表示输出特征图。

接着，将 X_{proj} 在通道维度上分割为两个并行的子张量 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$，如式(3)所示。

$$\boldsymbol{X}_1, \boldsymbol{X}_2 = \mathrm{chunk}(X_{\mathrm{proj}}) \tag{3}$$

随后，这两个分支都通过相同的独立路径进行处理。路径包含一个 DWConv、一个 GELU 激活函数以及一个残

差连接,旨在学习各自空间内的复杂非线性变换,如式(4)、(5)所示。

$$X'_1 = \mathrm{GELU}(\mathrm{DWConv}(\boldsymbol{X}_1)) + \boldsymbol{X}_1 \tag{4}$$

$$X'_2 = \mathrm{GELU}(\mathrm{DWConv}(\boldsymbol{X}_2)) + \boldsymbol{X}_2 \tag{5}$$

式中:X'_1 和 X'_2 分别表示 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 分支的输出特征图。

最后采用逐元素相乘将两个处理后的特征图 X'_1 和 X'_2 进行融合,如式(6)所示。

$$Y_{\mathrm{gated}} = X'_1 \odot X'_2 \tag{6}$$

式中:Y_{gated} 表示经过逐元素相乘后的输出特征图。

这一步是 IGU 设计的关键,与传统的相加和拼接操作不同,乘法融合在结构上直接模拟了 Retinex 理论中 $I = L \times R$ 的物理模型。这种设计引入了强大的归纳偏置:为了让模型在反向传播中最小化损失,优化器将被激励去让这两个并行分支学习功能上互补的表征。X'_1 将学习更稳定的、代表物体固有属性的内容特征,而 X'_2 则学习一个动态的门控权重。因此,逐元素相乘中 X'_2 将根据感知到的局部光照信息对 X'_1 中的特征进行加权。

这种门控使得网络能够自适应地放大那些对光照变化不敏感的、更能代表物体本质的核心特征,同时抑制那些由阴影、高光引起的噪声特征。最后,通过一个 1×1 卷积对融合后的特征进行整合,并将其投影回原始的通道维度,得到 IGU 的最终输出。

2.2　小目标跨尺度特征融合

在 RT-DETR 模型中,颈部网络采用的 CCFF 结构能够融合来自骨干网络不同层级的特征图,这种跨尺度融合机制对目标检测任务至关重要。

然而,对于无人机航拍小目标检测任务,CCFF 存在以下问题:首先是 CCFF 中简单的特征拼接(Concat)操作虽然保留了所有信息,但它平等地对待了所有尺度的特征。在小目标检测任务中,这会导致一个严重问题:来自深层、低分辨率的强语义特征,在数量上会压倒来自浅层、高分辨率的微弱、精细的纹理细节,导致小目标检测性能不佳。其次是 CCFF 忽略了分辨率最高、纹理和边缘信息最精细的 P2 层,导致小目标信息丢失。

因此,本文在原模型 CCFF 的基础上做出一系列改进,提出 SO-CCFF,其结构如图 4 所示。

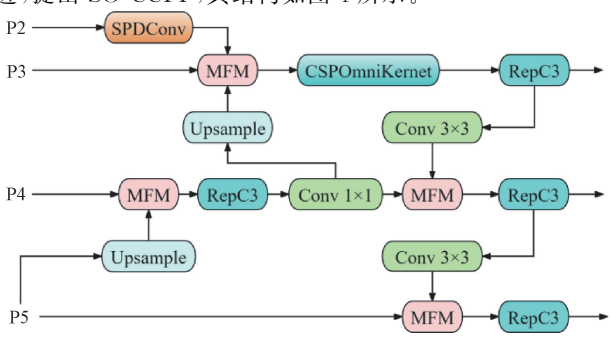


图 4　SO-CCFF 结构

Fig. 4　SO-CCFF structure

首先,为了解决 Concat 操作的问题,本文引入了调制特征融合模块(modulation fusion module, MFM)[20],其结构如图 5 所示。

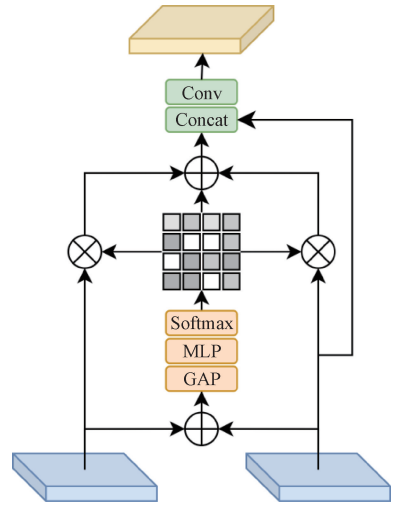


图 5　MFM 结构

Fig. 5　MFM structure

MFM 首先将输入的两个特征图 F^1_{in}、F^2_{in} 相加,再依次经过 GAP、MLP 和 Softmax 操作后得到系数矩阵 $\boldsymbol{A}_{\mathrm{mid}}$。$\boldsymbol{A}_{\mathrm{mid}}$ 中的元素值反映了 F^1_{in} 和 F^2_{in} 中的各特征的重要程度,通过 $\boldsymbol{A}_{\mathrm{mid}}$ 的调节可以突出各自重要特征,如式(7)所示。

$$G = \boldsymbol{A}_{\mathrm{mid}} \odot F^1_{\mathrm{in}} + \boldsymbol{A}_{\mathrm{mid}} \odot F^2_{\mathrm{in}} \tag{7}$$

其中,G 表示输出输出特征图。

随后,通过一个残差连接将 G 和 F^1_{in} 拼接起来增强梯度流并保留原始细节。最后,拼接结果通过卷积层处理,得到最终输出。

MFM 不再是平等地拼接所有特征图,而是学习了两个输入特征图之间的互补关系,并以此对二者进行自适应加权。这种机制能够放大关键的微小目标细节,同时抑制来此其他层级的冗余信息,从而在特征融合过程中保留了小目标的微弱特征。

随后,本文设计了一个小目标增强金字塔结构,将 P2 特征图输入到尺度置换深度卷积(scale-permuted depthwise convolution, SPDConv)[21] 中进行处理。SPDConv 处理低分辨率图像和小目标对象时非常有效,可以增强对小目标重要细节的提取能力。随后这些提取出的信息在与 P3 充分融合之后输入到 CSP-Omnikernel 模块中进行特征整合,其网结构如图 6 所示。

Omni-Kerel 由 3 个分支组成,即全局分支、大分支和局部分支,从而更有效地学习整体到局部的特征,提高小目标检测性能。其中,全局分支包括双域通道注意力模块和频率空间注意力模块,负责对全局域进行感知;大分支采用了不同尺寸的大核深度卷积获取上下文,更好的捕捉大尺度信息。局部分支则使用简单的 1×1 深度卷积,旨在捕捉局部信息,补充处理局部细节。

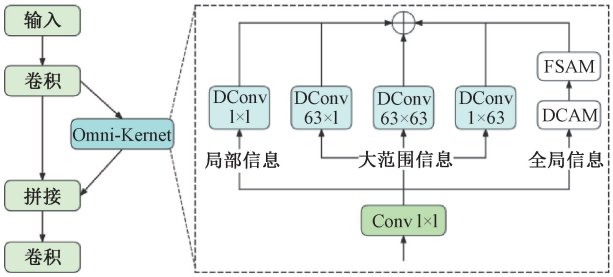


图 6　CSP-Omnikernel 结构

Fig. 6　CSP-Omnikernel structure

通过引入 MFM 取代了传统的特征拼接(Concat)，实现了对不同尺度特征的自适应加权，避免了浅层高分辨率信息的稀释；同时，通过 SPD-Conv 提取 P2 高分辨率特征层得到富含小目标信息的特征用于后续融合，并利用 CSPOmni-kernel 高效整合了从全局到局部的多尺度上下文信息。本文设计的 SO-CCFF 结构使得模型能够更准确地定位和识别小目标，提高小目标的检测能力。

2.3　高斯特征增强模块

在无人机视角下，物体边缘、轮廓和纹理等高频结构信息是区分微小物体与复杂背景的关键。这些信息随着分辨率的降低而模糊，但目标特征往往呈现出以特定点为中心的高斯分布。为了更好地捕捉这些目标特征，设计了 GFEM，其结构如图 7 所示。

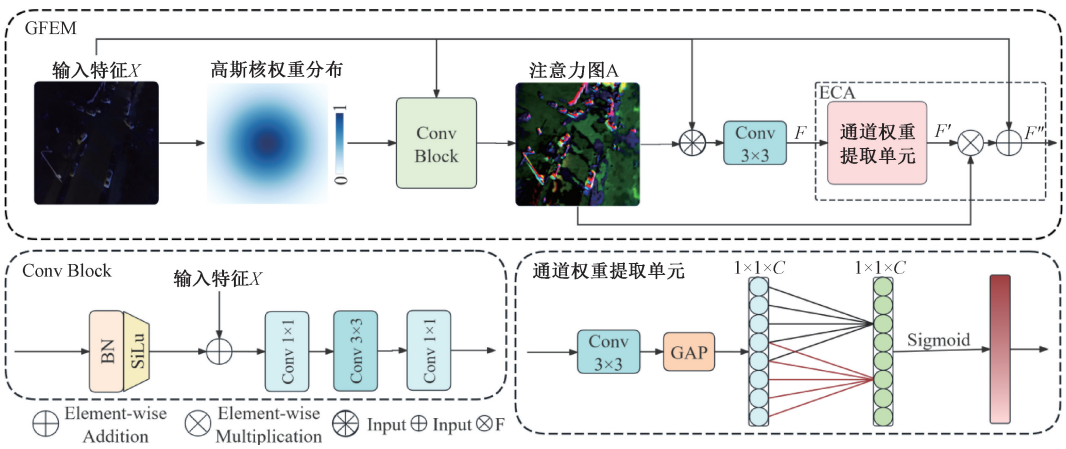


图 7　GFEM 结构

Fig. 7　GFEM structure

GFEM 首先利用一个固定的二维高斯核对输入特征图 X 进行深度可分离卷积，其权重被初始化为离散高斯核并保持固定。二维高斯核函数定义如式(8)所示。

$$G(x,y)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{8}$$

式中：(x,y)是相对于核中心的空间坐标，σ 是标准差，设置为 1.0。

高斯核作为一个经典的低通滤波器，其输出与原始特征图结合后，能够高效地提取出作为残差的高频成分。它有助于平滑特征变化并抑制噪声，从而保留关键的目标结构，防止其在网络前向传播过程中退化。然而，并非所有高频信号都有用，例如背景中的道路和楼房等。因此，提取出的高频信号图会进一步被送入一个卷积层 ConvBlock 中。该卷积层在训练过程中学习区分哪些结构与真实目标相关，哪些是无关噪声。它会输出一张经过学习和优化的空间注意力图 A，图中每个像素的值代表了该空间位置对于检测任务的重要性。获得注意力图 A 后，通过注意力图 A 与特征图 X 之间的逐元素乘法，对原始输入特征图 X 进行自适应的逐元素重加权。随后再与原始输入特征图 X 相加再经过一个 3×3 卷积得到输出 F，如式(9)所示。

$$F=\mathrm{Conv}_{3\times3}(X+(X\otimes A)) \tag{9}$$

最后为了突出关键通道，采用了一种高效通道注意力(ECA)策略，ECA 表示如式(10)、(11)所示。

$$F'=\mathrm{Sigmoid}(\mathrm{Conv}^{k}_{1\mathrm{D}}(\mathrm{GAP}(F))) \tag{10}$$

$$F''=\mathrm{Norm}((F'\otimes A)+X) \tag{11}$$

式中：GAP 表示全局平均池化，$\mathrm{Conv}^{k}_{1\mathrm{D}}$ 表示卷积核大小为 k 的一维卷积。

通过精确的特征重加权和突出关键通道，模块极大地提升了特征图的信噪比，使得潜在的目标在数值上与复杂的背景产生区分度，为后续的特征处理和最终的检测任务提供了更高质量的输入。

2.4　损失函数的优化

传统的损失函数普遍存在以下问题：首先是忽略了不同样本难易程度差异对结果的影响；其次是在训练初期，锚框会先扩大自身的尺寸以求与目标框产生重叠，然后再缓慢调整到正确位置和大小，而不是直接向目标框移动，这大大降低了模型收敛速度。针对以上问题，本文提出了一种改进的损失函数——Focaler-Powerful-loU，结合了

Focaler-IoU[22]和 Powerful-IoU[23]的思想,实现自适应聚焦难易样本并且解决锚框不合理扩大问题以提高模型性能。

Focaler-IoU 方法通过线性区间映射重构 IoU 值,使模型能够灵活聚焦于不同难度的回归样本,从而提升定位精度。Powerful-IoU 方法则通过设计一个对目标尺寸自适应的惩罚项和一个基于锚框质量的非单调梯度调整函数,来解决回归过程中不合理的锚框扩大问题,从而引导更直接高效的收敛路径,提升回归速度与检测精度。Focaler-Powerful-IoU 的定义如图 8 所示。

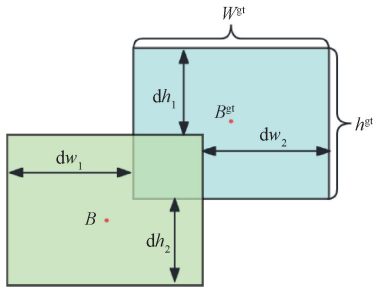


图 8　Focaler-Powerful-IoU 定义

Fig. 8　Focaler-Powerful-IoU definiton

Focaler-Powerful-loU 的计算公式如式(12)～(15)所示。

$$IoU^{focaler}=\begin{cases}0, & IoU<d\\ \dfrac{IoU-d}{u-d}, & d\leqslant IoU\leqslant u \quad 0\leqslant d,u\leqslant 1\\ 1, & IoU>d\end{cases} \tag{12}$$

$$L_{Focaler-IoU}=1-IoU^{focaler} \tag{13}$$

$$P=\left(\frac{dw_1}{w^{gt}}+\frac{dw_2}{w^{gt}}+\frac{dh_1}{h^{gt}}+\frac{dh_2}{h^{gt}}\right)/4 \tag{14}$$

$$L_{PIoU}=L_{IoU}+1-e^{-p^2},0\leqslant L_{PIoU}\leqslant 2 \tag{15}$$

式中:IoU 是原始的交并比值,L_{IoU} 表示原始的交并比损失,$IoU^{focaler}$ 是对 IoU 重构后的值,L_{PIoU} 为 Powerful-IoU 的损失函数。B 和 B^{gt} 分别为预测框和真实框;dw_1 是边界框 B 和 B^{gt} 左边界的水平距离;dw_2 是边界框 B 和 B^{gt} 右边界的水平距离;dh_1 是边界框 B 和 B^{gt} 上边界的垂直距离;dh_2 是边界框 B 和 B^{gt} 下边界的垂直距离;w^{gt} 为真实框的宽度。h^{gt} 为真实框的高度。

其中 d 和 u 是两个可调节的超参数,它们共同定义了一个有效学习区间。降低 u 的值,会使得更多 IoU 较高的样本梯度被抑制,梯度将主要由 IoU 低于 u 的样本贡献,从而实现了对困难样本的聚焦。提高 d 的值,会使得更多 IoU 较低的样本梯度被抑制,梯度将主要由 IoU 高于 d 的样本贡献,从而实现了对简单样本的聚焦。

惩罚项 P 的分子直接反映了预测框与目标框 4 条边的对齐程度,分母只与真实框的尺寸有关。当锚框中心点不变而仅扩大尺寸时,分子会增加或不变,但惩罚项 P 不会因此减小,这从数学上彻底消除了模型通过扩大锚框来降低损失的动机。最终的 Focaler-Powerful-IoU 损失函数结合了 Focaler-IoU 与 Powerful-IoU 两者的优势,以实现更精确的边界定位,其定义如式(16)所示。

$$L_{Focaler-Powerful-IoU}=2-e^{-p^2}-IoU^{focaler} \tag{16}$$

3　实验与结果分析

3.1　数据集与实验环境

选取国际无人机视觉领域权威数据集 VisDrone2019[24]作为实验验证对象。VisDrone2019 数据集由 288 个视频片段组成,包括 10 209 幅静止图像,每张图像平均包含 53 个实例对象,场景涵盖了不同的飞行高度、光照条件、天气状况等。数据集中有 10 类对象,每个类别分布如图 9 所示。

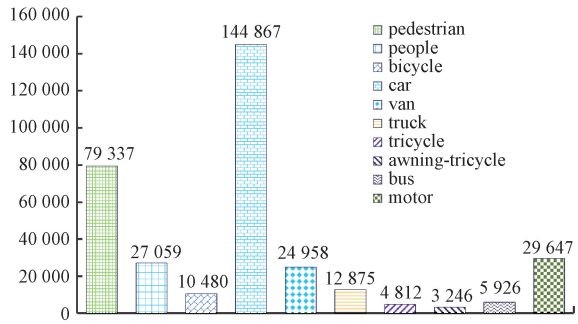


图 9　目标类别分类

Fig. 9　Target category distribution

VisDrone2019 数据集的图像中大部分实例对象尺寸长宽都小于 50 个像素,实例尺寸分布如图 10 所示。

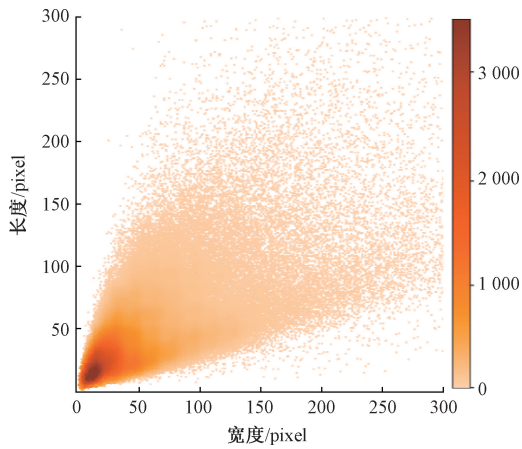


图 10　目标尺寸分类

Fig. 10　Target size distribution

在硬件方面,使用了 i5-12600KF 处理器,10 核 16 线程,主频 3.70 GHz,32 G 运行内存,图形处理器 RTX4060 Ti, 16 GB 显存;在软件方面,使用了 Pytorch2.2.2 和 Torchvision0.17.2;系统为 Windows11 专业版。具体的实验环境参数设置如表 1 所示。

表 1　实验环境参数设置表

Table 1　Setting table of experimental environment parameters

参数	设置
epochs	300
Batch	8
Imgsz	640
Workers	0
optimizer	AdamW
Lr0	0.000 1
Momentum	0.9
Weigh_decay	0.000 1

3.2　消融实验

实验采用常见评估指标包括准确率(precision，P)、召回率(recall，R)、平均精度(average precision，AP)、平均精度均值(mean average precision，mAP)。为评估每项改进的有效性，依次添加这些增强模块进行了一系列消融实验。实验结果如表 2 所示，表明每项修改均带来了性能的逐步提升。实验结果表明，用 CS-IGU 模块替换 backbone 中的 basicblock 后，模型的 mAP_{50} 和 $mAP_{50:95}$ 分别上涨了 1%和 0.7%，而参数量和计算量分别下降了 30.3%和 17.7%，充分表明了 CS-IGU 在增强模型性能、降低模型复杂度的积极作用。在单独加入 SO-CCFF 和 GFEM 后，虽然模型的参数量计算量都有小幅增加，但是模型的 mAP、准确率和召回率都有明显上升，说明了改进有效，模型对小目标的定位效果更佳，漏检的情况得以改善。最后单独加入 Focaler-Powerful-IoU，在模型参数量、计算量、FPS 都基本不变的情况下，mAP_{50} 和 $mAP_{50:95}$ 都上涨了 0.9%。

表 2　消融实验

Table 2　Ablation experiments

RT-DETR	CS-IGU	SO-CCFF	GFEM	LOSS	mAP_{50}/%	$mAP_{50:95}$/%	P/%	R/%	Param/($\times10^6$)	GFLOPs	FPS/fps
√					47.6	29.1	62.4	45.8	19.8	57.0	79.6
√	√				48.6	29.8	63.9	46.6	13.8	46.9	72.2
√		√			48.8	30.1	63.5	48.0	20.3	63.0	71.6
√			√		48.7	30.1	63.8	47.2	22.6	68.6	69.7
√				√	48.5	30.0	61.7	47.3	19.8	57.0	76.7
√	√		√		50.2	31.1	63.8	48.3	16.6	58.5	67.9
√	√			√	49.9	31.0	63.9	48.5	13.8	46.9	71.6
√		√		√	49.8	30.8	64.0	47.7	20.3	63.0	68.6
√			√	√	49.1	30.6	63.2	47.1	22.6	68.6	66.4
√	√	√	√		50.3	31.1	63.9	48.7	17.1	64.5	64.8
√	√	√	√	√	50.6	31.4	64.4	49.7	17.1	64.5	63.9

最终，通过将 4 种改进策略相互结合进行综合性的改进，相较于原始的 RT-DETR 基准模型，所构建的改进模型能够充分融合 4 种改进策略的优势。准确率和召回率相较于基准模型分别提升了 2.0%和 3.9%，mAP_{50} 和 $mAP_{50:95}$ 分别提升了 3%和 2.3%。从实验结果可见 IF-DETR 改进策略的有效性。

3.3　对比实验

消融实验已证实 IF-DETR 的有效性，为了深入探究改进算法的先进性，将 IF-DETR 与其他算法进行实验比较。在 VisDrone2019 数据集上，基于相同训练环境、相同数据集，并为每个模型使用推荐的超参数。选取目前主流的 FasterRCNN[25]、YOLOv8[26]、YOLOv11[27]和 YOLOv12[28]系列算法和参数量计算量更大的 Deformable-DETR[10]、RTDETR-R34 以及同领域模型 UAV-DETR[17]、SO-DETR[18]进行了一系列深入的比较。实验结果如表 3、4 所示，加粗部分为该类别在所有算法中的最优值。

通过表 3 可以得出，IF-DETR 在 VisDrone 数据集上综合性能优异。mAP_{50} 达到了 50.6%，超越了所有参与比较的 YOLO 系列模型以及同领域基于 RT-DETR 的改进模型。此外，本文算法达到了最高 63.2%的准确率和最高 49.7%的召回率，减少了漏检的可能性。但是在追求高精度过程中，本文算法在计算量方面相对于基准模型有所上升，导致 FPS 有所下降。本文算法通过增加少量计算量实现精度的显著提升，并且能够满足实时性的要求，整体性能表现最出色。

通过表 4 可以得出，IF-DETR 虽然在卡车(truck)、面包车(van)、公交车(bus)、遮阳篷三轮车(awning-tricycle)这类大物体检测精度上逊色于最优值，但仍优于基线模型。同时在数量最多的关键大物体汽车(car)检测精度最高。在行人(pedestrian)、人(people)、自行车(bicycle)、三轮车(tricycle)、摩托车(motor)这 5 个能代表小物体挑战的类别的检测任务当中表现出色，同时 P 和 R 均取得了最

表 3　不同模型在 VisDrone2019-DET-Val 数据集上对比实验

Table 3　Comparision experiments of different modules on the VisDrone2019-DET-Val dataset

模型	mAP_{50}/%	$mAP_{50:95}$/%	P/%	R/%	Param/($\times10^6$)	GFLOPs	FPS/fps
Faster R-CNN	33.3	17.1	43.6	35.7	41.3	206.4	15.4
YOLOv8n	35.2	20.5	46.1	34.9	3.0	8.1	183.2
YOLOv8s	40.1	24.0	52.8	39.0	11.1	28.5	118.3
YOLOv8m	44.0	26.9	56.0	42.5	25.8	78.7	82.4
YOLOv11n	34.9	20.4	44.8	34.7	**2.5**	**6.3**	185.5
YOLOv11s	41.3	24.9	52.4	39.9	9.4	21.3	120.2
YOLOv11m	46.8	28.8	58.4	44.4	20.0	68.2	85.4
YOLOv12n	35.2	20.7	46.4	34.8	2.5	6.3	177.8
YOLOv12s	41.6	25.2	52.4	40.1	9.2	21.2	125.5
YOLOv12m	46.0	28.5	55.8	44.9	20.1	67.2	83.3
Deformable-DETR	43.1	27.1	46.5	39.2	40.1	186.8	14.1
RT-DETR-R18	47.6	29.1	62.4	45.8	19.8	57.0	79.6
RT-DETR-R34	48.2	29.5	61.6	47.5	31.1	88.8	60.7
UAV-DETR-R18	49.7	30.6	63.0	47.6	21.3	72.5	36.5
SO-DETR-R18	49.4	30.6	61.1	47.7	20.6	65.3	67.5
IF-DETR	**50.6**	**31.4**	**63.2**	**49.7**	17.2	64.5	63.9

表 4　不同模型在 VisDrone2019-DET-Test 数据集上各类别对比实验

Table 4　Comparision experiments of different modules on the Visdrone2019-DET-Test dataset　(%)

模型	AP_{50}										mAP_{50}	P	R
	Pedestrian	People	Bicycle	Car	Van	Truck	Tricycle	Awning-tricycle	Bus	Motor			
Faster R-CNN	21.3	15.7	6.6	51.8	29.4	19.1	14.9	8.8	31.3	20.8	21.7	30.5	23.2
YOLOv8n	23.5	12.5	6.1	68.1	32.3	35.8	16.8	16.1	52.8	24.6	28.9	40.6	30.7
YOLOv8s	29.0	15.8	11.4	72.5	36.8	40.7	19.6	18.7	57.4	30.9	33.3	47.4	34.7
YOLOv8m	31.6	17.0	12.8	74.9	38.8	45.0	20.9	18.7	59.8	34.5	35.4	48.6	37.3
YOLOv11n	24.0	13.0	6.4	68.2	33.4	34.9	14.2	15.1	53.1	25.6	28.8	40.7	31.2
YOLOv11s	29.0	14.9	10.6	73.2	38.5	42.0	20.7	19.3	57.6	30.0	33.6	47.1	35.2
YOLOv11m	33.4	17.9	14.6	76.6	**42.5**	**48.2**	25.2	20.7	**61.6**	35.4	37.6	49.6	39.4
YOLOv12n	24.4	12.4	8.9	69.1	32.3	36.7	14.8	16.1	53.8	25.5	29.4	41.3	31.7
YOLOv12s	28.7	14.8	11.2	73.5	38.5	43.0	21.1	19.8	58.2	31.1	34.0	46.6	35.9
YOLOv12m	33.3	17.3	15.4	76.5	42.4	46.4	26.6	**23.9**	61.5	36.3	38.2	51.9	38.8
Deformable-DETR	38.7	23.8	12.6	70.5	34.7	30.9	20.9	15.1	40.4	39.1	32.6	47.3	33.8
RT-DETR-R18	38.9	28.7	14.4	76.7	35.4	41.0	23.9	17.4	57.1	40.7	37.4	54.3	39.0
RT-DETR-R34	39.4	28.5	11.7	76.9	37.3	41.6	24.4	17.2	57.2	40.1	37.5	54.2	39.3
UAV-DETR-R18	39.6	29.3	14.2	76.6	40.0	43.7	26.6	19.6	60.2	43.4	39.3	55.8	40.7
IF-DETR	**41.2**	**29.3**	**17.1**	**77.9**	39.5	44.5	**26.7**	20.2	59.4	**44.1**	**39.9**	**56.0**	**41.5**

优值。这充分说明了 IF-DETR 不仅在小目标检测上取得了明显的精度提升,在关键大目标汽车的表现也有明显优势。综上,相较于其他常用的检测方法,IF-DETR 在航拍图像中的目标检测综合精度上占优。

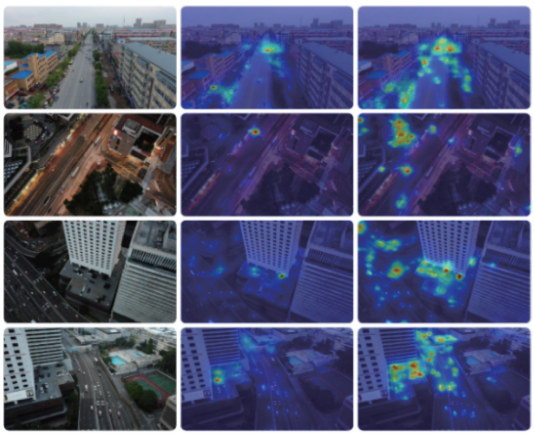
3.4　实验结果可视化

1) 模型热力图

本文采用 LayerCAM[29] 可视化技术直观地展示改进模型的性能。LayerCAM 用于显示深度神经网络中特征

映射对预测结果的贡献，通过激活图中某一区域的亮度来表示该区域在预测输出过程中所占的权重，颜色越鲜亮，表示预测输出的关注度越高。

本文选取了 4 张小目标多，检测难度大的图片作为示例，原图如图 11(a)所示。RT-DETR 模型和 IF-DETR 模型热力图结果如图 11(b)、(c)所示。



(a) 原图　(b) RT-DETR　(c) IF-DETR

(a) Original picture

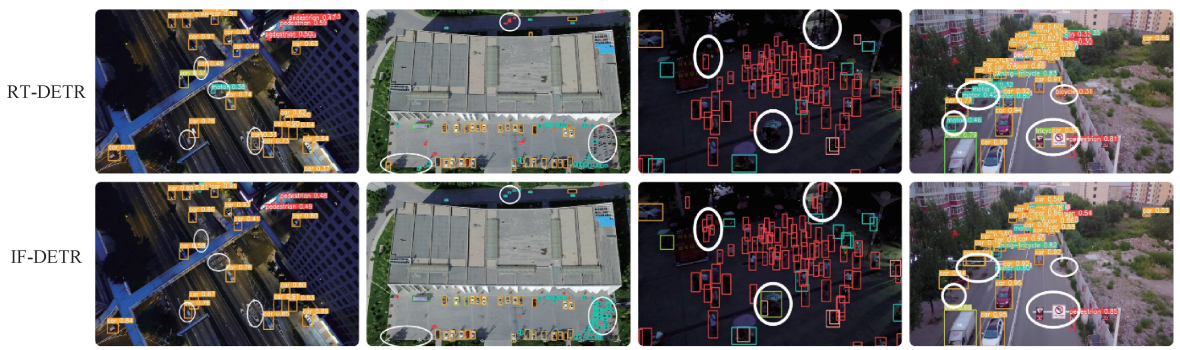
图 11　热力图可视化对比结果

Fig. 11　Comparison results of heat map visualization

观察可见，RT-DETR 输出的热力图高亮区域稀疏且分布不均，反映了模型在航拍场景下的目标检测能力不足，漏检情况严重。IF-DETR 输出的热力图高亮区域密集且分布均匀。在原图中的阴影区域，IF-DETR 较基线模型识别到了更多目标，说明光照门控单元发挥了作用，使改进模型能够识别到物体的本质特征，提高检测精度。同时，IF-DETR 成功检测到了大量原本被基线模型忽略的微小目标(如远处的行人和车辆)，说明 SO-CCFF 有效保留了原本丢失的小目标特征，减少了漏检现象。此外，基线模型部分高亮区域落在无意义的背景上，说明对背景噪声的抑制能力较弱。改进模型通过 GFEM 模块对高频结构信息增强，使得模型能够将密集排列的小目标从复杂的背景纹理识别出来，抑制了背景噪声，减少了粘连误检现象。综上，热力图结果证明了改进方法的有效性，有助于提高遥感目标检测的精度，且作用明显。

2) 检测结果实例

为了直观呈现本文模型在航拍图像目标检测这一任务中的表现，且更具说服力。本研究选取了 4 张不同场景下的图片进行检测效果的展示。如图 12 所示，上排图像为 RT-DETR 模型的检测效果，下排为本文模型的检测效果。



(a) 低光稀疏场景　(b) 远景小目标场景　(c) 低光密集场景　(d) 复杂背景场景

(a) Low-light sparse scene　(b) Distant small targets scene　(c) Low-light dense scene　(d) Complex background scene

图 12　4 种不同场景下检测可视化对比结果

Fig. 12　Comparison results of detection visualization under four different scenarios

图 12(a)中本文模型找到了基线模型漏检的车辆，同时避免了基线模型将道路交通标线错检为车辆的错误。图 12(b)中本文模型找到了基线模型漏检的密集微小的摩托车群。图 12(c)中本文模型在密集人群中正确检测出了基线模型漏检的行人。图 12(d)中本文模型找出了被遮挡的目标，同时避免了误检。可以看出本文模型分别在低光稀疏、远景小目标、低光密集、以及复杂背景干扰这 4 个复杂场景上都展现出了更好的效果，极大提高了航拍复杂场景下目标的检测能力，同时显著降低了原模型的漏检以及误检问题。本文模型在航拍目标检测任务中表现出较强的能力，在对目标精准定位的同时能够准确对其类别进行区分，展现了出色的鲁棒性和精确度。

4　结　　论

本研究提出了一种基于改进 RT-DETR 的航拍目标检测模型 IF-DETR。针对航拍图像中存在的复杂光照条件、小目标密集、复杂背景干扰等问题，通过在主干网络设计跨阶段光照门控单元强化模型对复杂光照条件下对目标的检测精度，并且在颈部网络构建小目标跨尺度特征融合结构进一步提高小目标的识别精度；此外，算法在颈部引入高斯特征增强模块增强目标结构信息，抑制复杂背景噪声；最后设计 Focaler-Powerful-IoU 损失函数提高对目标的定位精度。实验结果表明，IF-DETR 与 RT-DETR 相比，在 Visdrone2019 数据集上的 mAP_{50}、$mAP_{50:95}$、准确率

和召回率分别提升了 3.0%、2.3%、2%和 3.9%,同时在 Visdrone2019 中的小目标类别检测性能提升明显,展现了优于同类主流算法的检测效果,对无人机高精度目标检测任务具有较高的实际意义与应用价值。

参考文献

[1] 钟帅,王丽萍.无人机航拍图像目标检测技术研究综述[J]. 激光与光电子学进展, 2025, 62(10): 5-19.
Zhong Shuai, Wang Liping. Review of Research on Object Detection in UAV Aerial Images[J]. Laser & Optoelectronics Progress, 2025, 62(10): 5-19.

[2] 李琼, 考月英, 张莹,等.面向无人机航拍图像的目标检测研究综述[J]. 图学学报, 2024, 45(6): 1145-1164.
Li Qiong, Kao Yueying, Zhang Ying, et al. Review on object detection in UAV aerial images[J]. Journal of Graphics, 2024, 45(6): 1145-1164.

[3] 郭润泽,孙备,孙晓永,等.无人机弱光条件下多模态融合目标检测方法[J].仪器仪表学报, 2025, 46(1): 338-350.
Guo Runze, Sun Bei, Sun Xiaoyong, et al. Multimodal fusion object detection method for UAVs under low light conditions[J]. Chinese Journal of Scientific Instrucment, 2025, 46(1): 338-350.

[4] 孙晓永,孙备,郭润泽,等.面向多无人机协同的多模态目标检测方法[J].仪器仪表学报, 2025, 46(2): 209-220.
Sun Xiaoyong, Sun Bei, Guo Runze, et al. Multimodal target detection method for multi-UAV coordination [J]. Chinese Journal of Scientific Instrucment, 2025, 46(2): 209-220.

[5] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//IEEE Conference on Computer Vision and Pattern Recognition. Columbus, 2014: 580-587.

[6] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[7] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers [C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 213-229.

[8] 刘亚蒙,赵友全,孙振涛,等. 构建改进 RT-DETR 算法检测隐形眼镜环状波纹缺陷[J].电子测量与仪器学报, 2024, 38(5): 1-9.
Liu Yameng, Zhao Youquan, Sun Zhentao, et al. Constructing an enhanced RT-DETR algorithm on for detecting annular ripple defects in contact lenses[J]. Journal of Electronic Measurement and Instrucmentation, 2024, 38(5): 1-9.

[9] 周景,刘心,唐振洋,等. SDH-DETR 轻量化绝缘子缺陷检测算法[J].电子测量技术, 2025, 48(11): 88-104.
Li Ming, Liu Xin, Tang Zhenyang, et al. SDH-DETR lightweight insulator defect detection algorithm[J]. Electronic Measurement Technology, 2025, 48(11): 88-104.

[10] Zhu X Z, Su W J, Lu L W, et al. Deformable DETR: Deformable transformers for end-to-end object detection[PP/OL]. V4. (2021-03-18). https://doi.org/10.48550/arXiv.2010.04159.

[11] Meng D P, Chen X K, Fan Z J, et al. Conditional DETR for fast training convergence[C]//IEEE/CVF International Conference on Computer Vision, 2021: 3651-3660.

[12] Wang Y M, Zhang X Y, Yang T, et al. Anchor detr: Query design for transformer-based detector [C]// AAAI Conference on Artificial Intelligence, 2022, 36(3): 2567-2575.

[13] Liu S L, Li F, Zhang H, et al. DAB-DETR: Dynamic anchor boxes are better queries for DETR[PP/OL]. V4. (2022-03-30). https://doi.org/10.48550/arXiv.2201.12329.

[14] Zong Z F, Song G L, Liu Y. Detrs with collaborative hybrid assignments training [C]//IEEE/CVF International Conference on Computer Vision, 2023: 6748-6758.

[15] Zhang H, Li F, Liu S L, et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection[PP/OL]. V4. (2022-07-11)[2025-11-29]. https://doi.org/10.48550/arXiv.2203.03605.

[16] Zhao Y A, Lyu W Y, Xu S L, et al. DETRs beat YOLOs on real-time object detection[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 16965-16974.

[17] Zhang H X, Zhang H, Liu K, et al. UAV-DETR: efficient end-to-end object detection for unmanned aerial vehicle imagery [C]//2025 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2025: 15143-15149.

[18] Zhang H X, Zhang H, Mei A, et al. SO-DETR: leveraging dual-domain features and knowledge distillation for small object detection [C]//2025 International Joint Conference on Neural Networks (IJCNN). IEEE, 2025: 1-8.

[19] Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN [C]//IEEE/CVF Conference on Computer Vision and Ppattern Recognition Workshops, 2020: 390-391.

[20] Zhang Y F, Zhou S, Li H F. Depth information assisted collaborative mutual promotion network for single image dehazing[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 2846-2855.

[21] Sunkara R, Luo T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects [C]//Joint European conference on machine learning and knowledge discovery in databases. Cham: Springer Nature Switzerland, 2022: 443-459.

[22] Zhang H, Zhang S J. Focaler-IoU: More focused intersection over union loss[PP/OL]. V1. (2024-01-19) [2025-11-29]. https://doi.org/10.48550/arXiv.2401.10525.

[23] Liu C, Wang K G, Li Q, et al. Powerful-IoU: More straightforward and faster bounding box regression loss with a nonmonotonic focusing mechanism[J]. Neural Networks, 2024, 170: 276-284.

[24] Du D W, Zhu P F, Wen L Y, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results [C]//IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[25] Ren S Q, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015, 28.

[26] Yaseen M. What is YOLOv8: An in-depth exploration of the internal features of the next-generation object detector[PP/OL]. V1. (2024-08-28)[2025-11-29]. https://doi.org/10.48550/arXiv.2408.15857.

[27] Khanam R, Hussain M. YOLOv11: An overview of the key architectural enhancements[PP/OL]. V1. (2024-10-23)[2025-11-29]. https://doi.org/10.48550/arXiv.2410.17725.

[28] Tian Y J, Ye Q X, Doermann D. YOLOv12: Attention-centric real-time object detectors [J]. Advances in Neural Information Processing Systems, 2026, 38: 78433-78457.

[29] Jiang P T, Zhang C B, Hou Q B, et al. Layercam: Exploring hierarchical class activation maps for localization [J]. IEEE Transactions on Image Processing, 2021, 30: 5875-5888.

作者简介

杨洋，硕士研究生，主要研究方向为图像处理、目标检测等。

E-mail：1693138601@qq.com

魏为民（通信作者），硕士生导师，主要研究方向为信息安全、图像处理、数字取证等。

E-mail：wwm@shiep.edu.cn

郑书情，硕士研究生，主要研究方向为图像篡改检测、目标检测等。

E-mail：zsq4510@163.com

马凯楠，硕士研究生，主要研究方向为图像篡改检测、目标检测等。

E-mail：13965328271@163.com

张劲阳，硕士研究生，主要研究方向为图像处理、目标检测等。

E-mail：15255693499@163.com