DOI:10.11918/202511105

融合混合池化与局部注意力的无人机目标检测

马保全[1,2],范文龙[1],杨煜晨[3],韩昱哲[4],郝博文[2],武晓文[2],王甜竹[2]

(1.北京理工大学 计算机学院,北京 100081; 2.华北计算机系统工程研究所,北京 100000;
3.南京理工大学 紫金学院,南京 210023; 4.中国民航大学 空中交通管理学院,天津 300300)

摘 要:为解决无人机航拍图像中目标尺度变化剧烈、背景复杂,以及小目标占比高导致现有YOLO系列模型单阶段检测算法小、目标检测性能不足的问题,提出一种融合混合池化与局部注意力机制的无人机目标检测方法HP-LS-YOLO。该方法以YOLOv11模型为基础框架,首先,设计空间金字塔混合池化快速策略(SPHPF),在单一尺度上并行融合最大池化与平均池化特征,引入多感受野上下文信息,并通过卷积自适应筛选增强复杂背景下的特征表达能力。其次,构建融合小波卷积与局部注意力机制的C2PLSA模块,其中小波变换注意力在保持模型轻量化的前提下有效扩大感受野,捕获小目标周围的大尺度场景语义;局部自注意力块采用Top-K稀疏注意力机制,突出小目标的局部边缘与纹理特征,同时抑制冗余背景干扰。最后,在多个公开数据集上开展对比验证实验,通过与原模型以及同类型轻量化检测算法的对比,全方位检验所提改进策略对无人机航拍密集小目标检测效果的提升作用。结果表明:与YOLOv11模型相比,所提方法在MSCOCO数据集验证集上的$AP^{val}_{50:95}$、AP^{val}_{50}、AP^{val}_{S}分别提升0.9%、0.9%、1.4%,在VisDrone和UAVDT数据集上的额外实验进一步验证了该方法在复杂无人机场景下的鲁棒性,所提出的HP-LS-YOLO模型能够在保持模型轻量化的同时,有效提升无人机航拍图像中密集小目标的检测性能。

关键词:YOLOv11;无人机航拍图像;小目标检测;混合池化;局部注意力机制

中图分类号:TP391.41 **文献标志码**:A **文章编号**:0367-6234(2026)08-0154-10

Unmanned aerial vehicle object detection with hybrid pooling and local attention

Ma Baoquan[1,2], Fan Wenlong[1], Yang Yuchen[3], Han Yuzhe[4], Hao Bowen[2], Wu Xiaowen[2], Wang Tianzhu[2]

(1. School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China;
2. North China Computer System Engineering Research Institute, Beijing 100000, China; 3. Zijin College, Nanjing University of Science and Technology, Nanjing 210023, China; 4. College of Air Traffic Management, Civil Aviation University of China, Tianjin 300300, China)

Abstract: To solve the problem of insufficient small object detection performance of existing YOLO series single-stage detection algorithms caused by severe object scale variation, complex backgrounds, and a high proportion of small objects in unmanned aerial vehicle aerial images, this paper proposed an unmanned aerial vehicle object detection method fusing hybrid pooling and local attention mechanism, named HP-LS-YOLO. Based on the YOLOv11 framework, this method first designed a spatial pyramid hybrid pooling fast (SPHPF) strategy to introduce multi-receptive field contextual information at a single scale by parallelly fusing max pooling and average pooling features and enhanced the feature representation capability under complex backgrounds through convolution-based adaptive filtering. Secondly, a C2PLSA module integrating wavelet convolution and local attention mechanism was constructed, in which the wavelet transform attention effectively enlarged the receptive field while keeping the model lightweight and introduced large-scale scene semantics around small objects. The local self-attention block adopted a Top-K sparse attention mechanism to highlight the local edge and texture features of small objects, while suppressing redundant background interference. Finally, comparative validation experiments were carried out on multiple public datasets to comprehensively verify the enhancement effect of the proposed improvement strategies on dense small object detection in unmanned aerial vehicle aerial images by comparing with the baseline model and similar lightweight detection algorithms. The results indicate that compared with YOLOv11, the proposed method improves $AP^{val}_{50:95}$, AP^{val}_{50}, AP^{val}_{S} on the validation set of the MSCOCO dataset by 0.9%, 0.9%, 1.4%, respectively. Additional experiments on VisDrone and UAVDT datasets further verify the robustness of this method in complex

收稿日期:2025-11-30;**录用日期**:2026-01-05;**网络首发日期**:2026-05-29
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20260529.0946.002
基金项目:中国教育部基础学科与交叉学科突破计划(JYB2025XDXM108、JYB2025XDXM201);国家自然科学基金(62225203、62532001);北京市自然科学基金(L241010)
作者简介:马保全(1982—),男,博士,正高级工程师
通信作者:马保全,baoquan.ma@buaa.edu.cn;范文龙,2646911874@qq.com

unmanned aerial vehicle scenarios. The proposed HP-LS-YOLO can effectively improve the detection performance of dense small objects in unmanned aerial vehicle aerial images while keeping the model lightweight.
Keywords: YOLOv11; unmanned aerial vehicle aerial image; small object detection; hybrid pooling; local attention mechanism

近年来,无人机(unmanned aerial vehicle, UAV)技术的飞速发展极大地推动了无人机在交通流量监测、灾后损失评估、野生动物栖息地保护,以及精准农业管理等多个领域的实际应用。在这些场景中,从航拍图像中进行实时目标检测是一项关键任务:其性能直接影响后续决策的可靠性,例如高压输电线路巡检中的微缺陷识别[1]、通过车辆和行人检测进行交通流量分析[2],以及精准农业中作物病虫害的早期诊断[3]。

然而,无人机航拍图像的固有特性给目标检测任务带来了独特且严峻的挑战。尺度变化极端现象普遍存在:由于飞行高度和拍摄距离的动态变化,目标尺寸差异巨大,从仅占几个像素的远距离行人到跨越数百像素的近距离车辆,其尺度范围远超自然场景图像。航拍场景中密集遮挡和复杂背景较为常见,重叠目标(如人群或集群车辆)和杂乱背景(如植被、建筑物或地形纹理)常常模糊目标边界,降低特征辨识度。此外,小目标(如鸟类、小型动物或远距离行人)占比高,这类目标仅由 10 ~ 30 个像素组成,判别特征稀疏,易被噪声淹没。这些特性共同导致传统目标检测算法性能严重下降,尤其是在小目标检测方面,高漏检率和误检率极大地限制了无人机系统的实用价值。

过去几年,YOLO 系列等单阶段检测器因其速度和精度的平衡在目标检测领域备受青睐[4-16]。然而,将其直接应用于无人机航拍图像时,存在明显缺陷:特征提取过程中的多次下采样容易丢失小目标的细微特征,且强烈的复杂背景干扰使得可靠检测(尤其是小目标检测)难以实现。

为解决这些问题,本文基于 YOLOv11 模型设计了混合池化局部注意力 YOLO(hybrid pooling local self-attention-YOLO, HP-LS-YOLO)模型,旨在提高无人机图像目标(尤其是小目标)的检测精度。首先,设计了空间金字塔混合池化快速策略(spatial pyramid hybrid pooling fast, SPHPF)。该策略融合最大池化(保留突出边缘和纹理)和平均池化(提供背景上下文和亮度信息)的特征,随后通过卷积模块进行自适应滤波,提取关键特征组合。其次,通过将小波卷积(wavelet transform convolution, WTConv)与高效局部注意力机制(efficient local attention, ELA)串行连接,构建了小波注意力(wavelet transform-powered attention, WTPA)特征增强结构。WTConv 在多级小波分解与下采样后,在较低分辨率子带上施加小核卷积并经逆小波重建,相当于在不显著增加参数量的前提下扩大了每个位置的有效感受野,为小目标提供其周围大尺度场景的语义线索;ELA 在水平与垂直方向自适应建模空间注意力,增强与目标相关的响应并抑制部分背景噪声。最后,提出了融合局部注意力机制(local self-attention block, LSAB)的局部注意力(partial local self-attention, PLSA)干扰抑制模块。该机制针对卷积操作进行定制,为感受野内的每个像素生成动态权重,在卷积过程中增强与目标相关(尤其是小目标相关)关键区域的特征,同时削弱无关背景干扰。该机制在局部窗口内仅保留与当前像素最相关的 TopK 邻域特征,减少天空、地面纹理等无关区域对小目标响应的干扰,并在卷积过程中对目标区域给予更高权重。

1　相关工作

1.1　实时目标检测器

深度卷积神经网络的兴起将实时目标检测从两阶段框架(如 R-CNN 系列[17-19])转向了 YOLO 等端到端单阶段架构。YOLO 系列模型不断演进:YOLOv1 模型[4]首次在一次前向传播中实现目标定位和分类,但在小目标检测方面表现不佳。YOLOv2 模型[5]添加了锚点机制和 DarkNet19,提高了精度和速度。YOLOv3 模型[6]升级为带残差连接的 DarkNet53,支持多尺度检测。YOLOv4 模型[7]整合了CSP、SPP 和 PANet,实现了更好的速度-精度平衡。YOLOv5 模型[8]采用模块化设计和数据增强,便于部署。YOLOv6-YOLOv8 模型[9-11]优化了骨干网络/颈部网络,适用于轻量化边缘部署。YOLOv9 模型[12]引入了可编程梯度和高效层聚合。YOLOv10 模型[13]实现了无 NMS 的端到端处理。YOLOv11 模型[14]通过 C3k2/C2PSA 单元进行改进。YOLOv12 模型[15]添加了 RELAN 和区域注意力。YOLOv13 模型[16]提出了 HyperACE 和 FullPAD 以进一步提升性能。

随着这些进展,研究人员开发了专门的 YOLO 变体,以应对无人机航拍场景中的关键挑战——小

目标、复杂背景,以及有限计算能力。基于 YOLOv8 模型的 Drone-YOLO[20] 整合了 RepVGG 下采样层、3 层 PAFPN 结构,以及定制的 Sandwich-Fusion 模块,以提升多尺度小目标检测性能。RSG-YOLO[21] 通过 RFCAConv 注意力机制和轻量化 C2GELAN 单元重建骨干网络,实现了参数压缩和丰富梯度。UAV-YOLOv8[22] 采用 WIoU ×3 动态损失、BiFormer 稀疏注意力,并添加 FFNB 模块构建五尺度检测头,大幅降低了小目标漏检率。这些变体凸显了为无人机场景量身定制 YOLO 架构的必要性和价值。

1.2 大感受野卷积

作为神经网络的核心组件,卷积模块设计对模型性能至关重要。为在控制参数的同时提升特征提取能力,扩大感受野以捕捉更丰富的上下文信息已成为关键研究方向。近年来,出现了多种创新卷积模块:RFAConv[23] 引入感受野注意力机制,打破参数共享限制;FADC[24] 在局部频域动态调整扩张率;WTConv[25] 利用小波分解和小核卷积,实现参数对数增长与感受野指数扩张。这些模块通过多种策略为大感受野卷积提供了高效解决方案。

1.3 注意力机制

在计算机视觉领域,注意力机制对于提升深度神经网络性能至关重要。ELA[26] 是一种轻量化局部注意力方法,结合一维卷积和组归一化增强特征。它在不降低通道维度的情况下捕捉感兴趣区域的位置信息,适用于图像分类和目标检测等任务。

同样,SCSA[27] 整合了多语义空间注意力和渐进式通道自注意力,融合全局上下文依赖和通道间相关性,在视觉任务中表现优于先进注意力方法。RAGAT[28] 通过区域划分、独立注意力计算,以及自适应权重调整解决区域特征差异问题,平衡局部细节和全局特征。DSA[29] 基于输入内容动态分配空间注意力,在包含小目标的复杂背景中突出目标并抑制干扰。

尽管这些机制在实现方式和关注重点上存在差异,但均通过优化特征注意力,为计算机视觉发展提供了支持。

2 方法

2.1 整体架构

HP-LS-YOLO 模型的整体结构如图 1 所示。该模型保留了 YOLOv11 模型的架构框架用于目标提取,主要由 3 个核心组件组成:骨干网络实现多尺度特征提取、颈部网络进行多尺度特征融合,以及检测头完成目标位置和类别的最终提取。为实现高效特征提取,HP-LS-YOLO 模型用所提的 SPHPF 和 C2f –局部部分自注意力模块(C2f-partial local self-attention, C2PLSA)替换了原始 YOLOv11 模型中的相应模块。

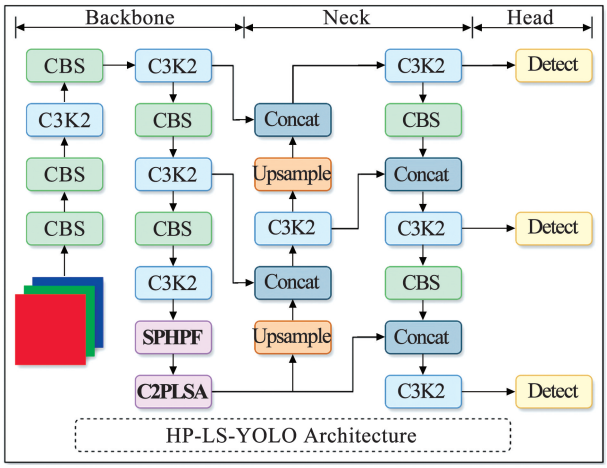


图 1 HP-LS-YOLO 模型的网络架构

Fig. 1 Network architecture of HP-LS-YOLO model

2.2 空间金字塔混合池化快速策略(SPHPF)

原始快速空间金字塔池化(spatial pyramid pooling-fast, SPPF)模块仅通过最大池化捕捉显著特征。在无人机场景中,这种方式忽略了平均背景信息,导致特征表示单一,且在复杂高空环境中对噪声更为敏感。

为解决无人机场景中小目标特征稀疏、背景干扰强的问题,本文提出 SPHPF 模块,在原有 SPPF 结构基础上,引入混合池化块(hybrid pooling block, HPB),以替代单一最大池化操作。HPB 通过并行分支同时引入最大池化与平均池化,并利用后续卷积对两类特征进行融合建模。从特征响应机理来看,最大池化侧重于保留局部区域内的极值响应,有利于突出目标的边缘、角点,以及纹理等判别性结构;平均池化则对局部窗口内的响应进行统计汇总,更接近于对区域整体亮度与纹理能量的估计。在无人机航拍图像中,小目标通常仅占据局部窗口中的少量像素,其余区域多为天空、地面,以及建筑背景,因此局部平均池化在统计意义上可视为对背景上下文信息的有效表征。将两种池化结果进行并行融合后,后续卷积能够对显著结构信息与上下文统计信息进行自适应加权,从而在单一尺度特征上同时编码“目标判别性结构”和“背景环境约束”。需要注意的是,目标尺度信息仍主要由骨干网络和颈部网络中的多尺度特征图提供,SPHPF 的作用并非替代多尺度建模,而是在单尺度特征层面引入多感受野上下文补充,以增强多尺度特征在复杂无人机场景下,尤其是小目标检测任务中的鲁棒性。SPHPF 模块的详细结构如图 2 所示。

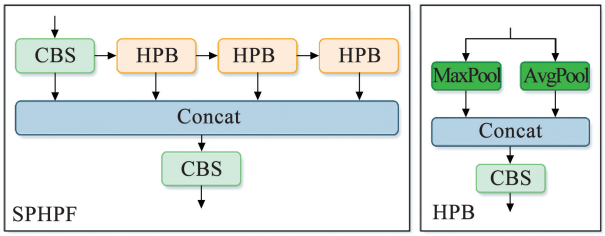


图 2　SPHPF 模块与 HPB 模块的详细架构

Fig. 2　Detailed architecture of SPHPF and HPB modules

SPHPF 模块以 HPB 为核心。对于输入特征 $\boldsymbol{X}\in\mathbf{R}^{B\times C\times H\times W}$,计算公式如下:

$$\boldsymbol{Y}=\mathrm{MaxPool}(\boldsymbol{X})\parallel_{\dim=c}\mathrm{AvgPool}(\boldsymbol{X}) \tag{1}$$

$$\mathrm{HPB}(\boldsymbol{Y})=\mathrm{Conv}(\mathrm{BN}(\mathrm{SiLU}(\boldsymbol{Y}))) \tag{2}$$

式中:$\parallel$ 为拼接操作,MaxPool(·)为最大池化操作,AvgPool(·)为平均池化操作,Conv(·)为 2D 卷积操作,BN(·)为批量归一化操作,SiLU(·)为激活函数。

2.3　C2PLSA 模块

在 C2PLSA 模块中,本文在初始卷积模块后添加了 WTPA 模块,并将 YOLOv11 模型中原始的 PSABlock 替换为 PLSA 模块。具体结构如图 3 所示。在 WTPA 模块中,本文将 WTConv 和 ELA 模块级联,并融入残差连接,这有助于模型扩大感受野。在 PLSA 模块中,基于 PSABlock 模块,本文在注意力模块操作后添加了 LSA 机制。该机制用于捕捉窗口内强相关的局部特征,从而增强特征表达。

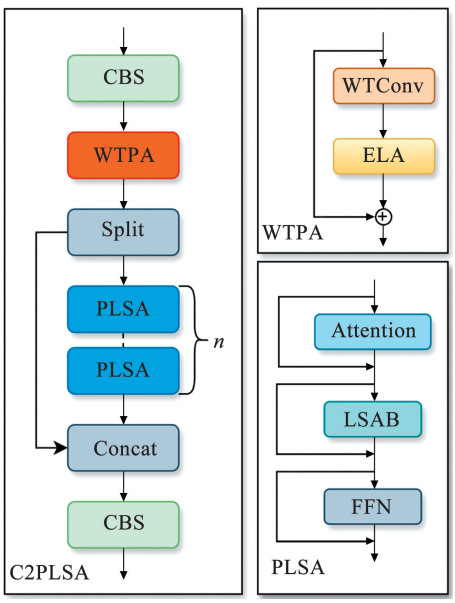


图 3　C2PLSA 模块的详细架构

Fig. 3　Detailed architecture of C2PLSA module

WTPA 模块通过残差连接将 WTConv 与 ELA 串行组合,用于在保持模型轻量化的同时增强特征的多尺度表达能力。WTConv 首先利用小波变换对输入特征进行多尺度频域分解,将特征映射至不同频段的子带空间。其中,低频子带主要包含目标及其周围的大尺度结构和语义上下文信息,高频子带则保留边缘与纹理等局部判别性细节。在子带域内对低分辨率特征施加小核卷积,相当于在不显著增加参数量的前提下实现感受野扩展与跨尺度信息交互。随后通过逆小波变换将处理后的特征重构回空间域,从而为后续检测头提供更加丰富且连续的大尺度场景语义。ELA 模块进一步从空间建模角度对特征进行增强,其通过条带池化与轻量一维卷积分别建模水平方向和垂直方向上的长程依赖关系,并生成对应的空间注意力权重,用以突出与目标相关的响应,抑制无关背景区域。通过将 WTConv 的频域多尺度上下文建模能力与 ELA 的空间方向注意力机制相结合,WTPA 模块能够在大尺度语义感知与局部细节保持之间取得平衡,从方法机理上解释了其在复杂无人机场景下对不同尺度目标,尤其是小目标检测性能的提升以及对背景干扰的有效抑制。以下详细阐述 WTConv 和 ELA 机制。

WTConv[25]是一种基于小波变换的层,旨在无缝替代深度卷积。其核心优势在于在不引入过多参数的情况下扩大感受野。具体而言,WTConv 首先通过级联小波变换(wavelet transform, WT)将输入特征分解为多个频段,然后对每个频段进行小核卷积,最后通过逆小波变换(inverse wavelet transform, IWT)将处理后的特征重建为输出。核心计算表达式为

$$\boldsymbol{Y}=\mathrm{IWT}(\mathrm{Conv}(\boldsymbol{W},\boldsymbol{WT}(\boldsymbol{X}))) \tag{3}$$

式中:$\boldsymbol{X}$ 为输入张量,$\boldsymbol{W}$ 为深度核的权重张量,$\boldsymbol{WT}$ 为用于将 $\boldsymbol{X}$ 分解为频率分量, IWT 为将卷积后的频率特征重建为输出张量 $\boldsymbol{Y}$。

ELA 是一种用于 CNN 的轻量化注意力模块,在不降低通道数的情况下增强空间特征表示,适用于多种视觉任务。ELA 计算包括 3 个步骤。

Step 1　对输入特征图 $\boldsymbol{X}_c$ 应用水平和垂直条带池化,生成位置向量 $z_c^h(h)$ 和 $z_c^w(w)$。该步骤的数学表达式如下:

$$z_c^h(h)=\frac{1}{H}\sum_{i=0}^{H-1}\boldsymbol{X}_c(h,i) \tag{4}$$

$$z_c^w(w)=\frac{1}{W}\sum_{i=0}^{W-1}\boldsymbol{X}_c(w,i) \tag{5}$$

Step 2　对位置向量应用一维卷积 F_h,F_w 和组归一化 $\boldsymbol{G}_n$,然后通过 σ 函数激活生成注意力图 $\boldsymbol{f}^h$、$\boldsymbol{f}^w$ 分别为:

$$\boldsymbol{f}^h=\sigma(\boldsymbol{G}_n(F_h(z_h))) \tag{6}$$

$$\boldsymbol{f}^w=\sigma(\boldsymbol{G}_n(F_w(z_w))) \tag{7}$$

Step 3 通过逐元素乘法将水平注意力权重 $\boldsymbol{f}^h$ 和垂直注意力权重 $\boldsymbol{f}^w$ 与输入特征 $\boldsymbol{X}_c$ 融合，生成增强后的输出 $\boldsymbol{Y}$ 为

$$\boldsymbol{Y}=\boldsymbol{X}_c\odot\boldsymbol{f}^h\odot\boldsymbol{f}^w \tag{8}$$

LSAB 通过局部窗口和 Top-K 选择机制，仅关注每个像素的 K 个最相关邻域特征。这可以有效过滤噪声和无关背景（如天空或地面纹理），同时突出小目标的局部细节（如边缘或纹理）。在无人机航拍图像中，小目标（如远距离无人机、行人等）往往只占据图像或特征图上的一个很小区域，像素数量远少于周围大面积的天空、地面，以及建筑背景（小目标通常仅由十几到几十个像素组成）。与这些背景像素相比，小目标内部像素在特征空间中更为相似，而与背景特征存在明显差异。LSAB 在局部窗口内仅保留与当前像素最相关的 Top-K 邻域特征，使得目标区域在特征聚合时主要与同一目标上的像素交互，而背景像素被赋予较低权重，从而实现对噪声和无关背景的抑制，并进一步突出小目标的边缘和纹理细节。结合标准注意力的全局视角，LSAB 有助于防止小目标在复杂背景中被大尺度特征淹没。具体结构如图 4 所示。

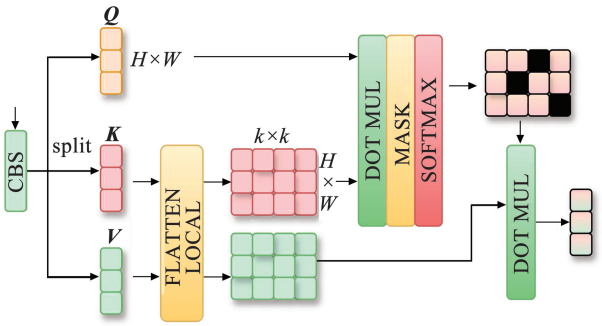


图 4 LSAB 模块的实现细节

Fig. 4 Implementation details of LSAB module

首先，输入 $\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$ 经过卷积模块处理后，被划分为 3 个矩阵，即 $\boldsymbol{Q}$，$\boldsymbol{K}$ 和 $\boldsymbol{V}$，如以下方程所示：

$$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}=\mathrm{Conv}(\mathrm{BN}(\mathrm{SiLU}(\boldsymbol{X}))) \tag{9}$$

式中：$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbf{R}^{H\times W\times\dim}$，dim 为特征通道数。

其次，对于矩阵 $\boldsymbol{K}$ 和 $\boldsymbol{V}$，在 $k\times k$ 窗口内，将其包含的特征展开并扁平化，得到 $\boldsymbol{K}_f$ 和 $\boldsymbol{V}_f$，即

$$\boldsymbol{K}_f=\mathrm{Flatten}(\boldsymbol{K}),\boldsymbol{V}_f=\mathrm{Flatten}(\boldsymbol{V}) \tag{10}$$

式中，$\boldsymbol{K}_f,\boldsymbol{V}_f\in\mathbf{R}^{H\times W\times k\times k\times\dim}$。

然后，计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}_f$ 的点积，得到局部注意力系数矩阵 $\boldsymbol{A}_l$。通过掩码操作抑制相关性低的特征，剩余系数通过 Softmax 函数进行归一化，即

$$\boldsymbol{A}_l=\mathrm{Softmax}(\mathrm{Mask}(\boldsymbol{K}_f\cdot\boldsymbol{Q})) \tag{11}$$

式中，$\boldsymbol{A}_l\in\mathbf{R}^{H\times W\times k\times k}$。

最后，矩阵 $\boldsymbol{A}_l$ 和矩阵 $\boldsymbol{V}_f$ 进行内积运算，得到最终输出结果 $\boldsymbol{Y}$，该结果捕捉了 $\boldsymbol{V}_f$ 中局部特征的加权组合，即

$$\boldsymbol{Y}=\boldsymbol{V}_f\cdot\boldsymbol{A}_l \tag{12}$$

式中，$\boldsymbol{Y}\in\mathbf{R}^{H\times W\times\dim}$。

3 实 验

3.1 数据集和参数设置

微软 COCO（MS COCO[30]）数据集是目标检测领域广泛认可的基准数据集。该数据集包含两个关键子集：Train2017（训练子集）约有 11.8 万张图像，Val2017（验证子集）约有 5 000 张图像。这些子集共涵盖自然场景中 80 个常见目标类别。在本实验中，模型在 Train2017 子集上进行训练，在 Val2017 子集上进行验证，旨在评估其在该基准数据集上的目标检测性能。此外，为了进一步展示 MS COCO 数据集中不同类别的分布情况，本文对 Train2017 子集的 80 个目标类别的实例数量进行了统计，并绘制了如图 5 所示的直方图。

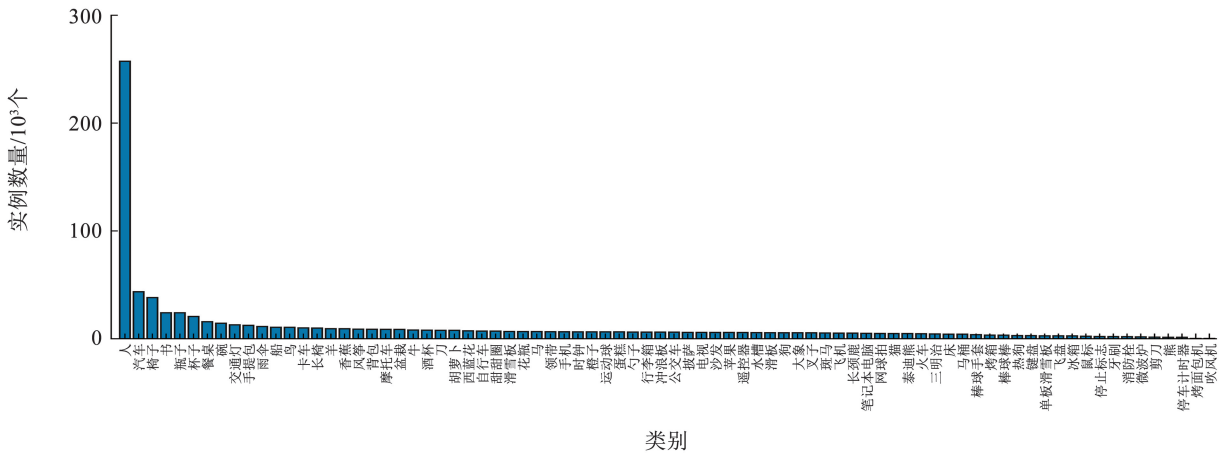


图 5 COCO Train 2017 数据集中各类别目标实例数量统计直方图

Fig. 5 Histogram of instance counts for each category in COCO Train 2017 dataset

VisDrone[31]数据集由无人机视角采集的大量真实场景图像构成,是当前无人机视觉领域广泛采用的权威基准数据集。该数据集涵盖城市道路、居民区、广场、校园等多种复杂环境,具有视角高、尺度变化大、目标密度高、遮挡严重等特点,对目标检测模型的鲁棒性提出了更高要求。VisDrone 检测任务包含 10 个目标类别,主要包括行人、车辆、非机动车等典型城市交通参与者。VisDrone 数据集包含多个子集,其中 VisDrone2019-DET-train(训练集)包含约 6 471张图像,VisDrone2019-DET-val(验证集)包含 548 张图像。所有图像均带有精细标注,包括边界框位置、类别标签,以及遮挡与可见性信息。在本实验中,模型在训练集上进行训练,并在验证集上进行评估,以衡量其在无人机视角下复杂场景中的目标检测性能。

UAVDT 是一个由无人机视角采集的真实交通场景数据集,广泛用于无人机视觉领域中的目标检测研究。该数据集主要覆盖城市道路、高速公路,以及交通路口等场景,具有视角高、目标尺度变化大、目标密集,以及遮挡严重等特点,对目标检测模型的鲁棒性提出了较高要求。在目标检测任务中,UAVDT 主要包含 Car、Truck,以及 Bus 等车辆类别。数据集由无人机航拍视频逐帧抽取得到,并为每一帧提供精确的目标边界框与类别标注。在本实验中,本文对 UAVDT 数据集进行了划分,其中训练集包含 53 676 张图像,测试集包含 24 143 张图像。模型在训练集上进行训练,并在测试集上进行评估,以验证其在无人机航拍复杂交通场景下的目标检测性能。

本实验的系统环境配置如下: Ubuntu 22.04.5 LTS 操作系统、RTX5090 图形处理器(GPU)、32 GB 内存、CUDA12.8、Python3.12.3,深度学习框架采用 PyTorch 2.7.0。在模型训练阶段,实施了特定的超参数和策略:在 COCO 数据集和 UAVDT 数据集训练的时候批次大小设置为 128、在 VisDrone 数据集训练的时候批次大小设置为 64,模型训练 600 个 epoch。初始学习率固定为 0.01,优化器采用 SGD 算法。采用线性衰减调度器,并搭配前 3 个 epoch 的线性预热策略。此外,将训练图像调整为 640 × 640 分辨率,并应用 Mosaic 数据增强方法以提高训练样本的多样性。

3.2　对比实验

表 1 给出了本文所提 HP-LS-YOLO 模型与现有 YOLO 系列模型在 MS COCO 数据集上的定量对比结果。如前所述,所有参与对比的模型均在数据集和参数设置所述的相同实验条件下进行训练。从表 1可以看出,所提 HP-LS-YOLO 模型在所有测试模型中实现了最先进的性能,尤其在小目标检测精度方面具有显著优势。与 YOLOv11n 模型相比,所提方法在 $AP^{val}_{50:95}$ 上提升 0.9%,在 AP^{val}_{50} 上提升 0.9%;需要注意的是,在 AP^{val}_{S} 上实现了 1.4% 的提升。

实验结果与方法的分析一致:所提模型在目标检测精度方面取得了显著提升,其中小目标检测的提升更为明显。接下来,将通过消融实验验证每个所提模块的有效性。

表 1　COCO 数据集上对比实验的定量结果

Tab. 1　Quantitative results of comparative experiments on COCO dataset

模型	参数量/M	$AP^{val}_{50:95}$	AP^{val}_{50}	AP^{Pval}_{75}	AP^{val}_{S}
YOLOv5n[8]	2.6	35.1	50.5	38.4	18.0
YOLOv6n[9]	4.5	36.6	51.6	39.4	17.0
YOLOv8n[11]	3.2	37.0	52.4	40.2	<u>18.5</u>
YOLOv9t[12]	2.1	37.7	52.4	40.9	18.2
YOLOv10n[13]	2.7	<u>38.5</u>	54.0	41.6	18.4
YOLOv11n[14]	2.6	<u>38.5</u>	<u>54.4</u>	<u>41.7</u>	<u>18.5</u>
HP-LS-YOLO	2.8	**39.4**	**55.3**	**42.5**	**19.9**

注:黑体数值表示各评价指标的最优结果,下划线数值表示各评价指标的次优结果。

表 2 给出了本文所提 HP-LS-YOLO 模型与 YOLOv11n 模型在 VisDrone 数据集上的定量对比结果。从表 2 可以看出,在参数量基本相当的前提下,HP-LS-YOLO 模型在 UAV 视角的 VisDrone 验证集上取得了与 YOLOv11n 模型相同的 $AP^{val}_{50:95}$,同时在关键指标 AP^{val}_{50}、AP^{val}_{75} 上分别实现了 0.1、0.2 个百分点的提升。这表明,在复杂的无人机俯视场景中,所提 HP-LS-YOLO 模型在保持模型轻量化的同时,能够进一步提升中高 IoU 阈值下的检测精度,对小尺度目标和遮挡目标具有更好的鲁棒性。

表 2　VisDrone 数据集上对比实验的定量结果

Tab. 2　Quantitative results of comparative experiments on VisDrone dataset

模型	参数量/M	$AP^{val}_{50:95}$	AP^{val}_{50}	AP^{val}_{75}
YOLOv11n[14]	2.6	**21.6**	<u>35.1</u>	<u>22.2</u>
HP-LS-YOLO	2.8	**21.6**	**35.2**	**22.4**

注:黑体数值表示各评价指标的最优结果,下划线数值表示各评价指标的次优结果。

表 3 给出了本文所提 HP-LS-YOLO 模型与 YOLOv11n 模型在 UAVDT 数据集上的定量对比结果。从表 3 可以看出,在参数量基本相当的前提下,HP-LS-YOLO 模型在更具综合性的 $AP^{val}_{50:95}$ 指标上取得了 0.3 个百分点的提升,表明本文方法在不同 IoU 阈值下具有更稳定、全面的检测性能。同时,HP-LS-YOLO 模型在 AP^{val}_{50} 指标上提升了 2.3 个百分点,有效增强了对无人机航拍场景中密集小目标的检测能力。尽管在 AP^{val}_{75} 上略有波动,但整体结果表

明,本文方法在保持模型轻量化的同时,能够在复杂无人机交通场景下显著提升综合检测性能。

表 3 UAVDT 数据集上对比实验的定量结果

Tab. 3 Quantitative results of comparative experiments on UAVDT dataset

模型	参数量/M	$AP^{val}_{50:95}$	AP^{val}_{50}	AP^{val}_{75}
YOLOv11n[14]	2.6	14.3	23.9	**15.3**
HP-LS-YOLO	2.8	**14.6**	**26.2**	14.7

注:黑体数值表示各评价指标的最优结果,下划线数值表示各评价指标的次优结果。

3.3 消融实验

为验证所提 C2PLSA 和 SPHPF 模块的有效性,进行了消融实验,定量结果汇总见表 4。仅融入 C2PLSA 模块时,与基线方法相比,$AP^{val}_{50:95}$ 提升 0.3%,AP^{val}_{50} 提升 0.3%,AP^{val}_{75} 提升 0.2%,AP^{val}_{S} 提升 0.5%;仅添加 SPHPF 模块时,相对于基线方法,$AP^{val}_{50:95}$ 提升 0.6%,AP^{val}_{50} 提升 0.5%,AP^{val}_{75} 提升 0.7%,AP^{val}_{S} 提升 0.5%;当同时采用两个模块时,$AP^{val}_{50:95}$、AP^{val}_{50}、AP^{val}_{75}、AP^{val}_{S} 分别提升 0.9%、0.9%、0.8%、1.4%。这些消融实验证实,SPHPF 和 C2PLSA 模块均有助于提升模型的目标检测精度,与空间金字塔混合池化快速策略和 C2PLSA 模块的分析一致。

表 4 不同模块对本文提出的 HP-LS-YOLO 模型的影响

Tab. 4 Impact of different modules on proposed HP-LS-YOLO model in this paper

C2PLSA	SPHPF	$AP^{val}_{50:95}$	AP^{val}_{50}	AP^{val}_{75}	AP^{val}_{S}
		38.5	54.4	41.7	18.5
√		38.8	54.7	41.9	19.0
	√	39.1	54.9	42.4	19.0
√	√	**39.4**	**55.3**	**42.5**	**19.9**

注:√表示选择该模块,黑体数值表示各评价指标的最优结果,下划线数值表示各评价指标的次优结果。

为验证 SPHPF 模块中 HPB 所采用的各种混合策略的有效性,进行了专门实验,定量结果汇总见表 5。结果表明,采用不同的混合策略均有利于提升模型的特征提取能力和目标识别精度。在这些策略中,最大池化和平均池化的拼接被证明在帮助模型捕捉目标特征方面最为有效。

表 5 不同混合特征方法对 HP-LS-YOLO 模型的影响

Tab. 5 Impact of different mixed feature methods on HP-LS-YOLO model

HPB	$AP^{val}_{50:95}$	AP^{val}_{50}	AP^{val}_{75}	AP^{val}_{S}
MaxPool(YOLOv11n)	38.5	54.4	41.7	18.5
Full ‖ MaxPool	39.0	54.9	42.1	19.5
MaxPool + AvgPool	**39.4**	**55.5**	**42.6**	19.4
MaxPool ‖ AvgPool(Ours)	**39.4**	55.3	42.5	**19.9**

注:黑体数值表示各评价指标的最优结果,下划线数值表示各评价指标的次优结果。

为验证 LSAB 模块中的超参数 K 对模型性能的影响,进行了消融实验(图 6)。在模块窗口大小固定为 5×5 的情况下,将 K 设置为 3、6、9、12 和 15,并使用 $AP^{val}_{50:95}$、AP^{val}_{50}、AP^{val}_{S} 评估性能。模型在 $K=9$ 时取得最佳性能,因为该值平衡了关键特征(尤其是小目标细节)的有效保留和特征图冗余信息的高效过滤,避免了对后续处理的干扰。

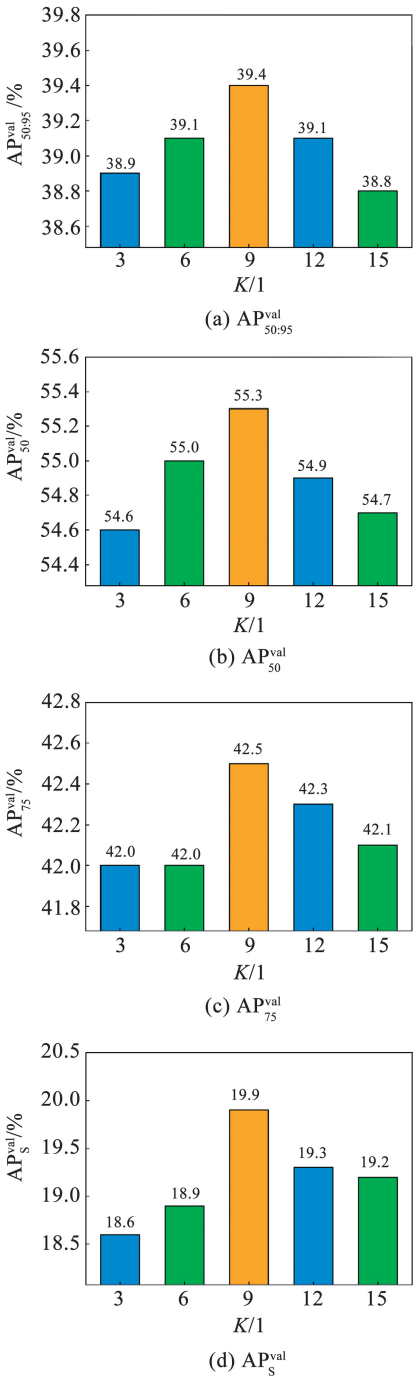


图 6 不同 K 下 LSAB 模块的性能对比(窗口大小为 5×5)

Fig. 6 Performance comparison of LSAB module under different K (window size 5×5)

当 $K=6$ 时,过于严格的筛选导致关键特征(如小目标边缘/纹理)丢失,降低了识别精度。当 $K=3$ 时,更严格的筛选剔除了更多有效特征(包括中目

标细节和小目标核心特征),严重损害了多尺度识别能力,导致整体性能下降。

当 $K=12$ 时,宽松的筛选未能抑制背景噪声(如地面纹理、天空杂波),掩盖了小目标的弱特征,降低了精度。当 $K=15$ 时,过于宽松的筛选引入了比 $K=12$ 更多的冗余信息:这进一步掩盖了小目标特征,并占用了大目标识别所需的资源,导致性能较 $K=9$ 时有显著下降。

3.4　可视化结果

图 7 为 YOLOv11 模型与 HP-LS-YOLO 模型在 COCOVal2017 上的两张图片的检测结果对比图。在上方的街市场景中,由于目标密集、尺度差异显著且存在遮挡,YOLOv11 模型对小尺度鸟类的检出率较低,仅识别到少量实例。相比之下,HP-LS-YOLO 模型能够更全面地检测到图中的鸟类目标,并在人员与牛类等较大目标上保持稳定表现,体现出更强的特征提取与小目标识别能力。在下方的海边场景中,鸟类目标尺寸更小、对比度更低,YOLOv11 模型仅捕获少数较为清晰的目标,而 HP-LS-YOLO 模型在远距离小目标检测上表现更优,能识别到更多飞鸟实例,同时保持对行人目标的稳定检测。总体而言,HP-LS-YOLO 模型在复杂背景与小目标场景下展现出显著更高的召回率和检测鲁棒性。

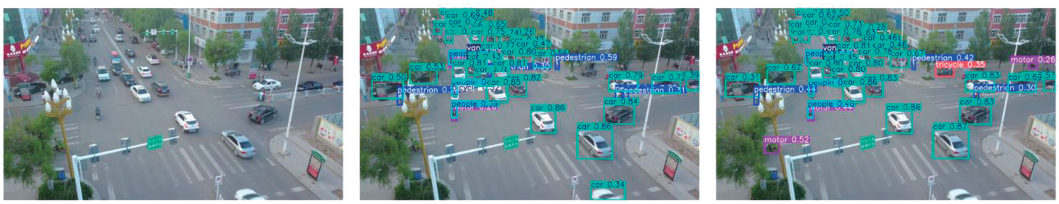


(a) 原始图像　　(b) YOLOv11模型检测结果　　(c) HP-LS-YOLO模型检测结果

图 7　YOLOv11 模型与 HP-LS-YOLO 模型在 COCOVal2017 上的两张图片的检测结果对比图

Fig. 7　Comparison diagram of detection results of two images on COCOVal2017 using YOLOv11 model and HP-LS-YOLO model

图 8 展示了一个典型的城市道路交叉路口场景(该图像来自 VisDrone 数据集的验证集)。在该城市路口场景中,车辆、行人,以及摩托车等多类交通参与者分布在多个方向,场景结构复杂。相比 YOLOv11 模型,HP-LS-YOLO 模型在小目标和远距离目标的检测上表现更优,能够识别更多的摩托车与行人实例,并在车辆类别上保持稳定检测。此外,HP-LS-YOLO 模型的目标置信度整体更高,使其在密集交通场景下展现出更好的鲁棒性与检测精度。

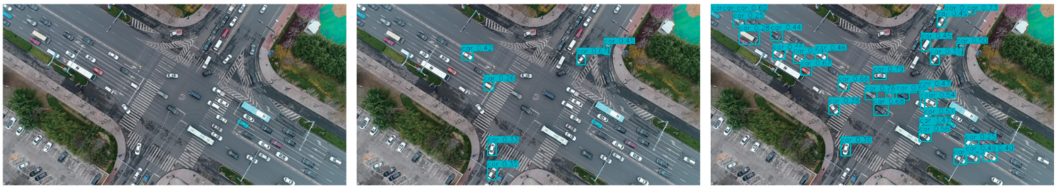


(a) 原始图像　　(b) YOLOv11模型检测结果　　(c) HP-LS-YOLO模型检测结果

图 8　典型的城市道路交叉路口场景

Fig. 8　Typical urban road intersection scene

如图 9 所示展示了一个典型的无人机航拍城市道路交叉路口场景(该图像来自 UAVDT 数据集的测试集)。该场景中车辆数量众多,目标尺度普遍较小,且存在一定程度的遮挡与密集分布,整体检测难度较高。相比 YOLOv11 模型,HP-LS-YOLO 模型在该场景下能够检测到更多的小尺度车辆目标,显著减少了漏检现象,尤其在道路远端和车辆密集区域表现更为明显。同时,HP-LS-YOLO 模型对车辆目标的置信度整体更高,检测结果更加完整稳定,表明本文方法在 UAVDT 数据集中复杂无人机航拍交通场景下具备更强的小目标感知能力和鲁棒性。

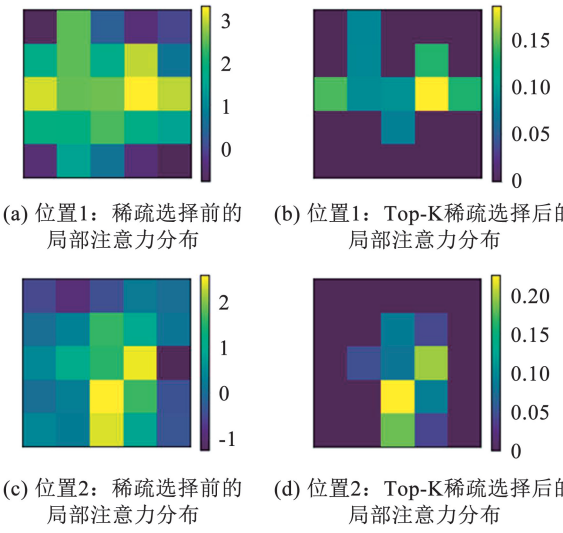


(a) 原始图像 (b) YOLOv11模型检测结果 (c) HP-LS-YOLO模型检测结果

图 9 UAVDT 数据集测试集中的典型无人机航拍交通场景

Fig. 9 Typical unmanned aerial vehicle aerial traffic scenes in test set of UAVDT dataset

为进一步验证 C2PLSA 模块中局部注意力机制的抑制效果,图 10 展示了两个典型空间位置的局部注意力分布。从图 10 可以看出,在未进行稀疏选择时,局部窗口内的相似度呈现较为分散的响应模式,背景纹理与目标区域同时获得较高响应,导致特征聚合过程中容易引入无关信息。



(a) 位置1：稀疏选择前的局部注意力分布 (b) 位置1：Top-K稀疏选择后的局部注意力分布

(c) 位置2：稀疏选择前的局部注意力分布 (d) 位置2：Top-K稀疏选择后的局部注意力分布

图 10 LSAB 局部注意力机制的特征抑制可视化结果

Fig. 10 Visualization results of feature suppression in LSAB local attention mechanism

经过 LSAB 的局部 Top-K 选择后,注意力分布显著收缩,仅与目标相关性较高的少量邻域被保留下来,而大部分背景区域的响应被有效抑制。形成的注意力结构在空间上更为集中,主要沿目标边缘及其局部关键位置分布,表明 LSAB 能够优先保留与目标语义一致的结构特征。

这种由分散到集中的注意力变化表明,LSAB 在复杂背景中能够显著削弱背景干扰,并突出小目标有限的边缘与纹理信息,从而为后续检测提供更加清晰、判别性更强的特征输入。这一现象也与前述量化实验中 C2PLSA 模块带来的性能提升相一致。

4 结 论

1)针对无人机航拍图像中目标尺度变化剧烈、背景复杂及小目标占比高等问题,本文提出了一种融合混合池化与局部注意力机制的无人机目标检测 HP-LS-YOLO 模型。该模型在 YOLOv11 模型框架基础上,引入空间金字塔混合池化快速策略(SPHPF)和融合小波卷积与局部注意力的 C2PLSA 模块,通过在单一尺度上引入多感受野上下文信息并结合局部 Top-K 稀疏注意力机制,有效增强了复杂背景下小目标特征的表达能力。

2)在 MS COCO 数据集上的实验结果表明,相比 YOLOv11n 模型,所提 HP-LS-YOLO 模型在 $AP^{val}_{50:95}$、AP^{val}_{50}、AP^{val}_{S} 指标上分别提升 0.9%、0.9%、1.4%,整体检测性能和小目标检测精度均得到提升。消融实验进一步验证了 SPHPF 与 C2PLSA 模块在缓解小目标特征稀疏、抑制复杂背景干扰方面的有效性,其中合理的 Top-K 参数设置在保留关键目标特征与过滤冗余背景信息之间取得了良好平衡。

3)在 VisDrone 和 UAVDT 等更具无人机场景特征的数据集上的实验结果表明,在参数量基本保持不变的前提下,HP-LS-YOLO 模型在中高 IoU 阈值指标和小目标检测性能上均优于基线模型,表现出较好的鲁棒性和泛化能力。上述结果表明,本文所提 HP-LS-YOLO 模型能够在保持模型轻量化的同时,有效提升复杂无人机航拍场景中密集小目标的检测性能,具有一定的工程应用价值。

参考文献

[1] Li Ziran, Zhang Yanwen, Wu Hao, et al. Design and application of a UAV autonomous inspection system for high-voltage power transmission lines[J]. Remote Sensing, 2023, 15(3): 865. DOI: 10.3390/rs15030865

[2] Byun S, Shin I K, Moon J, et al. Road traffic monitoring from UAV images using deep learning networks[J]. Remote Sensing, 2021, 13(20): 4027. DOI: 10.3390/rs13204027

[3] Bouguettaya A, Zarzour H, Kechida A, et al. A survey on deep learning-based identification of plant and crop diseases from UAV-based aerial images[J]. Cluster Computing, 2023, 26(2): 1297. DOI: 10.1007/s10586-022-03627-x

[4] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 779. DOI: 10.1109/CVPR.2016.91

[5] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 6517. DOI: 10.1109/CVPR.

2017. 690

[6] Redmon J, Farhadi A. YOLOv3: an incremental improvement[PP/OL]. V1. arXiv (2018-04-08) [2025-08-16]. https://doi.org/10.48550/arXiv.1804.02767

[7] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[PP/OL]. V1. arXiv (2020-04-23) [2025-08-16]. https://doi.org/10.48550/arXiv.2004.10934

[8] ULTRALYTICS. YOLOv5[EB/OL]. [2025-12-20]. https://github.com/ultralytics/yolov5

[9] Li Chuyi, Li Lulu, Jiang Hongliang, et al. YOLOv6: a single-stage object detection framework for industrial applications[PP/OL]. V1. arXiv (2022-09-07) [2025-08-16]. https://doi.org/10.48550/arXiv.2209.02976

[10] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023: 7464. DOI: 10.1109/CVPR52729.2023.00721

[11] ULTRALYTICS. YOLOv8[EB/OL]. [2025-12-20]. https://github.com/ultralytics/ultralytics

[12] Wang C Y, Yeh I H, Mark Liao H Y. YOLOv9: learning what you want to learn using programmable gradient information [C]// Computer Vision-ECCV 2024. Cham: Springer, 2025: 1. DOI: 10.1007/978-3-031-72751-1_1

[13] Wang Ao, Chen Hui, Liu Lihao, et al. YOLOv10: real-time end-to-end object detection[PP/OL]. V2. arXiv (2024-10-30) [2025-08-16]. https://doi.org/10.48550/arXiv.2405.14458

[14] ULTRALYTICS. YOLOv11[EB/OL]. [2025-12-20]. https://github.com/ultralytics/ultralytics

[15] Tian Yunjie, Ye Qixiang, Doermann D. YOLOv12: attention-centric real-time object detectors[PP/OL]. V1. arXiv (2025-02-18) [2025-08-16]. https://doi.org/10.48550/arXiv.2502.12524

[16] Lei Mengqi, Li Siqi, Wu Yihong, et al. YOLOv13: real-time object detection with hypergraph-enhanced adaptive visual perception [PP/OL]. V1. arXiv (2025-06-21) [2025-08-16]. https://doi.org/10.48550/arXiv.2506.17733

[17] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 580. DOI: 10.1109/CVPR.2014.81

[18] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015: 1440. DOI: 10.1109/ICCV.2015.169

[19] Ren Shaoqing, He Kaiming, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137. 10.1109/TPAMI.2016.2577031

[20] Zhang Zhengxin. Drone-YOLO: an efficient neural network method for target detection in drone images[J]. Drones, 2023, 7(8): 526. DOI: 10.3390/drones7080526

[21] Lv Yanhui, Tian Bo, Guo Qichao, et al. A lightweight small target detection algorithm for UAV platforms [J]. Applied Sciences, 2025, 15(1): 12. DOI: 10.3390/app15010012

[22] Wang Gang, Chen Yanfei, An Pei, et al. UAV-YOLOv8: a small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios[J]. Sensors, 2023, 23(16): 7190. DOI: 10.3390/s23167190

[23] Zhang Xin, Liu Chen, Yang Degang, et al. RFAConv: receptive-field attention convolution for improving convolutional neural networks[PP/OL]. V6. arXiv (2024-03-28) [2025-08-05]. https://doi.org/10.48550/arXiv.2304.03198

[24] Chen Linwei, Gu Lin, Zheng Dezhi, et al. Frequency-adaptive dilated convolution for semantic segmentation[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024: 3414. DOI: 10.1109/CVPR52733.2024.00328

[25] Finder S E, Amoyal R, Treister E, et al. Wavelet convolutions for large receptive fields[C]//Computer Vision-ECCV 2024. Cham: Springer, 2025: 363. DOI:10.1007/978-3-031-72949-2_21

[26] Xu Wei, Wan Yi. ELA: efficient local attention for deep convolutional neural networks[PP/OL]. V1. arXiv (2024-03-02) [2025-08-05]. https://doi.org/10.48550/arXiv.2403.01123

[27] Si Yunzhong, Xu Huiying, Zhu Xinzhong, et al. SCSA: exploring the synergistic effects between spatial and channel attention[J]. Neurocomputing, 2025, 634: 129866. DOI: 10.1016/j.neucom.2025.129866

[28] Liu Yuxi, Yang Wenhan, Bai Huihui, et al. Region-adaptive transform with segmentation prior for image compression [C]// Computer Vision-ECCV 2024. Cham: Springer, 2025: 181. DOI: 10.1007/978-3-031-72952-2_11

[29] Zhou Meng, Xie Jiadong, Xu Mingsheng. Dynamic attention mechanism in spatiotemporal memory networks for object tracking [PP/OL]. V1. arXiv (2025-03-21) [2025-08-05]. https://doi.org/10.48550/arXiv.2503.16768

[30] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[C]//Computer Vision-ECCV 2014. Cham: Springer, 2014: 740. DOI: 10.1007/978-3-319-10602-1_48

[31] Du Dawei, Zhu Pengfei, Wen Longyin, et al. VisDrone-DET2019: the vision meets drone object detection in image challenge results [C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul: IEEE. 2019: 213. DOI:10.1109/iccvw.2019.00030

(编辑　张　红)