http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2024.0548

多核上下文特征引导下的无人机航拍图像可信检测算法

张志豪，杜丽霞*，郝紫微，侯越
(兰州交通大学 电子与信息工程学院，兰州 730000)

摘 要：针对复杂背景下无人机（UAV）航拍图像目标尺度分布不均匀导致检测精度低的问题，提出一种用于无人机小目标检测的多核上下文特征可信检测算法（MCCF）。在主干网络部分设计多核上下文特征引导模块，该模块利用多尺度卷积核和全局上下文注意力以实现局部细节和全局结构的特征信息融合，从而有效抑制背景噪声；提出可学习位置编码器，该编码器通过动态位置编码自适应调整目标特征的位置表示，降低了采样过程中小目标的空间位置信息损失；为有效识别复杂区域的微小目标，提出多尺度细节融合模块，融合局部空间位置信息和全局结构语义信息，以有效捕捉不同尺度目标的细粒化特征。在 VisDrone2019-DET 数据集和 Drone 数据集的实验结果表明：相较于 RT-DETR 模型，MCCF 模型的 mAP0.5 分别提升了 2.1% 和 6.8%，有效改善了无人机航拍图像目标漏检和误检的问题。

关 键 词：无人机小目标检测；全局上下文注意力；多核上下文特征；编码器；特征融合
中图分类号：TP391.41
文献标志码：A **文章编号**：1001-5965(2026)11-4051-12

近年来，以卷积神经网络(convolutional neural network, CNN)为基础的检测技术表现出优异的性能，基于卷积神经网络的无人机目标检测已被广泛应用于各行各业，如军事打击[1]、交通管制[2]、行人检测[3]及交通监视检测等[4]。然而，由于无人机飞行高度不定、拍摄位置和角度具有特殊性[5]，致使航拍图像存在背景复杂、目标尺度分布不均匀等问题[6-7]，从而影响检测算法的准确性。因此，需要设计一种适用于无人机航拍图像的目标检测模型，以满足实际应用场景需求。

基于卷积神经网络的检测器分为2类：单阶段检测器[8]和双阶段检测器[9]。首先，双阶段检测器通过候选区域定位，其次，细化目标分类以进行准确检测。双阶段检测算法主要包括基于区域的卷积神经网络(region-based convolutional neural networks, RCNN)[10]、Faster-RCNN[11]和 Mask-RCNN 系列[12]，但上述方法由于生成大量边界框导致检测效率降低，难以有效满足无人机目标检测实时性的要求。相较于双阶段检测算法，单阶段检测算法具有较高的实时性，能有效满足无人机小目标检测的需求，主要包括YOLO 系列[13-15]、SSD 系列[16]、Efficient 模型[17]和 Refinedet 模型[18]等。其主要通过物体特征直接对目标类别进行识别和分类，使检测模型无需选择候选区域直接完成端到端的目标检测，从而提升模型的检测效率。目前，已有研究人员将单阶段目标检测算法应用于无人机航拍图像领域。Wang 等[19]基于 YOLOv5 检测框架提出了无人机图像目标检测器 SPB-YOLO，通过条形瓶颈模块和路径聚合网络提高了单阶段检测器在无人机密集目标检测任务中的准确性。薛珊等[20]提出无人机图像检测方

收稿日期：2024-07-18；**录用日期**：2024-09-30；**网络出版时间**：2024-11-01 13:31
网络出版地址：link.cnki.net/urlid/11.2625.V.20241101.1122.001
基金项目：国家自然科学基金 (62063014,62363020)；甘肃省自然科学基金 (22JR5RA365)
* **通信作者**. E-mail：dlx1228@mail.lzjtu.cn

引用格式：张志豪，杜丽霞，郝紫微，等. 多核上下文特征引导下的无人机航拍图像可信检测算法 [J]. 北京航空航天大学学报，2026，52（11）：4051-4062. Zhang Z H，Du L X，Hao Z W，et al. Multi-core contextual feature-guided algorithm for trusted detection of UAV aerial images [J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52（11）：4051-4062 （in Chinese）.

法YOLO-Ads,通过引入协调注意力模块以有效加强检测网络对空间和通道的关注度,提升检测模型的准确性和鲁棒性。Liu等[21]基于YOLOv3检测框架提出UAV-YOLO,通过增强浅层卷积以丰富空间信息,但对于小目标存在漏检和误检的问题。刘婷婷等[22]提出一种基于YOLOv3框架的无人机检测网络,将轻量化主干网络和场景上下文信息相融合以提升检测精度,但会遗漏目标关键特征信息。尽管单阶段检测算法在检测速度和检测精度上取得了一定的提升,但仍存在尺度不均匀目标检测精度低的问题。

鉴于无人机航拍图像中目标尺度分布不均匀,远距离目标特征难以有效提取的问题[23],深度学习建模理论中的扩大感受野[24]和自注意力机制[25],为实时、准确地提取目标全局上下文信息,提供了更为有效的解决方案。Yu等[26]利用扩张卷积以增加感受野范围,从而有效提取目标全局结构特征。Dai等[27]设计可变形卷积,通过卷积偏移学习重新定位空间采样位置,丰富空间特征信息,从而有效提取目标特征。Hu等[28]设计SENet,利用全局池化操作和注意力聚合全局上下文信息以进一步提高检测精度。Li等[29]设计LSKNet,利用大卷积核扩大感受野以有效提取目标全局上下文信息。Dosovitskiy等[30]设计ViT网络,通过分层窗口化捕获每个窗口内的长距离信息,并应用级联转换器提取不同窗口内的目标特征向量,但ViT无法有效提取局部细节信息。Peng等[31]设计Conformer混合网络结构,通过融合卷积运算和自注意机制以增强目标全局-局部信息表示,从而有效提升目标检测的准确性,但小目标信息交互时仍会遗漏细节信息。综上所述,扩大感受野通常会引入相当大的背景噪声,而自注意力则存在忽略目标局部特征表示的风险[32-33]。

针对上述问题,本文提出一种用于无人机航拍图像检测的多核上下文特征可信检测算法(multi-core contextual feature trusted detection network, MCCF)。该算法由多核上下文特征引导模块、可学习位置编码器和多尺度细节融合模块组成。多核上下文特征引导模块主要包括多核交互模块和全局上下文注意力。其中,多核交互模块主要捕获目标局部细节信息,而全局上下文注意力关注目标的全局结构特征,融合目标局部细节信息和全局结构特征以实现背景上下文关键特征信息的提取。在此基础上,提出基于可学习位置编码的编码器,该编码器通过可学习的位置编码方式以有效提取目标全局位置信息,进而增强模型的特征表达能力,同时缓解了小目标空间信息丢失问题。最后,针对区域内尺度不均目标检测精度低的问题,本文设计了多尺度细节融合模块,通过全局上下文引导模块与局部提取模块结合的方式,增强目标局部细节信息和全局结构信息间的交互融合能力,从而提升小目标检测的准确性。此外,为验证模型在极端天气条件下的鲁棒性,本研究引入了雾霾、风雪、阴雨极端天气的干扰测试,实验证明多核上下文特征可信检测算法在极端天气条件下的有效性和稳健性性。为解释改进思路在模型决策中的过程,本文对模型的全局感受野、整体模型对小目标的关注进行了解释性分析,以突出本文模型改进点对决策结果的影响。本文提出如下创新方法:

1)设计多核上下文特征引导模块,主要通过并行排列多个无膨胀率的多尺度卷积核在不同的感受野中提取密集的目标细节特征,其细节特征沿着通道维度自适应融合,从而聚合局部上下文信息,并行设计全局上下文注意力,利用全局平均池化和一维带状卷积以捕获远处目标特征间关系并增强局部区域内的特征,从而提高无人机航拍图像检测的准确性。

2)设计可学习位置编码器,通过动态位置编码学习到目标的最优空间位置信息,以弥补检测网络在特征提取过程中的信息损失。

3)设计多尺度细节融合模块,利用全局池化层和深度卷积核以融合局部位置信息和全局语义信息,从而更为精细地识别无人机航拍图像复杂区域的微小目标。

4)设计模型鲁棒性增强实验,通过引入极端天气扰动因素,以验证模型应对极端天气事件的稳健性。同时,为提升检测模型的可信性,对模型关键层特征分布、整体模型对小目标关注程度均进行特征可视化操作,以进一步突出改进点对决策结果的影响。

1 本文算法

本文提出基于RT-DETR模型的多核上下文特征可信检测算法,其网络结构如图1所示。首先,在主干网络中设计多核上下文特征交互模块(multi core context feature guidance module, MCF),利用多个无膨胀率的多尺度卷积核在不同感受野中收集目标的局部细节信息,并行设计全局上下文注意力机制以捕获全局结构信息。其次,设计可学习深度位置编码器(learnable position encoder, LPE),利用动态的学习位置编码自适应调整目标特征的空间位置信息,以有效弥补区域内目标特征丢失问题。最后,设计多尺度细节融合模块(multi scale detail

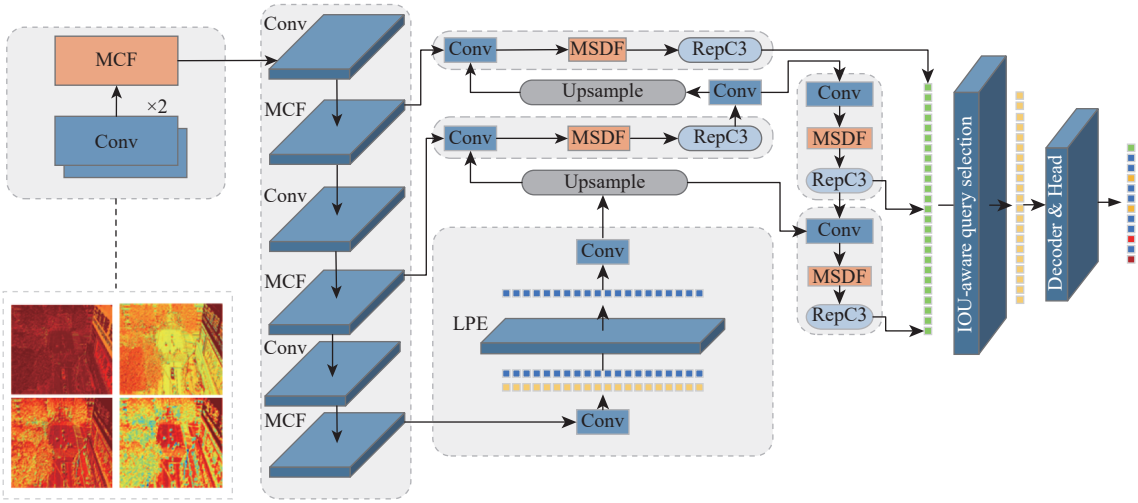


图 1　多核上下文特征可信检测算法网络整体架构

Fig. 1　Overall architecture of multi-core contextual feature trusted detection network

fusion module, MSDF), 通过融合目标局部细节和全局结构特征的方式实现对不同尺度目标的特征信息提取。

1.1　多核上下文特征引导网络

与常规目标检测不同, 无人机航拍图像目标检测常面临背景复杂和目标尺度不均匀的问题, 致使目标全局-局部交互特征难以有效表示, 进而导致目标关键特征无法有效提取。鉴于此, 本文提出多核上下文特征引导模块以有效筛选目标全局和局部关键特征信息, 其整体结构如图 2 所示。多核上下文特征引导模块由多核交互(multi core interaction, MCI)模块和全局上下文注意力(global contextual attention, GCA)组成。其中, 多核交互模块主要利用 1×1 深度卷积核来捕获图像细节信息, 然后利用一组并行设计的多尺度卷积核来捕获局部上下文信息。

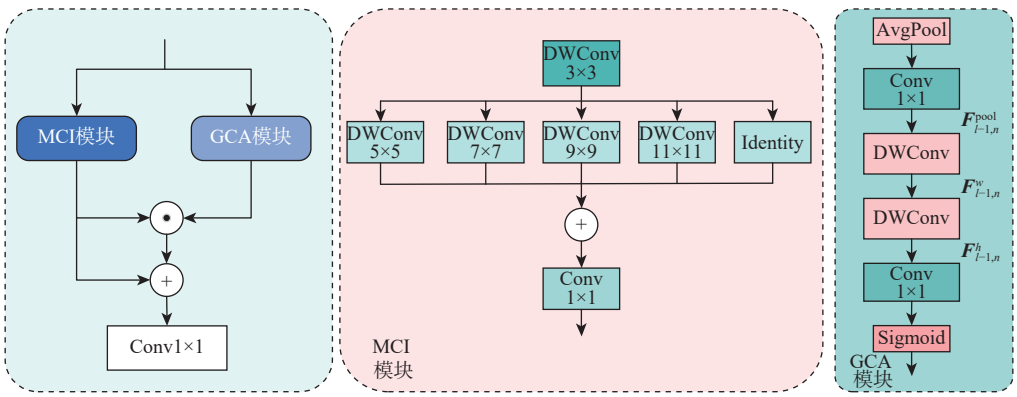


图 2　多核上下文特征引导模块

Fig. 2　Multicore contextual feature bootstrap module

多核交互模块第 1 阶段的第 n 个多核块内的卷积核具体表示如下:

$$\boldsymbol{L}_{l-1,n}=\mathrm{Conv}_{k_s\times k_s}(\boldsymbol{X}) \tag{1}$$

$$\boldsymbol{Z}_{l-1,n}^{(m)}=\mathrm{DWConv}_{k^{(m)}\times k^{(m)}}(\boldsymbol{L}_{l-1,n}) \qquad m=1,2,3,4 \tag{2}$$

式中: $\mathrm{Conv}_{k_s\times k_s}$ 表示 $k_s\times k_s$ 卷积神经网络, $k_s=3$ 表示 3×3 卷积; X 表示输入图像; $\boldsymbol{L}_{l-1,n}\in\mathbf{R}^{1/2C_l\times H_l\times W_l}$ 表示 $k_s\times k_s$ 卷积提取的局部特征; $\boldsymbol{Z}_{l-1,n}^{(m)}\in\mathbf{R}^{1/2C_l\times H_l\times W_l}$ 表示第 m 个 $k^{(m)}\times k^{(m)}$ 深度卷积; $\mathrm{DWConv}_{k^{(m)}\times k^{(m)}}$ 表示卷积核大小为 $k^{(m)}\times k^{(m)}$ 的逐通道深度卷积操作, $k^{(m)}=(m+1)\times 2+1$ 表示多尺度卷积核。

通过大小为 1×1 的卷积以融合局部特征和上下文特征, 表征各个通道之间的相互关系的具体操作如下:

$$\boldsymbol{P}_{l-1,n}=\mathrm{Conv}_{1\times 1}\left(\boldsymbol{L}_{l-1,n}+\sum_{m=1}^{4}\boldsymbol{Z}_{l-1,n}^{(m)}\right) \tag{3}$$

式中: $\boldsymbol{P}_{l-1,n}\in\mathbf{R}^{1/2C_l\times H_l\times W_l}$; $\mathrm{Conv}_{1\times 1}$ 卷积作为通道融合机制以集成具有不同感受野大小的特征。

多核交互模块利用多个无膨胀率的卷积核在有效捕捉不同尺度的特征空间信息的同时, 避免了过于稀疏的特征表示提取。最后将获取的尺度不均匀目标特征信息进行卷积融合, 使融合后的特征

空间具备丰富的局部上下文信息，从而提升无人机航拍图像在复杂背景下的检测精度。

为更好地提取无人机航拍图像小目标的全局上下文信息，弱化复杂背景信息的干扰。本文提出全局上下文注意力来捕获远距离目标之间的关系，并将全局上下文注意力集成到多核交互模块中以实现目标全局-局部信息之间的交互。该注意力利用全局平均池化和一维带状卷积来捕获远程目标之间的上下文相互依赖性并增强全局结构特征。其结构如图2所示。以第1阶段第 n 个多核上下文交互模块中的全局上下文注意力为例，本文采用平均池化和1×1卷积来获得全局目标特征：

$$\boldsymbol{F}_{l-1,n}^{\text{pool}} = \text{Conv}_{1\times1}(P_{\text{avg}}(\boldsymbol{X})) \tag{4}$$

式中：$\boldsymbol{F}_{l-1,n}^{\text{pool}}$ 为全局细节特征；$\boldsymbol{P}_{\text{avg}}$ 表示平均池化操作；$\boldsymbol{X}$ 表示输入图像。

$$\boldsymbol{F}_{l-1,n}^{w} = \text{DWConv}_{1\times k_b}(\boldsymbol{F}_{l-1,n}^{\text{pool}}) \tag{5}$$

$$\boldsymbol{F}_{l-1,n}^{h} = \text{DWConv}_{k_b\times1}(\boldsymbol{F}_{l-1,n}^{w}) \tag{6}$$

$$\boldsymbol{A}_{l-1,n} = \text{Sigmoid}(\text{Conv}_{1\times1}(\boldsymbol{F}_{l-1,n}^{h})) \tag{7}$$

$$\boldsymbol{F}_{l-1,\text{n}}^{\text{attn}} = (\boldsymbol{A}_{l-1,n}\odot\boldsymbol{P}_{l-1,n})\oplus\boldsymbol{P}_{l-1,n} \tag{8}$$

式中：$\boldsymbol{F}_{l-1,n}^{w}$ 和 $\boldsymbol{F}_{l-1,n}^{h}$ 表示不同深度方向的卷积输出；w 表示宽度；h 表示高度；$\text{DWConv}_{1\times k_b}$ 表示具有选择性深度方向的一维带状卷积；$\boldsymbol{A}_{l-1,n}$ 表示注意力图的输出；Sigmoid激活函数保证注意力图在(0,1)的范围内；$\boldsymbol{F}_{l-1,\text{n}}^{\text{attn}}$ 表示多核上下文特征引导模块的特征输出；⊙表示逐元素乘法；⊕表示逐元素求和。

1.2 可学习位置编码器

在多核上下文特征交互模块利用多尺度卷积核和全局上下文注意力进行全局-局部目标特征信息交互时，由于图像尺度变化，易导致远处目标的空间位置信息丢失，从而影响无人机航拍图像的检测性能。鉴于此，本文设计了一种基于可学习位置编码的编码器以获取目标的空间位置信息和特征信息。可学习位置编码器结构如图3所示，图3在编码器架构下将固有位置编码替换为可学习位置编码。该编码器能够在目标全局-局部特征信息的指导下，通过嵌入层(Embedding)学习每个输入图像的位置表示，使得编码器可以在训练过程中更新图像特征的位置矩阵，以实现每个目标位置信息的最优表示。然后，将无人机航拍图像采样的位置表示馈送至多头自注意力关注以捕捉不同位置表示的目标关键特征，有助于识别和定位图像中的多个目标。由此，可学习位置编码器通过可学习性的位置编码自适应地选择目标空间位置，并动态调整目标空间位置的数量和分布，从而有效解决了远处目标尺寸变化导致的空间位置信息丢失问题。

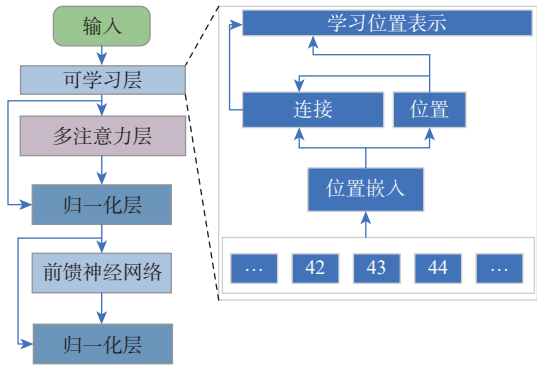


图3 可学习位置编码器架构

Fig. 3 Learning position encoder architecture

在训练过程中，可学习嵌入矩阵 $\boldsymbol{E}$ 的具体表达式可以表示为

$$\boldsymbol{E} = \alpha\cdot\nabla_{\boldsymbol{E}}\boldsymbol{L} \tag{9}$$

式中：α 为学习率，控制了每次更新位置的大小；$\nabla_{\boldsymbol{E}}$ 为损失函数；$\boldsymbol{L}$ 为关于嵌入矩阵 $\boldsymbol{E}$ 的梯度。

1.3 多尺度细节融合模块

针对复杂区域尺度不均匀航拍目标难以有效检测的问题，本文提出一种多尺度细节融合模块以实现局部位置信息与全局语义特征信息的融合，如图4所示。该模块由全局上下文引导模块和局部提取模块组成。其中，全局上下文引导模块通过全局平均池化层和全连接层捕获全局结构信息，使小目标全局结构特征更为显著。继而采用局部提取模块对目标全局上下文信息进行深度卷积操作，获得目标局部细节信息，进而增强目标局部细节信息和全局结构信息间的交互融合能力，从而效捕捉不均匀尺度目标的细粒化特征。为减少空间使用率，降低计算复杂度，本文将全局上下文注意力和局部提取模块串联操作，在降低多尺度细节融合模块复杂度的同时，避免遗漏了不同特征间的细节信息。

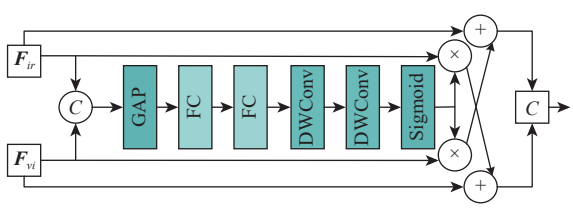


图4 多尺度细节融合模块架构

Fig. 4 Architecture diagram of the multi-scale detail fusion module

多尺度细节融合模块中的深度卷积和全局上下文引导模块对输入图像的增强特征表示如下：

$$\tilde{\boldsymbol{F}}_{ir}^{i}=\boldsymbol{F}_{ir}^{i}\oplus(\boldsymbol{F}_{vi}^{i}\otimes\delta(\mathrm{DWConv}^{n}(\boldsymbol{A}_{\mathrm{tten}ir})))\tag{10}$$

$$\boldsymbol{A}_{\mathrm{tten}ir}=\mathrm{GAP}(C(\boldsymbol{F}_{ir}^{i},\boldsymbol{F}_{vi}^{i}))\tag{11}$$

$$\tilde{\boldsymbol{F}}_{vi}^{i}=\boldsymbol{F}_{vi}^{i}\oplus(\boldsymbol{F}_{vi}^{i}\otimes\delta(\mathrm{DWConv}^{n}(\boldsymbol{A}_{\mathrm{tten}_{vi}})))\tag{12}$$

$$\boldsymbol{A}_{\mathrm{tten}_{vi}}=\mathrm{GAP}(C(\boldsymbol{F}_{ir}^{i},\boldsymbol{F}_{vi}^{i}))\tag{13}$$

式中: $\tilde{\boldsymbol{F}}_{ir}^{i}$和$\tilde{\boldsymbol{F}}_{vi}^{i}$分别表示局部空间位置信息和全局结构语义信息的输出; C表示连接 2 个位置的信息; $\boldsymbol{F}_{ir}^{i}$和$\boldsymbol{F}_{vi}^{i}$分别表示浅层位置信息和深层语义特征; $\oplus$表示逐元素求; $\otimes$表示逐元素乘法; DWConv^{n}表示n个深度卷积层; $(\cdot)$表示通道维度上的串联操作; $\delta(\cdot)$和$\mathrm{GAP}(\cdot)$分别表示Sigmoid函数和全局平均池化。

2　实　验

2.1　实验数据集

在无人机航拍中, 小目标占用的面积不到总面积的 1%～5%, 小目标的大小通常在 10～100 像素。同时, 随着无人机飞行高低和摄像头角度之间距离的变化, 目标的大小也会发生变化, 导致目标变小, 容易被忽视。因此, 为了验证本文模型在复杂环境中检测小目标的准确性, 本文在 2 个不同的数据集上进行了检测实验。

本文实验所采用数据集为 VisDrone2019-DET 无人机图像目标检测公开数据集和某市无人机实际采集的 Drone 数据集, 其详细介绍如下。

1) VisDrone2019-DET: 是一个基于无人机视觉的公开目标检测和跟踪数据集, 由 6 471 张训练图像、548 张验证图像和 3 190 张测试图像组成, 共计 10 209 张静态图像, 每张航拍图像的分布和比例不同, 使得航拍目标的实际尺寸远小于其他背景物体或其他类别的目标, 并存在密集小目标和小目标遮挡的现象。该数据集包含几个场景, 本文选择 10 个预定义的类别: 行人、人、汽车、卡车、公共汽车、卡车、机动车、自行车、遮阳三轮车和三轮车。

2) Drone 数据集: 在某市无人机监控捕获的实际数据集上验证了本文算法, 共包含 8 851 张训练集、2 213 张验证集, 像素维持在 1 920×1 920。每张航拍图像的分布和比例不同, 使得远处车辆目标的实际尺寸远小于其他靠近无人机摄像头的目标, 并且远处存在密集小目标和小目标遮挡的现象。预定类别为汽车、行人、摩托车、面包车、公交车、卡车、三轮车、自行车、摩托车、送货车、施工车。

2.2　评估指标和基准模型

实验平台使用 Ubuntu22.04 操作系统, 以 Pycharm 平台为编码工具, 结合 Pytorch1.8 构建模型, GPU 为 NVIDIA GeForce RTX 3080Ti。为了公平对比算法的性能, 所有算法采用相同的超参数设置进行训练、验证, 初始模型的参数设置为运行尺寸为 6, 运行轮数为 100, 比例尺寸为 0.5, 动量为 0.9, 权重衰减为 0.005。为验证本文算法的有效性和先进性, 本文使用召回率本文算 R、准确率 P、mAP0.5、mAP0.5∶0.95 和 mAP 的度量作为模型性能评估指标。其中, 平均精度均值(mean average precision, mAP)表示对所有类别的 AP 值求平均值, mAP0.5 为 IOU=0.5 时的 mAP, mAP0.5∶0.95 表示设置 10 个 IOU 阈值, 从 0.5~0.95, 步长为 0.05 时的 mAP。AP 和 mAP 的计算式如下:

$$P=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}}\tag{14}$$

$$R=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}}\tag{15}$$

$$A_{\mathrm{P}}=\int_{0}^{1}P(R)\,\mathrm{d}(R)\tag{16}$$

$$m_{\mathrm{AP}}=\frac{\sum_{i=1}^{N}A_{\mathrm{P}_i}}{N}\tag{17}$$

式中: N为物体类别个数; N_{FP}为假正样本, 即误检; N_{FN}为假负样本, 即漏报; N_{TP}为真正样本, 即正确检测。为了在无人机航拍图像检测任务中验证本文算法, 本文比较了几种最先进的检测框架, 即 Faster-RCNN、RetinaNet、CenterNet、SSD、ATSS-FPN-DyHead、TOOD、DINO、YOLOX-Tiny、GFL、RTMDet、YOLOv3、YOLOv4、YOLOv5、YOLOv7 和 RT-DETR。

在实际生活中, 由于单阶段检测框架实时性表现较好, 因此, 在 Drone 数据集中, 将本文与单阶段检测算法进行比较, 以 YOLOv3、YOLOv5、YOLOv6、YOLOv8、YOLOv8-Decoder 及 RT-DETR 为基准线。

2.3　消融实验

为验证本文多核上下文特征交互模块、可学习位置编码器和多尺度细节融合模块的有效性, 选用 RT-DETR 模型为消融实验的基准模型, 依次加入各个模块进行消融实验, 所有消融实验均在 VisDrone2019-DET 数据集上进行。首先, 在基准模型 RT-DETR 中加入多核上下文特征交互模块, 并将其命名为 G-MCF, 在此基础上, 加入可学习位置编码器, 将其命名为 G-LPE, 最后, 加入多尺度细节融合模块, 将其命名为 G-MCDF, 由此进行消融实验分析, 具体结果如表 1 所示。

表 1 给出了不同改进模型的微小目标识别性能。由表 1 可知, 基础 RT-DETR 模型的 R、P、mAP0.5、

表1 VisDrone2019-DET数据集中不同改进模块的客观指标对比

Table 1 Comparison of objective metrics for different improvement modules in VisDrone2019-DET dataset %

模型	*R*	*P*	mAP0.5	mAP0.5∶0.95
RT-DETR	40.8	56.5	42.5	25.7
G-MCF	41.0	57.6	43.0	26.0
G-LPE	41.5	58.0	43.6	26.4
G-MCDF	43.3	58.2	44.6	26.9

mAP0.5∶0.95分别为40.8%、56.5%、42.5%、25.7%,其目标识别精度不高。G-MCF模型的*R*、*P*、mAP0.5、mAP0.5∶0.95评价指标分别为41.0%、57.6%、43.0%和26.0%,相较于RT-DETR模型,其性能分别提升了0.2%、1.1%、0.5%和0.3%,这表明多核上下文特征交互模块能够有效提取目标的局部细节和全局结构,从而减少背景噪声的影响。G-LPE模型在G-MCF模型的基础上,增加了可学习位置编码器,其*R*、*P*、mAP0.5、mAP0.5∶0.95评价指标分别是41.5%、58.0%、43.6%和26.4%,相较于G-MCF模型,识别精度进一步提升,其指标分别提升了0.5%、0.4%、0.6%和0.4%,证明可学习位置编码器有效解决了目标特征信息丢失问题。G-MCDF模型在G-LPE模型基础上,增加了多尺度细节融合模块,其*R*、*P*、mAP0.5、mAP0.5∶0.95评价指标分别是43.3%、58.2%、44.6%和26.9%,其特征提取和融合能力进一步提升,其性能分别提升了1.8%、0.2%、1.0%和0.5%,意味着多尺度细节融合模块可以很好地处理尺度分布不均匀目标特征提取和融合问题。

2.4 类别错误消融实验

传统消融实验主要选用mAP作为模型性能评估指标。然而,mAP通常不透明,无法具体指明模型改进的细节,如性能的具体提升点或优化的错误类型。因此,需要更细致的分析方法来解释模型的改进,在本文消融实验中,实现了准确的类别错误检测,具体错误类型包括分类错误(Cls)、定位错误(Loc)、重复检测框错误(Dupe)、背景误差(Bkg)、漏检(Miss)、错误预测为负样本的数量(FN),以及错误预测为正样本的数量(FP)。这种细致的错误分析方法不仅揭示了模型在哪些方面需要改进,而且展示了改进模块在提高无人机巡检图像分析准确性方面的具体贡献。

从表2可以看出,RT-DETR模型的Cls、Loc、Dupe、Bkg、Miss、FP和FN分别是20.98、3.82、0.21、2.12、13.31、10.92和37.54,其模型检测性能较差。首先,加入多核上下文特征交互模块(G-MCF)后,G-MCF模型的Loc、Dupe、Bkg、Miss和FP分别是3.67、0.19、2.09、13.24和10.86,相较于RT-DETR模型,其筛选目标全局-局部特征信息的性能表现较优,其Loc、Dupe、Bkg、FP分别降低了0.15、0.02、0.03、0.07和0.06,表明本文设计的多核上下文特征引导模块有效捕捉到目标全局-局部关键信息,使模型聚焦于目标全局上下文区域。其次,添加可学习位置编码器后,G-LPE的Cls、Bkg和FP分别是20.45、2.07和10.44,相较于G-MCF模型,G-LPE的Cls、Bkg和FP分别降低了0.72、0.02和0.42,由此看出,可通过弥补复杂区域内丢失目标特征信息的方式,有效提升检测模型的准确性。最后,添加多尺度细节融合模块,G-MCDF的Dupe、Miss和FN分别是0.19、13.05、37.72,相较于G-LPE模型,其Dupe、Miss和FN分别降低了0.02、0.59和1.20,表明上下文特征融合模块有效提取了全局目标结构信息。综上所述,在无人机航拍图像检测领域,本文算法明显优于RT-DETR模型。

表2 VisDrone2019数据集中不同模型的性能比较

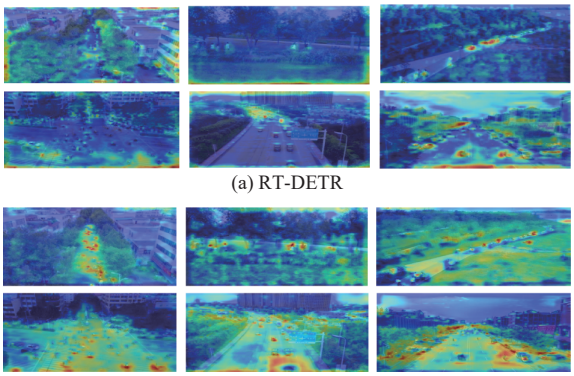
Table 2 Performance comparison of different models in the VisDrone2019 dataset

模型	Cls	Loc	Dupe	Bkg	Miss	FP	FN
RT-DETR	20.98	3.82	0.21	2.12	13.31	10.92	37.54
G-MCF	21.17	3.67	0.19	2.09	13.24	10.86	37.66
G-LPE	20.45	3.84	0.21	2.07	13.64	10.44	38.92
G-MCDF	20.98	3.92	0.19	2.24	13.05	10.90	37.72

2.5 可解释性实验

为提升改进模型的可信性,本节对本文算法进行可视化解释,如图5所示。其中,蓝色的区域为网络模型着重关注区域,颜色越深代表关注度越高。通过这种方法,可以更好地理解模型如何识别和处理图像中的不同特征,从而提高对改进模型的信任度和接受度。

由图5可知,本文算法的红色关注特征区域相较于RT-DETR更加广泛和准确,这意味着本文算法能够关注到全局上下文信息,并利用目标局部细节信息和全局结构之间的交互性来突出目标显著特征。而无人机航拍图像中的目标尺度分布不均



(a) RT-DETR

(b) 本文算法

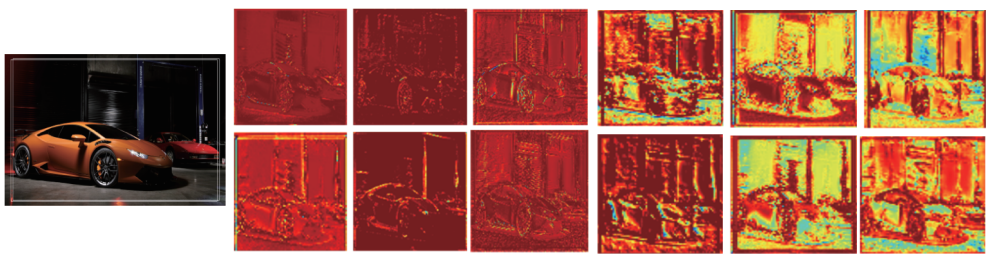
图 5　本文算法的模型决策关注可视化过程

Fig. 5　Model decision making in the proposed algorithm focuses on visualisation processes

匀, RT-DETR 模型的特征关注点较为有限, 无法有效识别远距离目标特征, 尤其是在树木遮挡区域, RT-DETR 模型无法对遮挡目标细节特征进行有效关注, 易导致小目标漏检和误检的问题。本文算法的优势不仅体现在其对目标局部细节和全局上下文结构的理解上, 还在于其能够有效利用动态位置编码以实现对目标最优位置的特征表示, 并利用细节融合模块有效识别遮挡区域和密集区域的小目标, 对于在复杂场景下的无人机航拍图像检测至关重要。

2.6　可视化实验

在目标检测领域, 理解模型如何做出决策至关重要。因此, 本文通过分析 CNN、Transformer 和本文算法的特征层变化, 揭示模型可解释性的决策过程。图 6 展示了这 3 种模型在处理图像时的差异: 基于 CNN 的 YOLOv5 检测模型能够识别具有显著特征的局部区域, 如图 6(a)、图 6(d)中的车辆局部细节特征, 但其忽略了反射在镜子中的红色车头, 远距离目标特征难以有效表示, 表明基于 CNN 的单阶段检测器在提取目标全局结构特征信息时存在一定的局限性。相比之下, 基于 Transformer 的 RT-DETR 模型虽然能够捕捉到图像全局结构的基本轮廓, 但在局部细节上表现不佳, 如图 6(b)、图 6(e)中, 如反射在镜子中的红色车头特征难以有效表征, 意味着基于 Transformer 的 RT-DETR 模型擅长处理提取远距离目标特征, 但缺乏对局部信息细节的捕捉。而本文算法通过全局和局部上下文信息的交互, 成功地提取了完整的局部对象特征和全局结构关系, 如图 6(c)、图 6(f)黄色车辆和红色车头的特征。该方法分析不仅展示了模型的决策过程, 而且强调了全局部上下文信息对于多核上下文特征网络的重要性, 由此进一步表明本文算法可实现较高精度检测。



(a) 原图　(b) YOLOv5 (浅层)　(c) RT-DETR (浅层)　(d) 本文算法 (浅层)　(e) YOLOv5 (深层)　(f) RT-DETR (深层)　(g) 本文算法 (深层)

图 6　CNN(YOLOv5)、Transformer(RT-DETR) 和 MCCF 的特征图比较

Fig. 6　Comparison of feature maps of CNN (YOLOv5), Transformer (RT-DETR), and MCCF

2.7　鲁棒性实验

近年来, 极端天气的频繁出现对无人机交通巡检的准确性构成了挑战。为了探索无人机航拍图像检测网络在复杂天气条件下的性能, 本文通过调整 VisDrone2019-DET 数据集中图像的室外环境, 模拟极端天气情况来验证本文网络的鲁棒性。在本文算法中设计了数据增强方法, 增加阴雨天气、雾霾天气、风雪天气作为干扰因素, 测试模型在极端天气中的适用性。不同极端天气中的部分航空目标检测结果如表 3 所示, 表中, 加粗数值为不同模型时对比下的最优数值。本文算法对于鲁棒性的评价指标为 mAP0.5。

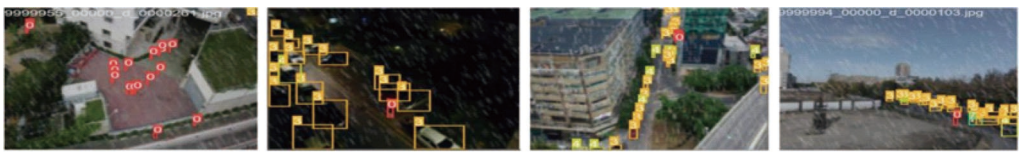
恶劣天气下的鲁棒性实验主要是验证多核上下文特征引导下的无人机航拍图像可信检测算法在极端天气条件下的鲁棒性能, 并确保其在复杂环境中仍然保持出色的检测能力。图 7 显示了阴雨天气、雾霾天气、风雪天气的可视化。

在阴雨天气的模拟场景中, RT-DETR 模型、YOLOv5-Decoder 模型和 YOLOv8-Deocder 模型的 mAP0.5 分别为 32.6%、21.5% 和 20.9%, 而本文算法在 640 像素×640 像素输入尺寸下的性能在所有模型中最为优越, mAP0.5 为 33.6%, 相较于 RT-DETR 模型、YOLOv5-Decoder 模型和 YOLOv8-Deocder 模型的 mAP0.5 分别提升了 1%、12.1%、12.7%。即使调整各个模型的输入尺寸, 本文算法仍然优于大多数其他模型。这表明, 本文算法通过利用自适应跨

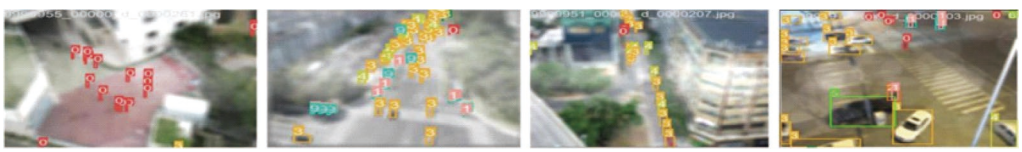
表3 VisDrone2019-DET数据集进行极端天气鲁棒性实验

Table 3 Conducting extreme weather robustness experiments on the VisDrone2019-DET dataset

算法	输入尺寸/(像素×像素)	mAP0.5/%		
		阴雨天气	雾霾天气	风雪天气
YOLOv5-Decoder	640×640	21.5	16.4	18.9
YOLOv5-Decoder	608×608	20.0	16.2	20.0
YOLOv8-Decoder	640×640	20.9	20.3	23.5
YOLOv8-Decoder	608×608	19.6	19.1	22.3
RT-DETR	640×640	32.6	30.3	34.1
RT-DETR	704×704	35.3	33.4	36.7
本文算法	**640×640**	**33.6**	**31.9**	**37.8**



(a) 阴雨天气



(b) 雾霾天气



(c) 风雪天气

图7 极端天气的可视化

Fig. 7 Visualization of extreme weather conditions

层注意力交互模块,有效地提升了目标特征地感受野并增强了检测模型对于目标关键特征的提取,而RT-DETR本身的自注意力机制,善于捕捉图像的全局结构信息,在阴雨天气无法有效识别目标的局部细节,特别是使局部特征密集区域。YOLOv5-Decoder模型和YOLOv8-Deocder模型擅长捕捉目标的局部细节,但无法有效识别远距离目标的关键特征。在降雨影响视野的情况下,MCCF有效地识别了图像中的目标并提升了检测模型的准确率。

在雾霾天气下,输入尺寸为640像素×640像素的RT-DETR模型、YOLOv5-Decoder模型和YOLOv8-Deocder模型的mAP0.5分别为30.3%、16.4%和20.3%,而输入尺寸为640像素×640像素的MCCF模型的性能在所有模型中最为优越,mAP0.5达到31.9%,相较于RT-DETR模型、YOLOv5-Decoder模型和YOLOv8-Deocder模型的mAP0.5分别提升了1.6%、15.5%、11.6%。即使输入不同尺寸的图像,MCCF模型也仍然优于大多数其他模型。在风雪天气下,输入尺寸为640像素×640像素的RT-DETR、YOLOv5-Decoder模型和YOLOv8-Deocder模型的mAP0.5分别为34.1%、18.9%和23.5%,本文算法在所有输入尺寸为640像素×640像素的图像中准确度最优,mAP0.5达到37.8%,相较于RT-DETR模型、YOLOv5-Decoder模型和YOLOv8-Deocder模型的mAP0.5分别提升了3.7%、18.9%、14.3%。这表明,本文算法能够通过其自适应跨层注意力交互模型和可变形编码器提取机制有效地扩大目标感受野并提取目标的深层语义信息。特别是在低分辨率和遮挡背景下,本文算法的多尺度交叉融合模块可以有效地收集目标的全局-局部交互特征信息和频域信息,以实现准确的目标特征识别和融合,这表明该算法在处理视觉噪声和光照变化方面具有出色的鲁棒性。

2.8 算法总体分析

为了验证本文算法的有效性,在VisDrone2019-DET数据集上将本文算法与目前主流的检测模型对比,即Faster-RCNN、RetinaNet、CenterNet、SSD、ATSS-FPN-DyHead、TOOD、DINO、YOLOX-Tiny、GFL、RTMSet和YOLO系列。如表4所示,本文算

表 4　各算法在 VisDrone2019-DET 数据集上总体性能比较

Table 4　Overall performance comparison of different algorithms on the VisDrone2019-DET dataset

算法	基准网络	mAP/%	mAP0.5/%	参数量
Faster-RCNN	ResNet50	21.5	35.7	41.700×10^6
RetinaNet	ResNet50	16.1	27.3	21.370×10^6
CenterNet	DLA-34	12.4	22.7	25.600×10^6
SSD	VGG-16	8.6	16.6	4.230×10^6
YOLOv3	MobileNet	8.1	17.7	3.750×10^6
YOLOv4	ResNet50	18.6	25.4	6.060×10^6
YOLOv5	CSPDarkNet	19.1	29.8	7.050×10^6
YOLOv7	CSPDarkNet	22.9	34.5	37.260×10^6
ATSS-FPN-DyHead	ResNet50	20.4	33.8	38.910×10^6
TOOD	ResNet50	20.4	33.9	32.400×10^6
DINO	DETR	25.3	44.5	47.560×10^6
YOLOX-Tiny	CSPDarkNet	14.8	27.8	5.035×10^6
GFL	CSPDarkNet	19.3	32.1	32.279×10^6
RTMDet	CSPDarkNet	18.4	31.2	4.876×10^6
RT-DETR	ResNet18	32.6	42.5	21.300×10^6
本文算法	ResNet18	**33.7**	**44.6**	$\mathbf{17.200\times10^6}$

法相较于其他几种经典的通用目标检测算法中具有良好的检测性能，其中，本文算法的 mAP 和 mAP0.5 分别是 33.7% 和 44.6%，相较于 Faster-RCNN 算法和 SSD 算法，mAP 和 mAP0.5 分别提升了 12.2%、8.9% 和 25.1%、28%；且本文算法的参数远低于 Faster-RCNN；相较于 ATSS-FPN-DyHead 和 TOOD，mAP 和 mAP0.5 分别提升了 13.3%、10.8% 和 13.3%、10.7%，而且本文算法的参数量远小于 ATSS-FPN-DyHead 和 TOOD；相较于 DINO、GFL 和 RTMDet，mAP 和 mAP0.5 分别提升了 8.4%、14.4%，15.3%、0.1% 和 12.5%、13.4%；相较于 Retina-Net 算法和 CenterNet 模型，mAP 和 mAP0.5 分别提升了 17.6%、21.3% 和 17.3%、21.9%；相较于 YOLOX-Tiny，mAP 和 mAP0.5 分别提升 18.9% 和 16.8%；与 YOLOv3 模型相比，mAP 提升了 25.6%，mAP0.5 提升了 26.9%；与 YOLOv4 模型相比，mAP 提升了 15.1%，mAP0.5 提升了 19.22%，相较于 RT-DETR 模型，mAP 提升了 1.1%，mAP0.5 提升了 2.1%。与上述模型相比，本文算法依然具有良好的检测能力。

图 8 显示了本文算法在 VisDrone2019-DET 数据集的实际应用场景可视化。从可视化图可以看出，本文算法能够高度准确地检测微小尺度目标，尤其是密集行人特征识别。本文使用 TP、FP 和 FN 作为指标，其中，绿色框（TP）表示正确的检测和真实的标签；红色框（FN）为漏检，未检测到；蓝色框（FP）表示错误检测。在实际场景可视化中，绿色边界框偏多，尤其是在识别远距离目标特征和密集目标特征时，绿色边界框的数量多于蓝色边界框和红色边界框。

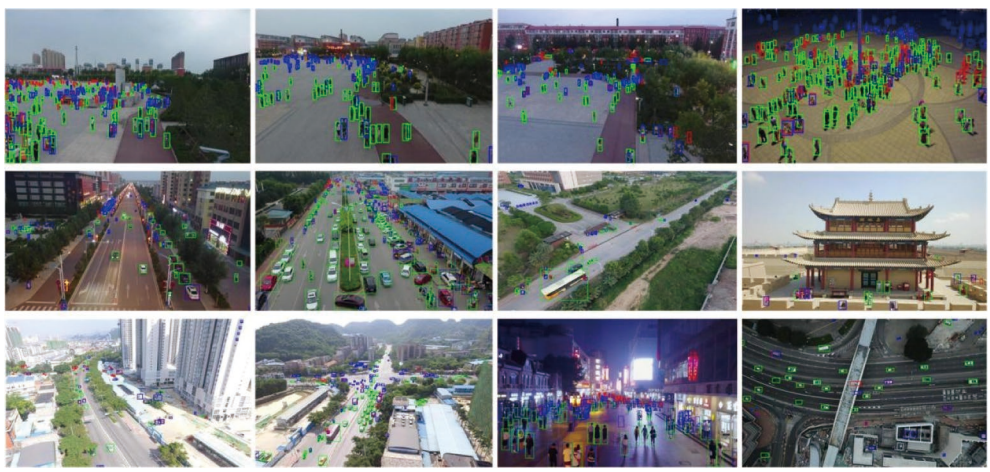


图 8　VisDrone2019-DET 数据集的实际应用场景可视化

Fig. 8　Visualisation of real-world application scenarios for the VisDrone2019-DET dataset

在某市实际无人机航拍图像数据集中测试了本文算法，以评估在实际场景中的泛化能力，并与 YOLOv3、YOLOv5、YOLOv6、YOLOv8、YOLOv8-Decoder 和 RT-DETR 进行对比，实验结果如表 5 和图 9 所示。本文算法的 R、P、mAP0.5、mAP0.5∶0.95 分别是 79.7%、86.0%、82.9% 和 60.2%，而 YOLOV3 的 R、P、mAP0.5、mAP0.5∶0.95 分别是 41.4%、82.6%、52.1%、30.9%，相较于 YOLOv3，本文算法的 R、P 分别提升了 38.3% 和 3.4%，mAP0.5 和 mAP0.5∶0.95 分别提升了 30.8% 和 29.3%；而 YOLOv5 的 R、P、mAP0.5、mAP0.5∶0.95 分别是 48.5%、88.1%、56.7%、38.6%，相较于 YOLOv5，本文算法的 R 提升了 31.2%，mAP0.5 和 mAP0.5∶0.95 分别提高了 26.2% 和 21.6%；YOLOv6 的 R、P、mAP0.5、mAP0.5∶0.95 分别是 40.4%、72.0%、45.2%、31.7%，相较于 YOLOv6 模型，本文算法的 R 和 P 分别提升了 39.3% 和 14%，mAP0.5 和 mAP0.5∶0.95 分别提升了 37.3% 和 28.5%。YOLOv8 的 R、P、mAP0.5、mAP0.5∶0.95

分别是47.4%、61.3%、52.1%、36.7%,相较于YOLOv8模型,本文算法的R和P分别提升了32.3%和24.7%,mAP0.5和mAP0.5∶0.95分别提升了30.8%和23.5%;YOLOv8-Decoder的R、P、mAP0.5、mAP0.5∶0.95分别是49.5%、76.7%、55.1%和37.2%,相较于YOLOv8-Decoder模型,本文算法的R和P分别提升了30.2%和9.3%,mAP0.5和mAP0.5∶0.95分别提升了27.8%和23.0%。RT-DETR的R、P、mAP0.5、mAP0.5∶0.95分别是64.9%、87.9%、76.1%和51.7%,相较于RT-DETR模型,本文算法的R提升了14.8%,mAP0.5和mAP0.5∶0.95分别提升了6.8%和8.5%。在表5中,与YOLO系列相比,多核上下文特征引导下的无人机航拍图像可信检测算法实现了比其他框架更高的mAP。综上所述,本文算法通过利用多核上下文特征引导模块和全局注意力模块,有效地捕捉全局和局部的上下文信息,使模型在处理复杂场景时可以有效实现局部细节与全局结构的特征信息融合。相比之下,YOLO系列模型主要依赖于标准的卷积神经网络结构,局限于局部区域特征,无法有效捕捉长距离依赖和全局特征信息。因此,本文算法能够更好地处理背景复杂、目标分布不均匀的无人机航拍场景。

表5 各算法在Drone数据集上的总体性能比较

Table 5 Overall performance comparison of different algorithms on the Drone dataset

Baseline	R/%	P/%	mAP0.5/%	mAP0.5∶0.95/%
YOLOv3	41.4	82.6	52.1	30.9
YOLOv5	48.5	88.1	56.7	38.6
YOLOv6	40.4	72.0	45.2	31.7
YOLOv8	47.4	61.3	52.1	36.7
YOLOv8-Decoder	49.5	76.7	55.1	37.2
RT-DETR	64.9	87.9	76.1	51.7
本文算法	79.7	86.0	82.9	60.2



图9 Drone数据集检测结果可视化

Fig. 9 Visualisation of Drone dataset detection results

图9显示了本文算法在Drone数据集的实际应用场景可视化。从可视化图可以看出,本文算法可以有效识别微小尺度目标,尤其是远距离目标特征。本文使用TP、FP和FN作为指标,其中,绿色框(TP)表示正确的检测和真实的标签;红色框(FN)为漏检,未检测到;蓝色框(FP)表示错误检测。在实际场景可视化中,近距离目标的绿色边界框偏多,尤其是在识别远距离目标特征时,绿色边界框的数量也多于蓝色边界框。

3 结 论

针对无人机航拍图像存在的背景信息复杂、目标尺度分布不均等问题,本文提出用于无人机小目标检测的多核上下文特征可信检测算法。首先,该算法在ResNet18主干网络部分设计多核上下文特征交互模块,通过全局上下文注意力和多尺度卷积核对全局上下文关键特征进行筛选和提取,从而缓解背景噪声问题。其次,构建可学习位置编码器,通过动态位置编码自适应调整目标空间位置信息,弥补目标丢失特征。最后,为有效识别关注区域内的目标特征,构建多尺度细节融合模块,利用全局上下文引导模块和深度卷积核捕获局部区域的目标全局-局部特征,以有效融合不同尺度的目标特征。

此外,为验证本文算法的鲁棒性,增加极端天气干扰因素,证明了本文算法的鲁棒性更优。实验结果表明,本文算法在不同场景中均表现出良好的检测性能。本文设计的多核上下文特征网络在探索计算成本和检测精度之间的平衡方面取得了很好的结果,但仍然存在一些限制。因此,期待着在

目标检测方面进一步改进本文算法。

参考文献（References）

[1] 姚雨, 宋春林, 邵江琦. 无人机航拍军事车辆实时检测及定位算法[J]. 兵工学报, 2024, 45(S1): 354-360.
Yao Y, Song C L, Shao J Q. Real-time detection and localization algorithm for military vehicles in drone aerial photography[J]. Acta Armamentarii, 2024, 45(S1): 354-360(in Chinese).

[2] 张河山, 谭鑫, 范梦伟, 等. 无人机高空航拍视角下小尺度车辆精确检测方法[J]. 交通运输系统工程与信息, 2024, 24(3): 299-309.
Zhang H S, Tan X, Fan M W, et al. Accurate detection method of small-scale vehicles from perspective of unmanned aerial vehicle high-altitude aerial photography[J]. Journal of Transportation Systems Engineering and Information Technology, 2024, 24(3): 299-309 (in Chinese).

[3] 张嘉辉, 赵威, 王子琛, 等. 基于检测和重识别的无人机行人跟踪算法[J]. 北京航空航天大学学报, 2024, 50(8): 2538-2546.
Zhang J H, Zhao W, Wang Z C, et al. UAV pedestrian tracking algorithm based on detection and re-identification[J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(8): 2538-2546(in Chinese).

[4] 冒国韬, 邓天民, 于楠晶. 基于多尺度分割注意力的无人机航拍图像目标检测算法[J]. 航空学报, 2023, 44(5): 268-278.
Mao G T, Deng T M, Yu N J. Object detection in UAV images based on multi-scale split attention[J]. Acta Aeronautica et Astronautica Sinica, 2023, 44(5): 268-278(in Chinese).

[5] 王殿伟, 胡里晨, 房杰, 等. 基于改进 Double-Head RCNN 的无人机航拍图像小目标检测算法[J]. 北京航空航天大学学报, 2024, 50(7): 2141-2149.
Wang D W, Hu L C, Fang J, et al. Small target detection algorithm based on improved Double-Head RCNN for UAV aerial images[J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(7): 2141-2149(in Chinese).

[6] 李超, 黄新宇, 王凯. 基于特征融合和自学习锚框的高分辨率图像小目标检测算法[J]. 电子学报, 2022, 50(7): 1684-1695.
Li C, Huang X Y, Wang K. Small object detection of high-resolution images based on feature fusion and learnable anchor[J]. Acta Electronica Sinica, 2022, 50(7): 1684-1695(in Chinese).

[7] 刘芳, 韩笑. 基于多尺度深度学习的自适应航拍目标检测[J]. 航空学报, 2022, 43(5): 463-474.
Liu F, Han X. Adaptive aerial object detection based on multi-scale deep learning[J]. Acta Aeronautica et Astronautica Sinica, 2022, 43(5): 463-474(in Chinese).

[8] 张娜, 戚旭磊, 包晓安, 等. 基于优化预测定位的单阶段目标检测算法[J]. 浙江大学学报 (工学版), 2022, 56(4): 783-794.
Zhang N, Qi X L, Bao X A, et al. Single-stage object detection algorithm based on optimizing position prediction[J]. Journal of Zhejiang University (Engineering Science), 2022, 56(4): 783-794(in Chinese).

[9] 曹磊, 王强, 史润佳, 等. 基于改进 RPN 的 Faster-RCNN 网络 SAR 图像车辆目标检测方法[J]. 东南大学学报 (自然科学版), 2021, 51(1): 87-91.
Cao L, Wang Q, Shi R J, et al. Method for vehicle target detection on SAR image based on improved RPN in Faster-RCNN[J]. Journal of Southeast University (Natural Science Edition), 2021, 51(1): 87-91(in Chinese).

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

[11] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of the Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2015, 28: 91-99.

[12] He K M, Gkioxari G, Dollar P, et al. Mask R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2961-2969.

[13] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]// Preceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 779-788.

[14] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]// Preceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6517-6525.

[15] Redmon J, Farhadi A. YOLOv3: an incremental improvement [EB/OL]. (2018-04-08)[2024-07-18]. https://arxiv.org/abs/1804.02767.

[16] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]// Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 21-37.

[17] Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection[C]// Preceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10778-10787.

[18] Zhang S F, Wen L Y, Bian X, et al. Single-shot refinement neural network for object detection[C]// Preceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4203-4212.

[19] Wang X R, Li W H, Guo W, et al. SPB-YOLO: an efficient real-time detector for unmanned aerial vehicle images[C]// Preceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication. Piscataway: IEEE Press, 2021: 99-104.

[20] 薛珊, 卢涛, 吕琼莹, 等. 基于多尺度融合和轻量化网络的无人机目标检测算法[J]. 湖南大学学报 (自然科学版), 2023, 50(8): 82-93.
Xue S, Lu T, Lyu Q Y, et al. Drone target detection algorithm based on multi-scale fusion and lightweight network[J]. Journal of Hunan University (Natural Sciences), 2023, 50(8): 82-93(in Chinese).

[21] Liu M J, Wang X H, Zhou A J, et al. UAV-YOLO: small object detection on unmanned aerial vehicle perspective[J]. Sensors, 2020, 20(8): 2238.

[22] 刘婷婷, 苗华, 李琳, 等. 融合场景上下文的轻量级目标检测网络[J]. 激光与光电子学进展, 2021, 58(20): 2010005.
Liu T T, Miao H, Li L, et al. Lightweight target detection network integrating scene context[J]. Laser & Optoelectronics Progress, 2021, 58(20): 2010005(in Chinese).

[23] 韩俊, 袁小平, 王准, 等. 基于 YOLOv5s 的无人机密集小目标检测算法[J]. 浙江大学学报 (工学版), 2023, 57(6): 1224-1233.
Han J, Yuan X P, Wang Z, et al. UAV dense small target detection

algorithm based on YOLOv5s[J]. Journal of Zhejiang University (Engineering Science), 2023, 57(6): 1224-1233(in Chinese).

[24] Yu F, Koltun V, Funkhouser T. Dilated residual networks[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 636-644.

[25] Aswani A, Shazeer N, Parmar N, et al. Attention is all you need[EB/OL]. (2023-08-02)[2024-07-18]. https://arxiv.org/abs/1706.03762.

[26] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[EB/OL]. (2016-04-30)[2024-07-18]. https://arxiv.org/abs/1511.07122.

[27] Dai J F, Qi H Z, Xiong Y W, et al. Deformable convolutional networks[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 764-773.

[28] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7132-7141.

[29] Li Y X, Hou Q B, Zheng Z H, et al. Large selective kernel network for remote sensing object detection[C]// Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2024: 16748-16759.

[30] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03)[2024-07-18]. https://arxiv.org/abs/2010.11929.

[31] Peng Z L, Huang W, Gu S Z, et al. Conformer: local features coupling global representations for visual recognition[C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 357-366.

[32] 吴一全, 童康. 基于深度学习的无人机航拍图像小目标检测研究进展[J]. 航空学报, 2025, 46(3): 174-200.
Wu Y Q, Tong K. Research advances on deep learning-based small object detection in UAV aerial images[J]. Acta Aeronautica et Astronautica Sinica, 2025, 46(3): 174-200(in Chinese).

[33] 陈金林, 吴一全, 苑玉彬. 无人机视角下目标检测的YOLO系列算法研究进展[J]. 北京航空航天大学学报, 2026, 52(10): 3509-3534.
Chen J L, Wu Y Q, Yuan Y B. Research progress of YOLO series algorithms for target detection under UAV viewpoint[J]. Journal of Beijing University of Aeronautics and Astronautics, 2026, 52(10): 3509-3534(in Chinese).

Multi-core contextual feature-guided algorithm for trusted detection of UAV aerial images

ZHANG Zhihao，DU Lixia*，HAO Ziwei，HOU Yue

(School of Electronic and Information Engineering，Lanzhou Jiaotong University，Lanzhou 730000，China)

Abstract：A multi-core contextual feature trusted detection network (MCCF) is developed for unmanned aerial vehicle (UAV) tiny target detection in order to address the issue of low detection accuracy caused by uneven target scale distribution in UAV aerial photos under complicated background. Firstly, a multicore contextual feature guidance module is designed in the backbone network part, which utilises multi-scale convolutional kernel and global contextual attention in order to achieve the fusion of feature information of local details and global structure, so as to effectively suppress the background noise. Secondly, in order to minimize the loss of spatial position information of small targets during the sampling process, a learnable position encoder is suggested that dynamically modifies the position representation of target characteristics. Finally, in order to effectively identify tiny targets in complex regions, this paper proposes a multi-scale detail fusion module, which fuses local spatial position information and global structural semantic information in order to effectively capture fine-grained features of targets at different scales. Finally, the experimental results on the VisDrone2019-DET dataset and the Drone dataset show that the mAP0.5 of the MCCF model is improved by 2.1% and 6.8%, respectively, compared to the RT-DETR model, which effectively improves the problems of target leakage and misdetection in UAV aerial images.

Keywords：unmanned aerial vehicle small target detection; global contextual attention; multi core contextual features; encoder; feature fusion

Received：2024-07-18；**Accepted**：2024-09-30；**Published Online**：2024-11-01 13：31

URL：link.cnki.net/urlid/11.2625.V.20241101.1122.001

Foundation items：National Natural Science Foundation of China (62063014,62363020); Natural Science Foundation of Gansu Province (22JR5RA365)

* **Corresponding author**. E-mail：dlx1228@mail.lzjtu.cn