2025 年 第 25 卷 第 13 期
2025, 25(13):05527-08

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2404266

引用格式:赵子琪, 李卫东, 李晓娟. 基于改进实时 Transformer 的航拍图像小目标检测算法[J]. 科学技术与工程, 2025, 25(13): 5527-5534.
Zhao Ziqi, Li Weidong, Li Xiaojuan. Small target detection algorithm in aerial images based on improved RT-DETR[J]. Science Technology and Engineering, 2025, 25(13): 5527-5534.

基于改进实时 Transformer 的航拍图像小目标检测算法

赵子琪[1], 李卫东[1,2*], 李晓娟[1,2]

(1. 河北经贸大学管理科学与信息工程学院, 石家庄 050062; 2. 河北省跨境电商技术创新中心, 石家庄 050062)

摘 要 针对无人机航拍图像中背景复杂、小目标样本多,难以提取有效特征等问题,提出一种改进实时 Transformer(real-time detection Transformer,RT-DETR)的无人机航拍小目标检测算法。首先,在特征融合网络中增加针对微小目标的特征融合结构,利用浅层特征图中丰富的位置信息来增强网络对小目标的检测能力,同时为了防止额外参数的增加,去除主干网络中最后一个残差结构;其次,设计一种多通道特征部分卷积模块(multichannel partial convolution,MCPConv),基于此重新构造了主干网络中的 BasicBlock 结构,命名为 MCP Block,减少通道特征冗余,提升多尺度细节特征的获取能力;引入具有学习能力的位置编码,获取更精确、更具表达能力的位置信息;最后引入归一化加权偏差(normalized weighted deviation,NWD)和平均精度驱动交并比(mean precision-driven IoU,MPDIoU)定位损失函数,降低对位置偏差的敏感性,加快模型收敛速度。实验结果表明,在 VisDrone2019-DET 数据集上,改进后的模型较原始模型参数量降低了 62%,检测精度 mAP50 提升了 3.9%,且 FPS 较改进前提升了 17%,对比其他主流检测模型具有更好的检测效果。

关键词 小目标检测; RT-DETR; 多通道部分卷积; 可学习位置编码

中图法分类号 TP391.41; **文献标志码** A

Small Target Detection Algorithm in Aerial Images Based on Improved RT-DETR

ZHAO Zi-qi[1], LI Wei-dong[1,2*], LI Xiao-juan[1,2]

(1. School of Management Science and Information Engineering, Hebei University of Economics and Business, Shijiazhuang 050062, China; 2. Hebei Cross border E-commerce Technology Innovation Center, Shijiazhuang 050062, China)

[**Abstract**] An algorithm has been proposed to detect small targets in unmanned aerial vehicle(UAV) aerial images. The algorithm is based on an improved real-time detection Transformer (RT-DETR) and aims to address the challenges posed by complex backgrounds and a large number of small target samples. To enhance the feature fusion network, a dedicated feature fusion structure for small targets has been incorporated, utilizing rich location information from the shallow feature map to improve the network's ability to detect small targets. Furthermore, the last residual block in the BackBone has been removed to prevent an increase in additional parameters. Additionally, the MCP Block, a reconstructed BasicBlock structure in the backbone network, has been designed, which includes a multi-channel feature partial convolution module (MCPConv) to reduce redundancy in channel features and enhance the acquisition of multi-scale detail features. Moreover, a location encoding mechanism with learning ability has been introduced to obtain more accurate and expressive location information. The normalized weighted deviation (NWD) and mean precision-driven IoU (MPDIoU) positioning loss functions have been incorporated to accelerate the convergence speed of the model and reduce sensitivity to position deviation. Experimental results on the VisDrone2019-DET dataset demonstrate that the improved model reduces parameters by 62% compared to the original model, increases mAP50 by 3.9%, and improves FPS by 17%. The improved model exhibits superior detection performance compared to other mainstream detection models.

[**Keywords**] small object detection; RT-DETR; multi channel partial convolution; learned position embedding

收稿日期: 2024-06-07 **修订日期**: 2025-01-16

基金项目: 河北省省级科技计划(SZX2020034);河北省高等学校科学研究计划(BJK2022041)

第一作者: 赵子琪(2000—),男,汉族,河北石家庄人,硕士研究生。研究方向:图像处理、目标检测。E-mail:583982413@qq.com。

* **通信作者**: 李卫东(1973—),男,汉族,河北石家庄人,博士,教授。研究方向:机器学习、深度学习。E-mail:14263859@qq.com。

随着计算机视觉和无人机技术的不断发展,结合无人机对航拍图像进行目标检测的方法越来越成熟。然而,无人机航拍图像存在背景复杂度高、目标尺寸小、外观模糊等问题。基于传统方法的目标检测技术对小目标检测精度低,对复杂背景下目标的适应性弱,容易发生漏检和误检。相比之下,基于深度学习的目标检测算法在检测速度和精度方面具有巨大的优势,能够显著提升目标检测的性能。

现阶段目标检测器有两种典型架构:基于卷积神经网络(convolutional neural networks,CNN)和基于 Transformer 的目标检测器。基于 CNN 架构的目标检测器近年来发展迅速,由最初的以区域卷积神经网络(region convolutional neural networks,R-CNN)等[1]双阶段算法发展到以 YOLO(you only look once)系列[2]为代表的单阶段算法。YOLO 系列在兼顾速度的同时仍能保持较高的检测精度,目前已广泛应用于各种工业场景中。

在无人机航拍领域,由于背景复杂,目标尺寸小,难以提取有效特征,会导致航拍场景下目标检测效果不佳,从而引发误检、漏检、目标定位不准确等问题。同时航拍目标检测任务对时间、空间复杂度都有很高的要求。针对这些问题,研究者们提出了很多改进策略。桑雨等[3]通过对预测头添加语义引导,并设计轻量级池化结构,强化小目标位置信息,减少了计算成本。刘晋川等[4]结合注意力机制构建特征挖掘模块,融入深层语义信息丰富小目标特征。吴稳稳等[5]针对小目标检测精度较低的问题,设计多级特征融合并引入注意力机制,对重要特征增强,提高特征图表达能力。康传利等[6]在网络中加入三维注意力机制并采用高斯距离判定误差,优化特征并优化损失函数。通过对模型的不断改进,单阶段目标检测器目前在航拍领域取得了较好的效果,但依然存在着无法解决的问题。

单阶段目标检测是“密集预测”思想的典型代表,单阶段目标检测器会在整个图像上密集地生成锚框[7],并对每个锚框进行目标分类和位置预测,而在众多锚框当中真正含有目标的锚框占比极低,正负样本数量的极大不平衡是单阶段目标检测网络始终面临的问题。

传统的基于卷积神经网络的目标检测离不开锚框设计、阈值筛选、非极大值抑制(non-maximum suppression,NMS)等前后处理步骤,并且其中超参数的设置极大影响模型的性能,这导致当前的目标检测架构难以做到真正意义上的端到端检测。直到 Transformer 应用到目标检测领域后,这一问题有了新的解决方案。Carion 等[8]提出基于 Transfomer 架构的新的端到端目标检测器(detection transformer,DETR),DETR 使用可学习的查询(query)来预测目标的存在,通过二分匹配来分配标签,解决了模型对锚框及非极大值抑制等后处理操作的依赖问题,将原本的“密集检测”变成了“稀疏检测”,实现了真正意义上的端到端检测。尽管有明显的优势,但 DETR 目前依然存在两个主要问题:训练收敛缓慢;查询过程难以优化。针对这两个问题目前也存在一些改进方案,Deformable DETR[9]通过提高注意力机制的效率[10]加速收敛过程,同时引入多尺度特征解决了小目标检测性能不足的问题。Conditional DETR[11]和 Anchor DETR[12]通过提高目标查询的定位能力加速模型收敛过程。DAB-DETR[13]引入 4D 参考点逐层迭代优化预测框。DN-DETR[14]通过引入查询去噪来加速训练收敛。DINO[15]沿用之前工作的思想并加以改进,取得了更好的效果,然而现有 DETR 算法还是无法满足实时性的要求。

在此基础上,Lü 等[16]提出了一个新的基于 Transformer 的实时目标检测算法 RT-DETR,相较于 YOLOv8,RT-DETR 以较少的训练时长在同等测试条件下展现出了更强的性能和更好的平衡,且检测速度不输 YOLOv8。在摆脱单阶段目标检测器对手工组件的依赖问题之后,DETR 系列模型经过不断的更迭,首次达到了实时检测的效果。这项研究为无人机航拍目标检测提供了新的思路。

RT-DETR 算法在 COCO 数据集[17]上取得了良好的结果。但在无人机航拍领域,尤其是小目标检测方面,仍然面临诸多挑战。具体包括图像中背景复杂、小目标样本多、尺度差异大以及难以提取有效特征等问题。此外,现有的损失函数对于不同尺度的空间敏感性差异较大,也会导致误检和漏检的现象。因此,现提出一种基于改进 RT-DETR 的航拍小目标检测方法,旨在增强其在航拍场景中的检测效果。

针对航拍场景下小目标样本多、难以提取有效特征的问题,构建针对小目标的多尺度特征提取与融合模块,显著提高模型对小目标的感知与捕获能力。同时优化特征融合结构,提升特征融合过程中小目标特征信息的比重,以增强小目标特征信息的表达能力。

针对航拍场景下目标尺度差异大的问题,提出一种多通道特征部分卷积模块,并基于此设计轻量化的 MCP Block。该模块通过捕捉来自不同尺度的特征信息,显著提升网络对多尺度特征的提取能力,从而增强模型对不同尺度目标的捕捉效果。与

传统的 Basic Block 相比，MCP Block 在提升特征提取能力的同时，有效减少复杂度和计算冗余。

针对航拍场景下背景复杂导致目标定位不准确的问题，提出使用可学习位置编码替代原始编码，使模型在复杂背景下能够对小目标进行更精准的定位。

针对航拍场景下误检、漏检目标过多的问题，结合归一化加权偏差(normalized weighted deviation，NWD)[18]和平均精度驱动交并比(mean precision-driven IoU，MPDIoU)[19]对损失函数进行重新设计，从而显著降低模型对小目标位置偏移的敏感度，提高模型对目标尺寸变化的鲁棒性，加快模型收敛速度。

1　算法介绍

1.1　整体框架

RT-DETR 是第一个实时端到端目标检测器，不仅在速度和精度上都优于目前最先进的实时检测器，而且不需要后处理，因此其推理速度无延迟，保持稳定。RT-DETR 共有 6 个版本：RT-DETR-R18、RT-DETR-R34、RT-DETR-R50、RT-DETR-R101、RT-DETR-L、RT-DETR-X，6 个版本的区别在于主干网络的选择，其中 RT-DETR-R18 模型网络小，运行速度快，对小目标应用场景的考虑，选择基于该模型进行改进。改进后的网络结构如图 1 所示。

RT-DETR 模型由主干网络、颈部网络以及解码器组成。使用 ResNet18[20]作为主干网络，利用主干的最后 3 个阶段的特征 S3、S4 和 S5 作为编码器的输入。S5 经过尺度内特征交互(attention-based intra-scale feature interaction，AIFI)后和 S3、S4 一起输入跨尺度特征融合模块(cross-scale feature fusion module，CCFM)将多尺度特征转化为图像特征序列。经过 IoU-aware 感知查询选择固定数量的图像特征作为解码器的初始对象查询。最后带有辅助预测头的解码器迭代优化对象查询以生成框和置信度分数。

1.2　多尺度特征提取

RT-DETR 模型中，输入图像经过主干网络后输出的 3 个特征图大小分别为 20×20，40×40 和 80×80，分别用来检测大、中以及小目标。针对航拍图像中小目标较多且分布集中的特点，提出在特征融合部分增加一层针对小目标的检测结构。由于浅层特征相比深层特征而言具有更丰富的位置信息及细节特征信息，更有利于小目标的检测，因此在主干

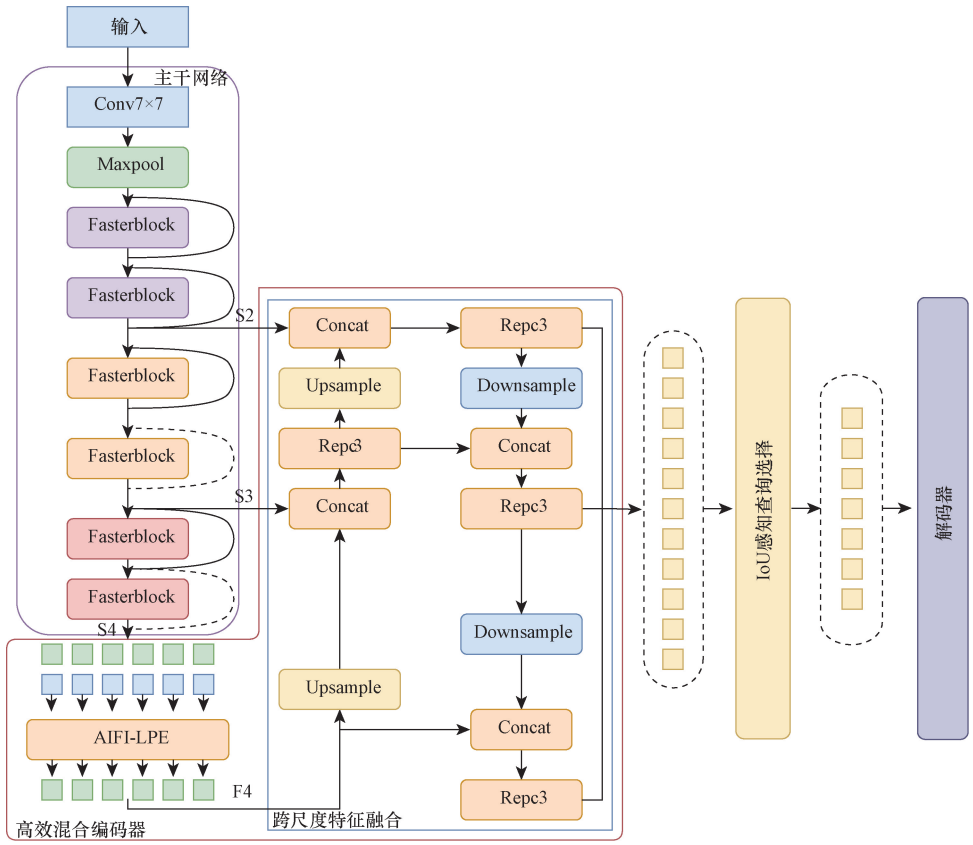
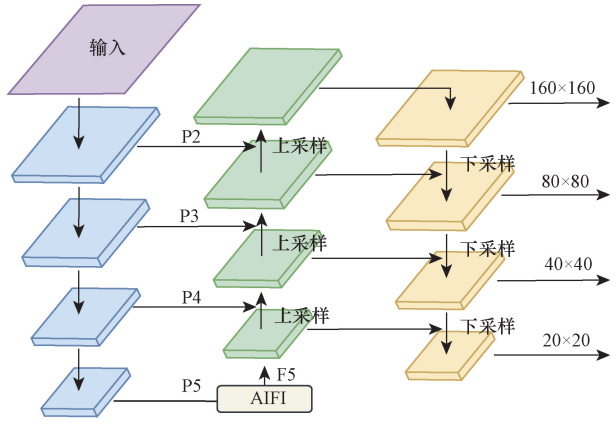


图 1　改进 RT-DETR 算法模型结构

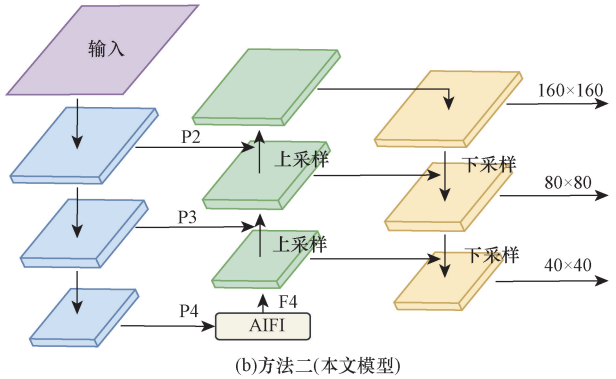
Fig. 1　Improved RT-DETR model structure

网络中提取出 160×160 大小的特征图并将其与更深层次的图像特征进行特征融合,以实现对小目标的精准定位与捕捉(方法一)。然而这种方法虽然在性能上有一定的提升,但网络的计算开销也随之增大,基于此,提出删除 20×20 大目标检测层及主干网络中对应的卷积下采样操作,同时在特征融合模块去除该层相关的特征融合操作,在极大降低参数量和计算量的同时仍能够使模型保持较好的检测效果(方法二)。

模型结构对比如图 2 所示,方法一为仅添加小目标检测层的模型,方法二为提出的模型,在去除 P5 相关层后将 P4 送入 Transformer Encoder 得到 F4,然后通过跨尺度特征融合模块增强特征的表达能力,提高小目标的检测性能和模型的收敛速度。模型效果对比如表 1 所示,可以看出提出的结构在



(a) 方法一



(b)方法二(本文模型)

图 2 模型结构对比图

Fig. 2 Model structure comparison

不影响模型速度的前提下参数量相比原模型结构获得了极大的降低,同时检测精度大幅提高。同时和方法一相比在相对检测精度仅降低 0.2% 的情况下极大减少了模型的参数和计算量,提高了模型检测速度,同时准确率也有了部分提升,实现了更高效的特征获取。

1.3 MCP Block

无人机航拍图像中目标尺寸小、分辨率低,由于主干网络下采样过程中卷积的不断堆叠,在深层特征图中虽然能获得更多的全局语义信息,但无法充分捕捉到小目标的细节特征。同时考虑到模型应用于无人机图像目标检测领域,在实际部署时速度与准确性同样重要。目前来看实时性较强的模型一般采用 GConv(group convlution)[21] 或者 DWConv(depthwise convolution)[22] 的方式实现更高的效率[23],如 MobileNets[24],ShuffleNets[25],GhostNet[26],EffcientNets[27]。DWConv 这种拆分通道并对每个通道分别进行卷积的方式虽然有效减少了 FLOPs(每秒浮点运算次数),却因无法捕捉通道间的关系而导致全局语义的丢失,因此通常以 DWConv + PWConv(pointwise convolution)的组合形式出现,PWConv 通过对生成特征图进行通道间的加权组合,生成新的特征图以更好的捕捉图像中的语义信息和特征关联性。然而在实践中通道数的增加又会导致更高的内存访问,降低整体计算速度。

在主干网络下采样过程中丰富多尺度的特征信息更有利于对不同尺寸目标进行特征提取,但同时会带来参数量和计算量的增加。观察到不同通道的特征图具有高度的相似性,存在大量冗余[23,26,28]。由此提出了一种新的 MCPConv,在减少特征冗余的同时提高模型对不同尺度特征的感知能力。首先对输入特征的 1/4 通道进行线性变换以进行空间特征提取,将该通道分别送入一个 3×3 的卷积核和一个 5×5 的卷积核(为了降低参数量及计算量这里使用两个 3×3 的卷积核替代并通过跳跃连接的方式实现),获取不同尺度的特征信息,融合多尺度的目标特征,经过 ReLU 激活函数,最后将经过特征提取后的通道与未经处理的 3/4 通道进行拼接得到最终输出。MCPConv 结构如图 3 所示。

表 1 模型效果对比

Table 1 Model performance comparison

方法	参数量/10^6	计算量	准确率/%	召回率/%	mAP50/%	mAP50:95/%	FPS/(帧·s^{-1})
RT-DETR	19.9	57.3	55.4	37.6	36.6	21.0	60
方法一	18.6	78.8	55.2	40.3	39.3	23.2	46
方法二	7.9	69.4	56.2	39.9	39.1	23.1	61

注:FFLOPs(giga floating-point operations per second),即每秒 10 亿(10^9)次的浮点运算数;mAP50 表示将交并比(IoU)设为 0.5 时,计算每一个类别下所有图片的平均精度(AP);mAP50:95 表示当 IoU 阈值在 0.5 ~ 0.95 逐步增加时的平均精度;FPS(frame per second)表示每秒钟填充图像的帧数。

假设 h、w 为输入数据的高度和宽度，c 为输入通道数，c_p 为部分输入通道数，k 为卷积核大小，标准卷积和多通道特征部分卷积的 FLOPs 及 Params 计算过程如下。

$$\text{FLOPs}_1 = hwk^2c^2 \tag{1}$$

$$\text{FLOPs}_2 = 2hwk^2c_p^2 \tag{2}$$

$$\text{Params}_1 = k^2c^2 \tag{3}$$

$$\text{Params}_2 = 2k^2c_p^2 \tag{4}$$

c_p 取 c 的 1/4，则 MCPConv 的 FLOPs 为标准卷积的 1/8，其 Params 为普通卷积的 1/8，有效地降低了模型所需的计算资源。

在 RT-DETR 模型中，主干网络的下采样过程是通过 BasicBlock 残差结构完成的，该结构由于卷积的堆叠导致冗余计算余较多，因此基于本文提出的 MCPConv，重新设计了 Basic Block，命名为 MCP Block，其结构如图 4 所示。

与原始的 Basic Block 相比，使用 MCPConv 对常规卷积进行替换，在其后添加两个 1×1 大小的卷积，并且只在中间层使用了 BN 以及激活函数，在保持特征多样性的同时保证较低的延迟。设计的 MCP Block 有效降低了计算量并丰富了下采样过程中的特征信息，实现了更高效的特征获取。

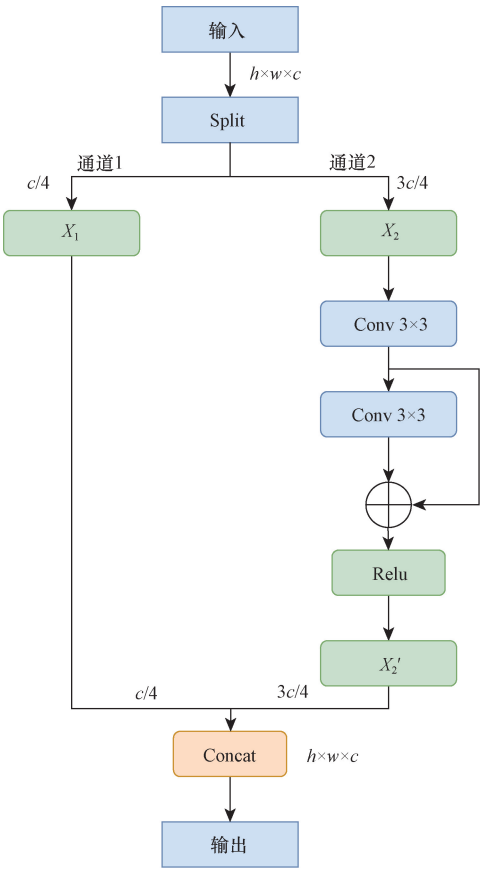


图 3　MCPConv

Fig. 3　Multichannel partial convolution

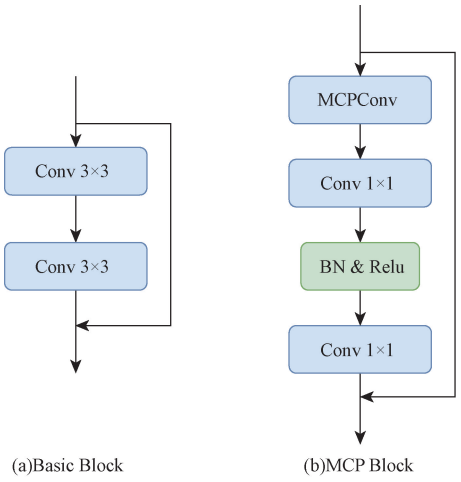


图 4　Basic Block 与 MCP Block 设计对比

Fig. 4　Comparison of Basic Block and MCP Block design

1.4　Learned Position Embedding

目前主流的 Transformer 方法主要依靠不可学习的相对/绝对位置编码描述像素或 patch 序列的先验位置关系。然而在小目标检测任务中目标往往在图像中占据的像素数量较少，因此其定位的准确度对检测结果至关重要。使用可学习位置编码能够使模型更准确地感知图像中目标的绝对位置信息，从而提高对小目标检测的准确性。

使用 1D Position Embedding，随机初始化生成可学习的位置编码，与输入特征图相加后输入网络中，通过反向传播算法，更新位置编码参数。

1.5　损失函数优化

由于 IoU 在小目标检测时对像素的位置偏差过于敏感，在目标尺度足够小时，轻微的边界框偏差就会导致 IoU 的急剧降低，从而导致对小目标的泛用性极差。因此引入 NWD 和 MPDIoU 损失相结合的方式替代 GIoU。

Wang 等[18]提出的 NWD 损失对不同尺度的物体不敏感，对边界框的重叠要求更低，甚至没有重叠也可以度量分布之间的相似性，因此更适用于小目标的检测。该方法先将边界框建模为 2D 高斯分布，然后使用归一化 Wasserstein 距离衡量高斯分布的相似性。首先，对于水平边界框 $R = (C_x, C_y, h, w)$，其中 (C_x, C_y)、w 和 h 分别表示中心坐标、宽度和高度。将其建模为高斯分布 $N(\boldsymbol{\mu}, \boldsymbol{\sigma})$，表达式为

$$\begin{cases} \boldsymbol{\mu} = \begin{bmatrix} C_x \\ C_y \end{bmatrix} \\ \boldsymbol{\sigma} = \begin{bmatrix} -\dfrac{w^2}{4} & 0 \\ 0 & \dfrac{h^2}{4} \end{bmatrix} \end{cases} \tag{5}$$

将边界框 A 和 B 之间的相似度转换为两个高斯分布之间的距离,两个二维高斯分布$\boldsymbol{\mu}_1 = N(\boldsymbol{m}_1, \boldsymbol{\sigma}_1)$ 和$\boldsymbol{\mu}_2 = N(\boldsymbol{m}_2, \boldsymbol{\sigma}_2)$,$\boldsymbol{\mu}_1$ 和$\boldsymbol{\mu}_2$ 之间的二阶 Wasserstein 距离为

$$W_2^2(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \| \boldsymbol{m}_1 - \boldsymbol{m}_2 \|_2^2 + \mathrm{Tr}[\boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_2 - 2(\boldsymbol{\sigma}_2^{1/2}\boldsymbol{\sigma}_1\boldsymbol{\sigma}_2^{1/2})^{1/2}] \tag{6}$$

对于由边界框 $A = \left(cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2}\right)$ 和 $B = \left(cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2}\right)$ 建模的高斯分布 N_a 和 N_b,最终简化为

$$W_2^2(N_a, N_b) = \left\| \left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2}\right]^{\mathrm{T}}, \left[cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2}\right]^{\mathrm{T}} \right) \right\|_2^2 \tag{7}$$

然而 $W_2^2(N_a, N_b)$ 是距离度量,不能直接用作相似性度量,对其进行归一化,得到最终的 NWD 损失表达式为

$$\mathrm{NWD}(N_a, N_b) = \exp\left(-\frac{\sqrt{W_2^2(N_a, N_b)}}{C}\right) \tag{8}$$

式(8)中:C 为常量。

NWD 损失提高了模型对小目标的检测精度,但会对模型的收敛速度产生一定影响。同步使用 MPDIoU,MPDIoU 简化了边界框之间的相似性比较,并且对边界框重叠与不重叠的情况都能很好处理,因此对小目标场景下检测的适用性较好,同时能够加快模型的收敛速度。MPDIoU 的计算过程如下。

$$\mathrm{MPDIoU} = \mathrm{IoU} - \frac{d_1^2}{h^2 + w^2} - \frac{d_2^2}{h^2 + w^2} \tag{9}$$

式(9)中:d_1^2 和 d_2^2 分别为预测边界框与真值边界框左上角点和右下角点坐标差值的平方之和,即左上角点和右下角点的距离的平方。

最终将 NWD 损失和 MPDIoU 结合得到损失函数表达式为

$$\mathrm{loss} = a(1 - \mathrm{MPDIoU}) + (1 - a)(1 - \mathrm{NWD}) \tag{10}$$

经实验对比,其中 a 取 0.7 时效果最优。

2 实验结果与分析

实验平台为 Ubuntu20.04 系统,内存为 24 GB,GPU 为 NVIDIA RTX4090,Python 版本为 3.8,PyTorch 版本为 1.13.1,CUDA 版本为 11.7。网络未使用预训练模型权重,在训练和测试阶段采用640×640 的图像输入大小,实验设置 200 个 epoch,batchsize 为 4,使用 adamw 优化器,学习率为0.000 1。同时为了保证实验的一致性,在测试阶段模型的 batchsize 均设置为 1。

2.1 数据集

采用天津大学机器学习与数据挖掘实验室 AISKYEYE 收集的 VisDrone2019 数据集进行实验,共包含 8 629 张图像,其中训练集包含 6 471 张图片,验证集 548 张,测试集 1 610 张。该数据集包含 10 个种类的目标,分别为行人、人、自行车、汽车、面包车、卡车、三轮车、遮阳棚三轮车、巴士、摩托车。

2.2 评估指标

实验采用 Params、GFLOPs、mAP50、mAP50:95 和 FPS 作为模型性能的评价指标。mAP、FPS 的计算公式如下。

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\% \tag{11}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\% \tag{12}$$

$$\mathrm{AP} = \int_0^1 P(r)\,\mathrm{d}r \tag{13}$$

$$\mathrm{mAP} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{AP}_i \tag{14}$$

$$\mathrm{FPS} = \frac{1\,000}{t} \tag{15}$$

式中:P 为准确率;R 为召回率;n 为检测类别数,TP、FP、FN、TN 分别为实际为正样本检测为正样本、实际为负样本检测为正样本、实际为正样本检测为负样本、实际为负样本检测为负样本;mAP50 为 IoU 阈值为 0.5 时所有类别的平均检测精度(mAP);FPS 为模型 1 s 内推理的图片数量,用来表示模型的推理速度,帧/s;t 为时间,s。

2.3 消融实验

为了验证所提每个改进部分的有效性,对每个改进分别在 VisDrone2019-DET 数据集上进行消融实验,并在测试集上验证其对小目标的检测性能,实验结果如表 2 所示。可以看出,在加入 MCP Block 后,由于残差结构中计算量减少并且过程中融合了多尺度的特征信息,显著降低了模型的参数量和计算量,并且检测精度提升 0.9%。AIFI-LPE 模块提供了可学习位置编码使模型能够更加准确的感知输入图像中的位置信息,使模型精度提升 1.2%。在引入 NWD 和 MPDIoU 后模型对小目标的定位效果更佳,检测精度提升 0.9%。最后加入提出的多尺度特征融合模块后,虽然模型计算量有少量增加,但参数量降低了 60%,并且精度提升 2.5%。从实验结果来看,本文模型在检测精度和网络体积上取得了恰当的平衡,更满足无人机航拍图像目标检测任务的要求。

2.4 对比试验

为了验证改进模型的检测性能,将提出的算法与

其他实时端到端目标检测器进行比较，实验结果如表3所示。在 VisDrone 数据集上，本文算法的 AP 达到 40.5%，FPS 达到 70 帧/s，相较于同规模甚至更大规模的 YOLO 系列算法，本文算法能够在很好地满足实时性（FPS >60 帧/s）的同时，极大降低了模型的参数量，且大幅提高了对小目标的检测精度，相较于其他最新的优秀模型具有更好的检测性能。

2.5　可视化试验

为了直观呈现本文模型在无人机航拍图像目标检测这一任务中的表现，将部分目标检测效果进行对比展示。如图 5 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，可以看出本文模型在远距离微小目标检测上展现出了更好的效果，极大提高了对目标的捕获能力，同时显著降低了原模型的漏检以及误检问题。本文模型在微小目标检测任务中表现出较强的能力，在对小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。

3　结论

针对无人机航拍图像中小目标检测中常见的漏检、误检等问题，提出了一种针对无人机航拍场景的小目标检测算法。首先，提出多尺度特征融合模块，并增加针对微小目标的检测层，以此增强网络对小目标的感知能力。同时为了防止模型过大，优化主干网络，去除了一个残差块。其次，在主干网络中使用 MCP Block 替换了传统的 Basic Block，使模型更加轻量化的同时，提升对小目标多尺度特征的提取能力，有效减少了特征冗余。此外，通过引入可学习位置编码，模型对复杂背景下小目标的精准定位能力得到了加强。最后，提出 NWD 和 MPDIoU 相结合的损失函数，在降低模型对位置偏差敏感度的同时加快了模型的收敛速度。实验结果表明，本文算法在 Visdrone2019 数据集上取得了良好的效果，领先于目前大多数主流目标检测算法，在实时无人机航拍目标检测领域有更大的优势。

表 2　消融实验

Table 2　Cryogenic experiment

序号	方法	参数量/10^6	计算量	准确率/%	召回率/%	mAP50/%	mAP50:95/%	FPS/(帧·s^{-1})
—	RT-DETR	19.9	57.3	55.4	37.6	36.6	21.0	60
A	RT-DETR + MCP Block	16.8	49.8	54.5	38.6	37.5	21.8	58
B	RT-DETR + AIFI-LPE	20.0	57.3	55.5	38.8	37.8	22.0	63
C	RT-DETR + NWD-MPDIoU	19.9	57.3	55.6	38.2	37.5	21.7	59
D	RT-DETR + 多尺度特征	7.9	69.4	56.2	39.9	39.1	23.1	61
E	A + B	16.9	49.8	56.2	39.1	38.0	22.1	60
F	A + B + C	16.9	49.8	56.0	39.4	38.5	22.4	61
本文	A + B + C + D	7.5	63.2	57.7	41.8	40.5	23.7	70

表 3　对比实验

Table 3　Comparative Experiment

方法	参数量/10^6	计算量	准确率/%	召回率/%	mAP50/%	mAP50:95/%	FPS/(帧·s^{-1})
YOLOv5-m	25.0	64.2	48.5	35.3	34.1	19.8	112
YOLOv5-l	53.1	135.0	48.4	37.1	35.5	20.8	99
YOLOv8-m	25.9	79.3	48.1	36.0	34.4	20.0	108
YOLOv8-L	43.6	165.7	49.2	36.8	35.3	20.7	101
本文	7.5	63.2	57.7	41.8	40.5	23.7	70

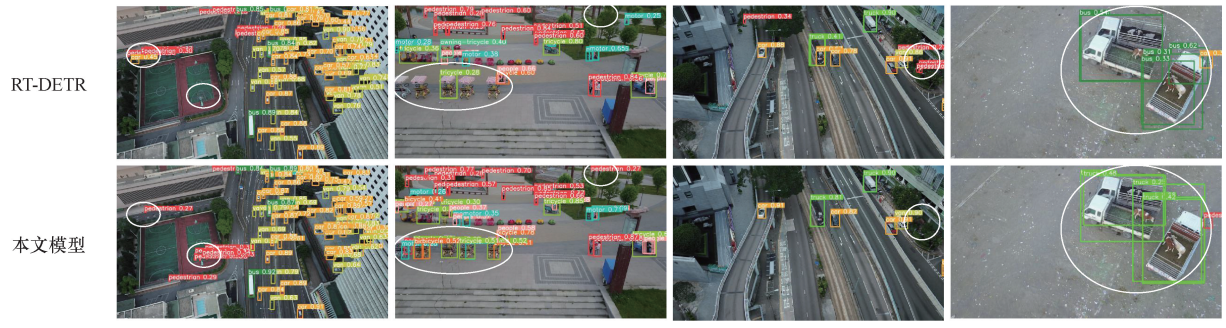


图 5　改进后的模型与原模型检测效果对比

Fig. 5　Comparison between the improved model and the original model

研究在无人机技术领域的发展中发挥了积极作用。通过技术创新,提出的算法显著提升了无人机航拍图像中小目标的检测精度和实时处理能力,同时实现了模型计算成本的降低。这使无人机在资源受限的条件下依然能够执行高效的目标识别任务。这对于农业现代化、地理测绘、环境监测等关键领域的发展具有重要的作用,同时也对加强国防安全具有积极影响,为中国无人机在复杂地理环境和多变气候条件下的应用提供了坚实的技术基础。

参 考 文 献

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

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

[3] 桑雨,李立权, 李铁. 轻量化 YOLOv7-tiny 的遥感图像小目标检测[J]. 科学技术与工程, 2024, 24(18): 7726-7732.
Sang Yu, Li Liquan, Li Tie, et al. Lightweight YOLOv7-tiny for remote sensing image small target detection[J]. Science Technology and Engineering, 2024, 24(18): 7726-7732.

[4] 刘晋川,黎向锋, 刘安旭, 等. 改进 RetinaNet 的无人机小目标检测[J]. 科学技术与工程, 2023, 23(1): 274-282.
Liu Jinchuan, Li Xiangfeng, Liu Anxu, et al. Improved RetinaNet for unmanned aerial vehicle small target detection[J]. Science Technology and Engineering, 2023, 23(1): 274-282.

[5] 吴稳稳,吴晓红, 刘强, 等. 基于全局注意力的多级特征融合目标检测算法[J]. 科学技术与工程, 2020, 20(27): 11185-11191.
Wu Wenwen, Wu Xiaohong, Liu Qiang, et al. Multi-level feature fusion object detection algorithm based on global attention[J]. Science Technology and Engineering, 2020, 20(27): 11185-11191.

[6] 康传利,张思瑶, 李玄皓, 等. 高斯 Wasserstein 距离改进轻量 YOLOv7 模型的遥感影像道路交叉口检测[J]. 科学技术与工程, 2024, 24(9): 3533-3542.
Kang Chuanli, Zhang Siyao, Li Xuanhao, et al. Gaussian Wasserstein distance improvement of lightweight YOLOv7 model for remote sensing image road intersection detection[J]. Science Technology and Engineering, 2024, 24(9): 3533-3542.

[7] Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks[J]. ArXiv, 2019: 1905.11946.

[8] 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.

[9] Zhu X, Su W, Lu L, et al. Deformable detr: deformable transformers for end-to-end object detection[J]. ArXiv Preprint ArXiv, 2020: 2010.04159.

[10] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 764-773.

[11] Meng D, Chen X, Fan Z, et al. Conditional detr for fast training convergence[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 3651-3660.

[12] Wang Y, Zhang X, Yang T, et al. Anchor DETR: query design for Transformer-based detector[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2022, 36(3): 2567-2575.

[13] Liu S, Li F, Zhang H, et al. DAB-DETR: dynamic anchor boxes are better queries for DETR[J]. ArXiv Preprint ArXiv, 2022: 2201.12329.

[14] Li F, Zhang H, Liu S, et al. DN-DETR: accelerate DETR training by introducing query denoising[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 46(4): 13619-13627.

[15] Zhang H, Li F, Liu S, et al. Dino: DETR with improved denoising anchor boxes for end-to-end object detection[J]. ArXiv Preprint ArXiv, 2022: 2203.03605.

[16] Lü W, Xu S, Zhao Y, et al. DETRs beat yolos on real-time object detection[J]. ArXiv Preprint ArXiv, 2023: 2304.08069.

[17] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[C]//Computer Vision-ECCV 2014: 13th European Conference. Zurich: Springer International Publishing, 2014: 740-755.

[18] Wang J, Xu C, Yang W, et al. A normalized Gaussian Wasserstein distance for tiny object detection[J]. ArXiv Preprint ArXiv, 2021: 2110.13389.

[19] Siliang M, Yong X. Mpdiou: a loss for efficient and accurate bounding box regression[J]. ArXiv Preprint ArXiv, 2023: 2307.07662.

[20] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Chengdu: IEEE, 2016: 770-778.

[21] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25.

[22] Sifre L, Mallat S. Rigid-motion scattering for texture classification[J]. ArXiv Preprint ArXiv, 2014: 1403.1687.

[23] Chen J, Kao S, He H, et al. Run, don't walk: chasing higher FLOPS for faster neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 12021-12031.

[24] Howard A, Sandler M, Chu G, et al. Searching for MobileNetV3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 1314-1324.

[25] Zhang X, Zhou X, Lin M, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6848-6856.

[26] Han K, Wang Y, Tian Q, et al. GhostNet: more features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 1580-1589.

[27] Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks[J]. ArXiv, 2019: 1905.11946.

[28] Zhang Q, Jiang Z, Lu Q, et al. Split to be slim: an overlooked redundancy in Vanilla convolution[J]. ArXiv Preprint ArXiv, 2020: 2006.12085.