文章编号:1671-4598(2026)05-0114-08 DOI:10.16526/j.cnki.11-4762/tp.2026.05.014 中图分类号:TP391.411 文献标识码:A

# 基于改进 RT-DETR 的轻量级小目标检测方法

郭致诚[1]，曹 静[2]，陈 曦[1]，胡 锴[1]，王国珲[1]
(1. 西安工业大学 光电工程学院，西安 710021；
2. 中国兵器工业试验测试研究院，西安 710116)

**摘要**：无人机航拍图像的小目标检测性能易受不规则拍摄角度、视角变化、遮挡及光照变化等因素影响而显著下降；为此，提出轻量级的无人机目标检测模型 LDRT-DETR；针对小目标与背景尺度差异显著的问题，引入多尺度边缘特征提取单元，从不同尺度获取特征并增强边缘信息，以提升模型对小目标的感知能力；设计上下文增强特征融合方法，以更有效整合上下文信息并减少融合过程中的细节损失，从而提高检测精度；针对无人机航拍图像中背景复杂、遮挡严重的特点，提出多尺度采样模块，该模块结合上采样与下采样机制，并通过多分支结构构建丰富的特征提取路径，从而增强特征表达能力；实验结果表明，LDRT-DETR 在 VisDrone2019 数据集上的 *mAP*50 达到 50.05%，较 RT-DETR 提升 2.82%，同时参数量和 GFLOPs 分别下降 24.7%和 10%。

**关键词**：目标检测；细节增强；特征融合；轻量化模型；无人机

# Detection Method for Lightweight Small Objects Based on Improved RT-DETR

GUO Zhicheng[1], CAO Jing[2], CHEN Xi[1], HU Kai[1], WANG Guohui[1]
(1. School of Optoelectronic Engineering, Xi'an Technological University, Xi'an 710021, China;
2. Norinco Group Testing and Research Institute, Xi'an 710116, China)

**Abstract**: The performance of small-object detection in UAV aerial images is significantly degraded by factors such as irregular shooting angles, viewpoint variations, occlusions, and illumination changes. To address these issues, a lightweight drone object detection model, LDRT-DETR, is proposed. For targets significantly different in scale from the background, a multi-scale edge feature extraction module is introduced to capture features across different scales and enhance their edge information, thereby improving the model's ability to perceive small objects. A context-enhanced feature fusion strategy is designed to more effectively integrate contextual information and reduce detail loss during the feature fusion, thus improving its detection accuracy. Additionally, to address complex backgrounds and severe occlusions in UAV aerial images, a multi-scale sampling module is developed, which incorporates both upsampling and downsampling mechanisms and constructs diverse feature extraction pathways through a multi-branch structure, thereby enhancing its feature representation. Experimental results on the VisDrone2019 dataset demonstrate that the LDRT-DETR achieves an *mAP*50 of 50.05%, with a increase of 2.82% over that of the RT-DETR, while reducing the number of parameters and GFLOPs by 24.7% and 10%, respectively.

**Keywords**: object detection; detail enhancement; feature fusion; lightweight model; UAV

## 0 引言

随着无人机技术的迅猛发展及低空经济的兴起[1]，无人机已广泛应用于智慧城市[2]、生态监测[3]、农业生产[4]等领域，尤其是无人机航拍技术因其低成本和高分辨率图像数据的优势，成为数据采集的核心工具。在农

**收稿日期**:2025-10-15； **修回日期**:2025-11-24。
**基金项目**:2024 年大学生创新创业训练计划项目(S202410702125);陕西省教育厅科学研究计划项目资助(22JY205)。
**作者简介**:郭致诚(2005-),男,大学本科。
**通讯作者**:王国珲(1980-),男,教授,博导。
**引用格式**:郭致诚,曹 静,陈 曦,等.基于改进 RT-DETR 的轻量级小目标检测方法[J].计算机测量与控制,2026,34(5):114-121.

业领域，农户通过获取高分辨率图像，结合人工智能算法实现作物长势分析；在生态环保中，航拍图像能够精确定位非法排污口、监测植被覆盖变化等。

尽管无人机航拍技术具有显著优势，基于无人机图像的目标检测任务仍面临诸多挑战。首先，目标往往仅占据图像的极小部分，且目标之间存在显著的尺度差异，增加了特征提取的难度。其次，航拍图像中常常包含建筑物、树木、道路等复杂背景，增加了模型对目标与背景的区分难度。再者，鸟瞰视角的随机性和光照条件的变化，进一步加剧了目标特征的不稳定性，降低了检测结果的鲁棒性。因此，提高无人机航拍图像中小目标的检测精度，仍是当前亟待解决的关键问题。

近年来，深度学习在图像处理领域取得了突破性进展，特别是卷积神经网络（CNN，convolutional neural network）在目标检测中的应用，显著提升了检测精度和实时性。现有基于深度学习的目标检测方法大体可分为两类：一阶段检测模型（如 YOLO[5]、SSD[6]）和两阶段检测模型（如 Faster R-CNN[7]）。两阶段检测模型精度较高，但其实时性较差；例如，文献［8］将改进的 Faster R-CNN 用于人脸识别，并拥有高效的检测性能。文献［9］采用全卷积架构进行目标检测，采用更深层的 ConvNet 作为特征分类器能实现更好的检测性能。一阶段模型虽提高了处理速度，但依赖非极大值抑制（NMS，non-maximum suppression），调优困难。文献［10］在 YOLOv4 模型中引入 MobileNetv2 轻量化骨干网络和 CIOU 损失函数优化小目标检测，其在无人机航拍数据集上的 *mAP*50 仅为 35%。文献［11］提出 UCM-YOLOv8 模型，通过设计任务动态对齐检测头与 Wise-Inner 损失函数提升无人机航拍图像小目标检测准确率，*mAP*50 为 44.7%。文献［12］提出 SCD-YOLO 模型，通过引入 CoordConv 增强位置特征和 Dy-Head 优化多尺度特征融合以提升模型特征表达能力，其在目标数据集上的 *mAP*50 为 43.8%，复杂度达 20.8 GFLOPs。文献［13］提出 HM-YOLO 模型，通过改进 C3k2 _ MSEE 模块，以突出图像边缘信息，并设计层次注意力融合模块（HAFB）强化上下文信息整合能力，*mAP*50 为 47.1%。综上，现有的深度学习模型在提升检测精度上相较于基线模型效果有限，且内部结构存在一定的冗余，使得推理速度较慢且计算成本较高。

在此背景下，Transformer 技术的引入为目标检测提供了新的方向。Transformer[14]模型凭借全局注意力机制和长距离特征建模能力，在自然语言处理领域取得了成功，并逐步应用于计算机视觉领域。基于 Transformer 的实时目标检测模型 RT-DETR[15]（Real-Time Detection，Embedding，and Tracking）省略了候选框生成和非极大值抑制环节，通过端到端架构大幅减少后处理时间，提升了检测效率。然而，RT-DETR 仍使用较大参数量的 ResNet[16]作为特征提取网络，这导致了模型在计算效率和实时性方面的瓶颈。尽管 ResNet 作为深度卷积网络在特征提取上具有较高的精度，但其参数量和计算复杂度使得该模型在处理高分辨率图像时，尤其是在实时应用场景中，难以满足低延迟的需求。此外，ResNet 的网络结构设计本身对多尺度特征的处理能力有限，这进一步影响了其在不同尺度目标上的表现。

为解决 RT-DETR 模型参数量大的问题，设计了一种名为 LDRT-DETR 的轻量级检测模型。该模型的主要目标是增强模型对目标特征的提取能力，大幅降低模型的参数量和复杂度。本研究的主要贡献总结如下：

1）在特征提取网络中引入多尺度边缘特征提取单元（MSEFE Unit，multi-scale edge feature extraction unit），采用自适应池化操作和边缘增强机制强化对小目标的敏感度。

2）在编码网络中引入上下文增强特征融合（CEF Fusion，context-enhanced feature fusion）模块，利用 SE 注意力机制实现上下文信息的自适应加权融合，有效提升特征表示的鲁棒性。

3）在编码网络中提出多尺度采样模块（MS module，multi-scale sampling module），一方面通过多尺度分支结构构建特征提取路径，丰富特征表达多样性；另一方面采用 Gate 机制对采样后特征生成自适应选择权重，动态筛选并强化关键信息。

## 1　RT-DETR 模型

RT-DETR 是基于 Transformer 的实时检测的创新模型。它可以同时并行处理多个目标，而且能够在 GPU 上实现高度优化的运算，实现对目标的快速检测。与 YOLO 系列模型相比，RT-DETR 未使用连续的卷积层来提取图像特征，而是采用 Transformer 中的自注意力机制来帮助模型学习不同特征之间的关系。此外，RT-DETR 在检测不同场景目标时，均表现出更高的效率和更好的鲁棒性，且保持了较快的检测速度。RT-DETR 模型由 3 个主要组件组成：主干提取网络、编码器和解码器。

## 2　LDRT-DETR 模型

本研究采用 RT-DETR 作为基准检测模型，首先在主干网络中嵌入多尺度边缘提取单元（MSEFE Unit），该单元通过分层捕捉小目标的边缘纹理与轮廓信息，强化不同尺度特征的细节表征能力。其次，在编码器网络中设计上下文增强特征融合模块（CEFF），该模块通过引入 SE 注意力机制对主干网络输出的特征进行权重分配与信息互补融合。此外，为进一步优化特征分辨率以匹配小目标尺度，本研究设计多尺度上采样模块

MSSM（UP）与多尺度下采样模块 MSSM（Down）进行协同整合，实现特征分辨率与语义信息的动态适配。LDRT-DETR 的结构示意图如图 1 所示。

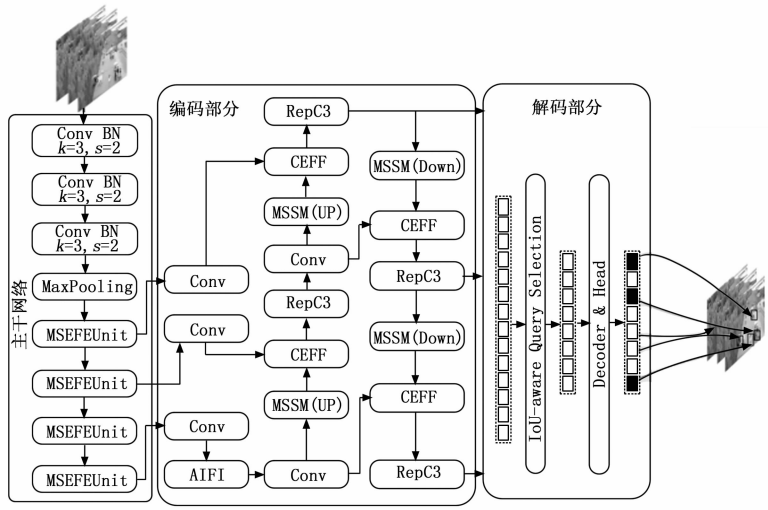


图 1 LDRT-DETR 模型结构

## 2.1 多尺度边缘特征提取单元

RT-DETR 通过 ResNet-18 网络逐级下采样有助于获取高层语义特征和大的感受野范围，但小目标在深层特征图中像素占比极低，纹理细节容易丢失，且导致 Transformer 注意力难以在稀疏的特征表示中有效捕获其语义信息。此外，单一尺度的卷积操作在提取全局语义时，易将目标边缘轮廓的高频信息与背景的低频信息混叠，从而降低检测精度。

针对上述挑战，本文引入多尺度边缘特征增强单元，如图 2 所示。该模块首先通过平均池化获取上下文内容，然后采用残差相减显式解耦边缘分量，并在 3×3、6×6、9×9、12×12 四个空间分辨率上并行提取；各支路经 1×1 压缩及分组 3×3 卷积后，由边缘增强模块（Edge Enhance Block）重新加权，再与局部细节支路在通道维拼接融合，从而增强小目标感知与定位能力。

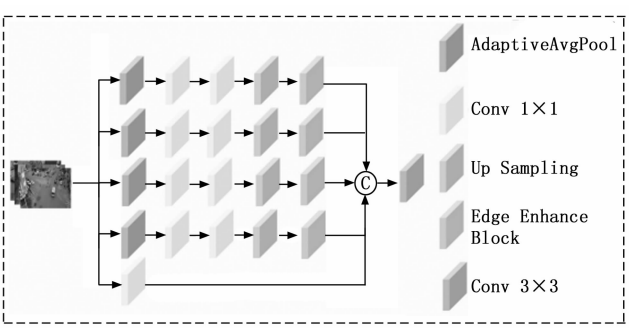


图 2 多尺度边缘特征增强单元结构

在多尺度边缘特征增强单元中，边缘增强模块为轻量级特征处理单元，其核心设计目标是提取并自适应强化特征图中的有效边缘信息（如目标轮廓、纹理边界），模块结构如图 3 所示。

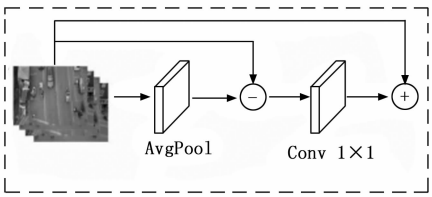


图 3 边缘增强模块结构

设输入特征图为 $x \in R^{C\times H\times W}$，$C$ 为通道数，$H$ 和 $W$ 分别为特征图的高和宽。首先通过多尺度自适应池化操作对特征图下采样，随后经过上采样、卷积和边缘增强模块实现对小目标特征的有效提取，具体流程如下所示。

2.1.1 多尺度特征分支

1）对输入特征 $x$ 进行不同尺度的自适应池化，生成多尺度特征：

$$x_{\text{pool}}{}^{b} = \text{AdaptiveAvgPool}_{3\times 3}(x) \tag{1}$$

其中：$b \in \{3, 6, 9, 12\}$，表示各分支的输出尺寸。

2）对池化后的特征进行通道降维与深度可分离卷积操作：

$$x^{b}_{\text{reduce}} = \text{Conv}(x^{b}_{\text{pool}}) \tag{2}$$

$$x^{b}_{\text{depth}} = \text{DepthConv}(x^{b}_{1}) \tag{3}$$

3）将处理后的多尺度特征插值回原始尺寸：

$$x^{b}_{up} = \text{Interpolate}(x^{b}_{\text{depth}}) \tag{4}$$

2.1.2 边缘增强模块

1）通过 3×3 平均池化以获取局部平滑特征：

$$\text{pool}(x) = \text{AvgPool}(x) \tag{5}$$

2）通过原始特征与平滑特征的残差获取边缘信息：

$$\text{edge}_{\text{raw}} = x - \text{pool}(x) \tag{6}$$

3）经 Sigmoid 激活生成边缘权重，并与原始特征融合：

$$\text{edge}_{\text{weight}} = \sigma[\text{Conv}(\text{edge}_{\text{raw}})] \tag{7}$$

$$x_{\text{enhanced}} = \text{x} + \text{edge}_{\text{weight}} \tag{8}$$

2.1.3 特征融合与输出

1）将原始特征的局部卷积结果与各分支边缘增强特征拼接：

$$x_{\text{local}} = \text{Conv}(x) \tag{9}$$

$$x_{\text{concat}} = \text{Concat}[x_{\text{local}} + \text{Edge Enhance Block}(x^{b}_{up})] \tag{10}$$

2）特征融合：

$$x_{\text{out}} = \text{Conv}(x_{\text{concat}}) \tag{11}$$

## 2.2 上下文增强特征融合模块

在 RT-DETR 模型中，传统特征图拼接与融合方法对不同来源的特征采用统一聚合策略，未根据语义表征重要性差异动态分配权重，导致背景噪声与冗余信息累

积，从而降低检测精度。具体而言，在多尺度拼接与融合过程中，浅层特征包含丰富的空间细节信息，但会引入更多的噪声与背景等干扰信息；而深层特征虽具备强语义表达能力，但空间分辨率较低。此外，现有方法多采用高层级语义特征单向增强低层级细节特征的模式，仅实现单方向信息流动，却忽略低层细节对高层语义的反馈作用，从而制约特征表达的完整性。针对上述问题，本研究设计的上下文增强特征融合模块，通过 SE 注意力[17]机制动态评估特征通道重要性以抑制无效信息，同时通过双向互补增强机制强化层级语义关联，模块结构如图 4 所示。

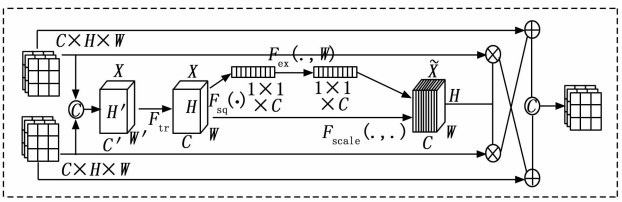


图 4　增强型上下文特征融合模块

增强型上下文特征融合模块以两路不同来源的特征作为输入。首先，通过通道维度拼接对两路特征进行初始信息聚合，并利用特征转换操作完成通道数等维度的适配；其次，通过 SE 注意力生成与特征通道重要性相匹配的动态权重。具体而言，SE（Squeeze-and-Excitation）注意力是一种通过动态调整特征通道权重来增强模型性能的方法。通过“压缩（Squeeze）”操作进行全局平均池化（Global Average Pooling），从每个通道中提取全局信息，生成该通道的描述。接着，使用“激励（Excitation）”操作，通过全连接层计算每个通道的加权系数，以便对特征图进行加权调整。与传统的拼接或加权求和等方法相比，SE 注意力通过动态生成加权系数，避免了固定规则的局限性，增强了特征融合的灵活性和适应性；随后，将上述权重通过逐元素乘积分别作用于对应的原始特征，实现通道维度的特征校准，以此强化关键特征信息并抑制无效干扰信息；最后，对校准后的两路特征进行拼接融合，输出包含多尺度信息且经过重要性加权的融合特征。

设输入特征为两种不同来源的特征图：$x_0 \in R^{C_0 \times H \times W}$（低层级特征）和 $x_1 \in R^{C_1 \times H \times W}$（高层级特征）。

1）通道调整：

当 $C_0 \neq C_1$ 时，通过 1×1 卷积对通道进行调整：

$$x'_0 = \mathrm{Conv}(x_0) \tag{12}$$

当 $C_0 = C_1$ 时，$x'_0 = x_0$。

2）特征拼接：

$$x_{\mathrm{concat}} = \mathrm{Concat}(x'_0, x_1) \tag{13}$$

3）通过 SE 注意力生成通道权重并作用于拼接特征：

$$z = \mathrm{AvgPool}(x_{\mathrm{concat}}) \tag{14}$$

$$s = \mathrm{Sigmoid}[W_2 \cdot ReLU(W_1 \cdot z)] \tag{15}$$

其中：$W_1$ 和 $W_2$ 表示全连接层的参数：

$$x_{\mathrm{attention}} = x_{\mathrm{concat}} \odot s \in {}^{2C_1 \times H \times W} \tag{16}$$

其中：⊙表示逐通道加权。

4）权重拆分与双向特征融合：

$$x_{0_\mathrm{weight}}, x_{1_\mathrm{weight}} = \mathrm{Split}(x_{\mathrm{attention}}) \tag{17}$$

$$x_0^{\mathrm{out}} = x'_0 + x_{1_\mathrm{weight}} \tag{18}$$

$$x_1^{\mathrm{out}} = x_1 + x_{0_\mathrm{weight}} \tag{19}$$

5）最终输出：

$$x_{\mathrm{out}} = \mathrm{Concat}(x_0^{\mathrm{out}}, x_1^{\mathrm{out}}) \in \mathbb{R}^{2C1 \times H \times W} \tag{20}$$

## 2.3　多尺度采样模块

编码网络是连接主干网络和解码网络的关键组件，核心功能是对主干网络输出的多尺度特征进行层级关联与融合增强，实现语义信息与细节特征的有效聚合，为解码网络提供高质量的特征输入。编码网络通过多层级特征交互架构，实现空间信息与语义信息的高效融合，所生成的融合特征在空间定位和类别判别上达到均衡优化。与 YOLO 系列模型采用的跨尺度拼接结构相比，RT-DETR 的编码网络因引入 Transformer 编码器与多阶段特征交互机制，其特征融合机制更复杂，虽提升了特征之间的关联度，但也增加计算开销。针对无人机航拍场景中大量存在的小尺度目标图像，RT-DETR 模型在多轮下采样与特征融合过程中，易丢失关键信息，且参数量较多，不利于模型部署的需求。因此，本研究提出了多尺度采样模块（包含上采样和下采样块），在实现模型参数量与计算量降低的同时，进一步提升检测精度。

图 5 中的上采样与下采样模块包含 3 条并行分支，各分支通过差异化的尺度变换为网络提供多尺度特征提取路径，从而丰富特征表达的尺度与语义多样性。在上

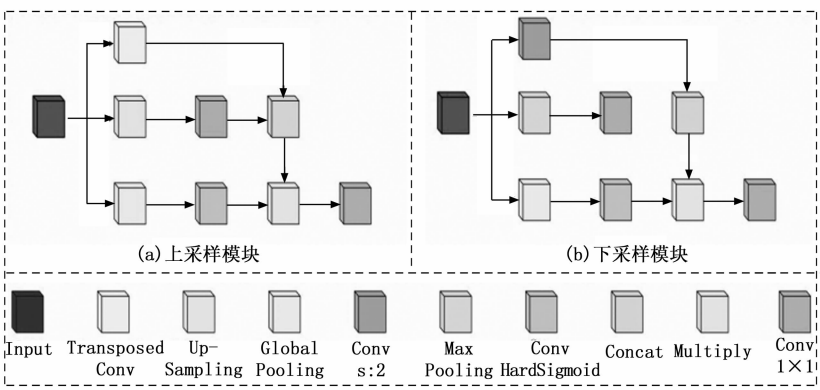


图 5　多尺度采样模块

采样模块中，首先通过 Transposed Conv 对输入特征图执行上采样以提升分辨率，恢复低层级细节信息；随后引入 Gate 机制（由 Conv 与 HardSigmoid 组成）对采样后的特征进行特征选择，强化关键语义信息并抑制无关特征，提升特征表达的有效性。在下采样模块中，通过 Global Pooling 对输入特征图进行下采样以压缩空间维度，提取全局语义信息；随后经卷积操作对全局语义特征进行优化。最终，上采样与下采样模块各分支处理后的特征通过拼接与逐元素相乘的组合操作完成融合，再经 1×1 卷积调整通道数，输出包含多尺度细节与全局语义的特征图。

## 3 实验结果与分析

### 3.1 实验环境

本实验平台硬件环境为 NVIDIA RTX 2080Ti；软件环境为 CUDA 12.0、Python 3.12.9、Pytorch 2.4.0。表 1 列出了训练阶段的参数配置，在训练时模型会将输入图像的尺寸统一缩放至 640×640。

### 3.2 数据集

本研究使用的数据集为 VisDrone2019[18]，该数据集共包含 10 209 张图像，涵盖了城市道路、居民区、工业区等多种环境，以及不同的拍摄高度、角度和光照条件。数据集包含十个目标类别，如行人、车辆、公交车、三轮车等，这些目标类型在不同场景中均有出现，能够有效评估模型在多样化场景中的表现。部分图像如图 6 所示。

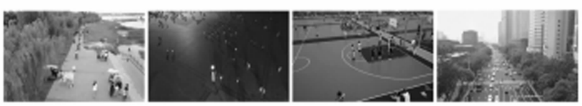


图 6 VisDrone2019 数据集示例

数据集拍摄的图像通过无人机在不同高度、拍摄角度及天气条件下采集获得。具体而言，VisDrone2019 数据集分为 6 471 张训练图像、548 张验证图像和 1 580 张测试图像，涵盖十类目标。各类别的分布情况如图 7 所示。

表 1 实验参数配置

| Parameters | Value |
|---|---|
| Epoch | 250 |
| Weight_decay | 0.000 1 |
| Batch_size | 4 |
| Momentum | 0.8 |
| Learning_rate | 0.000 1 |
| Optimizer | AdamW |

尽管数据集中包含多个目标类别，但部分类别（如“Car”）样本数量较多，部分类别（如“tricycle”或“awning-tricycle”）的样本较少，存在类别不平衡问题。为缓解这一问题，实验中采用了马赛克数据增强技术，提高了目标类别的多样性，从而增强了模型的鲁棒性。

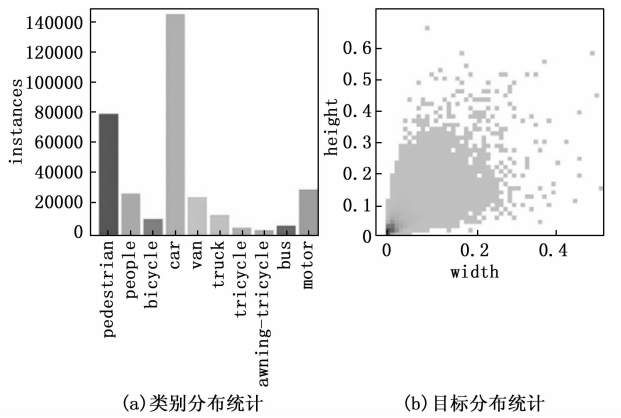


图 7 VisDrone2019 数据集类别分布

### 3.3 评价指标

本实验采用精准率（P，Precision）、召回率（R，Recall）、平均精度均值（mAP，mean average precision）、模型参数量（Params）和浮点运算次数（GFLOPs，giga FLOating-point operations）作为评价模型性能的综合指标，以充分验证所提方法在检测准确率与计算效率之间的平衡性能。相关计算公式如下：

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

$$Recall = \frac{TP}{TP + FN} \tag{22}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{23}$$

其中：$TP$ 表示模型正确预测的目标的样本数，$FP$ 表示模型错误预测目标的样本数，$FN$ 表示模型错误预测的正确样本数，$N$ 为数据集类别数。$mAP$ 表示模型能将图像中多少真正的目标检测出来。精准率和召回率用来衡量模型识别目标的能力，若精准率值越高，表明模型识别小目标的能力越强。

### 3.4 实验结果分析

表 2 展示了 LDRT-DETR 模型在 VisDro-ne2019 数据集上针对不同类别目标的检测结果。从总体性能来看，该模型的精准率和召回率分别为 62.63%和 49.08%。此外，$mAP50$ 为 50.05%、$mAP50:75$ 为 32.06%、$mAP50:95$ 为 31.04%，验证了模型对复杂航拍场景下多类别目标的检测适应性混淆矩阵可直观呈现模型对各类别样本的预测结果与真实标签的对应关系。图 8 为 LDRT-DETR 与 RT-DETR 模型的混淆矩阵对比图。通过分析矩阵中错误预测元素的分布可知，LDRT-DETR 在多个类别上的识别性能整体优于 RT-DETR，尤其体现在对关键交通目标的正确分类率更

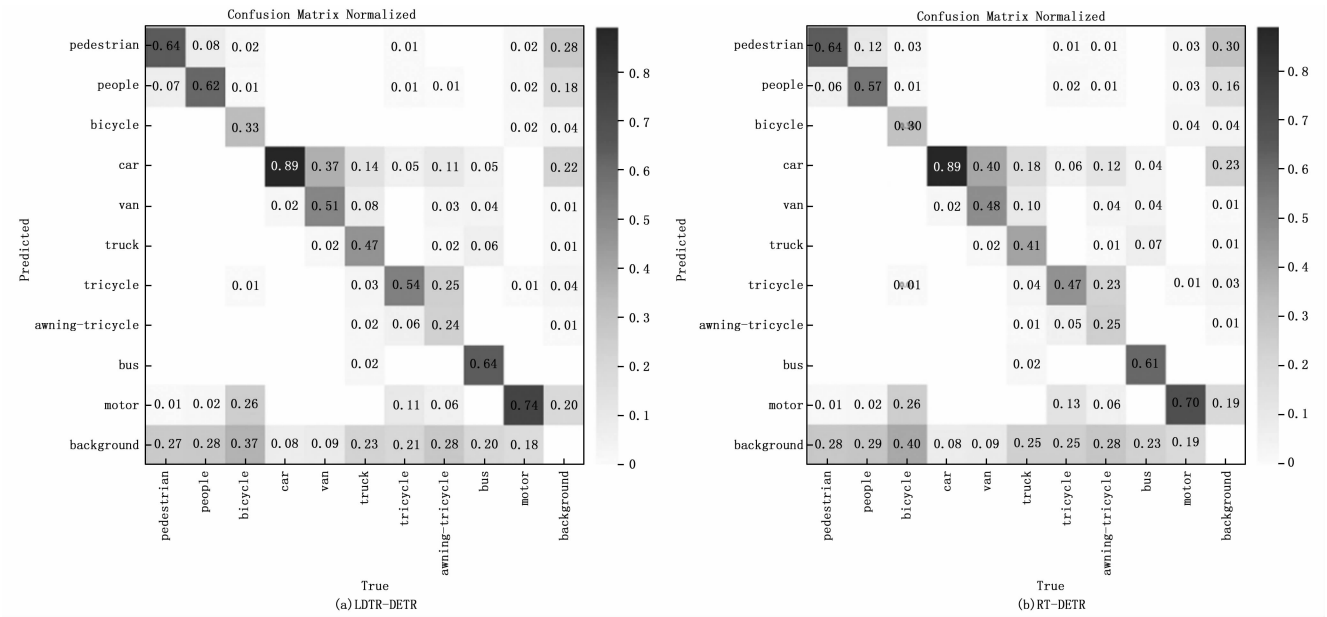


图 8　归一化混淆矩阵对比图

高且对易混类别的区分更加稳定。具体而言，LDRT-DETR 在主干类别上呈现出更高的对角线响应，同时对 motor 及 awning-tricycle 等外观相近的小目标类别的错分比例相对更低，表明其在细粒度类别区分方面具有更强的判别能力。此外，LDRT-DETR 对 background 的误检比例整体较低，说明其具备更优的背景抑制能力。相比之下，RT-DETR 在部分类别上仍存在较明显的混淆，且 background 类别中出现更多前景目标被误判的情况，反映出其特征表达对复杂场景的适应性仍有不足。

表 2　各类别在 VisDrone2019 数据集中的结果

| 类别名称 | 精准率/% | 召回率/% | *mAP*@50/% | *mAP*@50:75/% | *mAP*@50:95/% |
|---|---|---|---|---|---|
| pedestrian | 70.62 | 52.96 | 58.74 | 25.05 | 29.26 |
| people | 63.19 | 51.02 | 51.96 | 15.29 | 22.57 |
| bicycle | 48.03 | 25.95 | 24.54 | 9.30 | 11.66 |
| car | 81.85 | 83.79 | 86.33 | 71.90 | 63.24 |
| van | 67.91 | 49.62 | 53.09 | 46.32 | 39.90 |
| truck | 64.40 | 39.87 | 41.59 | 30.54 | 27.90 |
| tricycle | 49.87 | 43.44 | 38.15 | 25.05 | 23.17 |
| awning-tricycle | 40.36 | 18.80 | 18.85 | 13.23 | 12.46 |
| bus | 75.93 | 61.35 | 63.06 | 56.74 | 48.29 |
| motor | 64.10 | 64.00 | 64.21 | 27.19 | 31.90 |
| all | 62.63 | 49.08 | 50.05 | 32.06 | 31.04 |

为验证轻量级检测模型 LDRT-DETR 在 Visdrone2019 数据集上的综合性能，对模型的精准率（Precision）和召回率（Recall）曲线进行可视化分析，结果如图 9 所示。

从图 9 中可以看出，LDRT-DETR 在 Visdrone2019 数据集的可视化结果中展现出优异的性能，其 Precision-Recall 曲线整体位于基线模型 RT-DETR 曲线上方，曲线下面积更大。同时，该模型较高的精准率和召回率可以满足无人机视角下复杂场景的检测需求。

### 3.5　消融实验

为了验证 3 个协同模块改进效果，本研究创建 7 组消融实验进行对比。将多尺度边缘特征提取作为模块 *A*、上下文增强特征融合作为模块 *B*、多尺度采样作为模块 *C*。通过依次添加各个改进模块进行详细的对比，实验结果如表 3 所示。

表 3　消融实验对比结果

| 方法 | 精准率/% | 召回率/% | *mAP*50/% | *mAP*50:75/% | *mAP*50:95/% | GFLOPs | Params/M |
|---|---|---|---|---|---|---|---|
| RT-DETR | 61.94 | 46.24 | 46.85 | 29.74 | 28.74 | 57.0 | 19.8 |
| A | 62.60 | 47.80 | 49.05 | 30.80 | 30.02 | 48.4 | 14.5 |
| B | 61.05 | 46.62 | 47.74 | 29.98 | 29.25 | 57.0 | 20.0 |
| C | 61.63 | 47.02 | 48.20 | 30.05 | 29.19 | 51.3 | 14.9 |
| A+B | 62.43 | 48.94 | 49.90 | 31.83 | 30.88 | 48.4 | 14.6 |
| A+C | 63.48 | 47.37 | 49.32 | 31.48 | 30.42 | 51.3 | 14.8 |
| B+C | 62.32 | 46.84 | 49.07 | 31.27 | 29.65 | 57.6 | 19.3 |
| A+B+C | 62.63 | 49.08 | 50.05 | 32.06 | 31.04 | 51.3 | 14.9 |

表 3 的消融实验结果表明，A、B、C 三个模块对 RT-DETR 的性能提升具有显著作用。单模块引入：分别加入 A、B、C 模块后，模型 *mAP*50 分别提升 2.2%、0.89%和 1.35%，同时轻量化与精度提升实现平衡。两两模块组合：A+B、A+C、B+C 三种组合的精度均超过单模块，其中 A+B 组合的 *mAP*50、*mAP*50：75 和 *mAP*50：95 涨幅最高，体现出模块间的协同效应。三模块融合：当 A、B、C 模块全部嵌入模

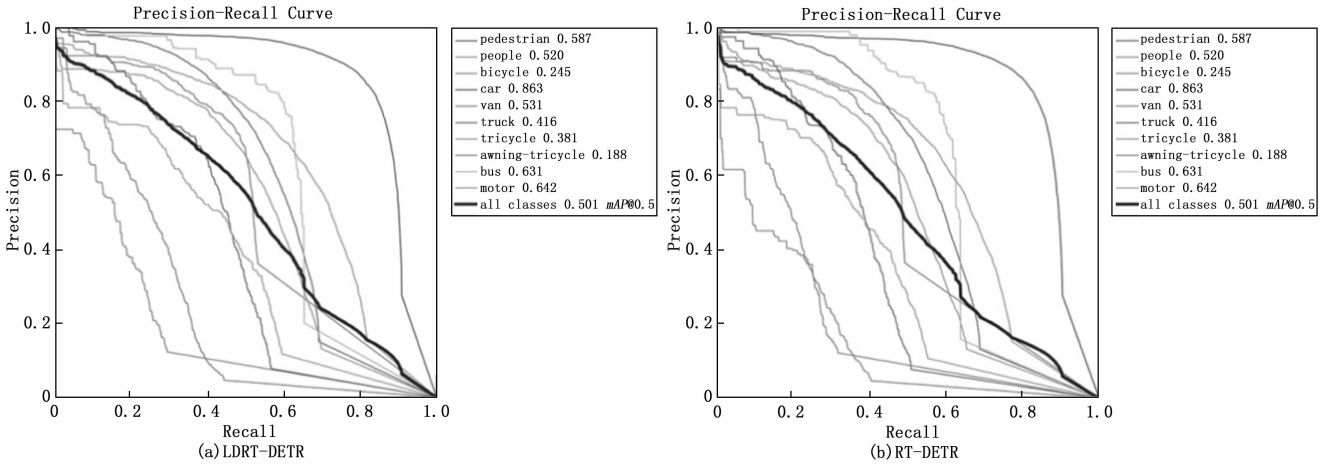


图 9 Precision-Recall 曲线对比图

型时，*mAP*50、*mAP*50：75、*mAP*50：95 分别达到 50.05%、32.06%和 31.04%，精准率和召回率分别提升至 62.63%和 49.08%，且 *mAP*50 较基线模型提升 2.82%，有效验证多模块融合对复杂无人机航拍场景检测的有效性。

### 3.6 对比实验

为验证 LDRT-DETR 模型的综合检测性能，本研究以 RT-DETR 为基线模型，并将其与 YOLO 系列等主流目标检测算法进行对比。实验从 *mAP*50、*mAP*50：75、*mAP*50：95 及 GFLOPs、Params 等多个维度进行量化比较，结果如表 4 所示。

表 4 各模型的对比结果

| 方法 | *mAP* 50/% | *mAP*50：75/% | *mAP*50：95/% | Params /M | GFLOPs |
|---|---|---|---|---|---|
| YOLOv5m[19] | 42.57 | 26.35 | 25.66 | 22.1 | 52.5 |
| YOLOv6m[20] | 40.72 | 25.31 | 24.64 | 51.3 | 158.3 |
| YOLOv8m[21] | 42.93 | 27.22 | 26.48 | 23.2 | 67.5 |
| YOLOv9m[22] | 44.16 | 27.28 | 26.73 | 16.7 | 60.5 |
| YOLOv10m[23] | 43.71 | 26.98 | 26.63 | 15.3 | 58.9 |
| YOLOv11m[24] | 45.82 | 28.55 | 27.91 | 20.0 | 67.7 |
| YOLOv12m[25] | 44.20 | 27.54 | 26.94 | 19.6 | 59.5 |
| RT-DETR[15] | 47.23 | 29.91 | 28.98 | 19.8 | 57.0 |
| LDRT-DETR | 50.05 | 32.06 | 31.04 | 14.9 | 51.3 |

由表 4 可知，LDRT-DETR 在 8 种主流检测算法中综合性能最优，相较于 YOLOv5m、YOLOv6m、YOLOv8m、YOLOv9m、YOLOv10m、YOLO11m、YOLOv12m、RT-DETR 模型，*mAP*50 分别提升 7.48%、9.33%、7.12%、5.89%、6.34%、4.23%、5.85%和 2.82%。相较基线 RT-DETR 模型，*mAP*50：75 和 *mAP*50：95 优势更突出，此外，Params 和 GFLOPs 分别下降 24.7%和 10%。该结果从多维度验证了 LDRT-DETR 在无人机航拍图像检测中的优势，表明其在复杂背景干扰下具有更强的鲁棒性和优越的检测性能。

### 3.7 可视化分析

图 10 通过 4 类典型场景的可视化对比，系统评价了 LDRT-DETR 与 RT-DETR 的性能差异。在密集目标场景中，LDRT-DETR 凭借增强的特征聚合机制，实现了更高的检测置信度与更低的漏检率，有效缓解了近距目标遮挡带来的歧义；RT-DETR 则出现多处漏检，部分预测框置信度偏低。稀疏目标检测中，LDRT-DETR 对单一目标的检测较为友好，而 RT-DETR 因特征提取不足导致出现误检现象。在低对比度与远距离目标场景下，LDRT-DETR 虽保持较高召回率，但背景抑制能力减弱，误检现象明显增加；RT-DETR 则展现出更强的判别稳健性，整体性能优于 LDRT-DETR。这表明，LDRT-DETR 在低信噪比条件下的抗干扰能力有待提升。

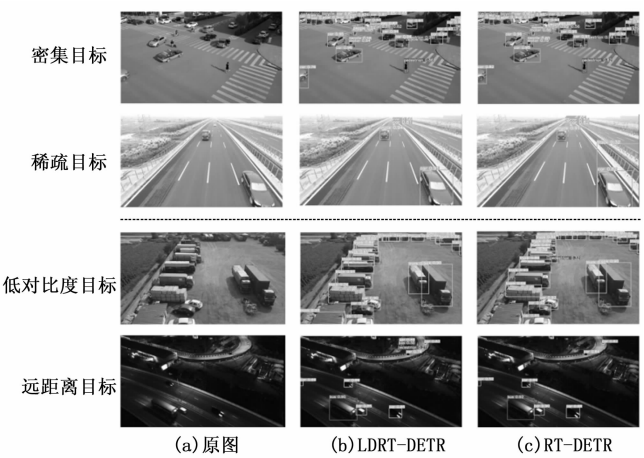


图 10 检测结果对比图

## 4 结束语

本研究设计的 LDRT-DETR 模型，旨在解决无人机航拍图像中小目标检测精度不足等关键问题。在主干网络部分，本研究引入多尺度边缘特征提取单元，通过对图像边缘纹理信息的捕获，实现对多级图像特征的高效细化，不仅使小目标检测 *mAP*50 提升至 49.05%，

还将模型参数量降低 26.8%。在特征融合阶段，设计的上下文增强特征融合模块通过跨尺度注意力权重分配，有效实现高分辨率与低分辨率特征的双向互补融合，显著增强模型对目标语义信息与空间信息的联合处理能力，*mAP*50 为 47.74%。此外，多尺度采样模块通过构建多条差异化特征提取路径，在丰富特征表达多样性的同时，进一步提升复杂场景下的检测精度 *mAP*50 为 48.20%。实验结果表明，LDRT-DETR 模型显著提高了基于无人机的监控系统在复杂场景中的应用有效性，尤其在拥挤城市环境中。未来计划扩展数据集，加入更多复杂天气（如雨天、雾天）和低光照环境等场景，以提升模型的泛化能力。

**参考文献：**


[1] 裘群禄，康泰云，陈超逸，等. 我国城市低空民用无人机管理的现状与思考 [J]. 中国军转民，2025 (16)：52－53.

[2] XU H，WANG L，HAN W，et al. A survey on UAV applications in smart city management：challenges，advances，and opportunities [J]. Selected Topics in Applied Earth Observations and Remote Sensing，IEEE Journal of (J-STARS)，2023，16：29.

[3] 林国敏，邓仕雄，刘继庚，等. 无人机多光谱遥感技术对城市水生态环境监测研究应用进展 [J]. 中国水运，2025，(18)：87－88.

[4] 杨少勇，杨浩宇. 无人机遥感技术在智慧农业中的应用与前景分析 [J]. 农业工程，2025，15 (6)：39－44.

[5] ZGOU R，GU M，HONG Z，et al. SIDE-YOLO：A highly adaptable deep learning model for ship detection and recognition in multisource remote sensing imagery [J]. IEEE Geoscience and Remote Sensing Letters，2025：1.

[6] LIU W，ANGUELOV D，ERHAN D，et al. SSD：Single Shot MultiBox Detector [C] // European Conference on Computer Vision (ECCV)，Amsterdam，Netherlands：Springer，2016：21－37.

[7] GIRSHICK R. Fast R-CNN [C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago，Chile：IEEE，2015：1440－1448.

[8] SUN X，WU P，HOI S C H. Face detection using deep learning：an improved faster RCNN approach [J]. Neurocomputing，2018，299 (JUL. 19)：42－50.

[9] YUN R，CHANGREN Z，SHUNPING X. Object detection based on fast/faster RCNN employing fully convolutional architectures [J]. Mathematical Problems in Engineering，2018：1－7.

[10] 胡　潇，潘申富. 基于改进轻量化 YOLOX 的无人机航拍目标检测算法 [J]. 计算机测量与控制，2024，32 (1)：57－63.

[11] 贾　亮，陈茂辉，王　琪，等. 改进 YOLOv8n 的航拍多尺度目标检测模型 [J]. 电子测量技术，2026，49 (1)：237－246.

[12] 罗文沛，李　军. 基于 YOLOv5s 的无人机视角下的小目标检测算法 [J]. 计算机应用，2025，45 (s1)：235－238.

[13] 李　珺，丁彬彬，史维娟，等. HM-YOLO：融合多尺度特征的轻量化航拍图像检测算法 [J]. 计算机工程与应用，2026，62 (1)：87－100.

[14] HAN K，XIAO A，WU E，et al. Transformer in transformer [J]. ArXiv Preprint ArXiv：2103.00112，2021.

[15] ZHAO Y，LV W，XU S，et al. DETRs beat YOLOs on real-time object detection [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024：16965－16974.

[16] HAN K，XIAO A，WU E，et al. ResNet in ResNet：generalizing residual architectures [J]. ArXiv Preprint ArXiv：1603.08029，2016.

[17] ZHONG Z，LIN Z Q，BIDART R，et al. Squeeze-and-attention networks for semantic segmentation [C] //2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)，IEEE，2020：1－10.

[18] DU D，ZHU P，WEN L，et al. VisDrone-DET2019：the vision meets drone object detection in image challenge results [C] //ICCV Visdrone Workshop，2019.

[19] JOCHER G，STOKEN A，BOROVEC J，et al. Ultralytics/YOLOv5：v3.0 [Z]. Zenodo，2020－08.

[20] LI C，LI L，JIANG H，et al. YOLOv6：a single-stage object detection framework for industrial applications [J]. ArXiv Preprint ArXiv：2209.02976，2022－09－07.

[21] SOHAN M，SAI RAM T，RAMI REDDY C V. A review on YOLOv8 and its advancements [C] //International Conference on Data Intelligence and Cognitive Informatics，Singapore：Springer，2024：529－545.

[22] WANG C Y，Yeh I H，LIAO H Y M. YOLOv9：learning what you want to learn using programmable gradient information [C] //Computer Vision-ECCV 2024：18th European Conference，Milan，Italy，September 29-October 4，2024，Proceedings，Part I. Cham：Springer Nature Switzerland，2024：1－21.

[23] WANG A，CHEN H，LIU L，et al. YOLOv10：real-time end-to-end object detection [C] //Advances in Neural Information Processing Systems. 2024，37：107984－108011.

[24] JOCHER G，QIU J. Ultralytics YOLO [EB/OL]. 2024. https：//github. com/ultralytics/ultralytics.

[25] TIAN Y，YE Q，DOERMANN D. YOLOv12：attention-centric real-time object detectors [J]. ArXiv Preprint ArXiv：2502.12524，2025－02－18.