

无人设备航拍行人识别算法研究

吴文逸¹, 陈睿智², 吴宪云², 刘 媛¹, 刘 勇¹,
樊新川¹, 杨晓辉¹, 田月全¹

(1. 北京航天测控技术有限公司, 北京 100041;

2. 西安电子科技大学 空天地一体化综合业务网全国重点实验室, 西安 710071)

摘要: 针对无人设备航拍图像模糊、噪点增多及高分辨率影像中行人目标像素占比少导致传统检测模型性能受限的问题, 开展航拍行人识别算法研究; 对检测模型架构与训练策略进行协同优化, 在 PP-YOLOE 基础上改进模型, 嵌入 Transformer 编码器构建长距离依赖感知网络, 增强复杂背景语义解耦能力, 并引入全局位置编码以优化空间定位; 采用基于 SAHI 滑动窗口的数据集切片策略, 配合 CLAHE 自适应直方图均衡及随机饱和度扰动等多尺度动态增强方法, 构建兼顾小目标特征保留与样本多样性的训练数据集; 经 VisDrone2019 数据集实验测试, 该方法有效识别多种检测环境下的行人目标, 较主流算法具有更高的检测精度与鲁棒性; 经实际应用验证, 满足无人设备智能巡检对高精度、高鲁棒性行人检测的需求。

关键词: 小目标检测; 航拍图像; PP-YOLOE; 注意力机制; Transformer

Research on Pedestrian Recognition algorithms for UAV Aerial Photography Equipment

WU Wenyi¹, CHEN Ruizhi², WU Xianyun², LIU Yuan¹, LIU Yong¹, FAN Xinchuan¹,
YANG Xiaohui¹, TIAN Yuequan¹

(1. Beijing Aerospace Measurement & Control Technology Co., Ltd., Beijing 100041, China;

2. State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China)

Abstract: To address the limitations of traditional detection models caused by image blur, increased noise, and low pixel occupancy of pedestrians in high-resolution UAV aerial images, research on a pedestrian recognition algorithm for aerial photography is conducted, which collaboratively optimizes the detection model architecture and training strategies. Based on the improved PP-YOLOE model, a Transformer encoder is embedded to construct a long-range dependency-aware network for enhancing semantic decoupling in complex backgrounds, and a global positional encoding is introduced to optimize spatial localization. A dataset slicing strategy based on the slicing aided hyper inference (SAHI) sliding window is adopted, combined with multi-scale dynamic augmentation methods such as the contrast limited adaptive histogram equalization (CLAHE) adaptive histogram equalization and random saturation perturbation, to construct a training dataset that balances small target feature preservation and sample diversity. Experimental tests on the VisDrone2019 dataset demonstrate that the method effectively identifies pedestrian targets under various detection environments, achieving higher detection accuracy and robustness compared to mainstream algorithms. Practical application verifies that the algorithm meets the requirements for high-precision and robust pedestrian detection in intelligent UAV inspection scenarios.

Keywords: small target detection; aerial image; PP-YOLOE; attention mechanism; Transformer

0 引言

计算机视觉 (Computer Vision) 作为人工智能的重

要分支, 在图像理解领域进展显著。目标检测作为其中的关键技术, 旨在定位图像目标的空间坐标^[1], 是图像分类、实例分割等任务的基础, 广泛应用于城市管理与

收稿日期:2025-12-10; 修回日期:2026-05-06。

作者简介:吴文逸(1993-),女,博士,工程师。

引用格式:吴文逸,陈睿智,吴宪云,等. 无人设备航拍行人识别算法研究[J]. 计算机测量与控制,2026,34(9):265-273.

安全监控中^[2]。近年无人设备平台凭借灵活的视觉感知能力,在智慧城市安防、交通监管等场景发挥重要作用^[3]。借助机载摄像头,无人设备能够突破传统视场限制,获取多角度航空图像^[4],为公共场所人群监测及异常预警提供关键支持。早期目标检测多采用滑动窗口扫描,特征提取依赖 SIFT、HOG 等手工设计特征^[5],并结合支持向量机或 AdaBoost 算法进行判别^[6]。文献^[7]提出的级联检测框架虽提升了效率,但此类手工特征对外观变化、光照波动及背景干扰敏感,泛化能力受限,难以满足复杂环境下的高精度检测需求。

随着卷积神经网络(CNN, convolutional neural network)的兴起,基于深度学习的目标检测方法逐渐取代传统方法,体现出更优的检测性能。2012 年提出的 AlexNet 模型,将图像分类准确率提升至 80% 以上,标志着基于深度学习的视觉识别进入快速发展阶段^[8]。典型的两阶段检测框架如 Faster R-CNN^[9]及 Mask R-CNN^[10]通过引入候选区域生成和池化结构等策略,显著提升检测精度。然而此类算法计算复杂度高、推理速度慢,难以适配对实时性与功耗要求严苛的无人设备嵌入式平台。与此同时以 YOLOv3^[11]、YOLOv4^[12]为代表的单阶段检测方法虽大幅提高检测速度,但其基于锚框(Anchor-based)的机制对超参数设置高度敏感。在无人设备高空俯视图场景下,行人目标尺度极小且长宽比多变,预设锚框往往难以精确匹配目标实际形状,导致背景误检或小目标漏检现象频发。相比之下,PP-YOLOE^[13]模型采用无锚框(Anchor-free)架构,摒弃了对先验框的依赖,通过变焦损失函数(VFL)与任务对齐学习(TAL)动态优化正负样本分配,有效解决小目标样本分布不均问题。结合其采用的 CSPRepResNet 主干网络^[14]与高效任务一致检测头(ET-Head),该模型在保持高推理速度的同时,展现出更优的特征提取效率与多尺度泛化能力,更适于解决航拍视角下的微小目标检测难题。

在卷积神经网络的基础上,引入注意力机制以提升模型对复杂场景的感知能力成为近年的研究热点。文献^[15]提出基于自适应协同注意力的检测模型,通过挖掘通道与空间特征关联提高密集小目标检测精度;文献^[16]设计 SNA-YOLOv5s 网络,引入空间转换卷积以保留小目标细节;文献^[17]融合多尺度特征与空间注意力模块,在无人设备行人检测任务中兼顾高精度与实时性。此外文献^[18]和文献^[19]在网络中嵌入卷积块注意力模块和空间金字塔池化(SPP, spatial pyramid pooling)结构,增强对遮挡行人的检测能力。

航拍图像易受传输干扰、光照变化、前景遮挡影响,数据质量不稳定,给现有算法带来较大挑战^[20]。本文在已有研究基础上,提出一种基于 PP-YOLOE 的

航拍行人检测优化模型。通过引入长距离依赖的 Transformer 编码器与全局位置编码模块增强上下文感知能力;设计滑动窗口切片与多尺度数据增强相结合的训练策略,以提高模型对小目标特征的提取能力。后续章节通过实验验证所提方法在 VisDrone2019 数据集上的性能,并与现有主流方法进行对比分析。

1 行人小目标数据增强策略

无人设备航拍图像兼具视场广阔与目标微小的矛盾特性,行人目标往往仅占据极少像素区域,且易受飞行高度波动、拍摄角度倾斜及复杂背景杂波的干扰。常规的数据预处理方式难以在保留高分辨率细节的同时适配检测网络输入要求。为此采用基于滑动窗口的高分辨率切片策略与多尺度动态增强方法,旨在最大化保留小目标纹理特征并提升模型对光照与尺度的泛化能力。

1.1 SAHI 滑动窗口切片

工业级无人设备采集的影像分辨率通常在 $1\,920 \times 1\,080$ 以上,而 PP-YOLOE 等主流检测模型的输入层通常被限制在 640×640 以平衡推理速度。若直接采用传统线性插值法进行降采样,图像中微小行人目标将因像素混叠而丢失关键纹理信息,导致特征提取层无法捕捉有效梯度。传统的图像切片方法多采用非重叠的固定网格裁剪,这种刚性分割机制极易导致位于切割边界的目标被截断。对于在航拍图像中仅占数个像素的行人目标而言,边界截断意味着关键几何特征的破坏,将直接导致漏检。

针对上述问题,引入 SAHI (Slicing Aided Hyper Inference)^[21]策略进行图像预处理。与传统方法不同,SAHI 通过设定重叠率维持目标的空间连续性。如图 1 所示,设定一个尺寸为 $W \times H$ 的滑动窗口,选定水平和垂直方向的重叠比例 α ,滑动窗口以步长 S_w 和 S_h 在图像上扫描,实现重叠切片。计算公式如下:

$$S_w = W(1 - \alpha), S_h = H(1 - \alpha) \quad (1)$$

该策略确保任意位置的行人目标至少完整出现在一个子图中,有效避免了边界截断效应。每个子图独立输入网络推理后,利用非极大值抑制方法(NMS, non-maximum suppression)将子图的局部检测框映射回原图的全局坐标系。NMS 依据交并比(IoU, intersection

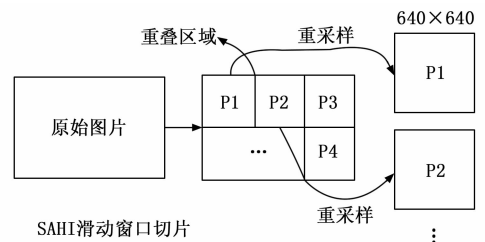


图 1 SAHI 数据切片原理示意

over union) 阈值筛选最优检测框, 对于两个检测框 B_i 和 B_j , IoU 计算方法如式 (2)。后续实验表明, 该方法在处理 VisDrone2019 等高分辨率数据集时, 能够显著缓解小目标特征在降采样过程中的耗散问题。

$$IoU(B_i, B_j) = \frac{B_i \cap B_j}{B_i \cup B_j}$$

(2)

1.2 多尺度训练与动态数据增强

1) 多尺度训练:

航拍视角的垂直变化导致同类目标在图像中的像素尺寸呈现剧烈波动。为解决单分辨率训练导致的模型尺度固化问题, 构建动态多尺度输入金字塔。设定尺寸缩放集合 $S = \{0.5, 0.75, 1.0, 1.25, 1.5\}$, 在训练迭代中随机选取缩放系数 $\lambda \in S$, 利用双线性插值重采样生成不同分辨率的输入张量, 以模拟无人设备不同飞行高度下的成像效果, 强迫特征提取网络学习具有尺度不变性的特征表达。

2) 动态环境适应增强:

针对航拍图像常伴随的强光斑、阴影及大气散射引起的低对比度问题, 传统全局直方图均衡化易放大背景噪声。为此引入限制对比度自适应直方图均衡算法 (CLAHE, contrast limited adaptive histogram equalization)^[22]。CLAHE 算法将图像划分为若干小块, 在每个块内独立进行直方图均衡并限制对比度过度增强, 最后通过双线性插值消除块间边界伪影。该方法在有效提升暗部行人轮廓清晰度的同时, 抑制平坦区域的噪声信号放大, 使模型能更好地适应光照剧烈变化的监测场景。

定义图像像素亮度阈值区间 $[\theta_{low}, \theta_{high}]$ 以及像素亮度方差阈值 σ_{std} , 当输入图像的平均亮度值 $\mu_l \notin [\theta_{low}, \theta_{high}]$ 以及亮度方差 $\sigma_l < \sigma_{std}$ 时, 表明图像可能存在明显的过曝或者过暗现象, 使用 CLAHE 算法对图像进行处理。平均亮度值 μ_l 和亮度方差 σ_l 的表达式如下:

$$\mu_l = \frac{1}{N} \sum_{x,y} I(x,y)$$

(3)

$$\sigma_l = \sqrt{\frac{1}{N} \sum_{x,y} [I(x,y) - \mu_l]^2}$$

(4)

无人设备航拍图像常为 8 位彩色图像, 使用 $[0, 255]$ 区间的整数值对图像像素点亮度进行量化表示。设定亮度阈值区间为 $[35, 225]$, 亮度方差阈值为 35, 使用 OpenCV 实现图像亮度统计和 CLAHE 算法处理。以图 2 (a) 为例, 计算图像的平均亮度 $\mu_l = 14.67$, 亮度方差 $\sigma_l = 24.61$, 图像存在过暗现象, 难以分辨行人小目标信息。图 2 (b) 为经 CLAHE 算法处理后的图像, 平均亮度 $\sigma_l = 36.25$, 亮度方差。可以看到图像有了明显的明暗区分度, 暗部细节得到了增强, 并抑制了处理过程中产生的噪声。

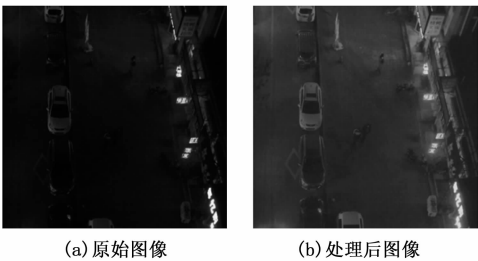


图 2 CLAHE 算法处理效果对比

此外结合随机饱和度扰动策略, 将图像转换至 HSV 空间, 并对饱和度通道 S 施加随机线性缩放 $\delta \sim U[0.5, 1.5]$, 以此模拟不同天气条件或传感器色偏导致的色彩失真, 迫使模型关注目标的纹理结构而非单一色彩特征, 从而提升在复杂背景下的辨识能力。色彩空间转换计算如式 (5) 所示。

$$\begin{cases} V = \max(R, G, B) \\ S = \frac{V - \min(R, G, B)}{V} \end{cases}$$

(5)

通过上述多尺度与动态增强联合策略, 训练数据集在保持小目标特征的同时, 获得了更多样化的样本分布。行人小目标数据增强流程如图 3 所示。原始图像经 SAHI 切片后生成子图, 子图在训练时随机调整分辨率, 并根据需要应用 CLAHE 和饱和度增强等操作。这样的训练集设计使得模型能够兼顾对高分辨率细节信息和多样化样本特征的学习, 提高了对复杂背景、小尺寸目标以及多变光照条件的检测能力。

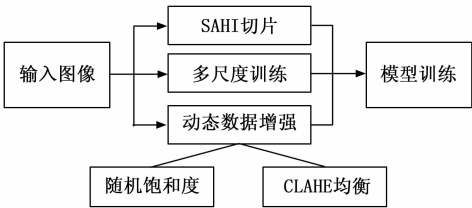


图 3 数据增强流程

2 行人小目标检测方法

PP-YOLOE 模型是百度飞桨团队基于 YOLO 系列优化而成的工业级目标检测算法, 核心架构如图 4, 包含改进的跨阶段局部残差网络 (CSPRepResNet)、路径聚合网络 (PAN, path aggregation network) 以及高效

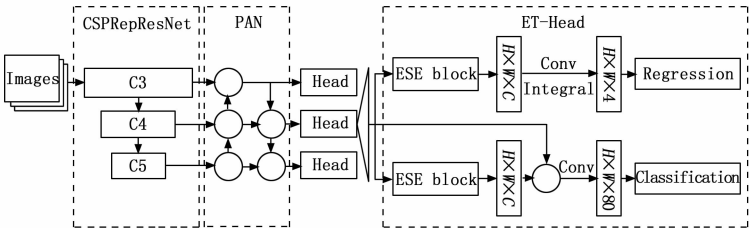


图 4 PP-YOLOE 模型结构

的任务一致检测头 (ET-Head, efficient task-aligned head) 三个部分。输入图像首先通过 CSPRepResNet 提取多尺度特征。该网络采用跨阶段部分连接结构 (Cross Stage Partial Connections) 将特征通道分割为两部分, 一部分直接传递至下一阶段, 另一部分经过残差堆叠处理后再融合。这种设计在降低计算量的同时保留梯度多样性, 其数学表达可简化为:

$$X_{\text{out}} = \text{Concat}[X_{\text{pass}}, \text{ResBlock}(X_{\text{process}})] \quad (6)$$

X_{pass} 表示直连通道, X_{process} 为经过残差块处理的通道。主干网络分别输出 C3、C4、C5 特征图, 尺度分别为原图的 1/8、1/16、1/32, 然后进入 PAN 进行跨层特征融合。PAN 通过自上而下和自下而上的双向连接, 将深层语义信息与浅层细节特征结合, 增强多尺度目标检测能力^[23]。融合后的特征输入至 ET-Head, 该模块采用轻量化设计并引入增强语义提取模块 (ESE, enhanced semantic extraction)^[24]。ESE 通过通道注意力机制对特征图全局平均池化生成通道描述向量, 再通过全连接层与 Sigmoid 函数计算各通道重要性系数:

$$\alpha_c = \sigma \left\{ W_2 \cdot \delta \left[W_1 \cdot \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W X_c(i, j) \right] \right\} \quad (7)$$

其中: W_1 、 W_2 为可学习参数, σ 为 Sigmoid 激活函数, 最终输出特征为 $X_{\text{out}} = \alpha_c \cdot X_c$ 。ET-Head 将分类与回归任务解耦为两个分支, 分别输出 80 维类别概率和 4 维边界框坐标, 并通过积分运算实现无锚框预测。

3 PP-YOLOE 算法优化

3.1 Transformer 注意力机制

Transformer 编码器利用自注意力机制与前馈网络的协同作用, 建模输入特征间的全局语义关联。如图 5 所示, 其基本单元由多头自注意力层和前馈网络堆叠构成, 通过残差连接与层归一化确保训练稳定性^[25]。自注意力机制用于计算特征空间中任意位置的关联权重, 建立长距离依赖关系。定义 n 为输入数据序列长度, d 为特征维度, 给定输入特征 $X \in \mathbb{R}^{n \times d}$, 通过线性投影生成查询矩阵 Q 、键矩阵 K 和值矩阵 V :

$$Q = XW^Q, K = XW^K, V = XW^V \quad (8)$$

其中: $W^Q, W^K, W^V \in \mathbb{R}^{d \times d_i}$ 为可学习的参数矩阵, 然后采用缩放点积注意力计算样本的特征关联, 使每个位置的特征能够聚合全局上下文信息。计算公式如下, 其中 Softmax 表示归一化指数函数。

$$\text{Attention}(Q, K, V) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (9)$$

使用前馈神经网络 (FFN, feed-forward neural network) 对特征进行非线性变换, FFN 由两层全连接层与高斯误差线性激活函数 (GELU, gaussian error linear unit) 构成, 每个子层的输出与输入进行残差连

接, 并通过层归一化抑制梯度消失, 使编码器既能建模全局依赖, 又能逐层提炼高层次语义特征。计算方式如式 (10) 所示:

$$\begin{cases} \text{FFN}(x) = \text{GELU}(xW_1 + b_1)W_2 + b_2 \\ x' = \text{LayerNorm}[x + \text{MultiHead}(x)] \\ x'' = \text{LayerNorm}[x' + \text{FFN}(x')] \end{cases} \quad (10)$$

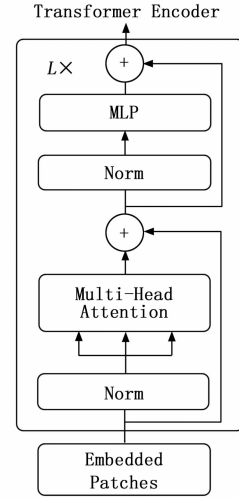


图 5 Transformer 编码器结构图

3.2 全局位置编码

经典卷积网络通过堆叠卷积核隐式捕获位置关系, 随着网络层数加深, 感受野扩大导致局部绝对位置信息被平滑, 难以在广域视野下精确建模微小目标的几何分布。此外常规的可学习位置嵌入 (Learnable Position Embedding) 依赖特定尺寸的网格参数, 在处理不同分辨率的航拍输入时易出现过拟合或插值误差。为此引入基于正弦与余弦函数的全局位置编码, 为网络输出的特征图添加显式的绝对空间坐标约束。给定特征图空间坐标 (x, y) , 其对应的位置编码向量 $PE_{x, y, 2i}$ 和 $PE_{x, y, 2i+1}$ 按通道维度交替生成, 计算依据如下:

$$\begin{cases} PE_{x, y, 2i} = \sin \left(\frac{x}{1000^{2i/C}} \right) \\ PE_{x, y, 2i+1} = \cos \left(\frac{y}{1000^{2i/C}} \right) \end{cases} \quad (11)$$

其中: i 表示当前维度索引, C 表示特征通道总数。得到的 PE 向量与原始特征图相加, 形成位置感知特征:

$$\tilde{X} = X + PE \quad (12)$$

该编码方式通过不同频率的三角函数构建连续的空间流形。理论上, i 较小的低频分量对应长波信号, 能够有效编码全局绝对坐标, 辅助模型在复杂的航拍大场景中区分不同区域的语义上下文; i 较大的高频分量对应短波信号, 对位置变化高度敏感, 能够保留微小行人的细粒度边界信息。这种多尺度频率特性与航拍图像中“视场广阔但目标微小”的空间分布特点高度契合。同

时正弦函数的平移不变性使得编码具备良好的外推能力, 当图像经过多尺度训练或随机裁剪等增强操作时, 位置编码仍能自适应维持相对空间关系的稳定性, 从而弥补 Transformer 结构本身缺乏归纳偏置 (Inductive Bias) 的缺陷, 提升模型对密集小目标的定位精度。

3.3 PP-YOLOE-ATT 改进模型

航拍图像中行人目标通常尺度微小、分布密集且背景复杂, 要求检测模型兼具全局语义感知与精细空间定位能力。原 PP-YOLOE 模型通过深层卷积获取较大感受野, 但受限于卷积核的局部性, 缺乏对长距离上下文依赖的捕捉能力, 且多次下采样导致细粒度位置信息丢失。为此提出 PP-YOLOE-ATT 改进模型, 在 CSPRepResNet 主干网络与 PAN 特征金字塔之间嵌入 Transformer 编码器与全局位置编码模块, 模型结构如图 6 所示。

为解决 Transformer 编码器对输入序列的维度要求与 CNN 输出三维特征张量不匹配的问题, 设计特征空间的维度重塑与还原机制。模型提取主干网络输出的深层特征图 $C_5 \in \mathbb{R}^{C \times H \times W}$, 其中 C 为通道数, H, W 为空间分辨率。在输入到 Transformer 编码器前执行空间展平操作 (Flatten), 将特征图转换为序列特征 $X_{\text{seq}} \in \mathbb{R}^{HW \times C}$ 。该序列特征与同维度的全局位置编码向量相加后, 输入到包含 4 层堆叠结构的 Transformer 编码器中。每层分配 8 个注意力头与 128 维键值空间; 隐藏层维度为 2 048, 前馈网络扩展比例为 4; FFN 层采用 GELU 激活函数与 Pre-LN 训练策略。编码器利用多头自注意力机制在序列维度上计算全局关联, 输出蕴含长距离依赖的上下文特征序列。随后引入逆变换操作 (Reshape), 将编码器输出的序列特征重新映射回 $\mathbb{R}^{C \times H \times W}$ 的三维空间结构, 使其能无缝接入 PAN 网络进行后续的多尺度特征融合。这种串行融合策略在保持原检测头 ET-Head 结构不变的前提下, 通过维度变换有效桥接了卷积网络与 Transformer 结构。模型的损失函数设计如下:

$$Loss = \frac{\alpha \cdot loss_{\text{VFL}} + \beta \cdot loss_{\text{GloU}} + \gamma \cdot loss_{\text{DFL}}}{\sum_i^{N_m} \hat{t}} \quad (13)$$

其中: $\hat{t}$ 表示目标的标准得分; VFL 函数用于对

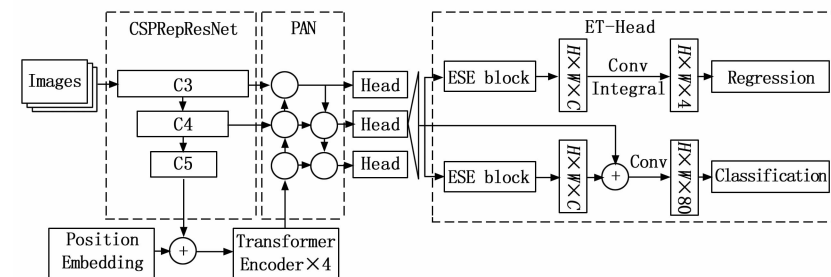


图 6 PP-YOLOE-ATT 模型结构

模型的分类任务优化, 减少简单负样本对训练的干扰; 分布焦点损失函数 (DFL, distribution focal loss) 和广义交并比损失函数 (GIoU, generalized intersection over union) 用于提升模型回归稳定性, 加快模型收敛, 提升小目标的检测精度。

4 实验与分析

4.1 数据集构建

实验选取 VisDrone2019 公开数据集作为基础数据源。该数据集由天津大学机器学习与数据挖掘实验室团队构建, 包含 10 209 张静态图像, 场景涵盖城市道路、广场、农田等多种复杂环境, 包含行人 (pedestrian/people)、车辆 (car/van/bus/truck) 等十类目标, 充分反映无人设备航拍图像中存在的视角多变、光照不均及背景复杂等特性。为算法在运动模糊、目标遮挡、多尺度适应性等真实场景下的性能评估提供丰富样本。

1) 图像质量量化筛选:

航拍过程中无人设备的高速运动与气流抖动常导致图像产生运动模糊。采用拉普拉斯方差算法 (Variance of Laplacian) 对图像清晰度进行量化评估^[26]。定义拉普拉斯核算子:

$$\mathbf{K} = \frac{1}{6} \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & 1 \\ 0 & -1 & 0 \end{bmatrix} \quad (14)$$

对于离散图像使用拉普拉斯算子进行卷积运算, 并输出图像的方差值 D_{var} 来衡量像素分布的离散程度计算公式如式 (15)。其中 $L(x, y)$ 为卷积后的像素值, 计算公式如式 (16)。

$$D_{\text{var}} = \sum_x \sum_y [L(x, y) - L]^2 \quad (15)$$

$$L(x, y) = (I * K)(x, y) \quad (16)$$

设定清晰度阈值 $T_{\text{blur}} = 100$, 当 $D_{\text{var}} < T_{\text{blur}}$ 时, 判定图像存在严重运动模糊或离焦, 将其从数据集中剔除。同时移除分辨率低于像素的低质量图像, 以避免过度上采样引入的人工伪影。

2) 亮度异常判定与自适应增强:

基于 HSV 色彩空间对图像亮度分量 V 进行统计分析。依据 1.2 节所述标准, 计算图像平均亮度 μ 与亮度方差 σ^2 。设定正常亮度区间为 $[35, 225]$ 。对于 $\mu \in [35, 225]$ 且 $\sigma^2 \geq 35$ 的图像, 保留原始数据; 对于 $\mu \notin [35, 225]$ 或 $\sigma^2 < 35$ 的图像, 即存在严重过曝、过暗或对比度缺失, 应用 CLAHE 算法进行自适应直方图均衡化处理, 修复暗部细节并抑制高光溢出; 对于经处理后仍无法满足视觉识别要求的极端异常样本, 予以剔除。经筛选与预处理, 最

终构建包含 5 338 张训练集、519 张验证集及 1 194 张测试集的实验数据分布。

3) 小目标尺度统计与 SAHI 切片:

对 VisDrone2019 数据集中行人目标尺寸进行分析。原始图像平均分辨率为 $1\,518 \times 1\,009$, 而行人目标平均宽度仅占图像宽度的 0.9%, 平均高度仅占图像高度的 2.6%, 属于典型的小目标检测任务。为解决高分辨率图像直接缩放导致的信息丢失问题, 应用 1.1 节提出的 SAHI 滑动窗口切片策略。设定切片窗口大小 $W=H=640$, 重叠率 $\alpha=0.3$, 对应滑动步长为 448 像素。经切片处理, 共生成 39 300 张子图。统计结果如图 7 所示, 以行人目标平均尺寸占图像尺寸 4% 为基准, 小于 4% 的极小目标占比由原图的 89.7% 显著下降至 57.3%。该策略有效提升了行人目标在训练样本中的像素占比, 为模型学习细粒度纹理特征提供了高质量的数据基础。

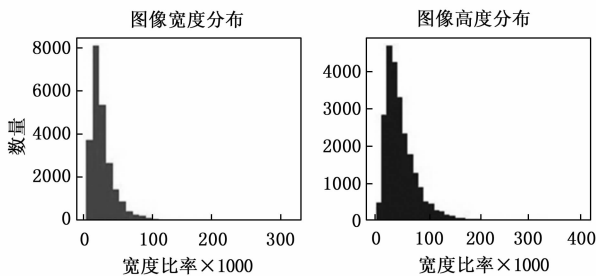


图 7 切片后子图行人目标尺寸占比分布

4.2 实验参数设置

本文使用的深度学习框架为 Paddle 和 Pytorch, 操作系统使用 Ubuntu20.04, 模型训练硬件平台选用一张 NVIDIA GeForce RTX 4090 24G 型号显卡, 推理平台选用一张 NVIDIA GeForce RTX 3060 Laptop 12G 型号显卡。

PP-YOLOE 模型与 PP-YOLOE-ATT 模型使用 ppyoloe_crn_l_obj365_pretrained. pdparams 模型作为网络的初始化权重。采用余弦退火策略动态调整学习率, 前 20 次迭代中逐步提升至 0.001 25, 批次大小为 8, 共计 80 Epoch。余弦退火策略的表达式为:

$$\eta = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left[1 + \cos\left(\frac{t}{T}\pi\right) \right] \quad (17)$$

其中: T 为退火周期, t 为当前迭代步数, η 为学习率。

对比实验选取 YOLOv5 模型及 RetinaNet 模型。YOLOv5 以其高推理速度成为轻量化检测的标杆, 但其小目标检测精度受限于锚框设计, 表现略有欠缺。RetinaNet 通过焦点损失 (Focal Loss) 缓解前景-背景样本不平衡问题, 在遮挡场景中表现稳定, 但计算复杂度较高^[27]。两类模型使用 Pytorch 作为深度学习框架, 训

练参数为默认设置, 输入图像分辨率为 640×640 。

模型评价指标使用 mAP_{50} 和 $mAP_{50:95}$ 。 mAP_{50} 侧重考察目标是否存在的判断能力, $mAP_{50:95}$ 侧重考察定位精度与模型鲁棒性。 mAP_{50} 和 $mAP_{50:95}$ 源于 mAP 指标。 mAP 计算基于每个类别的平均精度 AP , 如式 (18) 所示, 其中 k 表示类别数。 mAP_{50} 为 IoU 阈值为 0.5 时的 mAP 值, 即当 IoU 阈值设为 0.5 时, 预测框与真实框的重叠区域超过 50% 即判定为正确检测。 $mAP_{50:95}$ 设定 IoU 从 0.5 到 0.95 等间隔取 10 个递增阈值, 计算各阈值 AP 并取平均值得到。

$$mAP = \frac{1}{k} \sum_{j=1}^k AP_j \quad (18)$$

4.3 结果分析

4.3.1 模型性能评估

在 VisDrone2019 数据集上对四类模型进行性能测试, 结果如表 1 所示。实验重点验证 Transformer 融合策略对模型性能的提升效果。对比基准模型 PP-YOLOE 与改进后的 PP-YOLOE-ATT 模型数据可知, 在引入 Transformer 编码器及特征维度适配结构后, 模型 mAP_{50} 指标由 57.4% 提升至 68.3%, 增幅达 10.9%; $mAP_{50:95}$ 指标由 29.0% 提升至 34.7%。这一显著的精度增益证明, 所设计的“展平-编码-重塑”融合机制能够有效解决异构网络间的特征传递问题, 成功将 Transformer 的全局注意力优势引入检测框架。

表 1 不同算法的性能表现

检测模型	RetinaNet	YOLOv5	PP-YOLOE	PP-YOLOE-ATT
mAP_{50}	0.395	0.521	0.574	0.683
$mAP_{50:95}$	0.201	0.242	0.290	0.347
Speed/ms	10.63	5.6	23.41	27.18

尽管 PP-YOLOE-ATT 模型的推理时间增至 27.18 ms, 较原模型增加约 16.1%, 但仍显著优于 RetinaNet 模型, 且精度远超轻量化的 YOLOv5 模型 (mAP_{50} 为 52.1%)。表明该改进模型在可接受的计算成本增加范围内, 显著增强了对复杂背景下微小行人目标的感知能力, 验证了 CNN 与 Transformer 协同优化的有效性。

不同模型在数据集上的检测效果如图 8 所示, YOLOv5 模型能够识别大部分行人, 但对处于画面边缘、距离较远的行人识别能力较差, 如图 8 (a-2) 和图 8 (a-3), 这些行人在图像中尺度很小, 容易丢失特征。PP-YOLOE 模型对这些小目标行人的识别率有所提升, 但是对背景复杂或姿态变化的行人识别效果较差, 如图 8 (b-3) 中白色面包车右侧坐在站台上的行人没有被识别到。PP-YOLOE-ATT 模型则对上述存在的问题有着更好的识别效果, 如图 8 (c-2), 模型不仅能够识别清

晰完整的行人目标,对距离较远的行人目标和位姿变化的行人目标也有较好的识别能力。

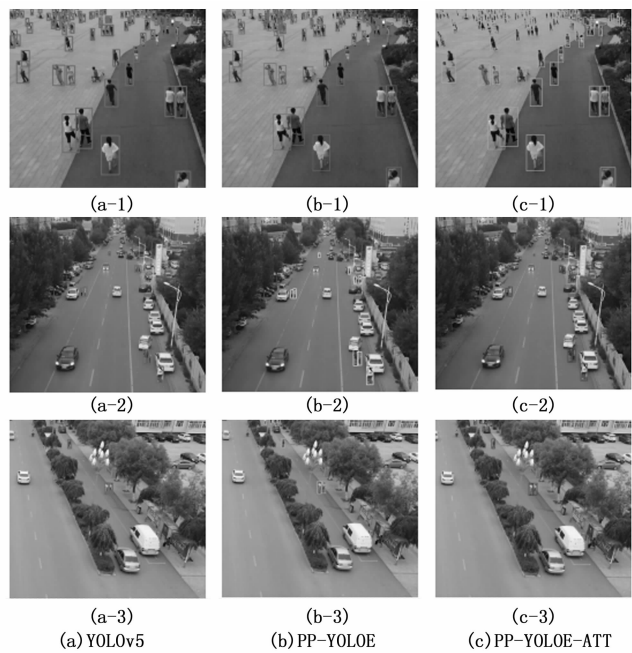


图 8 不同模型的检测效果

实验结果的差异充分验证了改进策略的有效性。检测效果的提升得益于全局位置编码与 Transformer 的协同作用。单纯的 CNN 结构在高层语义提取中易丢失微小目标的空间坐标信息,而引入基于频率分解的全局位置编码后,模型在特征交互阶段显式地保留了多尺度的空间约束。这使得检测头在回归微小行人边界框时,不仅依赖局部纹理,还能利用精确的全局位置先验,从而在复杂背景干扰下实现更紧致、准确的框定位,有效解决了航拍场景中常见的小目标定位漂移问题。

4.3.2 SAHI 切片性能增益

PP-YOLOE-ATT 模型在 VisDrone2019 切片数据集上的性能表现如表 2 所示。表 2 将数据集的训练方法分为三类,“不切片”表示直接使用原图进行训练和评估检测精度,“切片并评估原图”表示使用切片数据集训练模型但使用原图进行精度评估,“切片并评估子图”表示使用切片数据集训练并使用切片子图进行精度评估。

表 2 切片数据集对 PP-YOLOE-ATT 模型的性能增益			
训练方法	不切片	切片并评估原图	切片并评估子图
$mAP50$	0.683	0.620	0.748
$mAP50:95$	0.347	0.283	0.415
Speed/ms	27.18	27.86	22.57
FPS	37	36	44

实验表明,采用“切片并评估子图”的策略,其 $mAP50$ 达到 74.8%, $mAP50:95$ 达到 41.5%,较

“不切片”训练方式分别提升 12.8% 与 6.8%。这一显著增益表明 SAHI 切片能够有效克服传统降采样导致的信息耗散问题,其重叠冗余机制在保障高分辨率纹理输入的同时,极大提升模型对边缘区域微小目标的捕获能力。结合图 9 的检测效果进一步验证数据多尺度动态增强策略的鲁棒性。图 9 所示图片中存在密集且重叠的行人目标,采用不切片训练的 PP-YOLOE-ATT 模型因输入分辨率受限,在图 9 (a) 中难以区分粘连个体的边界特征;而引入 SAHI 切片训练后,如图 9 (b) 所示,模型在保留局部高频信息的基础上,成功在重叠区域解析出多个独立的行人目标。

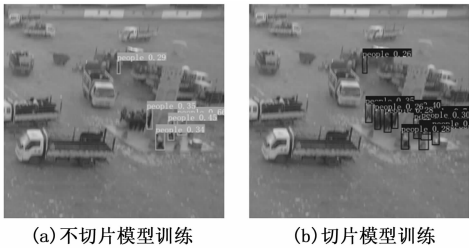


图 9 PP-YOLOE-ATT 模型在切片数据集上的性能表现

此外切片并评估子图的方法推理速度为 22.57 ms,平均帧率达 44 FPS,较不切片方法提升 20.4%。结果表明,提出的结合 SAHI 切片与多尺度动态增强的策略,不仅在理论上解决了小目标特征匮乏与环境适应性差的难题,在实际应用中亦满足工业级无人设备对 1 080 P、2 K 乃至 4 K 不同视频流的实时处理需求,在检测精度与推理效率之间取得了良好平衡。

4.4 消融实验

4.4.1 数据增强对性能的影响

为验证多尺度数据增强对模型检测精度的影响,使用未经增强处理的 VisDrone2019 数据集训练模型并统计检测精度,结果如表 3 所示。对比数据显示,在不使用数据增强方法的情况下,PP-YOLOE-ATT 模型 $mAP50$ 精度下降 3.1%,切片训练模型 $mAP50$ 精度下降 1.2%。这表明本文提出的多尺度动态增强及 CLAH-E 策略能够有效提升训练数据的纹理清晰度与对比度,有助于模型在复杂背景下捕捉行人目标特征。值得注意的是,数据增强对“不切片”训练方式下的模型性能提升更为显著,说明在图像降采样导致部分信息丢失的情况下,通过增强图像特征可有效补偿分辨率不足带来的性能损耗。

表 3 数据增强对 PP-YOLOE-ATT 模型性能影响				
数据集类型	预处理		未处理	
训练方法	不切片	切片	不切片	切片
$mAP50$	0.683	0.748	0.652	0.736
$mAP50:95$	0.347	0.415	0.324	0.400

4.4.2 模型性能验证

为评估模型在真实场景中的泛化能力,利用无人设备采集实际生活场景视频样本共计 30 分钟。依据光照条件将样本分为三组:第一组平均亮度约为 35,模拟夜间低照度巡检环境;第二组平均亮度约为 225,模拟强光过曝场景;第三组为正常光照场景。选取不同场景、不同视角的视频帧进行对比分析,检测效果如图 10 所示。无论在黑暗还是明亮场景下,PP-YOLOE-ATT 模型均能准确识别目标。如图 10 (c-1) 所示,模型对存在遮挡的行人目标亦具备良好的分辨能力。综合各项实验结果,经过多尺度数据增强与切片训练优化的 PP-YOLOE-ATT 模型展现出良好的泛化性能,能够适应实际的行人巡检任务需求。

然而模型针对困难样本的检测能力较简单样本仍有差距。如图 10 (a-3) 与图 10 (b-3),当行人与阴影区分度极低,或图像存在严重运动模糊导致纹理细节大量丢失时,会出现漏检现象。上述局限性为后续在极端环境下的模型优化提供了明确方向。

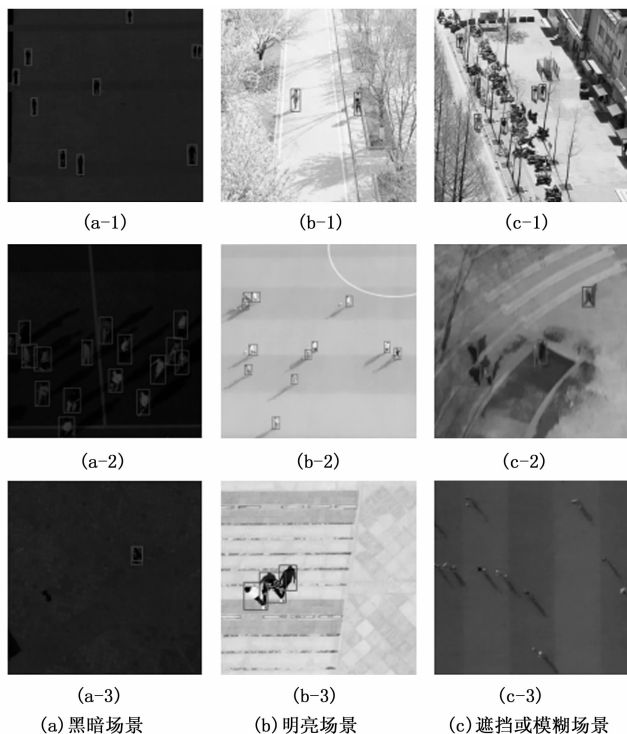


图 10 PP-YOLOE-ATT 模型性能验证

5 结束语

针对无人设备航拍场景下行人小目标难以高精度检测的问题,基于 PP-YOLOE 构建融合 Transformer 编码及全局位置编码的检测框架,并结合切片与数据增强策略提升远距离、小尺度与复杂背景条件下的特征表达能力,在多类复杂场景中均表现出更好的鲁棒

性和适应性。该类方法在智慧城市巡检、交通监控与应急救援等领域具有良好应用潜力,可为无人设备视觉感知系统的设计与工程落地提供有益参考。同时自注意力结构带来的计算负担、极端光照与强模糊场景下的可靠性下降以及泛化受限仍是影响实际部署的关键因素。后续可从模型轻量化、时序信息建模、多模态融合以及半监督学习等方向进一步完善无人设备视角下的行人检测技术。

参考文献:

- [1] LIU L, OUYANG W L, WANG X G, et al. Deep learning for generic object detection: a survey [J]. International Journal of Computer Vision, 2020, 128 (2): 261 - 318.
- [2] ZOU Z X, CHEN K Y, SHI Z W, et al. Object detection in 20 years: a survey [J]. Proceedings of the IEEE, 2023, 111 (3): 257 - 276.
- [3] MOHSAN S A H, KHAN M A, NOOR F, et al. Towards the unmanned aerial vehicles (UAVs): a comprehensive review [J]. Drones, 2022, 6 (6): 147.
- [4] KANELAKIS C, NIKOLAKOPOULOS G. Survey on computer vision for UAVs: current developments and trends [J]. Journal of Intelligent & Robotic Systems, 2017, 87 (1): 141 - 168.
- [5] 贾迪, 朱宁丹, 杨宁华, 等. 图像匹配方法研究综述 [J]. 中国图象图形学报, 2019, 24 (5): 677 - 699.
- [6] MA J, JIANG X, FAN A, et al. Image matching from handcrafted to deep features: A survey [J]. International Journal of Computer Vision, 2021, 129 (1): 23 - 79.
- [7] VIOLA P, JONES M J. Robust real-time face detection [J]. International Journal of Computer Vision, 2004, 57 (2): 137 - 154.
- [8] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60 (6): 84 - 90.
- [9] REN S, HE K, GIRSHICK B R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [C] // Advances in Neural Information Processing Systems, 2015: 91 - 99.
- [10] HE K, GKIOXARI G, DOLLAR P, et al. Mask R-CNN [C] // IEEE International Conference on Computer Vision (ICCV), 2017: 2980 - 2988.
- [11] REDMON J, FARHADI A. YOLOv3: An Incremental Improvement [Z]. ArXiv: 1804.02767, 2018.
- [12] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection [Z]. ArXiv: 2004.10934, 2020.

- [13] XU S, WANG X, LV W, et al. PP-YOLOE: an evolved version of YOLO [Z]. ArXiv: 2203.16250, 2022.
- [14] WANG C Y, LIAO H Y M, YE H I H, et al. CSPNet: a new backbone that can enhance learning capability of CNN [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020: 1571–1580.
- [15] 李子豪, 王正平, 贺云涛. 基于自适应协同注意力机制的航拍密集小目标检测算法 [J]. 航空学报, 2023, 44 (13): 244–254.
- [16] 张信佳, 王 芳. 基于多层次特征融合和注意力机制的无人设备图像小目标检测算法 [J]. 计算机工程, 2026, 52 (2): 148–157.
- [17] SHAO Z F, CHENG G, MA J Y, et al. Real-Time and accurate UAV pedestrian detection for social distancing monitoring in COVID-19 pandemic [J]. IEEE Transactions on Multimedia, 2022, 24: 2069–2083.
- [18] 常 青, 韩 文, 王清华, 等. 改进 YOLO 轻量化网络的行人检测算法 [J]. 系统仿真学报, 2022, 34 (1): 80–85.
- [19] XIANG N, WANG L, JIA C L, 等. 改进 YOLO 的遮挡行人检测仿真 [J]. 系统仿真学报, 2023, 35 (2): 286–299.
- [20] MITTAL P, SINGH R, SHARMA A. Deep learning-based object detection in low-altitude UAV datasets: a survey [J]. Image and Vision Computing, 2020, 104: 104466.
- [21] AKYON F C, ALTINUC S O, TEMIZEL A. Slicing aided hyper inference and fine-tuning for small object detection [C] //2022 IEEE International Conference on Image Processing (ICIP), 2022: 966–970.
- [22] CHANG Y, JUNG C, KE P, et al. Automatic contrast-limited adaptive histogram equalization with dual gamma correction [J]. IEEE Access, 2018, 6: 11782–11792.
- [23] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation [C] //IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 8759–8768.
- [24] LEE Y, PARK J. CenterMask: real-time anchor-free instance segmentation [C] //IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 13903–13912.
- [25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C] //Advances in Neural Information Processing Systems (NIPS), 2017: 5998–6008.
- [26] PECH-PACHECO J L, CRISTOBAL G, CHAMORRO-MARTINEZ J, et al. Diatom autofocusing in brightfield microscopy: a comparative study [C] //Proceedings 15th International Conference on Pattern Recognition (ICPR), 2000, 3: 314–317.
- [27] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C] //IEEE International Conference on Computer Vision (ICCV), 2017: 2980–2988.
- [28] KOURTESSIS P, et al. Impact of mobility on the performance of massive MIMO systems with channel aging [J]. IEEE Access, 2020, 8: 20965–20978.
- [15] CHOPRA A, MURTI Y, CHOCKALINGAM A. On the performance of zero-forcing and matched filter receivers in massive MIMO with channel aging [J]. IEEE Communications Letters, 2018, 22 (9): 1886–1889.
- [16] 孙 俊, 盛 伟, 屈家安, 等. 面向 6G 无人机通信网络的信号传输机理研究与分析 [J]. 无线电通信技术, 2025, 51 (1): 86–95.
- [17] NAM Y H, NG B L, SAYANA K, et al. Full-dimension MIMO (FD-MIMO) for next generation cellular technology [J]. IEEE Communications Magazine, 2013, 51 (6): 172–179.
- [18] ÖZDOĞAN R, BJÖRNSSON E, LARSSON E G. Massive MIMO with spatially correlated Rician fading channels [J]. IEEE Transactions on Communications, 2019, 67 (5): 3234–3250.
- [19] 3GPP TR 36.873. Study on 3D channel model for LTE (Release 12) [S]. 2015.
- (上接第 249 页)
- [9] NGO H Q, LARSSON E G, MARZETTA T L. Energy and spectral efficiency of very large multiuser MIMO systems [J]. IEEE Transactions on Communications, 2013, 61 (4): 1436–1449.
- [10] KHAWAJA W, GUVENC I, MATOLAK D W, et al. A survey of air-to-ground propagation channel modeling for unmanned aerial vehicles [J]. IEEE Communications Surveys & Tutorials, 2019, 21 (3): 2361–2391.
- [11] AL-HOURANI A, KANDEEPAN S, JAMALIPOUR A. Modeling air-to-ground path loss for low altitude platforms in urban environments [C] //IEEE GLOBECOM, 2014: 2898–2904.
- [12] XIAO Z, XIA P, XIA X G. Enabling UAV cellular with millimeter-wave communication: potentials and approaches [J]. IEEE Communications Magazine, 2016, 54 (5): 66–73.
- [13] TRUONG K T, HEATH R W. Effects of channel aging in massive MIMO systems [J]. Journal of Communications and Networks, 2013, 15 (4): 338–351.
- [14] PAPAFAFEIROPOULOS A, RATNARAJAH T,