•智能交叉科学与工程•

DOI:10.12454/j.jsuese.202400467

本刊网刊

基于改进YOLOv5和CombineSORT的车联网路侧视觉感知

李晓晖[1,2],杨　杰[1],夏　芹[1]

(1.中国汽车工程研究院股份有限公司,重庆 401122;2.中汽院(江苏)汽车工程研究院有限公司,江苏 苏州 215153)

摘　要:车路协同是中国实现智能汽车与智慧城市协同发展的重要战略,是弥补自动驾驶与欧美技术差距的核心路线。路侧视觉感知作为车路协同的关键技术手段,能够通过固定视角的摄像头实时监测交通目标,为智能网联系统提供高精度环境感知数据。然而,由于广角镜头下远端目标过小、复杂交通流下车辆相互遮挡频繁、低码率视频中目标运动模糊,以及高流量路口的多路数据同步处理需求,路侧视觉感知在实际应用中往往存在目标漏检、误检或目标ID变换问题。为此,本文兼顾检测精度与运算效率,提出一种基于改进YOLOv5模型和CombineSORT算法的图像识别及跟踪方法。在目标识别环节,通过引入多尺度特征增强模块优化YOLOv5的特征金字塔网络,结合超高效交并比损失函数与网络剪枝技术,显著增强了对小目标及遮挡目标浅层细节特征的提取能力,消融实验表明,在几乎不改变原模型大小的前提下,将mAP@90从0.894提升至0.937。在目标跟踪环节,通过在DeepSORT框架基础上集成Bot-SORT算法的强特征提取网络与StrongSORT算法的联合相似度矩阵,提出了CombineSORT算法,该算法以多项式拟合取代传统的卡尔曼滤波进行运动轨迹预测,舍弃了相机运动补偿,从而在复杂场景下实现了更平滑、更准确的跟踪。实验结果表明,在高流量十字路口场景下,召回率达到96.27%,多目标跟踪精度为0.900,且整体处理时间控制在80 ms以内,显著优于YOLOX、YOLOv7结合DeepSORT等主流组合,证明了其工程实用性。该方法采用轻量化设计,适配现有的边缘计算设备,可直接部署于车联网路侧单元,为智慧交通管理和高级别自动驾驶提供可靠的技术支撑,具有广阔的车路协同应用前景。

关键词:车路协同;路侧感知;图像识别;YOLOv5;CombineSORT

中图分类号:U495　　**文献标志码:**A　　**文章编号:**2096-3246(2026)02-0046-11

车路协同是中国实现智能汽车与智慧城市协同发展的重要战略,是弥补自动驾驶与欧美技术差距的核心路线[1]。视觉感知是车路协同的重要技术手段,也是当前人工智能与交通大模型方向的研究热点[2–3]。

视觉感知主要包含目标感知和目标跟踪两个环节,其中目标感知通常采用深度学习方法来实现对图像中目标的边界检测及类型识别,常用于智能网联汽车视觉感知的模型主要有RetinaNet[4]、Efficientdet[5]、YOLOv3～YOLOv7[6–10]系列模型,以及YOLO的旁系分支YOLOX[11–12]模型等。常见的多目标跟踪算法主要分为基于检测跟踪(DBT)和非检测跟踪(DFT)两大类,其中DBT算法的自主性和实时性更高,因此是业界目前研究的主流。常见的DBT算法包括基于纯边界框跟踪的SORT[13]和ByteTrack[14]算法,以及基于图像卷积特征的DeepSORT[15]、StrongSORT[16]和Bot-SORT[17]算法等,后者的跟踪精度更高,但对平台算力有更高的要求。路侧相机视角下的交通目标感知存在的主要问题有广角镜头下目标较小、车辆间相互遮挡及低码率视频中目标运动模糊等,这将导致上述算法在实际工作中出现目标漏检、误检或目标ID变换的问题。此外,对于一个十字路口,通常需要同步处理4路及以上的视频数据,且受环境所限,不能无限制地提升计算设备的性能,因此难以同时兼顾检测精度和运行效率。

针对以上问题,本文面向车路协同感知的实际工程需求,考虑路侧视觉感知的目标图像特征,对目前智能网联领域中成熟度更高、稳定性更好的YOLOv5模型进行改进,并结合多种主流跟踪方法的优点,提出CombineSORT算法对路侧交通目标进行识别和跟

收稿日期:2024－06－14　　**修回日期:**2024－12－05　　**网络出版日期:**2025－03－24

基金项目:江西省重点研发计划项目(20243BBG71033)

作者简介:李晓晖(1987—),男,工程师,博士. 研究方向:智能网联汽车. E-mail: lixiaohui3000@163.com

踪。实验数据表明，与其他常用方法相比，在车流量较大的十字路口，本文方法能同时具备较高的检测精度和较低的时间成本，因此更适合在现实场景中推广应用。

1　YOLOv5 改进方法概述

1.1　YOLOv5 概述

YOLOv5 的网络结构主要分为主干网（Backbone）、颈部（Neck）和检测器（Detector）3 个部分[18–20]。主干网采用 CSPDarknet53 网络提取输入图像特征，其中包含 Focus、CBL（卷积（conv）+批归一化（batch normalization，BN）+LeakyReLU 激活函数）、CSP（cross stage partial）、SPP（spatial pyramid pooling）等模块，可提供不同尺度的特征信息；颈部采用特征金字塔网络（FPN）+路径聚合网络（PAN）结构[21]，通过结合自上而下传递的语义信息和自下而上传递的定位信息实现多尺度特征融合，以提高模型对不同大小目标检测的准确性和鲁棒性；最后，利用检测器对特征图进行解码，由此判断图像中目标的类型和边界框。图 1 为 YOLOv5 的网络结构。

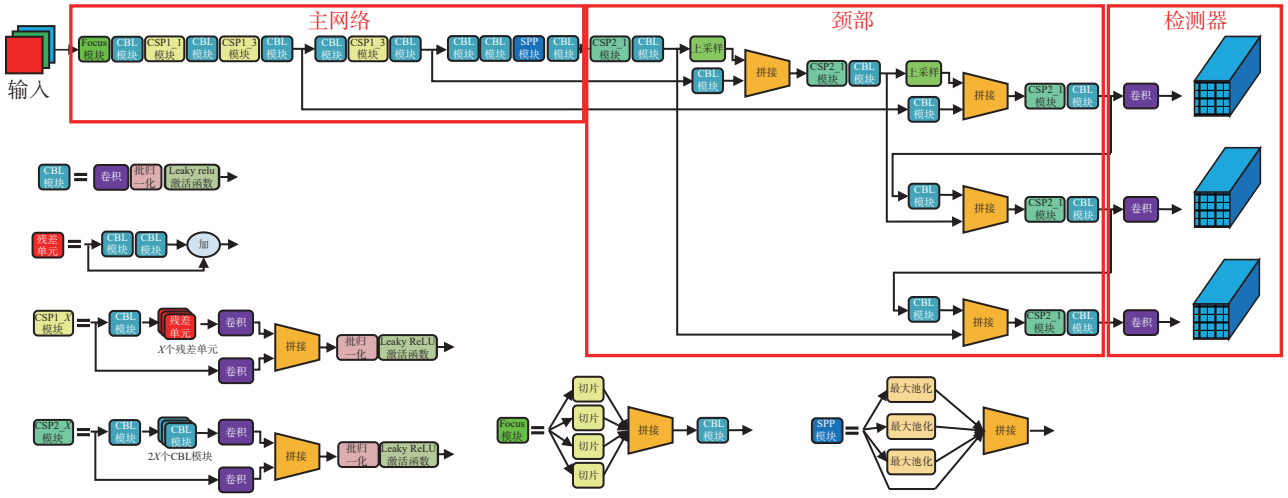

图 1　YOLOv5 的网络结构

Fig. 1　Network structure of YOLOv5

1.2　YOLOv5 的改进

路侧感知视角固定但视野较广，为了增强对遮挡目标及远端目标的识别能力，模型需要获取更多的目标浅层细节特征，同时需要在增改模型结构的同时保证模型的推理速率，因此对 YOLOv5 进行以下 3 点改进。

1.2.1　多尺度特征增强

采用一种多尺度特征增强（MFE）[22]方法优化 FPN 因维度下降带来的信息丢失问题。图 2 为 MFE 结构，它主要引入了尺度融合（scale fusion）、联合特征金字塔（CombineFPN）和像素区域注意力（PRA）3 个模块，其中 C2～C5、T3～T5、T3′～T4′、N3～N5、P3～P7、FP3～FP7、RP3～RP7 均为特征层。

图 2 中，尺度融合部分首先对 C3、C4、C5 执行卷积降维，生成 T3、T4、T5，同时对 C2 进行下采样，生成与 T3 相同分辨率的 T3′；然后，对 T3 和 T3′元素求和得到 N3，并将 T3′下采样以获得与 T4 具有相同分辨率的

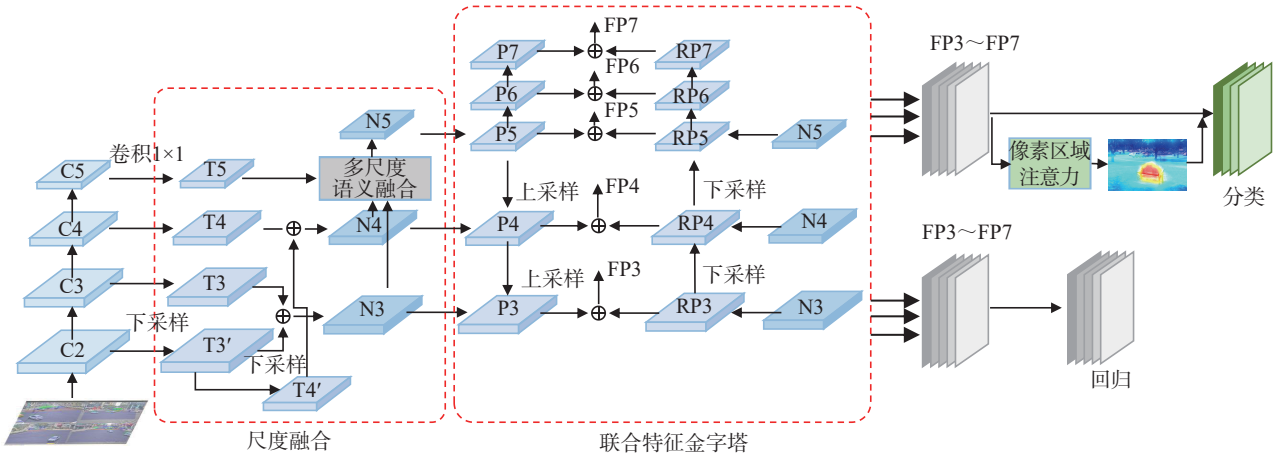

图 2　MFE 结构

Fig. 2　Structure of MFE

特征层 T4′，随即对 T4 和 T4′元素求和得到 N4；最后，采用多尺度语义融合（MSF）解决高低特征层语义不一致的问题，它通过融合 N3、N4、T5，间接将 C2 的浅层特征信息向上传递，最后生成新的特征层 N5。CombineFPN 模块的输入特征层 N3、N4、N5 来自尺度融合模块，P3、P4、P5 由自上而下结构生成，P6、P7 由 P5 连续两次下采样生成；自下而上结构与自上而下结构输入共享，先由 N3 直接得到 RP3，接着逐级依次下采样并与 N4、N5 元素求和直至生成 RP5，再对 RP5 连续两次下采样生成 RP6、RP7；最后，对 P3～P7 和 RP3～RP7 按对应层序号进行元素求和，得到 FP3～FP7。PRA 模块作用于模型检测头分类分支，以 CombineFPN 生成的特征层 FP3～RP7 作为输入，基于注意力机制计算每个像素与图像多区域的相关性，生成注意力图加权融合多尺度区域特征，并以残差方式输出增强特征图。该过程使每个像素融合远距离上下文信息，提升分类分支对复杂场景的判别能力。

1.2.2 损失函数优化

YOLOv5 的损失函数由分类损失、边界框损失和置信度损失构成。其中，分类损失和置信度损失采用二元交叉熵计算，边界框损失 L_{CIOU} 的计算公式为：

$$L_{\mathrm{CIOU}}=1-\left(B_{\mathrm{IOU}}-\frac{\rho^2(b,b^{\mathrm{gt}})}{c^2}-\alpha v\right) \tag{1}$$

式中，B_{IOU} 为检测框和真值框的交并比（IOU），b 和 b^{gt} 分别为两者的中心，$\rho^2(b,b^{\mathrm{gt}})$ 用于度量两者的中心距离，c 为两者最小外接矩形的对角线长度，v 用于表征两者宽高比的一致性，α 为关于 v 的权值因子。显然，以上计算方法只考虑了检测框的中心距离和宽高差异，为了增强边界框损失的收敛效果，采用进一步强化的超高效交并比损失函数 L_{SEIOU}[23] 对训练过程进行优化，其表达式为：

$$L_{\mathrm{SEIOU}}=1-\left(B_{\mathrm{IOU}}-\frac{\rho^2(b,b^{\mathrm{gt}})}{c^2}-\frac{\rho^2(w,w^{\mathrm{gt}})}{C_w^2}-\frac{\rho^2(h,h^{\mathrm{gt}})}{C_h^2}-\frac{\rho^2(A,A^{\mathrm{gt}})+\rho^2(B,B^{\mathrm{gt}})}{c^2}\right) \tag{2}$$

式中，w、w^{gt} 分别为检测框和真值框的宽，h、h^{gt} 分别为两者的高，$\rho^2(w,w^{\mathrm{gt}})$、$\rho^2(h,h^{\mathrm{gt}})$ 分别用于度量两者的宽、高差值，C_w 和 C_h 为两者最小外接矩形的宽和高，(A,A^{gt}) 和 (B,B^{gt}) 分别为两者的左上角点和右下角点坐标，$\rho^2(A,A^{\mathrm{gt}})$ 和 $\rho^2(B,B^{\mathrm{gt}})$ 为两组角点的距离。

1.2.3 模型剪枝优化

改进后的模型较原模型复杂程度更高，为了保证模型的推理效率，采用稀疏训练将模型的 BN 层稀疏化，并对卷积通道进行剪枝压缩。由于 BN 层通过缩放因子 $\boldsymbol{\gamma}$ 与卷积层每个通道关联，$\boldsymbol{\gamma}$ 的大小决定了卷积通道的激活程度，因此可通过稀疏正则化筛除激活度较小的通道，从而达到网络轻量化的目的[24]。

分别用 $\boldsymbol{z}_{\mathrm{in}}^{(l)}$ 和 $\boldsymbol{z}_{\mathrm{out}}^{(l)}$ 表示第 l 层 BN 层的输入和输出，则该 BN 层的作用方式为：

$$\hat{\boldsymbol{z}}^{(l)}=\frac{\boldsymbol{z}_{\mathrm{in}}^{(l)}-\boldsymbol{\mu}_{\mathrm{batch}}^{(l)}}{\sqrt{\left(\boldsymbol{\sigma}_{\mathrm{batch}}^{(l)}\right)^2+\varepsilon}},\ \boldsymbol{z}_{\mathrm{out}}^{(l)}=\boldsymbol{\gamma}^{(l)}\otimes\hat{\boldsymbol{z}}^{(l)}+\boldsymbol{\beta}^{(l)} \tag{3}$$

式中：$\hat{\boldsymbol{z}}^{(l)}$ 为计算过程的中间变量；$\boldsymbol{\mu}_{\mathrm{batch}}^{(l)}$ 和 $\boldsymbol{\sigma}_{\mathrm{batch}}^{(l)}$ 分别为第 l 层 BN 层每个批次（batch）的均值和标准差；$\boldsymbol{\gamma}^{(l)}$ 和 $\boldsymbol{\beta}^{(l)}$ 为第 l 层 BN 层的缩放因子向量和平移因子向量，$\boldsymbol{\gamma}^{(l)}=(\gamma_1^{(l)},\gamma_2^{(l)},\cdots,\gamma_C^{(l)})$，其中 C 表示第 C 个通道；$\otimes$ 表示逐元素乘法；ε 为一极小值常数。

由式（3）可知，$\gamma_C^{(l)}$ 决定了第 l 层第 C 个通道的激活幅度，$\left|\gamma_C^{(l)}\right|$ 越小，说明该通道对神经网络模型的贡献越低。因此，利用 L_1 正则化对所有 BN 层的缩放因子 $\boldsymbol{\gamma}$ 进行稀疏作用，结合式（2）中的损失函数 L_{SEIOU}，剪枝过程的损失函数 L 可写为：

$$L=\sum_{(\boldsymbol{x},\boldsymbol{y})}L_{\mathrm{SEIOU}}(f(\boldsymbol{x}_{\mathrm{input}},\boldsymbol{W}),\boldsymbol{y}_{\mathrm{output}})+\lambda\sum_{\gamma\in\Gamma}|\boldsymbol{\gamma}| \tag{4}$$

式中：$\boldsymbol{x}_{\mathrm{input}}$、$\boldsymbol{y}_{\mathrm{output}}$ 为模型的输入和输出；$\boldsymbol{W}$ 为除 Γ 外的所有训练权值；Γ 为所有 BN 层缩放因子向量的集合，$\Gamma=\left\{\boldsymbol{\gamma}^{(1)},\boldsymbol{\gamma}^{(2)},\boldsymbol{\gamma}^{(3)},\cdots\right\}$；$\lambda$ 为正则化强度系数，用以平衡任务精度与模型稀疏度。

模型将同时训练 $\boldsymbol{W}$ 和 $\boldsymbol{\gamma}$。训练完成后，对每一层的 $\boldsymbol{\gamma}^{(l)}$ 按绝对值从大到小进行排序，并按预设比例保留绝对值靠前的通道。该比例越小，剪枝力度越大，但对模型精度的影响也越大。剪枝完成后，需对压缩后的模型进行微调训练，以恢复因剪枝造成的精度损失。

1.3 模型训练

本文基于不同交通路口的 20 000 张图片进行模型训练，设定训练集和验证集的比例为 9∶1，图片的输入尺度为 864×864；设定训练轮次（epoch）总数为 300，其中前 285 个 epoch 将采用 Mosaic 和 MixUp 方法进行数据增强，以提高模型的鲁棒性，考虑到模型收敛和过拟合问题，最后 15 个 epoch 关闭数据增强；设定基于权重的图片样本抽取，即适度提高包含小基数目标图片的权值，适度降低包含大基数目标图片的权值，以尽可能保障不同种类目标数量的均衡。对新增的 3 个模块及剪枝操作进行消融实验，采用额外采集的 10 368 张图片对模型进行精度检验，并以模型参数量（params）和浮点运算次数（FLOPs）来表征模型的复杂程度。表 1 为消融实验结果。表 1 中，mAP@90 表示所有类别目标的真值框与检测框交并比在 90% 以上的总体召回率。由表 1 可见：YOLOv5 在增加不同的模块后，均有不同程度的精度

提升，但模型的参数量也随之增加；对集成 3 个模块的模型通过 L_1 正则化稀疏训练，并对 BN 层进行 40% 比例剪枝，重新训练后，保证了模型的总体召回率，但其复杂程度却大幅下降。

表 1　消融实验结果

Tab. 1　Results of ablation experiment

YOLOv5	尺度融合模块	联合特征金字塔模块	像素区域注意力模块	剪枝优化	mAP@90	params/10^6	FLOPs/10^9
√					0.894	21.2	49.0
√	√				0.912	24.4	50.8
√		√			0.923	25.3	52.0
√			√		0.916	24.1	50.3
√	√	√	√		0.939	31.0	65.4
√	√	√	√	√	0.937	22.6	51.7

注：“√”表示采用了相应的改进方法。

图 3 为模型改进前后的检测效果对比，其中分别用橙色、黄色和绿色的矩形框对检出的小汽车、公交车和货车进行标记。图 3(b)、(c)为图 3(a)红色虚线框内细节。由图 3 可见，改进后的模型较原模型更容易抓住目标的细节特征，因此在不同复杂程度路口场景下，对于远端及被遮挡目标的检出率及检出位置精度都显著提升。

(a) 原图　　(b) 改进前　　(c) 改进后

图 3　模型改进前后检测效果对比

Fig. 3　Comparison of detection effects before and after model improvement

2　CombineSORT 方法概述

2.1　几种典型跟踪方法

DeepSORT、StrongSORT 和 Bot-SORT 均是基于图像卷积特征的 DBT 算法，与其他大多采用重识别(ReID)的跟踪方法一样，需要提取目标的外观特征，并协同上下帧目标边界框的 IOU 进行轨迹追踪[25–26]。图 4 为 DeepSORT 基本跟踪框架。

DeepSORT 的基本跟踪框架综合考虑了目标的外观特征及位置变换。首先，以卷积神经网络提取检测目标的外观特征，通过前后帧目标特征的相似度矩阵进行目标与历史轨迹的匹配；接着，对其中特征相似度高的目标进行跟踪状态确认，其余目标则根据检测框与预测框的 IOU 进行第二轮匹配，并再次确认跟踪状态；最后，对于经过两轮匹配仍无法跟踪的目标，认定开始一条新轨迹。

DeepSORT、StrongSORT 和 Bot-SORT 均采用更加强大的 Bot 网络[27]来提取目标的外观特征，同时采用指数移动平均(EMA)和相机运动补偿来抵偿部分目标的特征突变。不同的是，StrongSORT 采用增强相关系数(ECC)进行相机补偿，并采用噪声尺度自适应(NSA)卡尔曼滤波器[28]；Bot-SORT 则采用了 GMC(global motion compensation)算法，同时利用一种将外观相似度和 IOU 相似度联合生成相似度矩阵来进行级联匹配的新方法[29]。

本文结合实际应用场景，提出 CombineSORT 算法。该算法基于 DeepSORT 基本框架，选择性集成 StrongSORT 和 Bot-SORT 的部分先进方法，同时用多项式滤波替代卡尔曼滤波，以提升跟踪精度并保证计算效率。

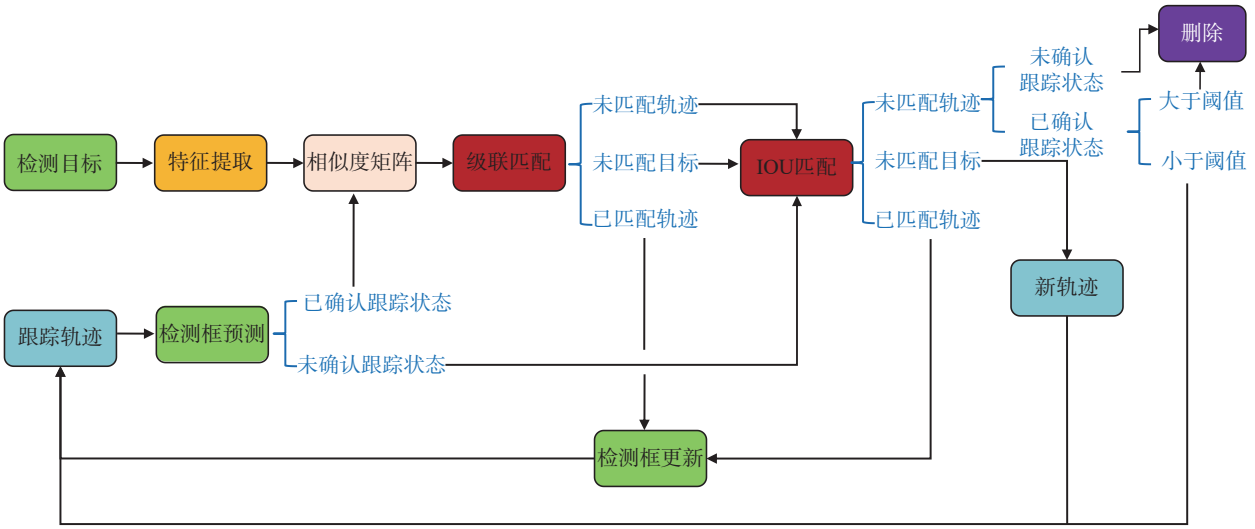

图 4 DeepSORT 基本跟踪框架

Fig. 4 Fundamental tracking framework of DeepSORT

2.2 CombineSORT 算法原理

通过 8 维向量$(\mu,v,w,h,\dot{\mu},\dot{v},\dot{w},\dot{h})$来描述检测框的运动空间，其中，$(\mu,v,w,h)$为检测框的中心位置和宽、高，$(\dot{\mu},\dot{v},\dot{w},\dot{h})$为上述 4 个维度的速度信息。在此基础上，针对图 4 中特征提取、检测框预测和相似度矩阵作如下改进。

2.2.1 基于剪枝的 Bot 网络

CombineSORT 采用跟 Bot-SORT 一样的主干网络来提取目标的外观特征，并通过 EMA 更新第 k 帧、第 i 个目标的外观特征 e_i^k：

$$e_i^k=\beta e_i^{k-1}+(1-\beta)f_i^k \tag{5}$$

式中，f_i^k 为当前检测并匹配的目标外观特征，通过一个动量因子 β 与第 k–1 帧、第 i 个目标的外观特征 e_i^{k-1} 进行融合。

与 StrongSORT 和 Bot-SORT 不同的是，CombineSORT 采用结构化剪枝对 Bot 网络进行优化，通过重新训练加入 L_1 正则化的损失函数，舍弃较小缩放因子对应的特征层通道，使 Bot 网络相较于 DeepSORT 的特征提取模型并不增加额外的时间开销。

2.2.2 多项式滤波

DeepSORT、StrongSORT 和 Bot-SORT 均基于卡尔曼滤波来预测目标的运动状态，然而假设所有目标都具有相同的观测噪声显然与实际情况不符。因此，即使 StrongSORT 引入了 NSA 协方差系数，但由于平面图像的透视关系，卡尔曼滤波并不足以准确描述某些复杂的运动轨迹。由于路侧固定视角下图像视野较为宽广且大部分车辆都具有相似的运动轨迹，CombineSORT 采用多项式滤波对目标的检测框进行轨迹拟合，同时舍弃了相机运动补偿。该算法通过构造 4 维向量(μ,v,w,h)进行多项式求解，由于基于目标的大量历史数据进行曲线拟合，因此预测结果能更好地逼近目标轨迹的拐点。

对于 p 阶多项式的求解问题，设向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 和 $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$ 分别表示 n 个观测样本的自变量与因变量取值，可通过多项式 $\hat{y}=a_0+a_1x+a_2x^2+\cdots+a_px^p$，$n\geqslant p$ 进行拟合，$(a_0,a_1,\cdots,a_p)$为多项式系数向量。则 $\boldsymbol{y}$ 中任意 $y_{i'}$ 及其拟合值的误差平方和 R^2 可表示为：

$$R^2=\sum_{i'=1}^{n}(y_{i'}-(a_0+a_1x_{i'}+a_2x_{i'}^2+\ldots+a_px_{i'}^p))^2 \tag{6}$$

根据最小二乘原理，在式(6)中对$(a_0,a_1,\cdots,a_p)$求偏导，可得：

$$\boldsymbol{X}_{\mathrm{Van}}^{n\times p}\cdot\boldsymbol{a}=\boldsymbol{y} \tag{7}$$

式中，$\boldsymbol{a}=(a_0,a_1,\cdots,a_p)^{\mathrm{T}}$，$\boldsymbol{X}_{\mathrm{Van}}^{n\times p}$ 为 $\boldsymbol{x}$ 的 p 阶范德蒙德矩阵。求解式(7)即可得到多项式的拟合系数。

令 $\boldsymbol{z}_i^k=(\mu_i^k,v_i^k,w_i^k,h_i^k)$ 为第 i 个目标在第 k 帧的检测框向量，$\boldsymbol{F}_i^{k-1}$ 为该目标检测框在第 k–1 帧的轨迹预测模型，$\boldsymbol{F}_i^{k-1}=(U_i^{k-1},V_i^{k-1},\Omega_i^{k-1},H_i^{k-1})$，其中，$U_i^{k-1}$、$V_i^{k-1}$ 分别为 μ_i^k、v_i^k 的二阶多项式函数，Ω_i^{k-1}、H_i^{k-1} 分别为 w_i^k、h_i^k 的一阶多项式函数，则该目标检测框的最优值 $\hat{\boldsymbol{z}}_i^k$ 可通过 $\boldsymbol{F}_i^{k-1}$ 及多项式拟合值 $\tilde{\boldsymbol{z}}_i^k$ 表示为：

$$\begin{cases}\tilde{\boldsymbol{z}}_i^k=\boldsymbol{F}_i^{k-1}(k),\\ \boldsymbol{F}_i^k=\mathrm{Polyfit}(\hat{\boldsymbol{z}}_i^{k-n+1},\cdots,\hat{\boldsymbol{z}}_i^{k-1},\varphi\boldsymbol{z}_i^k+(1-\varphi)\tilde{\boldsymbol{z}}_i^k),\\ \hat{\boldsymbol{z}}_i^k=\boldsymbol{F}_i^k(k)\end{cases} \tag{8}$$

式中，Polyfit(·)为对输入样本求多项式系数的函数，φ 为调节因子。

2.2.3 联合相似度矩阵

在交通目标中，时常出现同型号的车辆，其在相同的摄像头视角下往往表现出相近的外观特性，因此，DeepSORT 和 StrongSORT 将外观相似度和 IOU 相似度通过简单线性叠加构建的相似度矩阵会出现偏

差。本文选用 Bot-SORT 提出的联合相似度矩阵对级联匹配的第一层进行当前帧目标与历史轨迹的匹配，其表达式为：

$$\hat{d}_{i,j}^{\cos}=\begin{cases}0.5d_{i,j}^{\cos},(d_{i,j}^{\cos}<\theta_{\text{emb}})\wedge(d_{i,j}^{\text{iou}}<\theta_{\text{iou}});\\1,\text{其他}\end{cases}\tag{9}$$

$$C_{i,j}=\min\{d_{i,j}^{\text{iou}},\hat{d}_{i,j}^{\cos}\}\tag{10}$$

式(9)、(10)中，$d_{i,j}^{\text{iou}}$ 和 $d_{i,j}^{\cos}$ 分别为第 i 个目标轨迹当前第 j 个目标检测框的 IOU 距离和对应图像特征的余弦距离，$\hat{d}_{i,j}^{\cos}$ 为计算过程的中间变量，θ_{iou} 和 θ_{emb} 分别为 IOU 距离和余弦距离的预设阈值，$C_{i,j}$ 为最终联合相似度矩阵的第(i,j)项。

该方法充分考虑了前后帧目标外观相近但距离较远的情形，能够有效降低级联匹配过程中的错误配对概率。

2.3　CombineSORT 算法原理

综上所述，本文采用 CombineSORT 算法进行交通目标跟踪的过程如下：

1)根据图像识别得到的边界框提取目标的外观特征，同时初始化多项式滤波器，并将第一次出现的目标轨迹标记为未确认状态。

2)逐一比较当前帧目标的边界框和根据轨迹预测的边界框，若当前轨迹为确认状态，则按式(9)、(10)生成联合相似度矩阵，进入级联匹配，否则仅构建 IOU 相似度矩阵，进入 IOU 匹配。

3)级联匹配时，按时间回溯逐级构造联合相似度矩阵，并用匈牙利算法将当前检出目标与历史轨迹配对；将能够配对的轨迹结合对应的目标边界框按式(8)更新，将不能配对的轨迹和边界框构建 IOU 相似度矩阵，进入 IOU 匹配；值得一提的是，仅对级联匹配的第一层采用联合相似度矩阵，后续层将仅采用外观特征的余弦相似度矩阵。

4)IOU 匹配时，采用局部最大值法对 IOU 相似度矩阵进行处理，将能够配对的轨迹结合对应的边界框按式(8)更新，将不能配对的边界框初始化为一个新轨迹，并将其状态标记为未确认；当一个轨迹被连续配对 3 次以上时，将其状态转换为确认。

5)对于 IOU 匹配后剩下未配对的轨迹，若其连续未配对次数已超过设定的阈值，则将其删除；否则将其边界框按照拟合多项式进行更新。

6)重复以上步骤，直至当前帧所有测得的目标边界框和历史轨迹处理完毕。

3　实验结果及分析

为了验证以上算法，选择一个流量小于或等于 3 000 pcu/h(pcu/h 为交通工程领域中表示交通流量的标准单位，表示每小时通过该道路断面的交通量为 3 000 标准车当量)和两个车流量大于或等于 5 000 pcu/h 的十字路口来搭建实验环境。实验选用的边缘计算设备(MEC)集成 Intel® Core™ i7-10700T CPU 和 NVIDIA GeForce RTX3060 GPU，系统环境为 Ubuntu18.04，摄像头为海康 DS-2XD8A47F 型网络相机，该相机在帧率 25 f/s、分辨率 1 080 像素、码率 2 048 kb/s 下的平均时延约为 100 ms。

图 5 为路侧感知部署方案，将 4 台网络相机分别架设在红绿灯杆件的横臂上，通过千兆级光网交换机与 MEC 连接，设置路侧通信单元(RSU)为路侧感知系统的时钟源，网络相机与 MEC 均通过网络时间协议(NTP)与 RSU 进行时间同步。

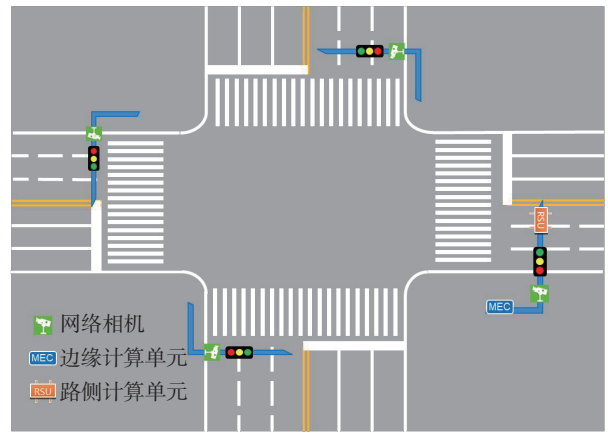

图 5　路侧感知部署方案

Fig. 5　Roadside perception deployment plan

通过改进 YOLOv5 及 CombineSORT 进行多路视觉感知的计算处理流程如下：

1)从 4 台网络相机中拉取视频流，将解码得到的图像转为单精度浮点数并缩放至 512×864，同时将多张图片组合成一个 batch；

2)利用改进的 YOLOv5 模型识别图像中的目标类型并提取检测框，通过剪枝的 Bot 网络提取检测框中目标的外观特征；

3)利用目标的外观特征和检测框信息构建联合相似度矩阵，通过 ComineSORT 的级联匹配和 IOU 匹配追踪目标轨迹并进行 ID 指派；

4)根据各摄像头的标定文件和视频流时间戳对各目标轨迹进行时空同步，在大地坐标系下对以上目标进行后融合处理。

由于车联网要求路侧感知上报频率为 10 Hz，考虑到网络相机的实际帧率，期望系统能在 80 ms 内完成一轮多通道视频数据的完整计算。为保障模型在实际工作中的实时性，采用以下方法对全套软件运行过

程进行优化：

1）基于多线程从网络相机获取视频流，采用计算统一设备体系结构（CUDA）硬解码技术将视频从H.264格式加速解算为RGB图像。

2）采用TensorRT（NVIDIA推出的一款用于深度学习模型高性能推理的软件开发工具包）将改进的YOLOv5模型和CombineSORT的Bot网络转换为.engine文件以在GPU上获得更高效的推理速度；同时，通过CUDA编程将图像归一化、检测器运算等处理模块移植到GPU上运行。

3）将跟踪过程设计为动态级联匹配，即图像中的目标总数超过一定阈值时，仅采用IOU匹配对行人和非机动车进行跟踪，同时按概率缩减在级联匹配过程中对机动车历史特征轨迹的回溯上限，由此限制跟踪过程中目标特征提取及特征匹配时间。

图6为十字路口路侧感知系统实时检测效果。其中，绿色和蓝色框分别为机动车和非机动车的检出标记。

(a) 十字路口1 (b) 十字路口2 (c) 十字路口3

图6 十字路口路侧感知系统实时检测效果

Fig. 6 Real-time detection performance of crossroads roadside perception system at crossroads

结合实际水平感知精度需求，将IOU为0.9设为所有目标的感知阈值（IOU低于0.9将同时计1次漏检和1次误检），统计图6中3个十字路口检测10 min的召回率、漏检数（FN）、误检数（FP）、ID切换次数（IDSW），并采用多目标跟踪精度（MOTA）[30]评估算法的跟踪效果。表2为十字路口检测结果统计。

表2 十字路口检测结果统计

Tab. 2 Statistics of crossroads detection results

十字路口编号	类别	样本数	轨迹数	召回率/%	FN	FP	IDSW	MOTA
1	行人	1 043	32	95.88	43	30	46	0.886
	非机动车	842	42	91.09	75	60	44	0.787
	小汽车	6 047	301	99.16	51	41	27	0.980
	公交车	289	8	94.81	15	14	1	0.896
	货车	256	7	94.92	13	12	0	0.902
	合计	8 477	390	97.68	197	157	118	0.944
2	行人	2 863	88	89.98	287	231	198	0.750
	非机动车	1 502	75	84.35	235	188	65	0.675
	小汽车	16 806	840	98.03	331	274	551	0.931
	公交车	2 667	76	94.98	134	102	27	0.901
	货车	1 225	35	95.27	58	55	9	0.900
	合计	25 063	1 114	95.83	1 045	850	850	0.890
3	行人	913	28	98.80	11	12	25	0.947
	非机动车	1 742	87	85.65	250	202	90	0.689
	小汽车	16 446	822	98.17	301	261	546	0.933
	公交车	3 192	91	95.39	147	116	36	0.906
	货车	805	23	95.03	40	41	7	0.891
	合计	23 098	1 051	96.76	749	632	704	0.910

表2中，样本数是所有帧图像中出现对应类别目标的数量，轨迹数是实际经过路口的对应类别目标数量。其中：行人和非机动车并不严格遵守交通规则，且由于体积较小，相互遮挡的情形时有发生，因此检测难度大，召回率和MOTA值都相对较低；而在某些视角下，厢式货车与公交车、三轮非机动车与小轿车具有极其相似的

外观特征，因此也拉低了对应目标的识别精度。

为了进一步验证本文方法的优势，采用不同识别模型和跟踪算法对 3 个路口的相同视频流进行交叉验证。表 3 为不同识别模型和跟踪算法结果对比。

表 3　不同识别模型和跟踪算法结果对比

Tab. 3　Comparison of results from different recognition models and tracking algorithms

路口类型	识别模型	跟踪算法	召回率/%	MOTA	识别计算时长/ms	跟踪计算时长/ms	总时长/ms
低流量路口（十字路口 1）	Efficientdet[5]	DeepSORT[15]	96.54	0.938	18	23	43
	YOLOv5[8]	ByteTrack[14]	97.63	0.940	24	2	27
	YOLOv5	DeepSORT	97.63	0.944	24	22	49
	YOLOX[11]	StrongSORT[16]	97.67	0.942	25	28	56
	YOLOv7[10]	本文跟踪算法	97.69	0.946	30	23	56
	本文识别模型	Bot-SORT[17]	97.68	0.945	25	27	55
	本文识别模型	本文跟踪算法	97.68	0.944	25	23	51
高流量路口（十字路口 2 和 3）	Efficientdet	ByteTrack	92.75	0.817	20	7	32
	Efficientdet	DeepSORT	92.76	0.882	19	46	70
	YOLOv5	ByteTrack	95.26	0.824	26	7	38
	YOLOv5	DeepSORT	95.29	0.887	26	47	78
	YOLOv5	StrongSORT	95.31	0.890	25	53	83
	YOLOX	DeepSORT	96.02	0.896	27	47	79
	YOLOX	Bot-SORT	96.03	0.897	28	59	92
	YOLOX	本文跟踪算法	96.02	0.896	28	46	79
	YOLOv7	Bot-SORT	96.28	0.901	32	59	96
	YOLOv7	StrongSORT	96.28	0.900	31	53	89
	YOLOv7	本文跟踪算法	96.28	0.900	32	46	83
	本文识别模型	Bot-SORT	96.28	0.901	28	59	92
	本文识别模型	StrongSORT	96.27	0.900	27	53	85
	本文识别模型	本文跟踪算法	96.27	0.900	27	46	78

表 3 中的总时长包含了从获取图像、解码、预处理到结果输出的完整流程耗时。由表 3 可见：在低流量的十字路口 1 场景下，各类方法的检测结果差异不大，若综合考虑计算资源消耗，轻量化的 Efficientdet 和 ByteTrack 反而更具实用性。而在高流量的十字路口 2 和 3，对比结果则显得相对复杂，在检测模型方面，YOLO 系列模型和 Efficientdet 相比能显著提升召回率，但识别计算时长也随之增加，其中 YOLOv7 的计算耗时最为显著；在跟踪算法方面，ByteTrack 虽跟踪计算时长极短，但其 MOTA 值明显偏低，难以满足高精度感知需求，而在基于外观特征的跟踪器中，StrongSORT 和 Bot-SORT 的跟踪计算时长均高于 DeepSORT，但其采用的运动补偿等优化机制主要针对视角变化的复杂运动场景，因此并不能显著提升固定视角下路侧感知的 MOTA 值。相较而言，本文通过将改进的 YOLOv5 模型与高效的 CombineSORT 相结合，在精度与效率之间取得了最佳平衡，总时长不超过 80 ms，证明了其在实际应用中的优势。

4　结　论

本文主要研究车联网路侧感知系统的视觉感知方法，针对路侧视角下的目标图形特征及运动轨迹特点，提出了 YOLOv5 的改进模型以及 CombineSORT 跟踪算法。通过实验和数据处理，得出以下结论：

1）YOLOv5 是目前智能网联领域中应用最广泛的图像识别模型之一，本文通过 MFE 优化 YOLOv5 的 FPN，采用超高效交并比损失函数 L_{SEIOU} 代替原损失函数训练模型，再基于稀疏训练对 BN 层稀疏化，对卷积通道进行剪枝压缩，进一步加强了模型对远端及被遮挡目标浅层特征的抓取，并在提升识别精度的同时保障了模型的推理效率。

2）CombineSORT 采用 DeepSORT 的基础框架，集成了 Bot-SORT 和 StrongSORT 的特征提取网络和联合相似度矩阵，同时引入多项式滤波改进预测模型，舍

弃了相机运动补偿，对于路侧固定视角的目标跟踪具备一定优势。

3）在多个十字路口进行实验，数据证明本文方法能基于路侧视角有效感知及跟踪路口的一般交通目标，且面对高流量十字路口时，较其他常用方法能同时具备较高的检测精度和较低的运算总时长，因此更适合面向车路协同场景推广。

本文模型仅针对路侧固定视角下微小及被遮挡目标的感知问题，对其他环境因素导致的算法瓶颈尚未深入探讨。未来可进一步研究恶劣天气对相机连续图像抓拍质量的影响，以突破路侧感知系统对沙尘、风雪等天气的性能瓶颈。

参考文献：

[1] Bao Xuyan,Yu Bingyan,Wang Jing.Research on development status and test method of roadside sensing system in Internet of vehicles[J].Mobile Communications,2021,45(6):43–47.[鲍叙言,余冰雁,王晶.车联网路侧感知系统发展现状及测试方法研究[J].移动通信,2021,45(6):43–47.]

[2] An Xin,Cai Baigen,Shangguan Wei.Vehicle road cooperative roadside perception fusion method[J].Measurement & Control Technology,2022,41(2):1–12.[安鑫,蔡伯根,上官伟.车路协同路侧感知融合方法的研究[J].测控技术,2022,41(2):1–12.]

[3] Long Xuejun,Tan Zhiguo,Gao Feng.Analysis of application status of multi-sensor fusion roadside sensing technology[J].China ITS Journal,2021(10):137–140.[龙学军,谭志国,高枫.多传感器融合路侧感知技术应用现状分析[J].中国交通信息化,2021(10):137–140.]

[4] Lin T Y,Goyal P,Girshick R,et al.Focal loss for dense object detection[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision(ICCV). Venice:IEEE,2017:2999–3007.

[5] Tan Mingxing,Pang Ruoming,Le Q V.EfficientDet:Scalable and efficient object detection[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR).Seattle:IEEE,2020:10778–10787.

[6] Redmon J,Farhadi A.YOLOv3:An incremental improvement[EB/OL].(2018–04–08)[2024–06–14].https://arxiv.org/abs/1804.02767v1.

[7] Bochkovskiy A,Wang C Y,Liao H M.YOLOv4:Optimal speed and accuracy of object detection[EB/OL].(2020–04–23)[2024–06–14].https://arxiv.org/abs/2004.10934.

[8] Kim J H,Kim N,Park Y W,et al.Object detection and classification based on YOLO-v5 with improved maritime dataset[J].Journal of Marine Science and Engineering,2022,10(3):377.

[9] Li Chuyi,Li Lulu,Jiang Hongliang,et al.YOLOv6:A single-stage object detection framework for industrial applications[EB/OL].(2022–09–07)[2024–06–14].https://arxiv.org/abs/2209.02976

[10] Wang C Y,Bochkovskiy A,Liao H M.YOLOv7:Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Vancouver:IEEE,2023:7464–7475.

[11] Ge Zheng,Liu Songtao,Wang Feng,et al.YOLOX:Exceeding YOLO series in 2021[EB/OL].(2021–08–06)[2024–06–14].https://arxiv.org/abs/2107.08430v2.

[12] Zhang Mingjiang,Wang Chengyuan,Yang Jungang,et al.Research on engineering vehicle target detection in aerial photography environment based on YOLOX[C]//Proceedings of the 2021 14th International Symposium on Computational Intelligence and Design(ISCID).Hangzhou:IEEE,2022:254–256.

[13] Bewley A,Ge Zongyuan,Ott L,et al.Simple online and realtime tracking[C]//Proceedings of the 2016 IEEE International Conference on Image Processing(ICIP). Phoenix:IEEE,2016:3464–3468.

[14] Zhang Yifu,Sun Peize,Jiang Yi,et al.ByteTrack:Multi-object tracking by associating every detection box[M]//Computer Vision-ECCV 2022.Cham:Springer Nature Switzerland,2022:1–21.

[15] Wojke N,Bewley A,Paulus D.Simple online and realtime tracking with a deep association metric[C]//Proceedings of the 2017 IEEE International Conference on Image Processing(ICIP).Beijing:IEEE,2018:3645–3649.

[16] Du Yunhao,Zhao Zhicheng,Song Yang,et al.StrongSORT:Make DeepSORT great again[J].IEEE Transactions on Multimedia,2023,25:8725–8737.

[17] Aharon N,Orfaig R,Bobrovsky B Z.Bot-SORT:Robust associations multi-pedestrian tracking[EB/OL].(2022–07–07)[2024–06–14].https://arxiv.org/abs/2206.14651v2.

[18] Wu Wentong,Liu Han,Li Lingling,et al.Application of local fully Convolutional Neural Network combined with YOLO v5 algorithm in small target detection of remote sensing image[J].PLoS One,2021,16(10):e0259283.

[19] Zhang Yu,Guo Zhongyin,Wu Jianqing,et al.Real-time vehicle detection based on improved YOLO v5[J].Sustainability,2022,14(19):12274.

[20] Wu Tianhao,Wang Tongwen,Liu Yaqi.Real-time vehicle and distance detection based on improved yolo v5 network[C]//Proceedings of the 2021 3rd World Symposium on Artificial Intelligence(WSAI).Guangzhou:IEEE,2021:24–28.

[21] Lin T Y,Dollár P,Girshick R,et al.Feature pyramid networks for object detection[C]//Proceedings of the 2017

IEEE Conference on Computer Vision and Pattern Recognition(CVPR).Honolulu:IEEE,2017:936–944.

[22] Zhang Zhenhua,Qin Xueying,Zhong Fan.MFE:Multi-scale feature enhancement for object detection[C]//Proceedings of the 32nd British Machine Vision Conference. [S.l.]: BMVA,2021:1–11.

[23] Cai Guanhong,Li Guoping,Wang Guozhong,et al.Lightweight traffic-light detection algorithm based on improved YOLOv5s[J]. Journal of Shanghai University(Natural Science Edition),2024,30(1):94–105.[蔡管鸿,李国平,王国中,等.基于改进 YOLOv5s 的轻量化交通灯检测算法[J].上海大学学报(自然科学版),2024,30(1):94–105.]

[24] Han Song, Pool J, Tran J, et al. Learning both weights and connections for efficient neural network[C]//Proceedings of the Advances in Neural Information Processing Systems 28(NIPS 2015).Montreal:NIPS,2015:708.

[25] Luo Wenhan, Xing Junliang, Milan A, et al. Multiple object tracking: A literature review[J].Artificial Intelligence,2021, 293:103448.

[26] Del Rosario J R B, Bandala A A, Dadios E P. Multi-view multi-object tracking in an intelligent transportation system: A literature review[C]//Proceedings of the 2017 IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management(HNICEM). Manila: IEEE, 2017:1–4.

[27] Luo Hao, Jiang Wei, Gu Youzhi, et al. A strong baseline and batch normalization neck for deep person re-identification[J]. IEEE Transactions on Multimedia, 2020, 22(10):2597–2609.

[28] Xu Wei, Du Xiaodong, Li Ruochen, et al. Attention-enhanced StrongSORT for robust vehicle tracking in complex environments[J].Scientific Reports,2025,15:17472.

[29] Li Tingting, Li Zhanbo, Mu Yuhong, et al. Pedestrian multi-object tracking based on YOLOv7 and Bot-SORT[C]//Proceedings of SPIE Conference on Third International Conference on Computer Vision and Pattern Analysis(ICCPA 2023).Hangzhou:SPIE,2023:369–374.

[30] Bernardin K,Stiefelhagen R.Evaluating multiple object tracking performance:The CLEAR MOT metrics[J]. EURASIP Journal on Image and Video Processing, 2008, 2008(1):246309.

Roadside Visual Perception in Internet of Vehicles Based on Improved YOLOv5 and CombineSORT

LI Xiaohui[1,2], *YANG Jie*[1], *XIA Qin*[1]

(1.China Automotive Engineering Research Institute Co., Ltd., Chongqing 401122, China;
2.Jiangsu CAERI Automotive Engineering Research Institute Co., Ltd., Suzhou 215153, China)

Abstract:

Objective Visual inspection is a critical technology for roadside perception in vehicle-road cooperative systems. However, in practical applications, achieving both high detection accuracy and computational efficiency simultaneously remains challenging due to limited computing resources. This study proposes a novel method based on an improved YOLOv5 combined with CombineSORT for image recognition and target tracking, which achieves strong detection performance while maintaining low computational time cost, as demonstrated through experimental results.

Methods Firstly, Multi-scale Feature Enhancement (MFE) was applied to the FPN of YOLOv5 to extract shallow target details. This module was mainly composed of Scale Fusion, CombineFPN, and Pixel–Region Attention. A super-efficient IOU (SEIOU) loss function and network pruning were applied to improve convergence and reduce model complexity. In this process, the loss was calculated based on differences in length, width, and diagonal between the detection boxes and the ground-truth boxes, while batch normalization (BN) layer sparsification was applied for convolutional channel filtering. Secondly, by combining DeepSORT, StrongSORT, and Bot-SORT, a new multi-target tracking method named CombineSORT was presented. In this approach, the basic framework of DeepSORT was adopted, and the BotNet with ResNet50 as the backbone was utilized to extract appearance features. Kalman filtering was replaced by polynomial fitting to improve trajectory smoothness, while the joint similarity matrix from StrongSORT was utilized to match targets with trajectories. Based on the operational procedure of the proposed algorithm, a series of experiments was designed to validate its effectiveness. Using images from real intersections, ablation tests verified the effectiveness and data volume contribution of each improved module. The algorithm was then compared to classical methods using intersection video streams with varying traffic volumes, all of which were executed on a mobile edge computer (MEC) with limited computing resources.

Results and Discussions Through ablation tests, the original YOLOv5 achieved an mAP@90 of 0.894 with a parameter quantity of 21.2 M. Scale Fusion, CombineFPN, and Pixel–Region Attention increased the mAP@90 of the original model to 0.91, 0.923, and 0.916, respectively, while the parameter quantity increased to 24.4, 25.3, and 24.1 M, respectively. The YOLOv5 model integrating all three modules achieved an mAP@90 of 0.939 with a parameter quantity of 31.0 M, after which network pruning reduced the parameter quantity to 6.6 M while maintaining an mAP@90

of 0.937. Through three groups of real intersection experiments, the average recall rates for Group 1 to 3 were 97.68%, 95.83%, and 96.76%, while the multiple object tracking accuracy (MOTA) values were 0.944, 0.890, and 0.910, respectively. Among all target categories, pedestrians and non-motorized vehicles exhibited relatively poor detection performance. Especially in Group 2, the recall rate and MOTA for pedestrians were 89.98% and 0.75, respectively, while those for non−motorized vehicles were as low as 84.5% and 0.675. This behavior occurred because these two target types had relatively small sizes and did not strictly follow traffic rules, which caused frequent occlusion and increased trajectory prediction difficulty. In addition, the recall rates of buses and trucks were nearly 3 percentage points lower than those of cars, especially in the group, where the recall rates were only 94.81% and 94.92%, respectively. This issue occurred because box trucks and buses exhibited similar appearance features, which increased the likelihood of misidentification from rear perspectives. When comparing the overall processing performance of different algorithms at low-volume intersections, the worst test result achieved a recall rate of 96.54% with a MOTA value of 0.938, while the best result achieved a recall rate of 97.69% with a MOTA value of 0.946. These results indicated that most algorithms achieved good detection performance under sparse target conditions, and lightweight models demonstrated advantages when considering computational resource constraints. However, for high-volume intersections, although the lightweight algorithm based on EfficientNet and ByteTrack exhibited the shortest computation delay, its recall rate and MOTA value were only 91.75% and 0.817, respectively. In contrast, algorithms based on YOLOv5, YOLOX, YOLOv7, and the improved YOLOv5 proposed in this study achieved recall rates ranging from 95.26% to 96.28%, while algorithms combined with DeepSORT, StrongSORT, Bot-SORT, and CombineSORT achieved MOTA values ranging from 0.887 to 0.901. However, most of these methods exhibited computation times exceeding 80 ms, which prevented real-time operation. Among algorithms with computation times below 80 ms, the proposed method based on improved YOLOv5 and CombineSORT achieved the best overall performance, with a recall rate of 96.27% and an MOTA value of 0.900, which confirmed its ability to balance detection accuracy and computational efficiency.

Conclusions This study focuses on traffic target perception from a fixed roadside perspective, and the results demonstrate the effectiveness and accuracy of the proposed algorithm. Compared to other commonly used algorithms, the proposed approach simultaneously achieves higher detection accuracy and lower time cost at high-volume intersections, indicating strong application potential in vehicle road collaboration scenarios. For improved engineering practices, further research can be conducted to enhance recognition and tracking performance based on continuous image sequences under adverse weather conditions.

Key words: vehicle-road cooperative; roadside perception; image recognition; YOLOv5; CombineSORT

（编辑　李轶楠）

引用格式: Li Xiaohui, Yang Jie, Xia Qin. Roadside visual perception in internet of vehicles based on improved YOLOv5 and CombineSORT[J]. Advanced Engineering Sciences, 2026, 58(2): 46–56. [李晓晖, 杨杰, 夏芹. 基于改进 YOLOv5 和 CombineSORT 的车联网路侧视觉感知[J]. 工程科学与技术, 2026, 58(2): 46–56.]