计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(3):63−72 [doi: 10.15888/j.cnki.csa.009410]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于车载环境的交通目标跟踪①

孟令辰, 孟　乔, 皇甫俊逸, 李　鑫

(青海大学 计算机技术与应用系, 西宁 810016)
通信作者: 孟　乔, E-mail: 250345481@qq.com

摘　要: 针对车载环境下小目标难以识别和相机动态移动造成的目标跟踪精度下降问题, 提出一种基于改进YOLOv5 与 ByteTrack 的交通目标跟踪方法. 首先, 引入 Transformer 与加权特征金字塔 (BiFPN) 结构的思想重构 YOLOv5 检测网络, 有效捕获了特征的全局依赖关系, 缓解了深层卷积小目标信息丢失问题, 改善了车载环境下的目标检测性能. 此后, 以 ByteTrack 为基础提出了添加相机移动补偿的 CMC-ByteTrack 跟踪方法, 更精准地描述了视频前后帧的数据关联关系, 提高了相机大幅位移时的跟踪精度. 实验结果表明, 改进 YOLOv5 的平均检测精度 (*mAP*) 达到了 82.2%, 相比原算法提高了 3.9%, 与 CMC-ByteTrack 结合后的跟踪准确性 (*MOTA*) 相比改进前的跟踪方法提高了 2.8%.

关键词: YOLOv5; 目标跟踪; Transformer; 特征融合; 相机移动补偿

引用格式: 孟令辰,孟乔,皇甫俊逸,李鑫.基于车载环境的交通目标跟踪.计算机系统应用,2024,33(3):63–72. http://www.c-s-a.org.cn/1003-3254/9410.html

Traffic Object Tracking Based on In-vehicle Environment

MENG Ling-Chen, MENG Qiao, HUANGFU Jun-Yi, LI Xin

(Department of Computer Technology and Applications, Qinghai University, Xining 810016, China)

Abstract: This study proposes a traffic object tracking method based on improved YOLOv5 and ByteTrack to address the problem of decreased tracking accuracy caused by the difficulty in recognizing small objects in the car environment and camera movement. Firstly, the study introduces the Transformer and weighted feature pyramid network (BiFPN) structure to reconstruct the YOLOv5 detection network. This effectively captures the global dependency relationships of features, alleviates the problem of information loss for small objects in deep convolutional layers, and improves the performance of object detection in vehicular environments. Subsequently, based on ByteTrack, the study proposes the CMC-ByteTrack tracking strategy that adds camera motion compensation. The method more accurately describes the data correlation relationship between the previous and subsequent frames of the video, improving tracking accuracy during significant camera displacement. Experimental results show that the improved YOLOv5 achieves mean average precision (*mAP*) of 82.2%, and 3.9% increase in comparison with the original algorithm. After integration with CMC-ByteTrack, the multiple object tracking accuracy (*MOTA*) is increased by 2.8% in comparison with that of the original tracking method.

Key words: YOLOv5; target tracking; Transformer; feature fusion; camera movement compensation

随着国内机动车保有量的不断攀升, 汽车已成为人们交通出行的主要方式. 然而, 汽车数量的增加使交通安全隐患问题与日俱增, 如何有效降低车辆与行人相碰的概率, 及时警示驾驶员可能与附近车辆发生碰

① 基金项目: 青海省自然科学基金 (2023-ZJ-989Q)
收稿时间: 2023-08-30; 修改时间: 2023-09-26; 采用时间: 2023-10-09; csa 在线出版时间: 2023-12-25
CNKI 网络首发时间: 2023-12-27

撞, 已经成为人们关注的重点. 基于车载相机的交通目标跟踪作为汽车辅助驾驶技术中的关键环节, 能够实现道路目标信息的精准感知, 辅助驾驶员做出合理的决策, 方便后续路径规划, 降低事故发生率[1]. 通常车载相机受安装位置和拍摄角度的限制, 其拍摄视野较小, 图像中的物体多为小目标, 并且相机常受到车辆运动或抖动的影响发生动态位移, 为目标的跟踪增加了难度, 是辅助驾驶领域具有挑战性的问题[2].

现阶段主流的跟踪技术主要为基于深度学习的目标跟踪技术, 其按工作方式可划分为两类. 一类是基于检测的跟踪范式 (tracking by detection, TBD). TBD 范式首先通过目标检测算法得到检测信息, 然后送入跟踪器, 经过轨迹预测和数据关联等步骤完成多目标跟踪. 例如, 2016 年, Bewley 等人[3]提出了一种快速在线的多目标跟踪方法 SORT (simple online and realtime tracking), 该算法使用卡尔曼滤波进行更新和预测过程, 再使用匈牙利算法实现目标匹配; Wojke 等人[4]基于 SORT 提出了 DeepSORT (simple online and realtime tracking with a deep association metric) 算法, 做了一次额外的级联匹配, 采用 Re-ID (Re-identification)[5]提取外观特征, 增加马氏距离作为运动信息的约束, 显著改善了目标 ID 次数切换频繁的问题. 另一类是联合检测与跟踪范式 (joint detection and tracking, JDT). JDT 范式的核心思想是在单个网络中同时完成目标检测和 Re-ID 任务, 代表算法有 TransTrack[6]、FairMOT[7]、CenterTrack[8]等. 彭嘉淇等人[9]于 2022 年提出的一种基于时空一致性的 FairMOT 跟踪算法, 使用空间相关计算相邻帧之间的运动偏移信息, 对前一帧的目标响应进行变换以应对帧间响应不一致的情况, 显著改善了多目标跟踪不一致问题.

目前, TBD 与 JDT 范式在各个场景中均能够表现出较好的鲁棒性, 但大多仍保留了 Re-ID 步骤, 在跟踪过程中会消耗大量运算时间, 制约跟踪性能. 针对此问题, Zhang 等人[10]提出了 ByteTrack 算法, 该算法仅使用运动信息来关联预测框与检测框, 移除了运算速度较慢的外观特征匹配过程, 大幅提升了跟踪速度. 因此, 本文采用检测算法中速度和精度均表现出色的 YOLOv5 作为检测网络并与 ByteTrack 多目标跟踪方法相结合, 在交通目标中常见的车辆与行人目标跟踪测试中取得了良好的跟踪效果. 同时, 为了解决车载环境下车辆与行人小目标难以检测、相机位置实时更新导致跟踪位置偏移等问题, 分别基于 YOLOv5 与 ByteTrack 方法进行改进, 主要工作内容如下.

1) 利用 Transformer 的思想构造 C3TR 模块改进骨干网络结构的同时将 BiFPN 结构引入特征融合层, 从网络的特征学习能力和多尺度特征融合两个层面改进 YOLOv5 网络, 强化卷积网络对小目标的适应性.

2) 针对相机运动造成的目标关联匹配不准确问题, 设计了一种基于相机移动补偿的 ByteTrack 跟踪算法, 能够准确地修正相机运动产生的错误预测框.

3) 在 BDD100K 公共数据集上设计消融实验, 分析各模块对 YOLOv5 模型的性能影响; 在 MOT17 公共数据集上设计对比实验对比其他跟踪方法并分析, 证明了本文所提改进方案的有效性.

1 相关原理

1.1 YOLOv5 方法概述

YOLOv5[11]是 Ultralytics 公司于 2020 年 6 月推出的目标检测算法, 其根据网络结构的宽度和深度可分为 YOLOs、YOLOm、YOLOl、YOLOx 这 4 个版本. 本文选择结构最简单、检测速度最快的 YOLOv5s 作为基础网络, 网络结构主要包括输入层 (input), 骨干网络 (backbone), 颈部网络 (neck) 和输出层 (output) 这 4 个部分.

backbone 由 CBS (Conv+Batch Normalization+SiLU)、C3、SSPF (spatial pyramid pooling fast) 等模块构成. CBS 包含尺寸为 6×6 的大卷积核, 与先前版本的 Focus 模块相比在 GPU 上的表现更为高效. C3 模块由先前版本的 BottleneckCSP[12]转变而来, 作为残差特征学习的主模块包含了 3 个 CBS 模块与多个 Bottleneck, 能够有效防止网络退化. SPPF 模块由 SPP 空间金字塔池化改进而来, 设计了带有多个小尺寸池化核级联代替原有 SPP 模块中单个大尺寸池化核, 在丰富特征图表达能力的同时进一步提高了运行速度, 降低了模型的计算量, 实现了自适应尺寸的快速输出. YOLOv5s 的 neck 部分采用 FPN (feature pyramid network)[13]+PANet (path aggregation network)[14]的架构对提取特征进行融合, 达到了深层与浅层语义信息相互传递的效果. Output 部分根据 neck 输出的 3 个不同尺度特征图分配 3 个不同宽高比的 anchor 以预测和回归目标.

1.2 ByteTrack 方法概述

ByteTrack 作为 TBD 范式的跟踪算法, 仅使用运

动信息来关联预测框与检测框, 算法核心思想是根据卡尔曼滤波[15]预测当前帧在下一帧的跟踪轨迹位置, 将预测框与实际框之间的 IOU (intersection over union) 作为两次匹配时的相似度, 再利用匈牙利算法[16]完成匹配.

在数据关联策略上, 以往基于检测的跟踪方法在数据关联过程中只选择高分框进行匹配, 对于低于检测阈值的低分框直接丢弃, 此做法会使某些遮挡严重、运动模糊的目标被直接过滤, 显然是不合理的. 而 ByteTrack 在跟踪过程中利用检测框与跟踪轨迹的相似性, 在保留高分检测结果的同时从低分检测结果中去除背景, 挖掘出因遮挡、形变、环境模糊等情况导致检测置信度较低的真实目标, 降低了漏检率的同时提高了轨迹的连贯性.

2 本文方法

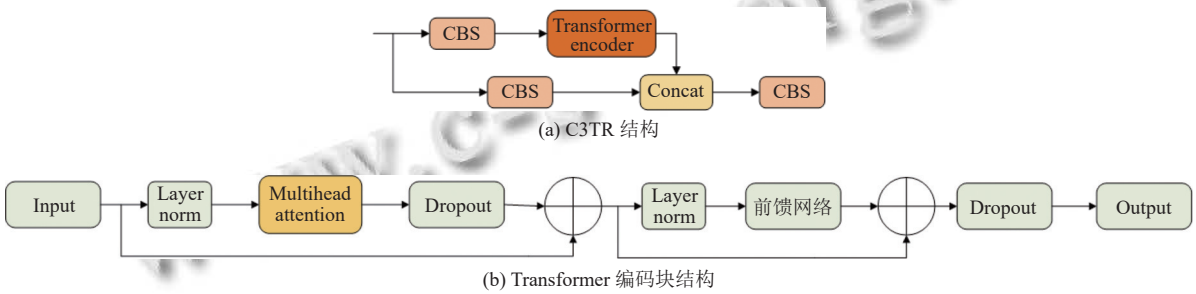
2.1 改进的 YOLOv5s 检测模型

2.1.1 构建 C3TR 模块改进骨干网络

基于卷积神经网络的目标检测模型在应用时通常存在一些问题, 卷积层的深入堆叠往往会扩大特征图的局部感受野, 不利于小目标检测. 相比之下, Transformer 在能够捕获丰富的全局信息的同时还可以关注到特征间的依赖关系. 鉴于驾驶环境中常存在远处目标尺度较小难以识别的情况, 同时考虑到 YOLOv5 存在网络末端特征图分辨率较低的问题, 本文借鉴了 Vision Transformer[17]的思想在骨干网络末端将 Transformer 中的多头自注意力机制 (multihead attention, MHA) 与卷积神经网络结合形成新的 C3TR 模块. 与原模块相比, 改进模块可以有效捕获全局信息以及上下文信息, 同时在本文实验数据集较大的前提下可以规避自注意力机制相比卷积神经网络缺少归纳偏置的缺陷, C3TR 结构如图 1(a) 所示.

C3TR 结构将原 C3 中的 Bottleneck 块替换为以多头自注意机制层为主的 Transformer 编码块, 编码块结构如图 1(b) 所示. 模块在多头自注意力层前应用层归一化, 层后应用 Dropout 与残差连接, 外加前馈连接层, 使网络收敛效果更好. 为处理平面图像特征, 在输入 Transformer 编码块前将二维特征展平为序列并映射到 query、key、value 这 3 个向量上, 之后利用自注意力机制将模型分为多个部分在网格单元间推理, 基于 query、key、value 这 3 个序列对每个网格单元取注意力加权求和并编码, 根据注意力权重决定图像中有价值的部分. 最后, 通过多头方式进一步完善了自注意力机制, 对每个头并行计算自注意力, 使模型能够在不同的关系子空间上工作. 自注意力计算方式如式 (1) 所示. 其中, $Attention(Q,K,V)$为自注意力特征, $Softmax$为激活函数, Q、K、V分别为查询向量、关键向量和值向量, d_k为缩放因子.

$$Attention(Q,K,V)=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{1}$$



(a) C3TR 结构

(b) Transformer 编码块结构

图 1 C3TR 结构示意图

2.1.2 引入跨尺度连接方法改进特征融合层

YOLOv5s 的颈部网络结构如图 2(a) 所示. 该网络利用自底向上的 FPN 和自顶向下的 PANet 两种特征融合方式丰富底层特征图所包含语义信息的同时增强了顶层特征图所包含的特征定位信息, 但此方式只融合高级信息, 骨干特征提取网络部分的浅层特征信息并没有被充分利用. 在车载相机拍摄物体中小目标特征占比较大的情况下, 采用 PANet+FPN 的特征融合策略难以有效识别该场景下的小目标特征. 而由 EfficientDet 算法[18]提出的加权特征金字塔结构 (BiFPN) 则可以更为高效地融合浅层信息, 防止特征丢失. 因此, 本文采用 BiFPN 改进颈部特征网络以提高小目标特征的识别能力.

BiFPN 结构如图 2(b) 所示, 该网络删除了那些对

特征网络贡献较小即只含一个输入边的节点, 在保留与 PANet 相同的双向信息传递方式的同时通过跨层连接来达到融合更多特性与保留深层次语义信息的效果. 原始 BiFPN 会根据不同输入特征的重要性对特征融合进行加权, 并将该结构作为一个整体反复使用以增强特征融合能力, 但是通过实验发现直接将带权重的 BiFPN 加入 YOLOv5s 会使训练出的模型参数量过多且检测效果不佳, 本文认为原因在于为输入特征设置权重与在特征层后添加注意力机制效果类似, 而注意力机制在不同场景下嵌入不同的网络位置对检测效果影响也都不同. 因此, 本文选择只使用其跨层特征融合结构替代原有特征融合结构而不采用其加权方法.

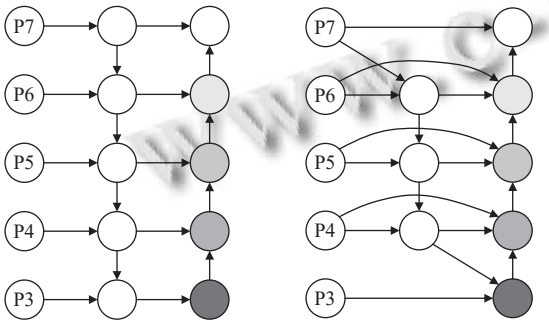


图 2 颈部网络结构对比

2.1.3 改进后的 YOLOv5s 网络结构

改进后的 YOLOv5s 网络结构如图 3 所示, 对比原 YOLOv5s, 使用本文所设计的 C3TR 模块替代原骨干网络末端最后两个 C3 模块, 之所以这样做是由于特征图分辨率越高输入数据量越大, 若在骨干网络其他位置使用会使模型参数量增加过多. 在颈部网络部分采用去除权重后的 BiFPN 跨尺度特征连接方法, 将骨干特征提取网络部分与 PANet 结构拼接, 使训练过程中始终有原始特征信息的参与, 达到融合更多特征信息的效果.

2.2 添加相机移动补偿的 ByteTrack 跟踪方法

ByteTrack 在跟踪过程中仅依赖目标的运动信息, 利用卡尔曼滤波的预测结果实现后续跟踪. 该做法的好处在于摈弃了外观特征匹配步骤, 大幅提高了跟踪速度, 但在驾驶环境中受刚性相机运动的影响图像平面前后变化显著, 这会使卡尔曼滤波预测边界框的位置失效, 导致目标跟踪框的位置偏移甚至丢失. 在缺乏相机运动的先验信息的前提下, 相邻帧间的图像配准可以近似相机的运动, 因此, 本文借鉴了 OpenCV 中相机稳定性模块的全局运动补偿 (global motion compensation, GMC)[19]策略设计视频稳定模块以改善相机运动导致假阴性错误较多的情况. 该技术本质上是将相机运动看作坐标系变换的过程, 通过求解坐标系的变换矩阵重新投影目标位置, 运动方向等信息, 有效揭示出背景运动情况并以此修正卡尔曼滤波的预测结果, 实现对当前帧预测框的位置信息矫正. 具体流程如图 4 所示.

首先, 应将画面中车辆及行人目标特征点滤除[20], 利用 SIFT 图像配准算法 (scale-invariant feature transform)[21]提取图像中的背景特征点并一一对应, 此后利用 RANSAC 算法 (random sample consensus)[22]即可得到平面坐标变换的仿射矩阵, 如式 (2) 所示. 式 (2) 中仿射矩阵Z可以将预测框从 k–1 帧坐标系转换到 k 帧坐标系, 矩阵包括M和T两部分, M为包含缩放与旋转部分的 2 维平面变化矩阵, 影响状态向量中所有元素, T为坐标系的平移变化向量, 只对观测目标的中心坐标位置产生影响. 在 ByteTrack 中卡尔曼滤波的状态向量表示为$[x, y, w, h, v_x, v_y, v_w, v_h]$, 其中$x$, y分别表示预测框的中心点横纵坐标, w和h表示预测框的宽度和高度, v_x, v_y, v_w, v_h分别表示前面 4 个变量的变化速度. 为了只用一次矩阵乘法完成中心点和宽高的仿射变换, 实现对卡尔曼滤波状态向量的一次性修正, 分别构造M_{k-1}^k 8 维矩阵和T_{k-1}^k 8 维向量, 如式 (4) 和式 (5) 所示. 最后通过式 (6) 即可完成卡尔曼滤波状态向量的更新修正, $\hat{x}_{k|k-1}$表示修正前的卡尔曼滤波状态向量, $\hat{x}'_{k|k-1}$表示修正后的卡尔曼滤波状态向量.

$$Z_{k-1}^k = [M_{2\times2}|T_{2\times1}] = \begin{bmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \end{bmatrix} \tag{2}$$

$$\hat{x}'_{k|k-1} = [x, y, w, h, v_x, v_y, v_w, v_h] \tag{3}$$

$$M_{k-1}^k = \begin{bmatrix} M & 0 & 0 & 0 \\ 0 & M & 0 & 0 \\ 0 & 0 & M & 0 \\ 0 & 0 & 0 & M \end{bmatrix} \tag{4}$$

$$T_{k-1}^k = [z_{13}\ z_{23}\ 0\ 0\ 0\ 0\ 0\ 0]^{\mathrm{T}} \tag{5}$$

$$\hat{x}'_{k|k-1} = M_{k-1}^k \hat{x}_{k|k-1} + T_{k-1}^k \tag{6}$$

2.3 车载环境下的交通目标跟踪方法

本文所提算法的整体流程概括如图 5 所示. 首先,

需要将车内相机拍摄下的视频拆分为帧序列并逐一输入改进后的 YOLOv5 检测器实现画面中车辆、行人目标的检测. 其次, 将检测所获取的置信度、类别和位置坐标等信息进行数据关联, 当输入视频第 1 帧时, 将所有检测结果加入跟踪轨迹集合, 从第 2 帧开始利用卡尔曼滤波预测已有轨迹目标在当前帧的目标位置, 经相机移动补偿模块修正后与当前帧中的目标检测框进行关联匹配. 最终, 通过关联匹配结果即可更新当前帧的轨迹集合, 获取交通目标的运动轨迹, 对没有成功匹配的已有跟踪轨迹将其保留 30 帧, 若超过 30 帧仍未成功匹配则将该轨迹从轨迹集合中移除, 图 5 中 Cnf 表示检测目标的置信度, Score 表示检测阈值.

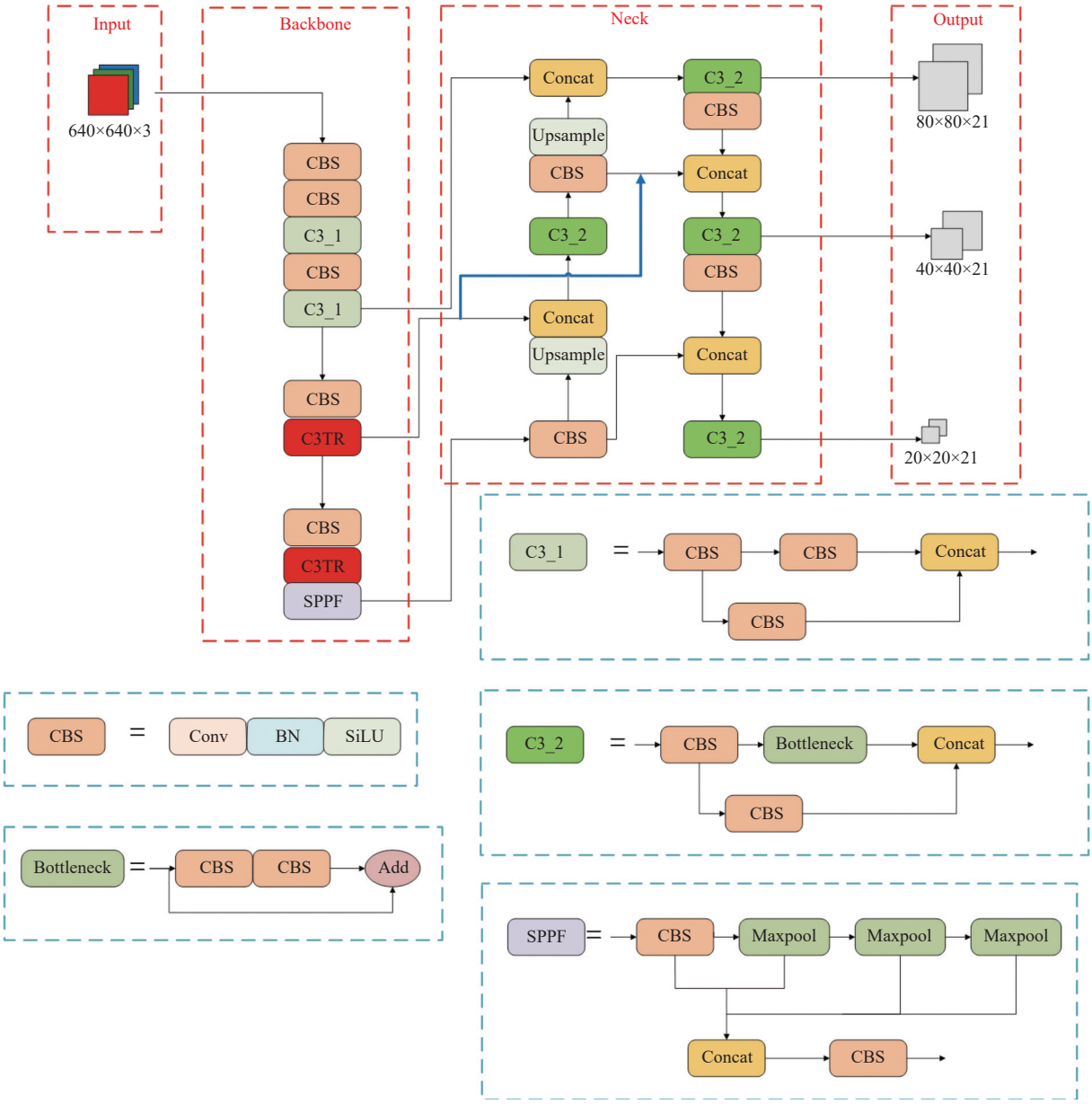


图 3 改进 YOLOv5s 网络

3 实验

3.1 数据集

选择 BDD100K 数据集[23]中车辆与行人两类用于检测模型的训练与消融实验结果分析. BDD100K 是一种大型自动驾驶场景检测数据集, 其中包含训练集 7 万、测试集 2 万、验证集 1 万共 10 万张图片, 可用于道路目标检测、车道线标记、全帧实例分割等任务.

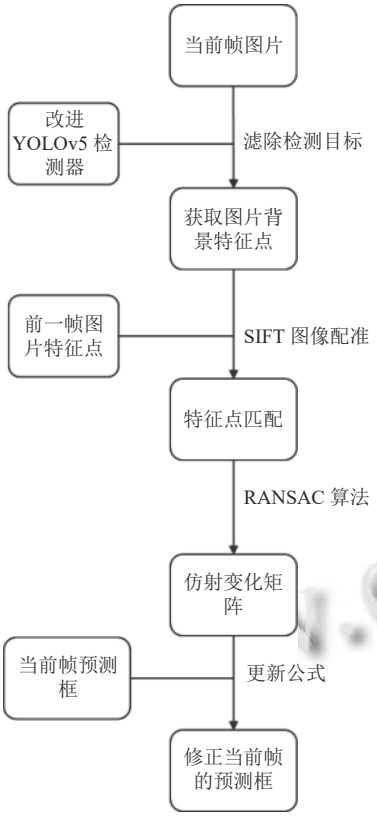


图 4 相机移动补偿策略

选择 MOT17 跟踪数据集[24]用于跟踪实验对比, MOT17 是一个用于评估多目标跟踪算法的基准数据集, 包含 120 个带有轨迹注释的视频片段, 共计超过 7 万帧, 涵盖了 7 个不同复杂度级别的不同场景.

3.2 实验环境

主要硬件环境如下, 操作系统: Linux 5.3.0 版本, Ubuntu 19.10, GPU: GeForce GTX1080ti, Cuda 版本 10.2. 编程语言采用 Python 3.8 版本, 结合 PyTorch 1.9.1 版本深度学习框架.

实验过程中设置初始学习率为 0.001, 批量尺寸大小为 16, 检测阈值为 0.4, 总训练轮次为 100 轮. 网络输入尺寸大小为 640×640, 选择 SGD 作为优化器.

3.3 性能评价指标

检测实验中, 在 BDD100K 数据集上对精确率, 召回率以及平均精度进行实验对比. 精确率 (P) 为预测为正样本的数据里预测正确的数量占比. 公式表达为:

$$P=\frac{TP}{TP+FP} \tag{7}$$

召回率 (R) 为真实为正样本的数据里预测正确的比例. 公式表达如式 (8).

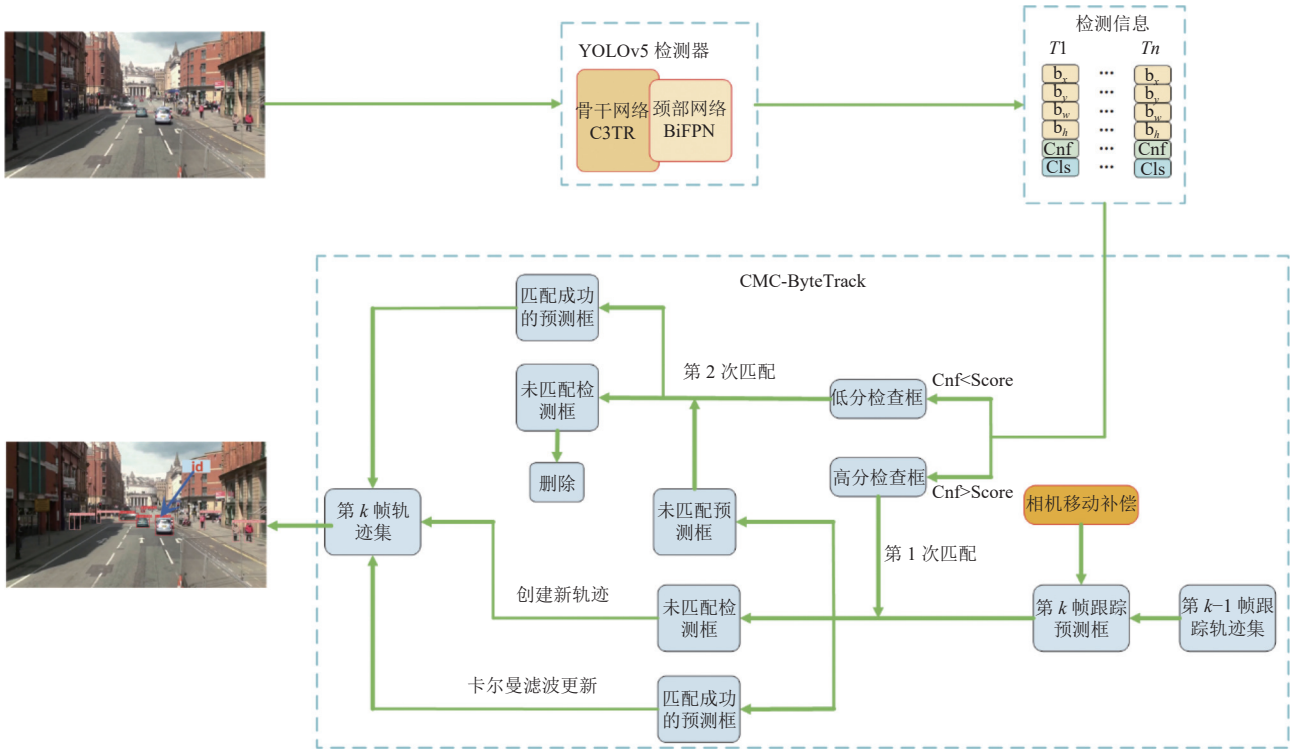


图 5 车载环境下的交通目标跟踪方法整体流程图

$$R=\frac{TP}{TP+FN} \tag{8}$$

平均精度 (mAP) 为所有类别预测精确度的平均值, 公式表达为:

$$mAP=\frac{1}{c}\sum_{i=1}^{c}AP_i \tag{9}$$

跟踪实验中, 在 MOT17 视频片段上评估目标的 IDP, IDR, $IDF1$分数, $MOTA$, $IDSW$, FPS 等多目标跟踪指标. 其中, IDP 与 IDR 表示目标 ID 的精确率与召回率, 把每个目标 ID 当作一类进行计算, 计算公式与上述检测部分相同. $IDF1$表示精确率与召回率的调和平均数, 其结果体现了视频中目标 ID 长时间稳定的能力. 公式如下所示:

$$IDF1=\frac{2IDTP}{2IDTP+IDFP+IDFN} \tag{10}$$

$MOTA$表示多目标跟踪准确度, 其数值结果体现了跟踪过程中确定目标的个数以及保持跟踪轨迹的性能, 公式如下:

$$MOTA=1-\frac{\sum(FN+FP+IDSW)}{\sum GT} \tag{11}$$

其中, $IDSW$ 是衡量跟踪算法稳定性的指标, 表示跟踪过程中目标 ID 的错误切换次数. FPS 是衡量跟踪算法实时性能的指标, 表示算法每秒能够处理的视频帧数, 公式表达为:

$$FPS=\frac{FrameNum}{ElapsedTime} \tag{12}$$

3.4 实验结果与分析

3.4.1 检测实验分析

1) 提取特征图效果

为验证提取网络结构的改进策略是否能够有效降低特征信息的丢失情况, 分别对改进前后的模型提取特征进行可视化结果对比, 效果如图 6 所示. 可以看出, 图 6(b) 中车辆线条提取的更明显, 车辆轮廓的保留完整度较图 6(a) 更高, 验证了改进策略对图像特征提取能力的有效提高.

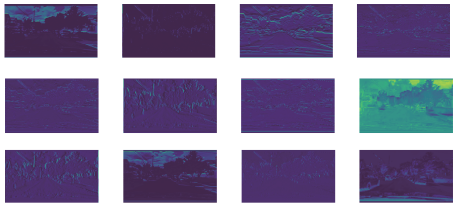
2) 消融实验

为了评估本文加入模块以及模块的组合对算法性能的优化程度, 基于 YOLOv5 构建消融实验, 从 BDD100K 中抽取 30 000 张图片作为训练集, 5 000 张图片作为验证集, 实验过程保持参数不变, 各类模型均训练 100 轮次, 实验结果如表 1 所示.

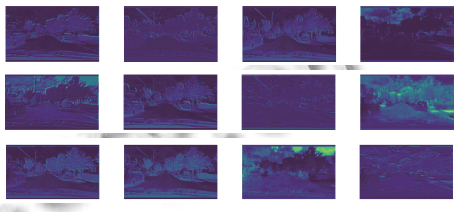
由表 1 可知, 在使用原始 YOLOv5s 网络结构进行模型训练能够得到 78% 的平均精度, 单独添加 BiFPN 和 C3TR 模块均能够提升模型平均精度, 分别可以提升 1.3 和 2.1 个百分点. 本文算法分别将 BiFPN、C3TR 融入 YOLOv5 模型最终在模型参数量只增加 4.9M 的情况下使 P 提高了 1.2%, R 提高了 4.3%, mAP 提高了 3.9%, 各项指标均达到了最优值, 说明主干网络提取质量的提高会更利于颈部网络的特征融合, 证明了改进策略的有效性.

3) 模型收敛性分析

为检验模型训练中是否存在过拟合的现象, 在图 7 中分别展示了 YOLOv5s 训练过程中在训练集和验证集上的损失值变化, 并与本文改进 YOLOv5s 对比. 图 7 中横坐标代表训练轮次, 纵坐标代表边界框损失值.



(a) 原特征提取效果



(b) 改进模型特征提取效果

图 6 特征图对比结果

表 1 消融实验

算法	P (%)	R (%)	mAP (%)	参数量 (M)
YOLOv5s	84.6	69.1	78.3	**13.7**
+C3TR	85.3	71.4	80.4	15.2
+BiFPN	84.1	70.9	79.6	14.8
本文方法	**85.8**	**73.4**	**82.2**	18.6

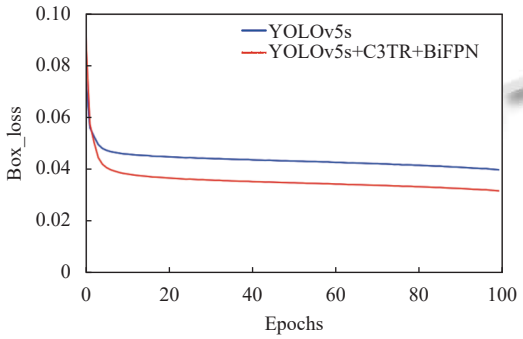
通过观察 Loss 曲线可知, 改进网络前后的损失曲线都比较平滑, 证明了在 BDD100K 数据集上训练出的模型具有良好的鲁棒性. 改进后的网络在训练集和验证集上取得了更快的收敛速度与更低的 Loss 值, 证明本文改进方法能够增强模型的泛化能力.

3.4.2 跟踪实验分析

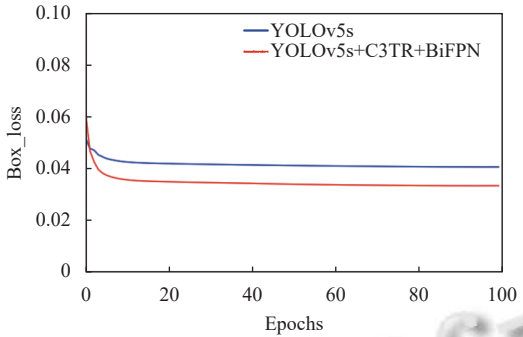
为进一步检验所提目标跟踪算法的有效性, 设计

实验将 CMC-ByteTrack 与原 ByteTrack 算法做对比, 测试结果如表 2 所示.

与原 ByteTrack 算法相比, *MOTA* 提升了 0.7%, *IDF*1 提高了 1.3%, 召回率提高了 5.1%. 可以看出, 召回率提升效果显著, 这是由于在原跟踪算法中, 随着车载相机在拍摄过程中发生位置或角度变化, 目标在图像中位置改变, 卡尔曼滤波预测下一帧图像中边界框的坐标位置会与实际位置不同, 导致实际检测框与预测框匹配效果较差甚至不能够匹配, 产生图中目标漏检或边界框位置不准确的现象, 而相机移动补偿模块可以很好地解决这一问题, 降低了跟踪框的丢失率.



(a) 训练集损失收敛图



(b) 验证集损失收敛图

图 7 模型收敛效果对比

表 2 跟踪改进效果

跟踪算法	*IDP* (%)	*IDR* (%)	*IDF*1 (%)	*IDSW*	*MOTA* (%)
ByteTrack	86.7	69.4	77.5	753	64.8
CMC-ByteTrack	83.9	74.5	78.8	842	65.5

对于 *IDSW* 次数增加以及 *IDP* 较改进前有所降低的现象, 本文认为原因如下.

加入相机移动补偿虽然能有效改善跟踪丢失问题, 但同时也需要对每一帧目标重新识别判断是否为之前追踪的同一个目标, 当图像中目标在前后帧图像配准中没有成功匹配时, 算法可能会将其认作不同的物体, 导致 ID 切换. 这也是为何 *IDR* 提升但 *IDSW* 次数却增加了. 而 *IDP* 的值表示已有跟踪结果中 ID 正确的轨迹所占的比例, 并不会考虑漏检或跟踪丢失的情况, 在 *IDSW* 次数增加的情况下 *IDP* 也会有所下降.

3.5 对比实验

对基于检测的跟踪算法而言, 跟踪性能不仅受跟踪算法影响, 且很大程度上依赖于检测算法的精确度. 为了评价改进检测算法是否对目标跟踪有提升效果以及改进跟踪策略是否具备有效性, 设计实验将改进前后的检测方法与当下性能表现较好的 DeepSORT, OCSORT 等跟踪算法相结合并在 MOT17 上进行对比展示, 结果如表 3 所示. 表 3 中 IM-YOLOv5 表示本文改进 YOLOv5 检测方法, CMC-ByTrack 表示增加相机移动补偿的 ByteTrack 跟踪方法.

表 3 对比实验结果

跟踪算法	*IDP* (%)	*IDR* (%)	*IDF*1 (%)	*IDSW*	*MOTA* (%)	*FPS* (f/s)
YOLOv5+ByteTrack	85.3	68.8	76.0	814	62.9	**39**
YOLOv5+DeepSORT	72.9	69.3	70.5	673	58.4	16
YOLOv5+OCSORT	82.5	70.6	76.4	874	62.6	37
IM-YOLOv5+ByteTrack	**86.7**	69.4	77.5	753	64.8	36
IM-YOLOv5+DeepSORT	74.6	70.1	72.3	**560**	61.5	16
IM-YOLOv5+OCSORT	81.2	72.0	76.5	931	62.8	34
IM-YOLOv5+CMC-ByteTrack	83.9	**74.5**	**78.8**	842	**65.5**	32

由表 3 中测试结果可知, 当跟踪方法保持不变时, 改进 YOLOv5 在 *MOTA*、*IDF*1 等多目标跟踪指标上均取得较明显的提升效果. 以 ByteTrack 为例, *MOTA* 能够提高 1.9%, *IDF*1 提高 1.5%, 但由于模型参数量的增加, 速度最终会降低 2–3 f/s, 在可以接受的范围内. 当检测方法保持不变时, ByteTrack 在速度与多目标跟踪准确度上要优于 DeepSORT 和 OCSORT 方法, 而 CMC-ByteTrack 在牺牲部分速度的前提下, 能够取得更高的跟踪精度. 得益于良好的检测性能与相机移动补偿对错误结果的有效修正, 本文所提方法在 *IDR*、*IDF*1 与 *MOTA* 上略优于其他方法, 相较基于原 YOLOv5+ByteTrack 的多目标跟踪方法分别提升 5.7%、2.8% 和 2.6%. 此外, 跟踪速度达到了 32 f/s, 能够满足跟踪任务中的实时性要求.

3.6 数据实测

为了更直观地展示本文改进算法相较原有算法的区别, 在 MOT17-14 视频上进行测试结果对比, 如图 8 所示. 其中, 第 1–3 行分别对应视频中第 22 帧、第 467 帧及第 572 帧. 置信度、ID 与类别信息均呈现于

目标检测框上方, 图 8 中蓝色实线表示目标运动状态下的跟踪轨迹.

由实验结果可知, 图 8(a) 中存在远距离车辆及行人漏检的情况, 而图 8(b) 则可以检测出图 8(a) 中难以识别到的目标并进行跟踪, 这是由于本文改进后的特征提取方法与多尺度特征融合方法能够有效关注到图像中的小目标特征, 实现小目标的精准检测. 从第 2 行的对比结果可以看出, 随着车辆转弯时相机的动态移动, 原跟踪方法中卡尔曼滤波的预测受到严重的干扰, 行人的跟踪预测边界框发生偏移, 而添加了相机移动补偿模块后, 即可通过估算相邻视频帧之间的相对运动进行补偿, 从而保持了运动一致性, 产生正确的跟踪预测结果.

(a) YOLOv5+ByteTrack 跟踪效果 (b) 改进 YOLOv5+CMC-ByteTrack 跟踪效果

图 8 车载环境下交通目标跟踪结果

4 结论与展望

本文以改进 YOLOv5s 作为检测器, 结合 ByteTrack 跟踪方法实现车载环境下交通目标的跟踪. 一方面, 将 Transformer 的思想融入 YOLOv5s 网络设计 C3TR 模块, 同时, 采用 BiFPN 结构替换原 FPN+PANet 结构, 有效提升了检测器的精确率, 改善了车载相机拍摄下小目标的漏检现象; 另一方面, 设计 CMC-ByteTrack 修正了刚性相机运动情况下的卡尔曼滤波预测框, 提高了相机位移时的跟踪准确度, 使跟踪算法能够更适应于车载相机下的交通目标跟踪场景. 实验结果证明, 改进后的算法在 *IDF*1 上提高了 2.8%, 在 *MOTA* 上提高了 2.6%, 能够应用于车载环境中的交通目标跟踪任务. 目前, 本文改进方法仍存在一些不足之处, 比如改进后的 YOLOv5 模型参数量较多、添加相机移动补偿的轨迹关联方法在 ID 切换次数上相较原方法有所增加等, 后续将针对此问题继续研究相关的改进方法以进一步提高跟踪稳定性.

参考文献

1 皇甫俊逸, 孟乔, 孟令辰, 等. 基于 GhostNet 与注意力机制

的 YOLOv5 交通目标检测. 计算机系统应用, 2023, 32(4): 149–160. [doi: 10.15888/j.cnki.csa.009048]
2 丁智. 车载视频监控运动目标检测与跟踪算法研究 [硕士学位论文]. 长沙: 湖南大学, 2016.
3 Bewley A, Ge ZY, Ott L, *et al*. Simple online and realtime tracking. Proceedings of the 2016 IEEE International Conference on Image Processing. Phoenix: IEEE, 2016. 3464–3468.
4 Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric. Proceedings of the 2017 IEEE International Conference on Image Processing. Beijing: IEEE, 2017. 3645–3649.
5 Zou Y, Yang XD, Yu ZD, *et al*. Joint disentangling and adaptation for cross-domain person re-identification. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 87–104.
6 Sun PZ, Cao JK, Jiang Y, *et al*. Transtrack: Multiple object tracking with Transformer. arXiv:2012.15460, 2020.
7 Zhang YF, Wang CY, Wang XG, *et al*. FairMOT: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision, 2021, 129(11): 3069–3087. [doi: 10.1007/s11263-021-01513-4]
8 Zhou XY, Koltun V, Krähenbühl P. Tracking objects as points. Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow: Springer, 2020. 474–490.
9 彭嘉淇, 王涛, 陈柯安, 等. 结合时空一致性的 FairMOT 跟踪算法优化. 中国图象图形学报, 2022, 27(9): 2749–2760. [doi: 10.11834/jig.220116]
10 Zhang YF, Sun PZ, Jiang Y, *et al*. ByteTrack: Multi-object tracking by associating every detection box. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 1–21.
11 周金涛, 高迪驹, 刘志全. 基于全景视觉的无人船水面障碍物检测方法. 计算机工程, 1–11. [doi: 10.19678/j.issn.1000-3428.0067238]
12 Srinivas A, Lin TY, Parmar N, *et al*. Bottleneck Transformers for visual recognition. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 16514–16524.
13 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944.
14 Liu S, Qi HF, Shi JP, *et al*. Path aggregation network for instance segmentation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8759–8768.
15 任其亮, 程昊东. 针对小车状态感知的卡尔曼滤波多传感器融合算法. 重庆理工大学学报 (自然科学), 2022, 36(11): 176–182.
16 姜燕, 王道波, 林飞, 等. 基于匈牙利融合遗传算法的多无人机不平衡目标分配. 电光与控制, 2023, 30(5): 6–10, 22. [doi: 10.3969/j.issn.1671-637X.2023.05.002]
17 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2010.
18 Tan MX, Pang RM, Le QV. EfficientDet: Scalable and efficient object detection. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2019. 10778–10787.
19 程松. 基于图像校正的无人机视频多目标跟踪方法研究 [硕士学位论文]. 长春: 吉林大学, 2023. [doi: 10.27162/d.cnki.gjlin.2023.001885]
20 Zhang GC, Vela PA. Good features to track for visual SLAM. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1373–1382.
21 李文举, 王子杰, 崔柳. 基于多特征融合和改进 SIFT 的目标跟踪算法. 郑州大学学报 (理学版), 2024, 56(1): 40–46. [doi: 10.13705/j.issn.1671-6841.2022268]
22 Schnabel R, Wahl R, Klein R. Efficient RANSAC for point-cloud shape detection. Computer Graphics Forum, 2007, 26(2): 214–226. [doi: 10.1111/j.1467-8659.2007.01016.x]
23 Yu F, Chen HF, Wang X, *et al*. BDD100K: A diverse driving dataset for heterogeneous multitask learning. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2018. 2633–2642.
24 Aharon N, Orfaig R, Bobrovsky BZ. BoT-SORT: Robust associations multi-pedestrian tracking. arXiv:2206.14651, 2022.

(校对责编: 牛欣悦)