

基于双重注意力机制、轨迹预测与无人机遥感技术的人车流量检测方法

曾张帆, 谢临风

(湖北大学 计算机与信息工程学院, 湖北 武汉 430062)

摘要: 人、车流量的精确检测统计对于公共安全和资源管理有重要意义。针对现有检测统计方法设备投入大、维护成本高、检测区域受限、易受环境因素影响等问题,提出一种基于双重注意力机制、轨迹预测与无人机遥感技术的人车流量检测方法。该方法在RefineDet网络的基础上加以改进,通过引入双重注意力机制、使用特征融合RNN替换TCB模块等方式增强对小目标的识别能力。同时通过自适应卡尔曼滤波预测轨迹,实现了对运动目标的实时跟踪,避免了因目标距离过远、行动方式特殊导致的错误统计。自建数据集训练网络模型,并在不同场景下进行测试。实验结果表明,所提改进模型相较对照模型的MOTA提高了2.3%~3.3%,在精度接近最新模型的同时检测速度更快,基本达到实时性需求;在多种场景下的综合评价指标F值均能达到95%以上,在多种实际场景检测中的误检率均低于0.3%。

关键词: 无人机;人车流量统计;双重注意力机制;轨迹预测;多目标跟踪

DOI:10.11907/tj.231524

中图分类号:TP391.4

文献标识码:A

开放科学(资源服务)标识码(OSID):

文章编号:1672-7800(2024)006-0075-10



Pedestrian and Vehicle Flow Detection Method Based on Dual Attention Mechanism, Trajectory Prediction, and Drone Remote Sensing

ZENG Zhangfan, XIE Linfeng

(School of Computer Science and Information Engineering, Hubei University, Wuhan 430062, China)

Abstract: Accurate detection and statistics of human and vehicle traffic flow are of great significance for public safety and resource management. A pedestrian and vehicle flow detection method based on dual attention mechanism, trajectory prediction, and drone remote sensing technology is proposed to address the issues of high equipment investment, high maintenance costs, limited detection areas, and susceptibility to environmental factors in existing detection and statistical methods. This method is improved on the basis of the RefineDet network by introducing a dual attention mechanism and replacing the TCB module with feature fusion RNN to enhance the recognition ability of small targets. At the same time, adaptive Kalman filtering is used to predict trajectories, achieving real-time tracking of moving targets, avoiding erroneous statistics caused by target distance and special action modes. Train network models with self built datasets and conduct tests in different scenarios. The experimental results show that the proposed improved model has a MOTA improvement of 2.3% to 3.3% compared to contrast models. The detection speed is faster while the accuracy is close to the latest model, basically meeting the real-time requirements; The comprehensive evaluation index F value can reach over 95% in various scenarios, and the false detection rate in various actual scene detection is less than 0.3%.

Key Words: UAV; pedestrian and vehicle flow detection statistics; dual attention mechanism; trajectory prediction; multi-object tracking

0 引言

随着智慧城市的建设与发展,公共安全需求不断增

长,商业、交通、安防等领域对人、车流量实时精确统计的应用需求越来越强烈,尤其是广场、主干道、风景名胜等大型开放性公共场所的人、车流量实时检测统计具有重要价值,可根据监测信息对人流、车流进行及时高效的疏导

收稿日期:2023-05-20

基金项目:国家自然科学基金项目(61902114, 61977021);湖北省自然科学基金面上项目(2021CFB503)

作者简介:曾张帆(1983-),男,CCF会员,湖北大学计算机与信息工程学院副教授、博士生导师,研究方向为软件工程、深度学习;谢临风(2001-),男,湖北大学计算机与信息工程学院学生,研究方向为图像处理、人工智能。本文通讯作者:谢临风。

与调控,以提高道路利用率、减少拥堵,以及预防踩踏和连环交通事故等危害公共安全事件的发生。

现有人、车流量监测方法可以分为接触式方法和非接触式方法两类^[1]。接触式方法包括在监测点安装闸机、压敏装置等,设备安装与运营维护成本高,闸机通过速度慢、效率低、易造成拥堵且会给通行者带来不舒适的体验;压敏装置在多人多车同时通过时易产生误判。非接触式方法主要包括基于红外线的检测技术和基于机器视觉的监控统计技术。红外检测对单个目标的检测准确率高,但需要限制道路通行,且当多个目标或行人携带大尺寸的行李物品等障碍物通过时易产生误判。依托于监控摄像头的机器视觉检测方法随着机器学习技术的进步而快速发展,但现有方法在面对小体积目标和遮挡物时误检率较高,且监控摄像头高成本、高能耗、检测区域受限、易受环境因素影响的缺点也不易克服。

为解决开放区域临时性短时间人车流量检测问题,如人流车流高峰期的广场、公园、街道、景区等区域,本文依托无人机进行人车流量的检测统计。无人机可以调整飞行高度和拍摄角度,减小遮挡、光照、烟雾等环境因素的影响,与传统方法相比部署更灵活、成本更低。

本文使用的无人机利用无线传输将视频发送到主机,在主机上利用检测网络与流量统计算法完成人车流量的检测统计。在目标检测部分,本文以 RefineDet 网络^[2]为基础,通过对 2 个卷积神经网络(Convolutional Neural Network, CNN)与 1 个循环神经网络(Recurrent Neural Network, RNN)进行并联,先寻找候选区域再由细定位模块精确获得目标位置,进而引入注意力机制,增强了对被遮挡小目标的定位能力,在检测精度相较 YOLOv5^[3]等 one-stage 方法显著提升的同时,也取得了优于 Faster RCNN^[4]等 two-stage 目标检测网络的时效性;在多目标跟踪部分,本文基于 Deep SORT 算法^[5]进行了改进,使用自适应卡尔曼滤波预测轨迹,利用运动匹配与外观匹配的线性加权作为度量匹配检测与预测结果,并引入级联匹配解决目标被长时间遮挡的问题。与端到端方法相比,本文方法具有以下优势:①灵活性。两阶段方法可以独立优化每个步骤,也便于针对不同场景和资源限制进行调整,例如在暗光等检测困难的环境下使用参数量更大的检测模型,分配更多计算资源给目标检测部分;②鲁棒性。两阶段方法可以更好地处理复杂情况,例如丢失目标时,目标跟踪阶段仍有可能正确地跟踪目标。

1 基于 RefineDet 网络的改进检测网络

1.1 RefineDet 网络简介

目标检测算法分为 one-stage 和 two-stage 两种。one-stage 方法使用一个 CNN 网络直接预测不同目标的类别与位置而单独寻找候选区域,典型方法包括 SSD (Single

Shot MultiBox Detector)^[6]、YOLO^[7]系列,其特点是速度快但准确度较低。two-stage 是指检测算法需要分两步完成:首先使用选择性搜索算法(Selective Search)或 CNN 网络产生候选框,然后在候选框上进行分类与回归,典型方法包括 R-CNN^[8]、Faster-RCNN 系列,其特点是准确度较高但速度较慢。

RefineDet 网络继承了以上两种方法的优点并一定程度上克服了它们的缺点。该网络使用 VGG16 或 ResNet101 作为特征提取网络,主要由 Anchor 细化模块(Anchor Refinement Module, ARM)和目标检测模块(Object Detection Module, ODM)组成,结构如图 1 所示。

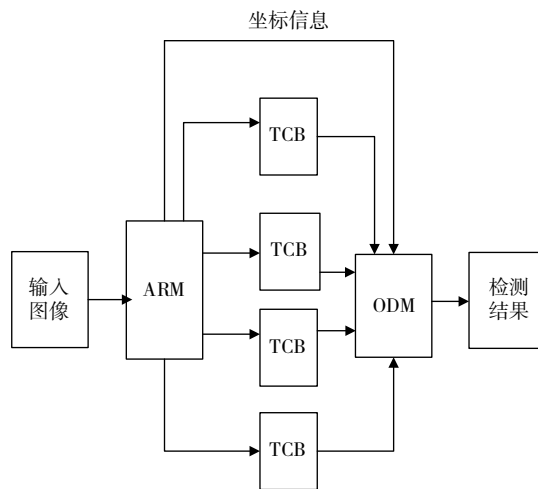


Fig. 1 Architecture of RefineDet

图1 RefineDet 网络结构

ARM 和 ODM 两个模块相互连接模仿了 two-stage 结构,ARM 生成 refined anchor boxes, ODM 在 refined anchor boxes 的基础上进行回归,从而继承了 two-stage 方法高准确率的优点。ARM 的主要作用是识别并移除 negative anchors 以减少分类器的搜索空间,同时粗调 anchor 的位置和大小,为后续回归器提供更好的初始化。ODM 的结构与 one-stage 的 SSD 网络类似,其以 refined anchor 为输入,回归精确的对象位置,并基于 refined anchor 预测多类别标签,从而继承了 one-stage 方法快速的优点。

为了连接 ARM 与 ODM, Zhang 等^[2]设计了传输连接块(Transfer Connection Block, TCB)传输 ARM 中的特征,将 ARM 中不同层的特征转换为 ODM 所需的形式,以便 ODM 可以共享 ARM 中的特征。同时为了匹配它们之间的维度,使用反卷积操作放大 high-level 特征图,并以逐元素的方式对它们进行求和。在求和后加入卷积层,以保证检测特征的可分辨性。RefineDet 的时间复杂度和空间复杂度均为 $O(N)$, 其中 N 为网络的总参数量。

1.2 改进检测网络

RefineDet 网络虽然在保证高效的同时实现了较高的检测精度,但对小目标的识别仍有欠缺。针对无人机拍摄

高度较高、目标较小, 以及原有 RefineDet 网络对小目标检测任务精度受限的问题, 本文提出一种适用于小目标检测的 RefineDet 改进算法, 网络框架如图 2 所示。具体流程为: ①将图片输入 ARM, 得到 refined anchor 和 feature map,

并去除部分负样本; ②将 ARM 输出的 refined anchor 传递给 ODM, 同时将 feature map 传递给 RNN 以融合高层与底层特征信息, 并转化成 ODM 所需的形式; ③由 ODM 回归精确的对象位置, 并基于 refined anchor 预测多类别标签。

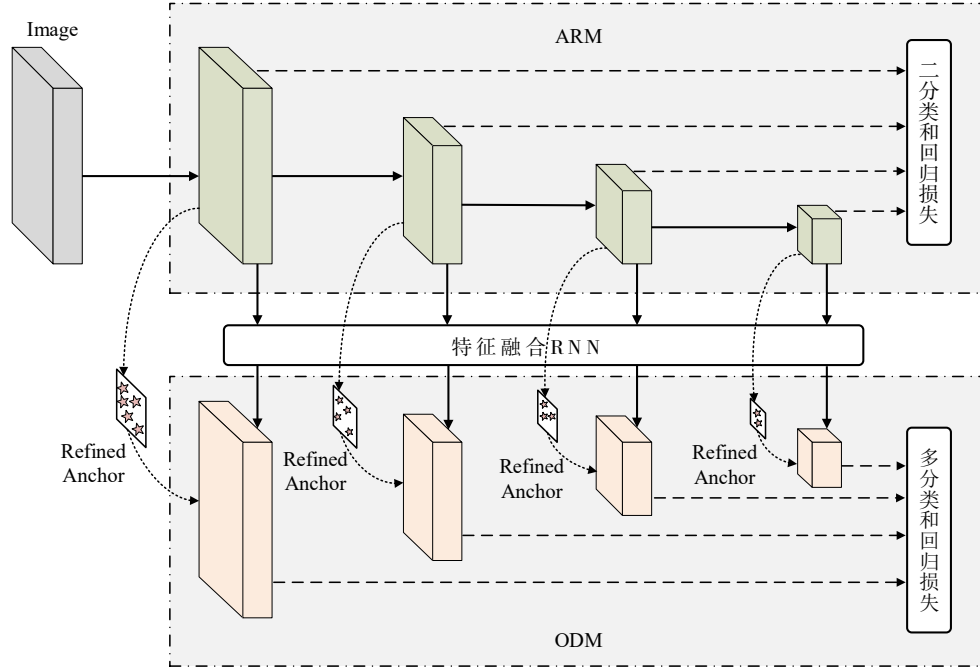


Fig. 2 Improved RefineDet network structure

图 2 改进 RefineDet 网络结构

1.2.1 引入注意力机制的 ARM

特征提取网络选取 ResNet101, ARM 部分 4 个灰色矩形块 (feature map) 的 size 分别为 40×40 、 20×20 、 10×10 、 5×5 , 其中前 3 个矩形块为 ResNet101 网络本身的输出层, 第 4 个矩形块为另外添加的残差模块。为增强对小目标的检测能力, 在 ARM 中引入双重注意力机制^[9]。注意力机制通过关注重要特征、抑制不必要特征来增强网络对特征较少的小目标的识别能力^[10]。在 ResNet101 的每个 ResBlock 后添加空间注意力模块 (Spatial Attention Module, SAM) 和通道注意力模块 (Channel Attention Module, CAM), 整体结构如图 3 所示。ResNet101 网络的时间和空间复杂度与 ResBlock 的数量和特征图大小成正比。

(1) SAM 模块关注“有意义的信息在哪里”, 利用任意两点特征之间的关联相互增强其特征的表达。结构如图 4 所示。

具体计算过程如下: 特征图 $A(C \times H \times W)$ 首先通过卷积降维获得特征图 B, C, D , 然后将其 reshape 为 $C \times N$, 其中 $N = H \times W$; 然后将 reshaped B 的转置与 reshaped C 相乘得到任意两点特征之间的关联强度矩阵, 再通过 SoftMax 操作获得 spatial attention map $S(N \times N)$; 接着将 S 中的响应值作为权值对特征 D 进行融合, 并 reshape 为原来形状, 最后与 A 相加得到输出 E 。

Spatial attention map 的计算公式为:

$$S \in R^{N \times N}: s_{ij} = \frac{\exp(B_i \cdot C_j)}{\sum_{i=1}^C \exp(B_i \cdot C_j)} \quad (1)$$

式中: s_{ij} 度量第 i 个位置对第 j 个位置的影响, 即第 i 个位置与第 j 个位置之间的关联程度。该值越大, 二者越相似。

Output 的计算公式为:

$$E \in R^{C \times H \times W}: E_j = \alpha \sum_{i=1}^N (s_{ij} D_i) + A_j \quad (2)$$

式中: α 表示尺度系数, 初始化为 0, 逐渐学习分配至更大权重; 每个位置的结果特征 E 为所有位置与原始位置的加权和。

(2) CAM 模块关注“哪些信息是有意义的”, 建模通道之间的关联从而生成通道注意力图, 使各个通道之间能产生全局关联。结构如图 5 所示。

具体计算过程如下: 与 SAM 相似, 本文分别对 A 进行 reshape ($C \times N$) 以及 reshape、transpose ($N \times C$); 将得到的两个特征图相乘获得任意两个通道间的关联强度, 再通过 SoftMax 得到 channel attention map $X(C \times C)$; 接着以 X 为权值对 reshaped A 进行融合, 再 reshape 为原来形状, 最后与 A 相加得到输出 E 。

Channel attention map 的计算公式为:

$$X \in R^{C \times C}: x_{ij} = \frac{\exp(A_i \cdot A_j)}{\sum_{i=1}^C \exp(A_i \cdot A_j)} \quad (3)$$

式中: x_{ij} 度量第 i 个通道对第 j 个通道的影响。

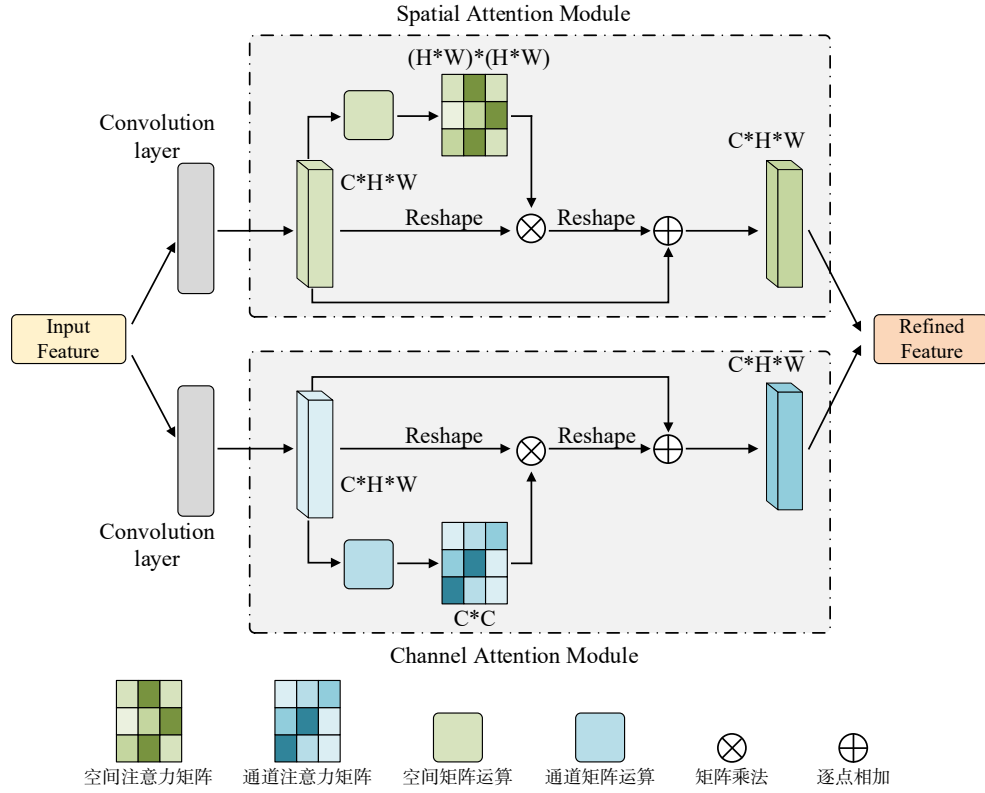


Fig. 3 Architecture of dual attention module

图3 双重注意力模块结构

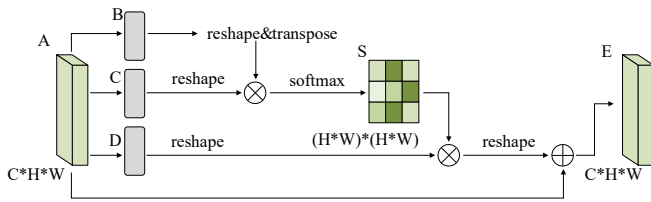


Fig. 4 SAM module structure

图4 SAM模块结构

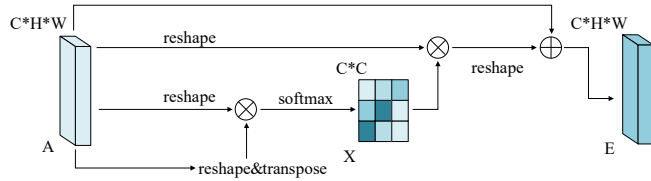


Fig. 5 CAM module structure

图5 CAM模块结构

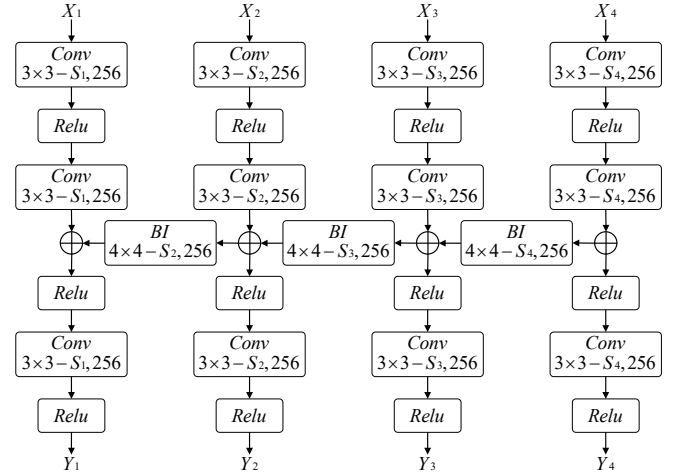
Output的计算公式为:

$$E \in R^{C \times H \times W}; E_j = \beta \sum_{i=1}^C (x_{ij} A_i) + A_j \quad (4)$$

式中: β 表示尺度系数,初始化为0,并逐渐学习分配到更大权重。每个通道的结果特征 E 为所有通道特征与原始特征的加权和。

1.2.2 特征融合与转化

采用1个RNN代替RefineDet网络中的多个TCB模块并加以改进,结构如图6所示。通过RNN将来自ARM中不同层的特征转化为ODM所需的形式,以便ODM能够共享ARM中的特征。同时,RNN也对大尺度的上下文进行



BI: Bilinear Interpolation,双线性插值

Fig. 6 Recurrent neural network structure

图6 RNN结构

整合,将高层特征传递到每个特征层并加入到 transferred features 中,形成大量相互关联的上下文信息,以提高对小目标物体的检测能力。原有TCB模块为了匹配不同层特征的维度,使用转置卷积的操作放大高层特征图,并逐元素求和。转置卷积虽然可以扩大尺寸、细化特征图,但如果卷积核大小不能被步长整除,转置卷积的输出结果就会不均匀重叠(Uneven Overlap),会使图像中某个部位的颜色比其他部位更深,即棋盘效应。为避免上述问题,本文首先使用双线性插值方法扩大尺寸,然后再进行卷积。

1.2.3 ODM

ODM 由 RNN 的输出和预测层组成, 用于回归精确的对象位置, 并基于 refined anchor 预测多类别标签。ODM 的结构与 SSD 类似, 但其浅层 feature map 首先融合了高层 feature map 信息, 然后基于每层 feature map 进行 bounding box 预测, 最后整合各层结果。同时, 与 SSD 直接使用 default box 进行回归不同, 本文模型采用两步策略, ODM 使用 ARM 精炼后的 refined anchor boxes 进行进一步检测, 使得检测效果更好, 尤其是提高了对小体积目标的检测能力。

1.2.4 难负例挖掘

大多数 box 的标签为背景标签, 即负样本, 虽然通过 ARM 能够过滤掉一部分负样本, 但其数量依然过多, 因此本文采用难负例挖掘 (Hard Negative Mining) 方法缓解极端的正负样本不平衡问题。将正负样本比例设定为 1:3, 选择具有最高损失值的负样本。

1.2.5 损失函数

损失函数包括 ARM loss 和 ODM loss 两种, 其中 ARM loss 包括 cross-entropy (二分类) 和 smooth-L1, ODM loss 包括 softmax-loss 和 smooth-L1。ARM 为每个 anchor 分配一个二分类标签 (分别表示是否为对象), 同时回归其位置和大小以获得 refined anchor。随后, 将负置信度小于阈值的 refined anchor 传递给 ODM, 进一步预测对象类别以及准确的对象位置和大小。

损失函数公式为:

$$\mathcal{L}(\{p_i\}, \{x_i\}, \{c_i\}, \{t_i\}) = \frac{1}{N_{ARM}} \left(\sum_i \mathcal{L}_b(p_i, [l_i^* \geq 1]) + \sum_i [l_i^* \geq 1] \mathcal{L}_r(x_i, g_i^*) \right) + \frac{1}{N_{ODM}} \left(\sum_i \mathcal{L}_m(c_i, l_i^*) + \sum_i [l_i^* \geq 1] \mathcal{L}_r(t_i, g_i^*) \right) \quad (5)$$

式中: i 表示一个 batch 中的第 i 个 anchor; l_i^* 表示 an-

chor i 的 ground truth 类别; g_i^* 表示 anchor i 的 ground truth 位置和大小; p_i 表示置信度; x_i 表示 ARM 中 anchor 的坐标; c_i 表示预测类别; t_i 表示 ODM 中的预测框坐标信息; N_{ARM} 和 N_{ODM} 分别表示 ARM 和 ODM 中 positive anchor 的数量; $\mathcal{L}_b$ 表示二分类损失; $\mathcal{L}_m$ 表示多分类损失; $\mathcal{L}_r$ 表示回归损失; $[l_i^* \geq 1]$ 表示如果负置信度大于阈值 θ 则返回 1, 否则返回 0。

1.2.6 推理

推理阶段, ARM 首先过滤掉负置信度大于阈值 θ 的 anchor, 然后 refine 剩余 anchors 的位置和大小; 然后由 ODM 接收这些 refined anchor, 并输出每幅图像前 400 高置信度的检测结果; 最后使用非极大值抑制算法 (Non Maximum Suppression, NMS), IoU 限定为 0.45, 并保留每幅图像前 200 个高置信度的检测结果作为最终结果。

2 基于多目标跟踪的人车流量统计算法

传统的人、车流量统计方法主要为单线统计法, 通过统计跨越参考线的目标数量计算流量^[11]。然而针对复杂路况, 传统方法有诸多弊端, 例如同一个目标在参考线附近徘徊将导致无效通过次数被计入, 镜头抖动会导致参考线变动等。为解决上述问题, 本文设计了基于多目标跟踪的流量统计算法, 主要包括预测、检测、关联、更新 4 个核心步骤, 具体为: 首先输入图像, 通过目标检测器获取检测框; 然后提取目标的表观特征与运动特征, 使用卡尔曼滤波预测轨迹; 其次利用运动匹配、外观匹配和级联匹配完成数据关联; 最后更新跟踪序列。具体流程如图 7 所示。在目标检测阶段处理当前帧时, 多目标跟踪方法可以同时处理前一帧检测到的目标, 两个步骤并行运行, 相较 DeepSORT 算法提高了效率。

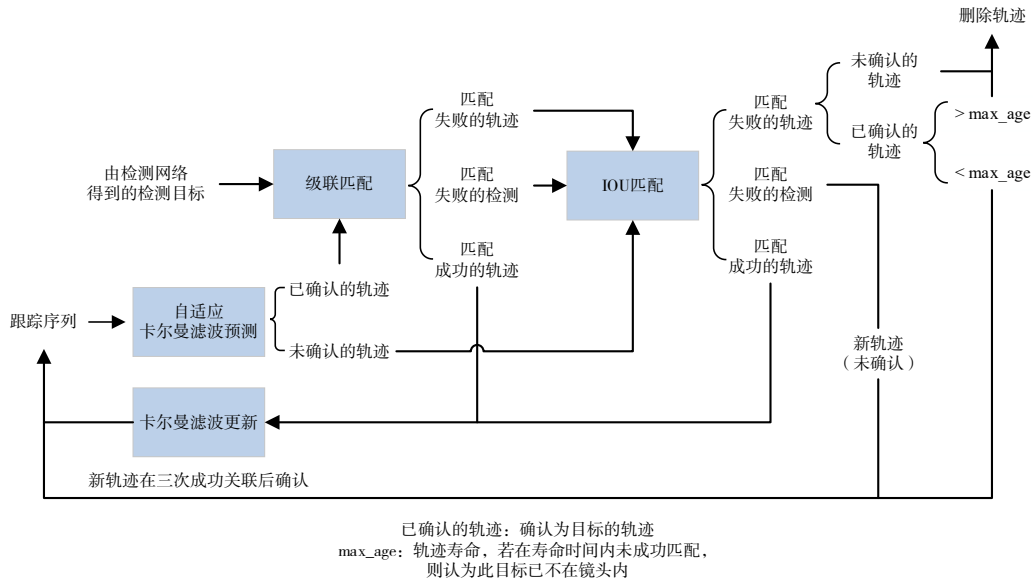


Fig. 7 Multiple object tracking method flow

图 7 多目标跟踪流程

2.1 基于噪声自适应卡尔曼滤波的轨迹预测

2.1.1 卡尔曼滤波

卡尔曼滤波是一种利用线性系统状态方程,通过系统输入输出观测数据,对系统状态进行最优估计的算法,主要分为预测和更新两个步骤。在相邻帧,人、车的位移较小,可近似看作线性运动,认为目标间运动独立,则运动目标的系统状态可表示为:

$$X = \left[x, y, \frac{w}{h}, h, dx, dy, d\frac{w}{h}, dh \right]^T \quad (6)$$

式中: x 和 y 分别表示 box 中心点的横纵坐标; w 和 h 分别表示 box 的宽和高; dx 、 dy 、 $d\frac{w}{h}$ 、 dh 分别表示其变化速度。

考虑到人、车目标检测中检测框高的变化幅度一般大于检测框宽的变化幅度,故给检测框高度变化赋予比宽度更大的影响权重,以使状态估计更加稳定。

状态方程表示为:

$$x_k = Ax_{k-1} + B\mu_{k-1} + \omega_{k-1} \quad (7)$$

式中: A 为状态增益矩阵,维度为 8×8 。

设定观测状态 z ,表示为:

$$z = (x_{\min} \quad y_{\min} \quad x_{\max} \quad y_{\max})^T \quad (8)$$

式中: $x_{\min}$ 、 $y_{\min}$ 、 $x_{\max}$ 、 $y_{\max}$ 分别表示检测框 x 、 y 坐标的最小值和最大值。

观测方程表示为:

$$z_k = Hx_k + v_k \quad (9)$$

式中:增益矩阵 H 维度为 4×8 。

噪声包括过程噪声 $Q(8 \times 8)$ 、观测噪声 $R(4 \times 4)$ 和先验估计误差 $P(8 \times 8)$ 。

预测(时间更新方程)步骤为首先计算先验估计(状态控制一般不设立)。公式为:

$$\hat{x}_k^- = A\hat{x}_{k-1} \quad (10)$$

然后计算先验估计协方差。公式为:

$$P_k^- = AP_{k-1}A^T + Q \quad (11)$$

更新(状态更新方程)步骤为首先计算卡尔曼增益。公式为:

$$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1} \quad (12)$$

再计算最优估计。公式为:

$$\hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-) \quad (13)$$

先验估计误差的协方差更新公式为:

$$P_k = (I - K_k H) P_k^- \quad (14)$$

2.1.2 观测噪声自适应设计

卡尔曼滤波的准确性依赖于系统模型性能和噪声设定。为解决多目标跟踪任务中噪声来源复杂、无法精确获取其统计特性的问题,本文参考文献[12]和[13]中的原理设计了基于残差的观测噪声自适应方法,不断在线估计和修正噪声统计特性以提高滤波精度。

k 时刻的新息 d_k 定义为滤波器实际观测值 Z_k 和预测观测值 $\hat{Z}_k$ 之差。表示为:

$$d_k = Z_k - \hat{Z}_k = H_k X_k + y_k - H_k \hat{X}_k^- = H_k \tilde{X}_k^- + y_k \quad (15)$$

式中: $\tilde{X}_k^- = X_k - \hat{X}_k^-$ 为一步预测误差。

新息估计方差表示为:

$$\hat{C}_{d_k} = \frac{1}{W} \sum_{i=k-W+1}^k d_i d_i^T \quad (16)$$

式中: W 为滑动数据窗口长度, $k \geq W$ 。

k 时刻的残差定义为滤波器的实际观测值 Z_k 与估计观测值 $\hat{Z}_k$ 之差。表示为:

$$r_k = Z_k - \hat{Z}_k = R_k C_{d_k}^{-1} d_k \quad (17)$$

残差方差表示为:

$$C_{r_k} = E(r_k r_k^T) = R_k C_{d_k}^{-1} R_k \quad (18)$$

滤波增益表示为:

$$K_k = P_k^- H_k^T C_{d_k}^{-1} = P_k^- H_k^T R_k^{-1} \quad (19)$$

基于残差的自适应 R_k 估计值为:

$$\hat{R}_k = \hat{C}_{r_k} + H_k P_k H_k^T \quad (20)$$

式中: $\hat{C}_{r_k}$ 为残差估计方程。公式为:

$$\hat{C}_{r_k} = \frac{1}{W} \sum_{i=k-W+1}^k r_i r_i^T \quad (21)$$

滑动估计窗口 W 长度越小,滤波器对动态变化的灵敏性越好,但采样点较少使得滤波器容易偏置和发散; W 越大则估计值偏差越小,但会使滤波器跟踪短期变化的能力变弱。经过实验验证,本文取 $W = 15$ 。

2.2 数据关联

传统的数据关联一般使用匈牙利算法(Hungarian Algorithm),但在目标密集、遮挡严重的场景下极易导致原有目标被识别为新目标(即ID switch),因此本文使用运动特征关联与外观特征关联的线性加权来计算相似度,并引入级联匹配减少被遮挡目标发生ID switch的次数。具体流程如图8所示。

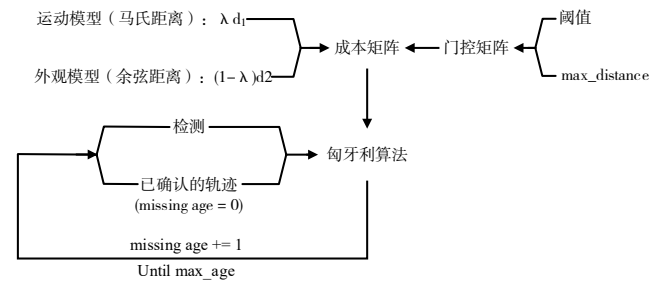


Fig. 8 Data association flow

图8 数据关联流程

2.2.1 运动特征关联

运动特征关联是通过马氏距离对目标的运动状态与预测结果进行相似性匹配。计算公式为:

$$d^{(1)}(i, j) = (d_j - y_i)^T S_i^{-1} (d_j - y_i) \quad (22)$$

式中: d_j 表示第 j 个检测框的位置; y_i 表示第 i 个跟踪器的目标预测位置; S_i 表示检测位置与平均跟踪位置之间的协方差矩阵。

利用 X^2 分布,如果马氏距离小于指定的阈值 $t^{(1)}$,则关联成功。表示为:

$$b_{i,j}^{(1)} = 1[d^{(1)}(i,j) \leq t^{(1)}] \quad (23)$$

式中: $t^{(1)}$ 为 X^2 分布 95% 置信度阈值,本文取 9.487 7。

2.2.2 外观特征关联

人、车运动的不确定性很高,仅使用马氏距离关联容易出现 ID switch,因此加入外观特征关联,采用最小余弦距离度量特征距离。对一段帧数内目标特征向量进行保留,由集合 R_k 表示为:

$$R_k = \{r_k^{(i)}\} (k \in [1, L_k]) \quad (24)$$

式中: L_k 为保留的目标特征数量。经实验验证,本文设定为 80。

最小余弦距离表示为:

$$d^2(i,j) = \min \{1 - r_j^T r_k^{(i)} | r_k^{(i)} \in R_i\} \quad (25)$$

式中: $d^2(i,j)$ 为第 i 个追踪器最近 80 个成功关联的特征集与当前帧第 j 个检测结果特征向量间的最小余弦距离。

与 $b_{i,j}^{(1)}$ 类似,设定阈值 $t^{(2)}$,判断是否关联成功。判断公式为:

$$b_{i,j}^{(2)} = 1[d^{(2)}(i,j) \leq t^{(2)}] \quad (26)$$

最后使用两种度量方式的线性加权作为最终度量(马氏距离不满足阈值时不计算余弦距离以减少计算量、提高速度)。表示为:

$$c_{i,j} = \lambda d^{(1)}(i,j) + (1 - \lambda) d^{(2)}(i,j) \quad (27)$$

式中: λ 为不同关联度量比例系数; $c_{i,j}$ 越小则第 i 个跟踪轨迹与第 j 个检测目标越相似。

外观特征提取算法 DeepSORT 采用 ResNet 输出 128 维目标特征向量,对于无人机视角下的小目标特征提取效果不佳,故本文使用各项性能更优的 FastReID^[14] 训练的 Reid 模型进行外观特征提取,其效果更好,运行速度更快。

2.2.3 级联匹配

若目标长时间被遮挡,卡尔曼滤波预测的不确定性将大大增加,状态空间内的可观察性也将大大降低。当两个跟踪器竞争同一检测结果的匹配权时,由于遮挡时间较长轨迹的马氏距离更小,使得检测结果更倾向于与其关联,从而产生错误结果。为此,引入级联匹配,对没有丢失的轨迹优先匹配。

为应对因画面抖动、遮挡导致某时刻目标没有观测状态($z = \text{None}$),而当此目标再次被观测时检测框相较最优估计距离过远而无法匹配的问题,本文在目标丢失时使用先验估计状态 $\bar{x}$ 作为其最优估计值 x ,基于历史状态更新跟踪框的位置,以有效降低小幅度抖动和小体积遮挡物的影响。停止跟踪经过 n 帧后仍没有被观测到的目标,视为离开区域,停止预测计算。参数 n 可根据具体场景下的抖动幅度、遮挡物大小、目标通过速度灵活调整。

3 实验方法与结果分析

3.1 数据集

本文通过自建数据集的方式进行网络训练。使用大疆 Mavic 3 Classic 无人机在武汉市内多处进行倾斜向下视角的图像采集,排除模糊失焦等低质量图片,最终选取 6 193 张图片,其中街道、公园、广场场景占比约为 2:2:1。同时从 VisDrone 数据集中选取拍摄高度适中的行人和车辆类别图片 3 807 张,共 10 000 张图片组成数据集,随机挑选 18 组连续拍摄的图片(即 18 段视频,共 1 000 帧)分别组成测试集与验证集各 500 帧,其他 9 000 帧作为训练集。

3.2 实验配置

实验训练平台配置为:Windows 10 专业版,CPU 为 Intel(R) Core(TM) i7-10700, RAM 为 DDR4 16.0 GB, GPU 为 NVIDIA GTX 3070Ti。实验环境为 Python 3.9 的 PyCharm 集成开发工具。

3.3 通信开销

使用大疆 Mavic 3 Classic 无人机拍摄并通过大疆 OcuSync3 图传方案将视频无线传输到服务器端,在支持 2.4 GHz/5.8 GHz 双频的地区最大传输速率为 120 Mbps^[15]。对于 1 920×1 080 分辨率,60 帧/s 的 RGB 视频,若采用 H.264 压缩算法,压缩比为 40:1,则通信开销约为 74.65 Mbps,在城市环境下的实地测试中可以实现实时传输(无人机与服务器端的距离小于 2 km)。

3.4 实验结果分析

3.4.1 不同模型比较分析

为使模型评价更加客观,本文采用多目标跟踪领域通用的评估指标进行评估,包括多目标跟踪准确度(MOTA)、跟踪器 ID 维持能力(IDF1)、MT (Mostly Tracked)、ML (Mostly Lost)、目标 ID 切换次数(IDS)等。

对本文模型与 RefineDet + DeepSORT 模型、主流模型 JDE^[16]、YOLOv5^[17]+DeepSORT 模型、最新模型 MeMOT^[18] 进行性能比较,分别在自建数据集上进行训练并测试,输入分辨率为 1 920×1 080。结果见表 1。

Table 1 Performance comparison of different models

表 1 不同模型性能比较

模型	MOTA ↑	IDF1 ↑	MT /% ↑	ML /% ↓	IDS ↓	Hz ↑
RefineDet + DeepSORT	77.4	73.9	48.2	14.6	1 175	24.3
JDE	76.4	68.1	47.3	16.2	1 481	43.4
YOLOv5+DeepSORT	77.2	74.4	45.8	14.2	1 242	26.3
MeMOT	81.3	78.5	55.2	13.1	772	12.8
本文模型	79.7	77.6	53.8	13.4	854	32.6

可以看出,本文模型在处理中距离俯视角多目标跟踪任务时在精度和速度方面均表现优秀。JDE 模型将重识别特征提取与目标检测器相结合,达到了最快速度 33.4 Hz,但精度较低,MOTA 仅为 76.4;最新的 MeMOT 模型基于

Transformer, 通过保留一个大的时空内存来存储被跟踪对象的特征向量, 使其精度最高, MOTA 达到了 80.3, 但计算量大, 速度仅为 12.8 Hz^[19]; 本文模型在 MOTA 仅比 MeMOT 低 0.6 的情况达到了 32.6 Hz 的速度, 该速度虽然比 JDE 低 10.8 Hz, 但基本满足实时性需求。

由于针对无人机视角下的小目标进行了多项优化, 在面对密度大、相互遮挡的目标以及距离远、像素较少的目标时, 本文模型较 JDE、YOLOv5+DeepSORT、MeMOT 等对照模型有明显优势, 示例见图 9、图 10、图 11(彩图扫 OSID 码可见, 下同)。为展示方便, 图中未标出 id。

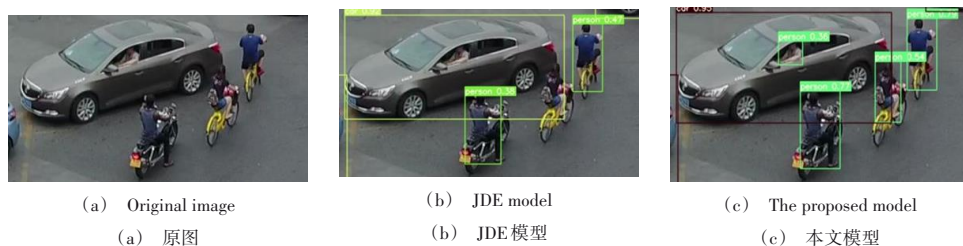


Fig. 9 Example 1

图 9 示例 1



Fig. 10 Example 2

图 10 示例 2

可以看出, 漏检、误检出现的主要原因为远处目标所占像素小和相互遮挡, 在流量统计时可以通过调整无人机高度、拍摄角度和设定流量统计线的方式减小其影响。

3.4.2 消融实验

为证明本文改进部分的必要性和有效性, 在自建数据集上进行消融实验, 结果如表 2 所示。以 RefineDet + Dee-

pSORT 为基准, 将目标检测部分更换为改进的 RefineDet 后, 得益于更高的目标检测精度, MOTA 提高了 1.7, 但由于引入了注意力机制, 增加了计算复杂度, 使得速度降低了 2.4 Hz; 将多目标跟踪部分替换为本文算法后, 得益于卡尔曼滤波的噪声自适应处理和 FastReID 更好的外观特征提取性能, IDF1 有较大提高, 虽然噪声自适应处理增加了计算量, 但采用轻量级网络结构的 FastReID 比原本的外观特征提取网络更快, 同时本文算法设计了目标检测与跟踪并行运行机制, 使得整体速度大幅提高。

3.4.3 人车流量统计实验分析

为定量评价本文人车流量统计方法的实用性, 选取 3 段时长为 10 min 的无人机拍摄视频: 视频 1 为郊区, 尚未安装摄像头; 视频 2 为城市十字路口, 虽然安装了交通摄像头但位置和角度固定, 无法覆盖所有行人; 视频 3 为马拉松活动现场, 没有监控摄像头且流量统计需求为临时性、短时间的。3 段视频中的人车流量和密度递增, 展示了传统方法难以应对的情况。通过人工统计计算出测试视频的正检数量、误检数量、漏检数量, 并以此计算精确率 P 、召回率 R 和综合评价指标 $F1$ 。本文模型对 3 段视频的检测效果如图 12 所示, 图中红线为单独设定的流量统计线, 每个 ID 首次跨越时流量加 1, 并记录跨越方向。目标检测性能指标如表 3 所示。

可以看出, 在 3 段视频中, 本文模型的精确率均高于 97%。视频 1 中流量较小、拍摄俯角较大、小目标较少, 召回率达到 98.75%; 视频 2 和 3 中拍摄俯角较小, 距离远、像素少、相互遮挡严重的小目标数量较多, 导致召回率比视频 1 低 6%。

然而, 在实际统计中, 将统计线设定在距镜头较近的位置即可避免远处低召回率的情况, 此时人、车流量统计结果如表 4、表 5 所示。可以看出, 通过合理设置流量统计线, 多种场景下本文模型的人车流量误检率均小于 0.3%, 准确性和实时性优秀, 具有较高的实用价值。

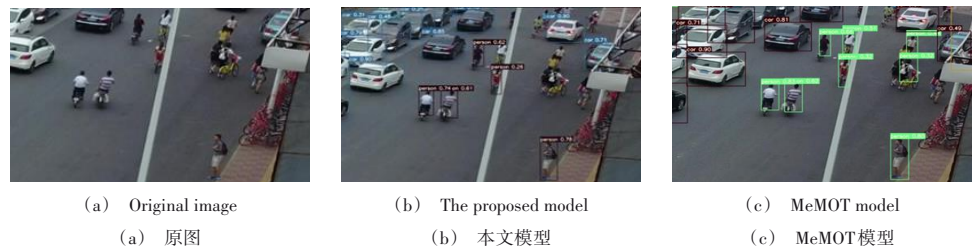


Fig. 11 Example 3
图 11 示例 3

Table 2 Ablation study result
表 2 消融实验结果

模型	MOTA ↑	IDF1 ↑	IDS ↓	Hz ↑
RefineDet + DeepSORT	77.4	73.9	1 075	24.3
引入注意力机制的 RefineDet+ DeepSORT	79.1	75.3	1 038	21.9
RefineDet + 本文多目标跟踪算法	78.3	76.4	977	36.3
本文模型	79.7	77.6	854	32.6

Table 3 Object detection performance
表 3 目标检测性能指标 (%)

视频编号	精确率 P	召回率 R	$F1$
1	98.38	98.75	98.56
2	97.85	92.53	95.12
3	98.74	92.81	95.68

Table 4 Results of pedestrian flow statistics
表 4 人流量统计结果

视频编号	实际人流量	统计人流量	误检率/%
1	122	122	0
2	874	876	0.228
3	2 876	2 871	0.174

Table 5 Results of vehicle flow statistics
表 5 车流量统计结果

视频编号	实际车流量	统计车流量	误检率/%
1	67	67	0
2	492	491	0.203
3	5	5	0

4 结语

本文通过建立基于双重注意力机制的 RefineDet 目标检测网络实现了对无人机遥感图像中小体积目标的实时准确识别,利用自适应卡尔曼滤波和重识别实现了基于轨迹预测的多目标跟踪,并利用级联匹配、以预测值更新跟踪器等方法缓解了因抖动、遮挡导致的目标丢失问题,提高了稳健性。多目标跟踪的人车流量统计方法解决了重复统计与无效通过问题,实现了高精度的实时流量统计。实验结果表明,本文方法在多种场景下效果优于其他主流模型与最新模型,对于解决开放区域的临时性、短时间人

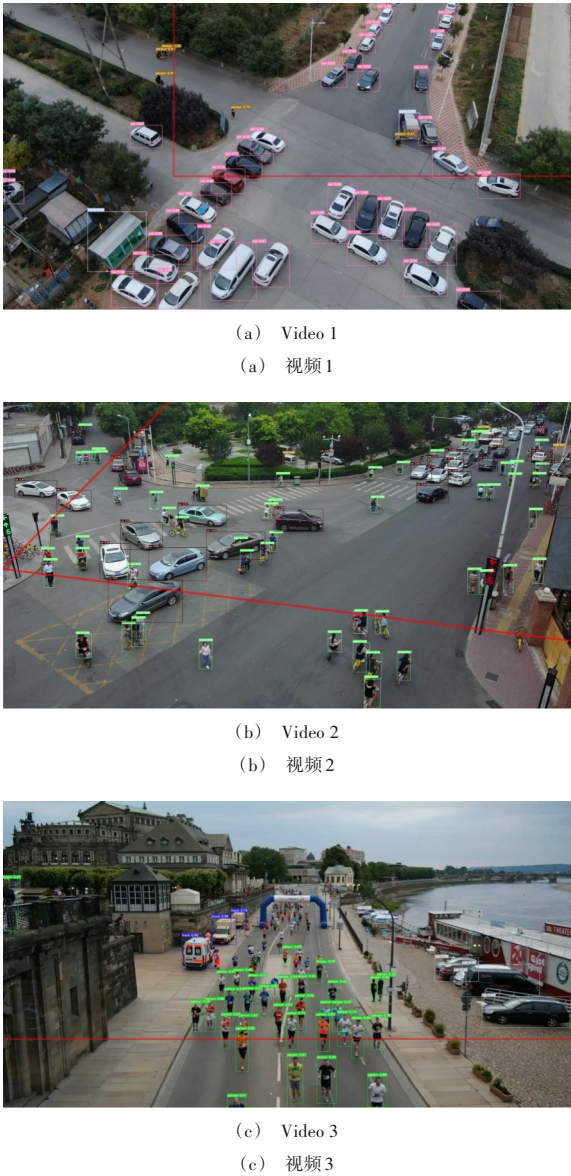


Fig. 12 Video detection effects of the proposed model in three real scenes
图 12 3 种实际场景下本文模型的检测效果

车流量检测统计问题具有较强的实用价值,尤其是在目标较小、遮挡严重等复杂场景中表现优异。

然而,本文方法仍然存在不足之处,例如车辆开窗时乘客会被识别为行人从而导致误差(例如图 9),但其特征与远距离小体积行人相似,无法通过增加数据集规模和提

高置信度等方式避免,若排除车辆识别框内的行人又会导致车侧行走的行人被漏检;部分低矮建筑物在俯视角拍摄下与载货卡车相近(例如图 12(a)左下角的灰顶建筑物),易出现误识别。同时,本文模型的二次定位在提高检测精度的同时也增加了计算量、降低了检测速度,可以考虑通过简化特征提取网络、减少参数等方式减少计算量。

参考文献:

- [1] YANG Z Y, JIN L, LIU C. Method of pedestrian flow monitoring based on received signal strength [J]. Modern Electronics Technique, 2020, 43(14):156-160.
杨志勇,金磊,刘灿. 基于 RSS 的人流量监测方法[J]. 现代电子技术, 2020, 43(14):156-160.
- [2] ZHANG S, WEN L, BIAN X, et al. Single-shot refinement neural network for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018:4203-4212.
- [3] SINGHAL N, PRASAD L. A comparative analysis of deep learning based vehicle detection approaches[J]. International Journal of Next-Generation Computing, 2023, 14(2):485-501.
- [4] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6):1137-1149.
- [5] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric [C]//2017 IEEE International Conference on Image Processing, 2017:3645-3649.
- [6] GIRSHICK R, DONAHUE J, DARRELL T, et al. Region-based convolutional networks for accurate object detection and segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1):142-158.
- [7] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector [C]//Computer Vision-ECCV 2016: 14th European Conference, 2016: 21-37.
- [8] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection [DB/OL]. <https://arxiv.org/abs/2004.10934>.
- [9] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019:3146-3154.
- [10] GUO M H, XU T X, LIU J J, et al. Attention mechanisms in computer vision: a survey [J]. Computational Visual Media, 2022, 8(3):331-368.
- [11] HAN Y, GAO X D, JI R J, et al. Systematic design and software development of campus traffic statistics based on face recognition [J]. Journal of Nantong Vocational University, 2022, 36(1):76-81.
韩燕,高晓东,季荣军,等. 基于人脸识别的校园人流量统计系统设计与实现[J]. 南通职业大学学报, 2022, 36(1):76-81.
- [12] FANG J C, YANG S. Study on innovation adaptive EKF for in-flight alignment of airborne POS [J]. IEEE Transactions on Instrumentation and Measurement, 2011, 60(4):1378-1388.
- [13] SHI N, CHEN Z, NIU M, et al. State-of-charge estimation for the lithium-ion battery based on adaptive extended Kalman filter using improved parameter identification [J]. Journal of Energy Storage, 2022, 45:103518.
- [14] HE L, LIAO X, LIU W, et al. Fastreid: a pytorch toolbox for general instance re-identification [DB/OL]. <https://arxiv.org/abs/2006.02631>.
- [15] IQBAL S. A study on UAV operating system security and future research challenges [C]//2021 IEEE 11th Annual Computing and Communication Workshop and Conference, 2021: 759-765.
- [16] WANG Z, ZHENG L, LIU Y, et al. Towards real-time multi-object tracking [C]//Computer Vision-ECCV 2020: 16th European Conference, 2020:107-122.
- [17] JOCHER G, CHAURASIA A, STOKEN A, et al. Ultralytics/yolov5: V7.0-YOLOv5 SOTA realtime instance segmentation [EB/OL]. <https://ui.adsabs.harvard.edu/abs/2022zndo...7347926J/abstract>.
- [18] CAI J, XU M, LI W, et al. MeMOT: multi-object tracking with memory [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 8090-8100.
- [19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [DB/OL]. <https://arxiv.org/abs/1706.03762>.

(责任编辑:尹晨茹)