2021 年 3 月 第 3 期
Vol. 42 No. 3 2021

无人车驾驶场景下的多目标车辆与行人跟踪算法

顾立鹏,孙韶媛,李 想,刘训华,宋奇奇

(东华大学 信息科学与技术学院,上海 201620)
E-mail: glp1224@163.com

摘 要: 考虑到现有的基于检测的多目标跟踪算法多会出现因目标漏检或数据关联算法冗余而造成的目标 ID 频繁切换、跟踪轨迹断开等问题,提出了无人车驾驶场景下的多目标车辆与行人跟踪算法. 首先,选取 CenterNet 网络作为目标检测器,并用嵌入了 1×1 卷积和 SE-Net 的 Res2Net 来替代网络原有的残差单元,以提升网络对空间信息和通道信息的提取能力,提高目标检测器性能. 接着,用孪生网络来提取目标所在区域的特征,进行关联概率度量,再用匈牙利算法对相邻帧目标进行关联. 最后,用区域推荐网络设计的辅助跟踪器对漏检或消失又出现的目标进行持续跟踪,并将可靠的跟踪结果合并到轨迹中. 实验结果表明,与已有的方法对比,所提方法在 KITTI 跟踪基准数据集上对于车辆与行人的跟踪具有竞争力.

关 键 词: 机器视觉; 目标检测; 孪生网络; 区域推荐网络; 多目标跟踪
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2021)03-0542-08

Multi-object Vehicle and Pedestrian Tracking Algorithm in Driving Scene of Unmanned Vehicle

GU Li-peng, SUN Shao-yuan, LI Xiang, LIU Xun-hua, SONG Qi-qi
(College of Information Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: Considering that the existing multi-object tracking algorithms based on tracking-by-detection framework, they often have the problems of frequent switching of object's ID and disconnection of tracking track caused by missing detection of object or redundancy of data association algorithm. Thus, this paper proposes a multi-object vehicle and pedestrian tracking algorithm in driving scene of unmanned vehicle. Firstly, CenterNet network is selected as the object detector, and res2net embedded with 1×1 convolution and SE-Net is used to replace the original residual unit in the network, so as to improve the network's ability to extract spatial information and channel's information and improve the performance of the object detector. Then, siamese network is used to extract the features of the region where the target is located, and the probability of association is measured. Then, the Hungarian algorithm is used to match the detected object of adjacent frame. Finally, the auxiliary tracker designed by region proposal network is used to track the missing or disappearing objects continuously, and the reliable tracking results are incorporated into the trajectory. Compared with the existing methods, the experimental results show that the proposed method is competitive for vehicle and pedestrian tracking on the KITTI tracking benchmark dataset.
Key words: machine vision; object detection; siamese network; region proposal network; multiple object tracking

1 引 言

多目标跟踪(Multi-Object Tracking, MOT)是计算机视觉领域的一个研究热点,在自动驾驶、机器人续航、视频监控与行为分析等领域发挥着重要的作用[1]. 相比于单目标跟踪,多目标跟踪主要是在输入的视频中定位多个目标,维持它们的 ID 不变,并形成各自的轨迹,因此更复杂、更具挑战. 近 10 年来,随着深度神经网络的迅速发展,基于检测的多目标跟踪算法受到了广泛的关注. 这类方法主要将多目标跟踪问题分为两个步骤: 第 1 步,使用目标检测网络检测出给定视频序列中每一帧存在的感兴趣的目标; 第 2 步,使用数据关联算法将检测到的目标随着时间推移,分配各自的 ID,并生成各自的轨迹[2].

尽管已经过多年的研究,多目标跟踪算法的性能仍还远未达到人类的水平. 当前该问题面临的挑战主要包括: 未知的目标类别及数量; 目标之间频繁的遮挡; 目标的漏检或误检等[1]. 针对上述问题,目前的解决方法大多集中在以下几个方面: 优化提升目标检测网络性能、设计更具表现力的目标特征模型、设计更高效的数据关联算法[3]. 对于优化提升目标检测器性能,Zhou[4] 等人提出了 CenterNet 网络,利用关键点估计来确定潜在目标的中心点,并回归出宽高尺寸、偏移量等目标,具有整体网络计算开销小、精度高且速度快等优点. Shang-Hua Gao[5] 等人提出了 Res2Net 模块,其可被便捷地嵌入到现有的目标检测网络中,在不增加网络整体计算开销的基础上,提升目标检测网络的性能. 针对设计更具表现力的目标特征模型,Bo Li[6] 等人提出的 SiamRPN 网络,通过将孪生网络和区域推荐网络结合到一起实现对初始帧的给定目标的

收稿日期: 2020-03-23 收修改稿日期: 2020-05-20 基金项目: 上海市科委应用基础研究项目(15JC1400600)资助. 作者简介: 顾立鹏,男,1996 年生,硕士,研究方向为目标检测、多目标跟踪等; 孙韶媛,女,1974 年生,博士,教授,研究方向为夜间机器视觉; 李 想,女,1995 年生,硕士,研究方向为场景预测和深度学习; 刘训华,男,1996 年生,硕士,研究方向为深度学习、目标检测等; 宋奇奇,男,1994 年生,硕士,研究方向为目标跟踪、计算机视觉等.

跟踪,前者用来提取目标在上一帧的区域和在当前帧的两倍区域的卷积特征,后者用来推理出该目标在当前帧的状态. 针对优化数据关联算法,Leal-Taixé[7] 等人使用孪生网络提取局部时空域特征,再根据两个检测到目标的时空域特征的响应之间的几何距离,得到两个检测到目标之间的关联概率,最后用匈牙利算法对相邻帧检测到的目标进行关联.

受上述启发,为了解决无人车驾驶场景下的多目标跟踪所面临的各种问题,本文从优化目标检测器和数据关联算法两个方面,提出了一种无人驾驶场景下的多目标车辆与行人跟踪算法: 1) 提出了 Res2Net_plus 模块,具体是在 Res2Net 模块中嵌入了 1×1 卷积和 SE-Net 模块,以融合空间信息和通道信息,提升网络对目标区域特征的提取能力; 2) 用 CenterNet 网络作为目标检测器,并用 Res2Net_plus 模块替代网络原有的残差单元,以进一步提升 CenterNet 网络对无人车驾驶场景下车辆和行人的检测精度; 3) 受 SiamRPN 网络启发,将其网络一分为二,孪生网络部分设计为关联概率网络,进行基于外观特征的关联概率度量,而区域推荐网络部分设计为辅助跟踪器,来对历史帧中的漏检或消失又出现的目标进行持续跟踪,并将可靠的跟踪结果合并到存在的轨迹中; 4) 设计了一种基于目标外观特征和位置信息融合的匹配策略,具体是孪生网络对检测到目标所在区域的外观特征进行提取,作为主要的匹配依据. 同时,目标的位置信息作为辅助匹配依据,用于剔除外观相似但在两帧中所处的位置较远的虚假匹配关系. 在 KITTI 跟踪基准数据集上的实验结果表明,与已有的方法对比,本文方法具有竞争力,尤其对于因目标检测器的漏检或目标的消失又出现所导致的跟踪轨迹不连续或目标 ID 频繁切换的问题有很大的改善作用.

2　多目标跟踪框架

本文提出的无人车驾驶场景下的多目标检测算法由目标检测网络、关联概率网络和数据关联模块 3 部分组成,算法整体框图如图 1 所示. 其中,①为输入视频序列; ②为目标检测网络检测每帧的车辆与行人; ③为提取目标外观和位置信息,并经过关联概率网络,得到关联概率矩阵; ④为经过数据关联模块,得到车辆与行人跟踪结果(目标的 ID 和包围框坐标) .

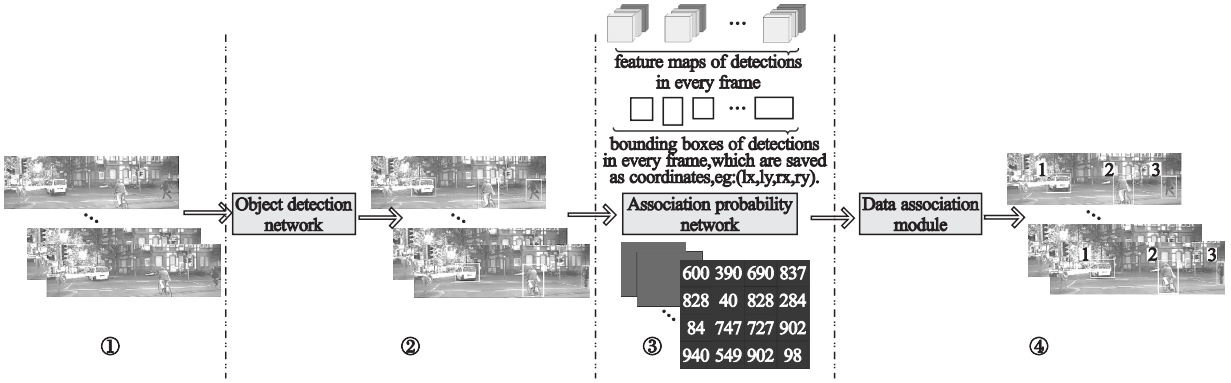

图 1　算法整体框图

Fig. 1　Algorithm block diagram

2.1　目标检测网络

目标检测网络对基于检测的多目标跟踪器的整体性能有着至关重要的影响,这是因为目标检测网络的误检或漏检将造成目标 ID 的频繁切换或目标轨迹的断开等问题[8]. 因此,一个兼顾精度和速度的目标检测网络对多目标跟踪算法十分重要.

2.1.1　CenterNet 网络

本文选取了 CenterNet 网络作为多目标跟踪的目标检测器. 不同于基于锚框的检测网络,CenterNet 网络将目标检测问题巧妙地转换成关键点估计问题,利用关键点估计来确定目标的中心点,同时在中心点处回归出该目标的其他属性,如宽高尺寸、中心点的偏移量等. 这使得该网络在整体计算开销相对较小的情况下,拥有很好地提取并利用目标内部的信息的能力,实现了潜在目标的检测,因此具有其精度高且速度快的优点,尤其对于无人车驾驶场景下数量多且有频繁遮挡的车辆与行人具有很好的检测能力,网络结构图如图 2 所示.

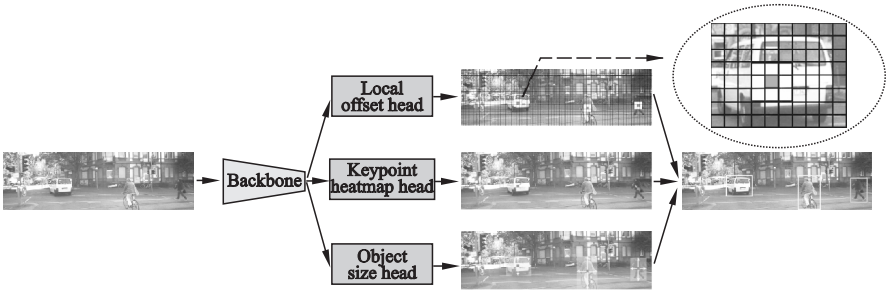

图 2　CenterNet 网络结构图

Fig. 2　Network structure of CenterNet

在图 2 中,右上角虚线圈内网格中的黑点是对该图片内的车辆进行偏移量估计的放大图,以实现对车辆位置的修正,

对其精确定位. 本文中所使用的 Backbone 为带有动态卷积的 DLA_34 网络,该网络是通过多级跳跃连接的,并以多次迭代的方式融合浅层与深层的信息,以获得更具表现力的特征. 在 Backbone 的输出端增加了 Keypoint heat head、Object size head 和 Local offset head,分别回归输入图像中潜在目标的关键点、宽高尺寸和中心点的偏移量,从而可以精准地检测到视频序列中的车辆与行人.

2.1.2 Res2Net_plus 模块

由于无人车驾驶场景下存在很多较小的目标或行人和车辆之间相互遮挡的情况,为了提高 CenterNet 网络对小目标和遮挡目标的检测效果,本文提出了 Res2Net_plus 模块. Res2Net_plus 模块通过结合 Octave Conv[9] 和 SE-Net[10] 的思想对原始 ResNet(Bottleneck) 模块进行改进. Octave Conv 的核心思想是将原始特征图按不同频率进行分解,对含有不同频率信息的特征图分开操作,从而可以加速卷积的计算和提高任务的性能. 而 SE-Net 的思想是引入了注意力的机制,用一个权重来表示输入特征图的通道在后续阶段的重要程度,以实现特征图的空间信息和通道信息的融合.

受此启发,本文在 Res2Net 模块中的以层级残差式风格连接 3×3 卷积前端分别加上一组 1×1 的卷积,以获得含有不同频率信息的特征图,如图 3(b) 中①号虚线框所示. 同时,在 Res2Net 模块中嵌入 SE-Net,以融合特征通道间的关系,进一步提升网络的空间特征表现力,得到更加有表现力的多尺度特征,如图 3(b) 中②号虚线框所示. Res2Net 模块和 Res2Net_plus 结构如图 3 所示.

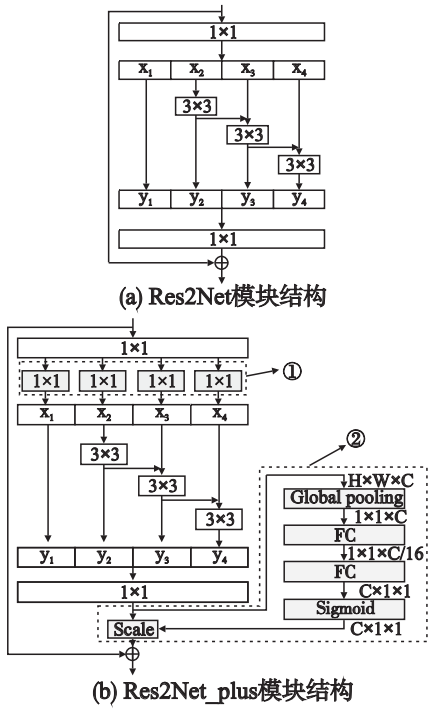

图 3 Res2Net 和 Res2Net_plus 模块结构

Fig. 3 Module structure of Res2Net and Res2Net_plus

在图 3 中,假设经过头部第一个 1×1 卷积后的尺寸为 $H \times W \times C$ 的特征图为 U',原来的 Res2Net 模块仅仅简单的将 1×1 卷积后的特征图按通道等分成 4 份,而 Res2Net_plus 是将 1×1 的卷积后的特征图分别经过 4 个 $1 \times 1 \times (C/4)$ 的卷积,生成 4 个 $H \times W \times (C/4)$ 特征图送入后续卷积层,以获得含有不同频率信息但通道数减为原来 4 倍的特征图. 同时,嵌入的 SE-Net 模块则是对经过尾部最后一个 1×1 卷积输出的尺寸为 $H \times W \times C$ 的特征图 U'' 先后进行 Squeeze 操作、Excitation 操作及 Scale 操作,以实现特征图的空间信息和通道信息的融合. 其中,Squeese 操作是使用全局池化,将大小为 $H \times W \times C$ 的输入特征图转为 $1 \times 1 \times C$ 的特征描述,计算方法如公式(1). Excitation 操作是将得到的 $1 \times 1 \times C$ 的特征描述经过两个全连接层和一个 Sigmoid 激活函数,得到 $1 \times 1 \times C$ 的通道间的权重 s. Scale 操作是按通道将获得 $1 \times 1 \times C$ 的权重 s 与原始输出的 $H \times W \times C$ 的特征图 U'' 通过简单的乘法进行融合,得到 $H \times W \times C$ 的特征图 U''',计算方法如公式(2).

$$z_c = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} U''_c(i,j) \tag{1}$$

其中,W 和 H 分别为特征图 U'' 的第一、二维的尺寸,U''_c 表示 U'' 在通道 c 处的感受野,z_c 为标量,表示经过 Squeeze 操作得到的 $1 \times 1 \times C$ 的特征描述在通道 c 处的值.

$$U'''_c = s_c \cdot U''_c \tag{2}$$

其中,U''_c 表示 U'' 在通道 c 处的感受野,s_c 为标量,表示通道间的权重 s 在通道 c 处的值,U'''_c 表示经过 Scale 操作得到的特征图 U''' 在通道 c 处的感受野.

2.2 关联概率网络

外观特征是目标检测领域中一种具有很好区分性的属性. 尤其是在目标之间相互遮挡或存在许多外观相似的目标时,外观特征可以被使用来对目标进行检测、识别和区分.

在早期的研究中,一些人工制作的特征常常被使用来表征物体的外观特征. 随着深度神经网络的发展,基于深度神经网络提取的物体的外观特征被广泛地使用在目标检测、跟踪等领域. 本文利用基于嵌入了 CIR 单元(cropping-inside residual units) 的 CIResNet_22 为主干网络的孪生网络提取目标的外观特征,并分别将其两两直接进行卷积计算如公式(3) 所示,得到关联概率值,值越高目标越相似,具体网络结构如图 4 所示.

$$k = U_1 * U_2 \tag{3}$$

其中,U_1 和 U_2 为两个不同目标经过 CIResNet_22 提取的尺寸一样的特征图,* 表示卷积计算操作,k 为一个标量,值越高,表示两个目标越相似,反之,则差异越大.

该关联概率网络的输入是目标检测网络对视频序列每帧检测到的目标的二维包围框的左上角和右下角坐标,(x_1, y_1) 和 (x_2, y_2). 然后根据二维包围框的坐标将每帧检测到的目标裁剪出来,把尺寸调整为 127×127,送入 CIResNet_22 网络中,来提取每帧中检测到的每个目标的外观特征,其尺寸为 $6 \times 6 \times 256$. 假设上一帧和当前帧分别代表第 t 帧和第 t+1 帧,且分别检测到目标的数量为 N_t 和 N_{t+1}. 然后,将第 t 帧的 N_t 个特征图与第 t+1 帧的 N_{t+1} 个特征图两两直接进行卷积计算,可以得到尺寸为 $N_t \times N_{t+1}$ 的关联概率矩阵,例如图 4 右上角虚线圈内所示.

图 4 中的虚线圈内的关联概率矩阵,其竖直方向和水平方向上的数字分别代表当前帧和上一帧检测到的第几个目标,且矩阵里面的数字代表关联概率值. 例如,2 行 3 列的数字为 828,代表上一帧的第 2 个物体与当前帧的第 3 个物体之

间的关联概率为 828. 在本文中,相似度值大于 30 的两个目标被认为是相似的.

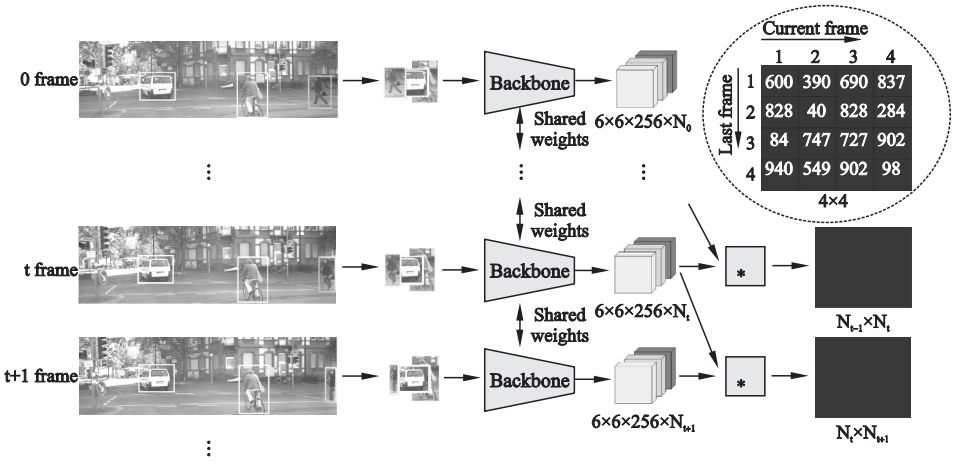

图 4　关联概率网络结构图

Fig. 4　Structure diagram of association probability network

2.3　数据关联模块

数据关联也是基于检测跟踪的多目标跟踪方法中十分关键的一步,这直接决定了所检测到的目标之间匹配的效率与最终跟踪效果. 本文设计了一种级联形式的数据关联方法,首先采用匈牙利算法来完成相邻帧之间检测到的目标的初步匹配,后续两步工作由基于区域推荐网络[11]的辅助跟踪器完成,以进一步提升目标跟踪能力.

2.3.1　数据关联

数据关联一共分 3 步进行,工作流程如图 5 所示.

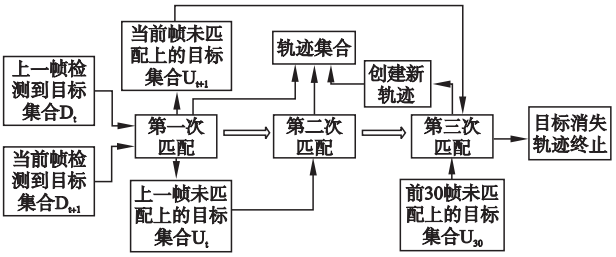

图 5　数据关联模块工作流程图

Fig. 5　Workflow of data association module

第 1 次匹配: 充分利用目标的外观特征,得到相邻帧的目标的初步匹配关系. 以“集合 D_t 和 D_{t+1}”为输入,通过关联概率网络计算得到相邻帧的关联概率矩阵. 然后,使用匈牙利算法[13]得到相邻两帧所检测到的目标的初步匹配关系. 接着,将满足下述条件 1 和条件 4 的当前帧的目标合并到存在的轨迹中; 反之,则将上一帧和当前帧的未匹配上的目标分别放入“集合 U_t”和“集合 U_{t+1}”.

第 2 次匹配: 利用基于区域推荐网络的辅助跟踪器对漏检的目标进行持续跟踪. 以“集合 U_t”为输入,使用辅助跟踪器对上一帧未匹配上的目标进行持续跟踪,得到当前帧的位置状态和跟踪得分. 接着,将满足条件 2、条件 3 和条件 5 的跟踪结果合并到存在的轨迹中; 反之,则认为该目标已消失,将其放入“集合 U_{30}”. 设置一个最大连续跟踪帧数 $N_{\max}$,当超过连续 $N_{\max}$ 帧时,该漏检的目标还未出现,则不再对此进行跟踪.

第 3 次匹配: 利用辅助跟踪器对前 30 帧已经消失的目标进行推理判断是否在当前帧再次出现. 以“集合 U_{t+1} 和 U_{30}”为输入,使用辅助跟踪器对前 30 帧未匹配上的目标进行持续跟踪,推理出在当前帧的状态. 接着,将满足下述条件 2、条件 3 和条件 5 的跟踪结果合并到存在的轨迹中,并将其从“集合 U_{30}”中删除. 最后,当前帧未匹配上的目标为新出现的目标,为其创建新的轨迹.

其中,条件 1 为关联概率大于阈值 30; 条件 2 为跟踪得分大于阈值 0.9; 条件 3 为预测得到的目标包围框距离图片边界大于 15 像素值; 条件 4 为两个目标的包围框的 IoU 值大于阈值 0.01; 条件 5 为预测得到的目标包围框与历史帧中该目标的包围框的 IoU 值大于阈值 0.01.

2.3.2　辅助跟踪器

本文中所使用的辅助跟踪器是基于区域推荐网络设计的,主要目的是改善因目标检测网络的漏检导致目标 ID 频繁切换或轨迹断开的问题,具体网络结构如图 6 所示.

其中,辅助跟踪器中使用的主干网络和前面关联概率网络的主干网络一样,都是使用 CIResNet_22 网络,且共享权重. 辅助跟踪器输入由两部分组成,分别为未匹配上的目标集合(包括特征图、ID 和包围框的坐标)和当前帧的图片. 其中,未匹配的目标的特征图作为模板帧的特征图,尺寸为 $6\times6\times256$. 而检测帧的特征图的提取通过两个步骤: 首先,以未匹配的目标的中心点在当前帧图片中裁剪出同样中心点的两倍区域; 然后用 CIResNet_22 网络提取出尺寸为 $22\times22\times256$ 的特征图; 接着,模板帧和检测帧的特征图复制双份,分别送入区域推荐网络的分类分支和回归分支中进行后续操作. 其中,分类分支,用于区分目标的前景和背景; 回归分支,用于对目标的候选区域进行微调.

在分类分支中,模板帧和检测帧的特征图,分别通过一个尺寸为 $3\times3\times(2\times256)$ 和尺寸为 $3\times3\times256$ 的卷积核进行卷积操作,产生了尺寸为 $4\times4\times(2\times256)$ 的模板帧的特征 $[\varphi(z)]_{cls}$ 和尺寸为 $20\times20\times256$ 的检测帧的特征 $[\varphi(x)]_{cls}$. 如公式(4),模板帧的特征 $[\varphi(z)]_{cls}$ 作为卷积核去卷积检测帧的特征 $[\varphi(x)]_{cls}$,产生了尺寸为 $17\times17\times2$ 的响应图 $A^{cls}_{w\times h}$.

其中,两个维度分别对应每个 Anchor 的目标的前景和背景的分类分数.

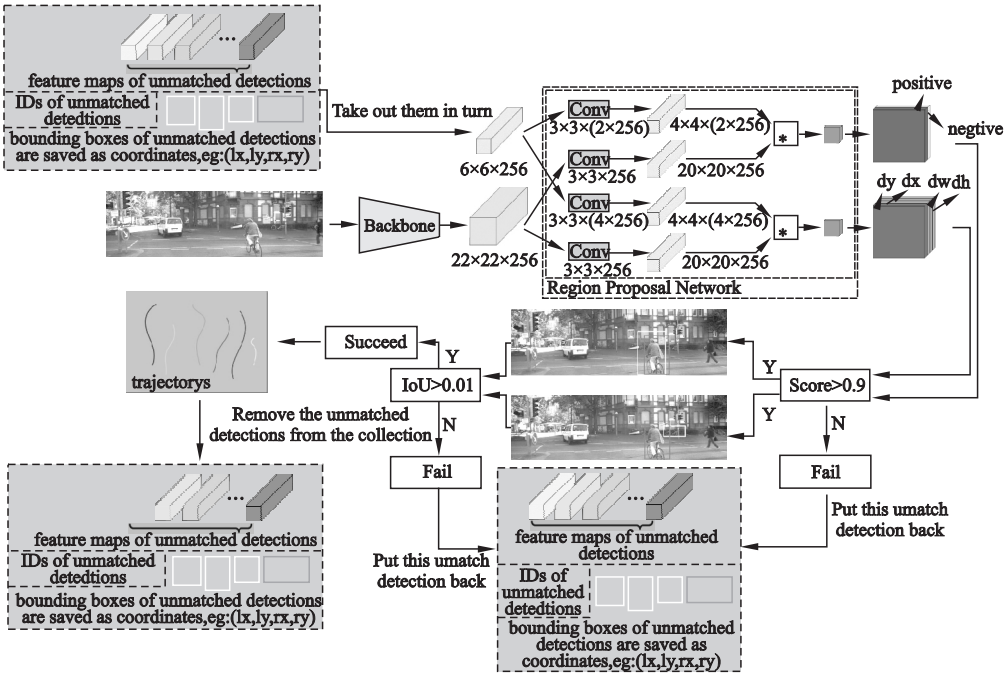

图 6 辅助跟踪器结构图

Fig. 6 Structure diagram of auxiliary tracker

归回分支和分类分支一样,分别通过一个尺寸为 $3\times3\times(4\times256)$ 和尺寸为 $3\times3\times(2\times256)$ 的卷积核进行卷积操作,产生了尺寸为 $4\times4\times(4\times256)$ 的模板帧的特征 $[\varphi(z)]_{reg}$ 和尺寸为 $20\times20\times256$ 的检测帧的特征 $[\varphi(x)]_{reg}$. 接着,如公式(5),产生了尺寸为 $17\times17\times4$ 的响应图 $A^{reg}_{w\times h}$. 其中,4 个维度分别对应每个 Anchor 的 (dx,dy,dw,dh),即目标的坐标及尺寸.

$$A^{cls}_{w\times h\times 2}=[\varphi(x)]_{cls}*[\varphi(z)]_{cls} \tag{4}$$

$$A^{reg}_{w\times h\times 4}=[\varphi(x)]_{reg}*[\varphi(z)]_{reg} \tag{5}$$

最后,使用 Softmax 损失函数优化分类分支得到的 $A^{cls}_{w\times h}$,并将得到的前景目标输入到归回分支得到的 $A^{reg}_{w\times h}$. 接着,计算候选区域与真实边框之间的差值,并对差值最小的候选区域进行微调,从而得到最终的跟踪状态结果. 最后,将可靠的跟踪结果合并到存在的轨迹中.

3 实验与结果分析

3.1 实验配置与数据集

本实验是使用 Pytorch 0.4.1 框架实现的,实验配置如表 1 所示.

表 1 实验配置

Table 1 Experimental configuration

Item	CPU	Computing memory	GPU	System
Content	Intel i5-9400F	8GB	NVIDIA GTX 1080Ti	Ubuntu16.04

数据集使用了公开的 KITTI 目标检测基准数据集和 KITTI 目标跟踪基准数据集,其中,目标检测数据集主要是对汽车和行人的检测,其训练集和测试集分别有 7481 和 7518 张图片,而目标跟踪数据集主要是对汽车与行人的跟踪,其训练集和测试集分别有 21 和 29 个视频序列. 在本实验中,将 KITTI 目标检测基准数据集原有标注的 8 个不同的类别合并为两个类别,具体是将 Car、Van 和 Truck 这 3 类合并为 Car 类,将 Pedestrian、Person_sitting 和 Cyclist 这 3 类合并为 Pedestrian 类,且仅保留 Car 和 Pedestrian 类. 同时,也将 KITTI 目标跟踪基准数据集原有标注的 Pedestrian 和 Person 类合并为 Pedestrian,将 Van 和 Car 合并为 Car 类,且仅保留 Car、Pedestrian 和 Cyclist 类. 同时,将 KITTI 目标检测基准数据集的训练集按 8∶1∶1 划分为训练集、验证集和测试集,用于对目标检测网络的训练和评估. 另外,考虑到目标检测数据集和目标跟踪数据集有部分图片是重合的,为了公平起见,首先将目标跟踪数据集的训练集中 21 个视频序列分别按 7∶3 的比例切分成训练集和验证集,分别用来重新训练目标检测网络和验证本文提出的多目标跟踪算法. 接着,将目标跟踪数据集的训练集全部用来重新训练目标检测网络,用于在该目标跟踪数据集的测试集上评估本文所提出的多目标跟踪算法.

3.2 训练过程

3.2.1 Centernet 网络训练过程

首先,本文为了评估 CenterNet 网络,先使用 KITTI 目标检测基准数据集. 然后,为了评估多目标跟踪算法,使用 KITTI 目标跟踪数据集重新训练 CenterNet 网络,训练过程都是未加载预训练模型,优化器都为 Adam,初始的学习率都为 1.25×10^{-4}. 在按比例划分好的目标检测数据集的训练集和目标跟踪数据集的训练集上的两次训练过程都是一样的,输入分辨率为 512×512,训练 140 个 epoch,batch 为 8,并分别在第 90 和 120 的 epoch 处,使学习率分别下降 10 倍. 而在目

标跟踪数据集的全部训练集上先后训练 3 次,第 1 次未加载预训练模型,第 2、3 次均以上次训练出来的模型为预训练模型来加载训练. 首先以 512 × 512 的输入分辨率,训练 230 个 epoch,分别在第 90 和 120 的 epoch 处,使学习率分别下降 10 倍. 接着,以 384 × 1280 的输入分辨率,训练 140 个 epoch,分别在第 90 和 120 的 epoch 处,使学习率分别下降 10 倍. 最后,以 384 × 1280 的输入分辨率,训练 40 个 epoch,分别在第 10 和 15 的 epoch 处,使学习率分别下降 10 倍.

3.2.2　关联概率网络和区域推荐网络训练过程

由于关联概率网络和基于区域推荐网络的辅助跟踪器是受 SiamRPN 网络启发而设计的,且其权重共享,因此关联概率网络和区域推荐网络的权重可由以 CIResNet_22 为主干网络的 SiamRPN 网络训练得到. 首先,SiamRPN 网络加载在 ImageNet 数据集上训练得到的预训练模型. 接着,在使用裁剪程序处理后的 VID 和 Youtu-BB 数据集上训练,训练 50 个 epoch. 裁剪程序为从历史帧中裁剪出目标模板区域,并将其尺寸变为 127 × 127,且以历史帧的目标的中心点在当前帧图片中裁剪出同样中心点的两倍区域,并将其尺寸变为 255 × 255,以组成成对的图片用于网络的训练.

3.3　评价指标

评估多目标跟踪算法采用的指标如表 2 所示,其中,MOTA 和 MOTP 对多目标跟踪算法总体性能进行评估,而

表 2　多目标跟踪算法的指标

Table 2　Metrics used for multiple object tracking

Metrics	Better	Perfect	Description
MOTA	higher	100%	Multiple Object Tracking Accuracy
MOTP	higher	100%	Multiple Object Tracking Precision
MT	higher	100%	Mostly tracked targets
ML	lower	0	Mostly lost targets
IDS	lower	0	Identity switches

Mostly Tracked(MT) ,Mostly Lost(ML) 、ID-Switch(IDS) 和 Fragmentations(FRAG) 对跟踪器在给目标分配正确的 ID 的效率进行评估. 另外,本文也做了关于目标检测网络性能的实验评估,其指标如表 3 所示.

表 3　目标检测算法的指标

Table 3　Metrics used for object detection

Metrics	Better	Perfect	Description
$AP_{0.50:0.95}$	higher	1	AP at IoU = 0.50:0.05:0.95
$AP_{0.50}$	higher	1	AP at IoU = 0.50
$AP_{0.75}$	higher	1	AP at IoU = 0.75
AP_S	higher	1	AP for small objects: area < 322
AP_M	higher	1	AP for small objects: 322 < area < 962
AP_L	higher	1	AP for large objects: area > 962

3.4　实验结果及分析

图 7 为 KITTI 跟踪基准数据集的测试集中视频序列 1 中连续的 4 帧视频序列的目标检测与跟踪的结果. 从图 7 中可以看出,在第 3 和第 4 帧中,目标检测器因汽车有部分遮挡,所以没有检测到这辆汽车,但是辅助跟踪器却能很好地跟踪这辆有部分遮挡的汽车,其 ID 为 7,没有发生改变. 图 8 为 KITTI 跟踪基准数据集的测试集中视频序列 0 中连续的 4 帧的目标跟踪的结果. 从图 8 中可以看出,本文算法在拥挤的停车环境中仍然能对车辆进行很好的跟踪.

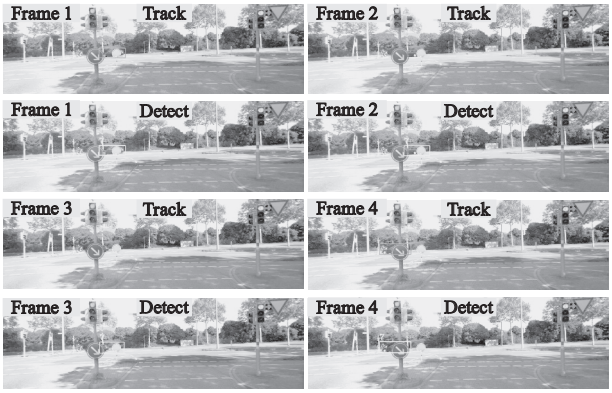

图 7　KITTI 跟踪基准测试集中序列 1 中连续 4 帧视频序列的检测与跟踪对比结果

Fig. 7　Comparison results of detection and tracking of four consecutive video sequences in video sequence 1 of KITTI tracking benchmark test set

为了验证本文提出的 Res2Net_plus 模块对 CerterNet 网络的影响,本文将 Res2Net_plus 模块拆分成 Res2Net、1 × 1 和 SE-Net 3 部分,在 KITTI 目标检测数据集上进行了对比实验,结果如表 4 所示. 另外,为了验证所提出的关联概率网络、辅助跟踪器对整体多目标算法性能的影响,本文在 KITTI 跟踪基准数据集中包含 21 个视频序列的训练集上进行了对比实验,结果如表 5 所示. 最后,为了与已有的多目标跟踪算法进行比较,本文还在 KITTI 目标跟踪数据集中包含 28 个视频序列的测试集上进行与其他多目标跟踪算法的对比实验,结果分别如表 6 和表 7 所示.

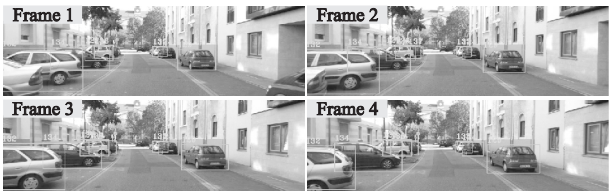

图 8　KITTI 跟踪基准测试集中序列 0 中连续 4 帧视频序列的跟踪的结果

Fig. 8　Video sequence tracking results of four consecutive frames in video sequence 0 of KITTI tracking benchmark test set

从表 4 中可以看出,Res2Net、1 × 1 和 SE-Net 模块可使 CenterNet 网络的各项指标均有提升,表明了所提出的 Res2Net_plus 模块可提升 CenterNet 网络对无人车驾驶场景下车辆和行人的检测精度.

从表 5 中可以看出,相较于仅使用目标的位置信息(相邻帧目标的 IoU 值) ,关联概率网络和辅助跟踪器均可以提高多目标跟踪算法对车辆目标的跟踪能力. 另外,基于区域推荐网络的辅助跟踪器也可以一定程度上提高多目标算法的性能,尤其是 IDS 和 FRAG 这两个指标都有明显的提升.

从表 6 和表 7 中可以看出,本文所提出的无人车驾驶场景下的多目标跟踪算法对于车辆与行人这两类,在大部分指标上都领先其他几个多目标跟踪算法. 尤其,本文提出的多目标跟踪算法在 MOTP 和 FRAG 这两个指标均领先其他算法许多,这也表明了所提出的多目标跟踪算法在无人车驾驶场景下具有很好的竞争力.

表 4 在 KITTI 检测基准数据集(训练集中划分出来的 748 张图片)上试验 Res2Net_plus 模块对 CerterNet 网络的影响

Table 4 Effects of Res2Net_plus module for CenterNet on the KITTI detecting benchmark dataset (748 pictures divided from the training set)

ResNet(Bottleneck)	Res2Net	1×1	SE-Net	$AP_{0.50:0.95}$	$AP_{0.50}$	$AP_{0.75}$	AP_S	AP_M	AP_L	Total	FPS
√	-	-	-	0.557	0.846	0.609	**0.411**	0.554	0.667	8448166	20
-	√			0.549	0.839	0.589	0.398	0.550	0.655	8206822	19
-	√	√		0.553	0.838	0.595	0.404	0.554	0.654	8870086	17
-	√	√	√	**0.558**	**0.847**	**0.614**	0.385	**0.563**	**0.671**	8980462	16

表 5 在 KITTI 跟踪基准数据集(21 个视频序列的训练集中划分出来的验证集)上试验各模块对多目标跟踪算法的影响

Table 5 Effects of each module for the multi-object tracking algorithm on the KITTI tracking benchmark dataset (the verification set divided by 21 video sequences of training set)

IoU	SimNet	Auxiliary tracker	Car						Pedestrian					
			MOTA	MOTP	MT	ML	IDS	FRAG	MOTA	MOTP	MT	ML	IDS	FRAG
√	-	-	67.21%	**84.66%**	48.00%	14.67%	17	127	44.96%	**77.84%**	36.20%	29.31%	14	49
√	√	-	67.25%	**84.66%**	48.00%	14.67%	15	125	45.49%	**77.84%**	**36.21%**	29.31%	5	46
√	√	√	**68.80%**	84.18%	**52.67%**	**13.33%**	**11**	**81**	**46.60%**	77.69%	**36.21%**	**25.86%**	**3**	**34**

表 6 在 KITTI 跟踪基准数据集的测试集上与其他多目标跟踪算法对比实验结果('Car'类)

Table 6 Comparison of experimental results with other multi-object tracking algorithm on test set of KITTI tracking benchmark dataset('Car' class)

Method	MOTA	MOTP	MT	ML	IDS	FRAG
Point3DT [13]	68.24%	76.57%	**60.62%**	12.31%	**111**	725
SASN-MCF_nano [14]	70.86%	82.65%	58.00%	**7.85%**	443	975
RMOT [15]	65.83%	75.42%	40.15%	9.69%	209	727
Ours	**71.74%**	**83.86%**	46.77%	17.08%	191	**518**

表 7 在 KITTI 目标跟踪数据集的测试集上与其他多目标跟踪算法对比实验结果('Pedestrian'类)

Table 7 Comparison of experimental results with other multi-object tracking algorithm on test set of KITTI tracking benchmark dataset('Pedestrian' class)

Method	MOTA	MOTP	MT	ML	IDS	FRAG
AB3DMOT [16]	36.36%	64.86%	14.09%	48.45%	142	773
NOMT-HM [17]	**39.26%**	71.14%	**21.31%**	41.92%	184	863
Complexer-YOLO [18]	16.46%	62.69%	2.41%	**38.14%**	527	1637
Ours	39.11%	**74.89%**	14.78%	44.33%	**79**	**565**

4 结 论

本文提出多目标算法在无人车驾驶场景下对车辆与行人具有很好的跟踪能力. 实验结果表明,提出的 Res2Net_plus 模块可以有效提高目标检测器对车辆与行人的检测精度,关联概率网络也能很好地构建目标的特征表达模型,从而显著提高多目标跟踪算法对目标的跟踪能力. 另外,辅助跟踪器也可以有效对漏检的目标进行持续跟踪,这样可以很好地改善因目标部分遮挡、目标检测器失效造成的目标漏检所导致的目标 ID 频繁切换或跟踪轨迹断开等问题,尤其可以从 IDS 和 FRAG 这两个指标看出. 但从实验结果可以看出,相较于其他的算法,对车辆的跟踪,本文提出的算法的 MOTA 和 MT、ML 这 3 个指标还不是很具有竞争力,还有提升的空间,后续的研究将进一步解决好对相互拥挤且外观相近的目标跟踪能力较弱的问题.

References:

[1] Baser E, Balasubramanian V, Bhattacharyya P, et al. FANTrack: 3D multi-object tracking with feature association network [C] //2019 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2019: 1426-1433.

[2] Porzi L, Hofinger M, Ruiz I, et al. Learning multi-object tracking and segmentation from automatic annotations [J]. arXiv preprint arXiv: 1912.02096, 2019.

[3] Zhang L, Li Y, Nevatia R. Global data association for multi-object tracking using network flows [C] //2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008: 1-8.

[4] Zhou X, Wang D, Krähenbühl P. Objects as points [J]. arXiv preprint arXiv: 1904.07850, 2019.

[5] Gao S H, Cheng M M, Zhao K, et al. Res2Net: a new multi-scale backbone architecture [J]. arXiv preprint arXiv: 1904.01169, 2019.

[6] Li B, Yan J, Wu W, et al. High performance visual tracking with si-

amese region proposal network [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8971-8980.

[7] Leal-Taixé L, Canton-Ferrer C, Schindler K. Learning by tracking: siamese CNN for robust target association [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016: 33-40.

[8] Vartanian O, Coady L, Blackler K. 3D multiple object tracking boosts working memory span: implications for cognitive training in military populations [J]. Military Psychology, 2016, 28(5): 353-360.

[9] Chen Y, Fan H, Xu B, et al. Drop an octave: reducing spatial redundancy in convolutional neural networks with octave convolution [C] //Proceedings of the IEEE International Conference on Computer Vision, 2019: 3435-3444.

[10] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[11] Ren S, He K, Girshick 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.

[12] Kuncheva L I. Full-class set classification using the Hungarian algorithm [J]. International Journal of Machine Learning and Cybernetics, 2010, 1(1-4): 53-61.

[13] Wang S, Sun Y, Liu C, et al. PointTrackNet: an end-to-end network for 3-D object detection and tracking from point clouds [J]. arXiv preprint arXiv: 2002.11559, 2020.

[14] Gündüz G, Acarman T. Efficient multi-object tracking by strong associations on temporal window [J]. IEEE Transactions on Intelligent Vehicles, 2019, 4(3): 447-455.

[15] Yoon J H, Yang M H, Lim J, et al. Bayesian multi-object tracking using motion context from multiple objects [C] //2015 IEEE Winter Conference on Applications of Computer Vision, IEEE, 2015: 33-40.

[16] Weng X, Kitani K. A baseline for 3d multi-object tracking [J]. arXiv preprint arXiv: 1907.03961, 2019.

[17] Choi W. Near-online multi-target tracking with aggregated local flow descriptor [C] //Proceedings of the IEEE International Conference on Computer Vision, 2015: 3029-3037.

[18] Simon M, Milz S, Amende K, et al. Complex-YOLO: real-time 3D object detection on point clouds [J]. arXiv preprint arXiv: 1803.06199, 2018.