

基于深度学习的多目标跟踪研究进展综述

姜来为¹, 王 策², 杨宏宇¹

(1. 中国民航大学 安全科学与工程学院, 天津 300300; 2. 中国民航大学 计算机科学与技术学院, 天津 300300)

摘 要: 本文首先指出多目标跟踪任务面临的挑战和传统方法的局限性, 其次对基于检测的跟踪和联合检测跟踪两类算法进行文献综述与分析, 然后对多目标跟踪算法常用的评估指标和公开的数据集进行归纳, 并对两类算法的性能进行比较分析, 最后立足于当前研究现状对有待解决的问题和未来研究重点做出预测和展望。

关键词: 计算机视觉; 多目标跟踪; 基于检测的跟踪; 联合检测跟踪; 深度学习

中图分类号: TP391 **文献标志码:** A **文章编号:** 1671-5497(2025)11-3429-17

DOI: 10.13229/j.cnki.jdxbgxb.20240149

Review of multi-object tracking based on deep learning

JIANG Lai-wei¹, WANG Ce², YANG Hong-yu¹

(1. School of Safety Science and Engineering, Civil Aviation University of China, Tianjin 300300, China; 2. School of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China)

Abstract: Firstly, this paper points out the challenges faced by the design of multi-target tracking algorithms and the limitations of traditional methods. Secondly, a literature review and analysis of two types of algorithms are conducted: detection-based-tracking and joint-detection-tracking. Then, the commonly used evaluation indicators and publicly available datasets in the multi-object tracking algorithms were summarized, and the performance indicators of the two types of methods were analyzed. Finally, based on the current research status, the predictions and outlooks on the problems to be solved and the focuses of the future researches are made.

Key words: computer vision; multi-object tracking; detection based tracking; joint detection tracking; deep learning

0 引 言

多目标跟踪 (Multi-object tracking, MOT) 是计算机视觉领域的重要研究方向, 其主要任务是在视频序列中实时、准确地检测出多个目标的位置, 然后赋予每个目标唯一的编号, 将连续帧中相同编号的目标连接形成轨迹。

近些年, 随着深度学习的快速发展, MOT 在视频监控、智慧交通、自动驾驶、体育广播等领域展现出巨大的潜力, 这些任务也都依赖于高质量的跟踪算法^[1,2]。根据视频流中 MOT 算法在处理当前帧时是否会用到后续帧的信息, 可以将其分为在线跟踪和离线跟踪两大类: 在线目标跟踪

收稿日期: 2024-02-05.

基金项目: 国家自然科学基金重点项目 (U2033210); 天津市教委科研计划项目 (2020KJ025).

作者简介: 姜来为 (1986-), 女, 讲师, 博士. 研究方向: 计算机视觉, 无线通信, 网络空间安全. E-mail: lwjiang@cauc.edu.cn

在进行目标身份推理的过程中,只能用到当前帧以及之前帧的信息;而离线目标跟踪除了用到当前以及之前的帧外,还会用到后续帧的信息,使用全局信息提高跟踪质量,把MOT看作一个全局优化问题^[3,4]。在线方法虽然相较于离线方法普遍性能差一些,但由于很多实际应用场景有着实时性的要求,因此,在线方法应用更加广泛,本文涉及的MOT算法为在线方法。

本文围绕MOT算法设计相关问题进行深入研究,并对现有算法进行分析比较,具体贡献如下:

(1) 本文涵盖了近些年最新的MOT方法,根据其跟踪过程中检测和跟踪是否独立分为基于检测的跟踪(Detection based tracking, DBT)和联合检测跟踪(Joint detection tracking, JDT)两大类,着重研究深度学习技术在MOT领域的广泛应用,并深入分析了这些方法的优点和局限性,这种分类方式有助于读者更好地理解 and 区分不同类型的MOT算法。

(2) 给出了基于深度学习的MOT算法常用的评估指标和数据集,同时探讨了每种指标评估的侧重点以及不同数据集的特点,这将有助于读者更有针对性地提升算法性能,为后续的研究提供参考。

(3) 对当前MOT算法在各大常用数据集上的结果进行了统计和性能比较,并分析了产生差异化原因。

(4) 对有待解决的问题和未来可能的研究重点做出预测和展望,以期吸引科研人员共同探索这一极具价值的研究方向,加速MOT算法的实际应用。

1 问题与挑战

1.1 多目标跟踪面临的主要挑战

MOT与单目标跟踪(Single object tracking, SOT)相比,不仅需要判别不同目标和背景,还需要同时长时间准确跟踪、保持多个目标身份,以上特点使得SOT算法难以直接应用于MOT中。在基于视觉信息的MOT中,最直观的特征就是目标的外观特征,复杂场景下外观特征有很多不确定性因素,包括目标部分遮挡、长时间跟踪下目标消失、目标变化、背景干扰、目标外观相似等,使得目标在视频前后帧之间的外观特征差距过大,给正确识别和跟踪目标带来挑战。此外,目标运

动模式不确定、算法实时性也给相关研究带来一定的困难,这也是研究者们关注的重点。图1为MOT常见干扰因素,总的来讲,多目标跟踪面临的主要挑战包括以下几个方面:

(1) 目标部分遮挡:目标部分遮挡是MOT算法面临的主要挑战之一,如图1(a)所示,当同一场景中出现多个目标时,它们可能会相互遮挡,导致目标在图像中出现部分遮挡的情况,这会使得目标的外观信息不够完整,进而造成目标外观模型表示不足,同时遮挡的部分也会引入无用的干扰信息。

(2) 长时间跟踪下目标消失:目标在运动过程中可能会暂时移出监控范围或者与其他物体重叠形成遮挡,即短暂的目标消失,如图1(b)中实线框为目标被正常检测和跟踪,图1(c)中虚线框为目标实际存在但因完全遮挡未被检测出来,这会导致跟踪器无法有效地更新,目标重新出现在监控范围内被识别为新的目标,导致跟踪失败。

(3) 目标变化:目标变化也是跟踪过程中面临的挑战之一,主要包括目标尺度变化、目标形变、目标旋转等。图1(d)(e)为目标形变样例,图中的人是典型的非刚性物体,短时间内不同帧之间同一目标形态可能存在较大差异,这为算法的后续跟踪带来了很大困难。

(4) 背景干扰:背景干扰也是目标跟踪相关研究中的常见问题,如何在复杂场景中将目标和背景有效分离出来,减少背景因素的干扰,是目标跟踪面临的挑战之一。在现实场景中,存在着天气、光照、背景等变化,如图1(f)(g)所示,夜晚复杂的光照会给提取目标外观特征带来干扰,降低算法识别跟踪的准确性。

(5) 目标外观相似:物体的外观特征包括颜色、纹理、形状等,如图1(h)所示,视频中存在多个外观相似的目标,如何提取具有判别性的物体外观特征也是研究的重点与难点。

(6) 目标运动模式的不确定性:运动特征也是MOT中的常用特征,区别于有明确道路指向的车辆跟踪等简单运动模式场景,在复杂场景下,目标的行为模式很难长久维持。由于目标众多,目标之间的相互影响也会影响目标的运动,因此,发掘其中隐含的运动模式十分具有挑战性。

(7) 算法的实时性:一些实际应用场景要求MOT算法能够实时地检测跟踪场景中感兴趣的目標,而现在很多MOT算法为了保证跟踪精度

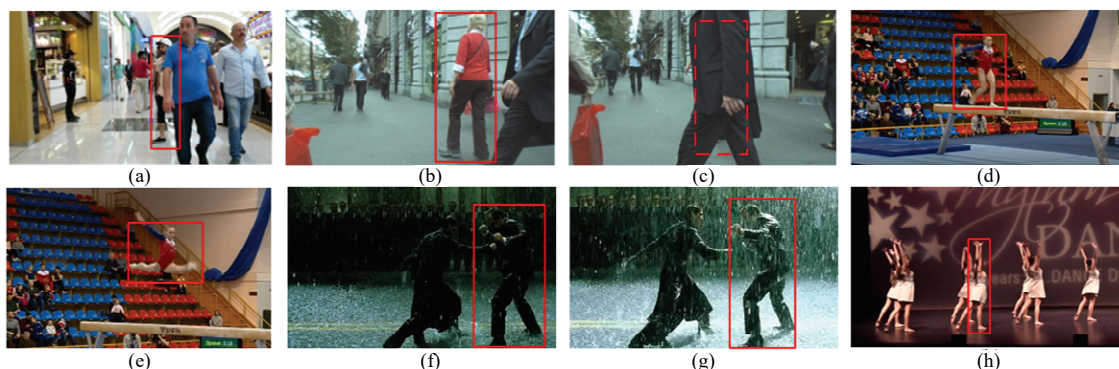


图 1 MOT 常见干扰因素

Fig. 1 Common interference factors in MOT

通常会用到复杂模型,难以保证算法的实时性。如何设计具有实时性、可以应用到现实场景中的 MOT 算法也是近年来研究的热点。

1.2 传统多目标跟踪及其不足之处

早期的 MOT 算法主要是通过对目标的外观进行建模,然后在后续的视频序列中找到最相似的特征进行目标定位,常用的特征包括尺度不变特征变换(Scale-invariant feature transform, SIFT)、加速稳健特征(Speeded-up robust features, SURF)、方向梯度直方图(Histogram of oriented gradients, HOG)特征、Harris 特征等。KLT (Kanade-Lucas-Tomasi)算法^[5]是一种经典的光流估计算法,用于计算图像序列中像素点在时间上的运动,它基于一组特征点的选取,通过计算这些特征点在相邻帧之间的位移估计像素的运动。KLT 算法的优点是计算效率高,适用于小范围的目标跟踪;然而,它对光照变化和视角变化比较敏感,且不适用于目标遮挡或大范围的目标跟踪。在复杂场景下,KLT 算法的效果并不理想。

随着研究的深入,人们发现基于目标建模的方法对整张图片进行处理实时性较差,且目标本身的外观变化具有随机性和多样性。如果仅通过单一的数学模型描述待跟踪的目标具有很大局限性,难以有效进行预测,不能很好地解决跟踪时出现的目标遮挡问题。后来研究者们开始将预测算法加入跟踪中,在预测值附近进行目标搜索,从而缩小搜索范围。基于卡尔曼滤波的跟踪方法^[6]通过状态方程和历史帧目标位置预测下一帧的目标位置,但是该类仅适用于线性运动。

为了对非线性运动进行更好的跟踪,Isard 等^[7]提出粒子滤波算法,其核心思想是使用一组粒子表示目标的状态分布,并通过不断地重采样和更新估

计目标的位置和运动。该方法由于缺少当前时刻的观测信息,在复杂场景下跟踪效果较差。Nummiaro 等^[8]通过融入颜色特征改善目标形变和遮挡情况,但当跟踪目标和背景颜色相似度较高时,会导致跟踪目标丢失。杨欣等^[9]提出一种基于多特征融合的自适应粒子滤波算法,将目标的颜色、边缘和纹理特征引到粒子滤波中,有效弥补了单一特征以及固定权值跟踪中的不足。采用核密度构建表观模型的方法可以缩小搜索范围,通常采用均值漂移对运动目标的位置进行评估,比如 Comaniciu 等^[10]将均值漂移算法用于目标跟踪,使用颜色直方图建立表观模型,通过梯度上升不断迭代求解密度函数的局部极值定位目标。针对目标被遮挡或发生形变的问题,Jeyakar 等^[11]提出了一种基于加权片段的目标跟踪方法,其权重来自片段与背景之间的颜色差异,通过将分块和多核融合结合起来进行跟踪,不仅对光照不敏感而且在处理目标大面积遮挡方面鲁棒性很好。

此外,基于相关滤波的 MOT 算法也是常用的一类方法。相关滤波源于信号处理领域,相关性用于表示两个信号之间的相似程度,其核心思想是寻找一个滤波模板,让下一帧的图像与滤波模板做卷积操作,响应最大的区域是预测的目标,其本质是求解一个多元二次多项式的回归问题。Bolme 等^[12]提出的最小化输出平方误差和(Minimum output sum of squared error, MOSSE)首次将相关滤波应用于目标跟踪领域,其核心思想是最小化输出平方误差和训练一个滤波器以实现目标跟踪,但是其得到的是一个线性回归的模型。Henriques 等^[13]通过引入核函数的方式将分类器变为非线性回归模型,解决了低维线性不可分的问题。MOSSE 使用单通道灰度特征对目标进行

表示,为了提高模型在复杂场景的鲁棒性,研究者们尝试引入使用不同类别的特征。Danelljan等^[14]将颜色属性引入相关滤波器并设计了一种自适应降维方法降低计算复杂度。Zhang等^[15]基于朴素贝叶斯框架建立目标物体与其局部上下文之间的时空关系,解决了目标位置模糊的问题。

上述传统MOT算法在目标运动相对缓慢、光照条件较好以及背景相对简单的情况下可以做到实时跟踪。但是这些使用手工设计特征的方法容易受到光照变化、目标遮挡、视角变化等因素的影响,对目标形状和背景干扰较为敏感,与此同时对目标运动模型和外观模型的先验知识有着较高要求,使得

算法在复杂场景下鲁棒性较差且泛化能力有限。

深度学习作为一种机器学习方法,自20世纪80年代开始出现。2012年,Krizhevsky等^[16]在ImageNet图像分类挑战赛中使用卷积神经网络(Convolutional neural networks, CNN)取得了显著优势,自此深度学习引起了广泛关注。本文对传统的MOT算法和基于深度学习的MOT算法的发展历程做了总结,对比了其各自的优缺点,并以时间为线索绘制了一些具有代表性的工作,如图2所示。从图2中可以看出,2014年以前,MOT方法以传统方法为主,自2014年以后深度学习方法逐渐引入MOT领域,并逐渐成为主流方式。

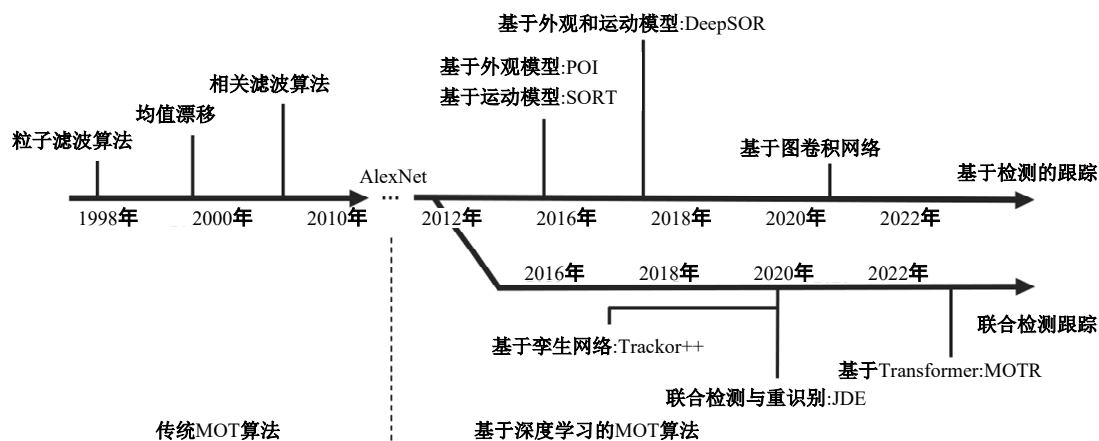


图2 不同阶段代表性MOT算法发展脉络图

Fig. 2 Context diagram of representative MOT algorithm development at different stages

2 基于深度学习的多目标跟踪算法相关概念及分类

2.1 基本定义

近些年,研究者们将深度学习应用于MOT领域取得了巨大进展。基于深度学习的MOT算法通常包括以下4个步骤:

(1) 目标定位(目标检测):输入的视频通过目标检测算法对视频中某一帧的目标进行定位,得到目标的边界框信息。

(2) 特征提取:检测完成后,通过特征提取算法提取当前帧每个检测目标的特征,常用的方法包括使用卷积神经网络提取目标的外观特征,使用卡尔曼滤波算法建模目标的运动特征,这些特征将用于检测目标和已存在轨迹的匹配。

(3) 相似度计算:相似度计算用于判断视频相邻不同帧两个目标是否属于同一个实体,衡量当前目标检测框与已存在轨迹的相似程度。

MOT算法经常用到外观相似度计算,通过比较目标的外观特征,如颜色直方图、纹理特征、深度特征等度量目标与轨迹之间的外观相似度,常用的度量方法包括欧氏距离、余弦相似度、相关系数等。此外,针对目标密集且相互遮挡的情况,外观特征不能很好地表征目标,常引入运动特征,使用外观特征和运动特征共同完成相似度的计算。

(4) 数据关联:根据相似度计算的结果关联检测目标与轨迹,将相同目标分配同一个编号匹配已存在的轨迹,将新出现的目标初始化为一个新轨迹,删除视线范围内消失的轨迹。常用的数据关联方法有匈牙利算法、多假设关联、概率数据关联等。

2.2 基于深度学习的多目标跟踪算法分类

基于深度学习的MOT算法根据检测和跟踪阶段是否独立,可以分为DBT和JDT两个类别。图3为DBT和JDT算法流程图,可以看出DBT算法主要分为检测和跟踪两个阶段。第一

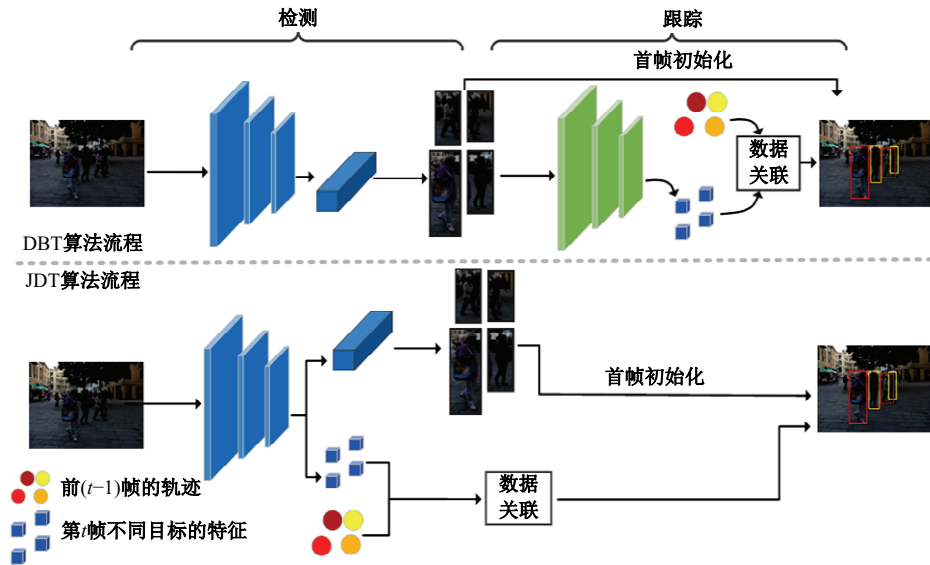


图3 DBT和JDT算法流程图

Fig. 3 DBT and JDT algorithm flowchart

阶段使用基于卷积神经网络的检测器对每一帧图像中感兴趣的目标用矩形框进行标注,第二阶段跟踪器使用检测器的输出结果再次提取目标的外观、运动等特征,并计算和 $(t-1)$ 帧中已存在轨迹特征之间的相似度以进行目标和轨迹的关联。JDT算法则是将检测和跟踪融合到一个

框架中,对检测器中提取的特征进行了复用,跟踪器可以直接利用检测器输出目标的外观特征,避免了跟踪器接收检测结果的过程。近些年很多MOT算法都是基于上述两种范式,图4为目前常见的基于深度学习MOT算法分类,下文将详细介绍。

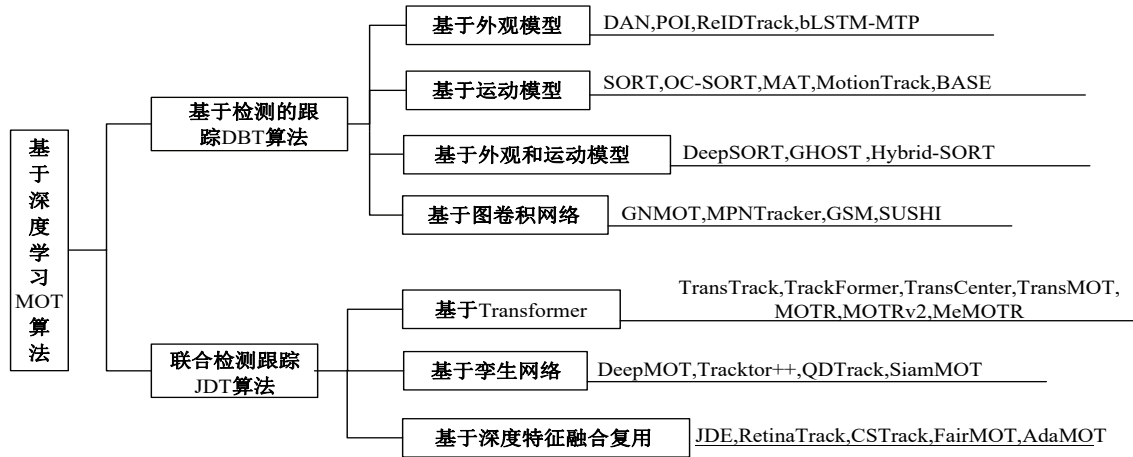


图4 常见的基于深度学习的MOT方法

Fig. 4 Common MOT methods based on deep learning

3 基于深度学习的DBT算法

DBT算法是非常常见的一类MOT算法,主要包括检测和跟踪两个阶段:首先通过基于深度学习的目标检测算法在每一帧中检测出可能的目标物体位置;然后跟踪阶段使用检测结果作为输入,通过一定的关联策略建立帧与帧之间目标的轨迹,下面对算法的各个阶段展开详细论述。

3.1 检测阶段

基于检测的MOT将检测和跟踪分隔开,检测由检测器完成,然后跟踪器使用检测器输出的结果进行特征关联,从而形成每一个目标的轨迹。因为跟踪器的输入全部来源于检测器的输出,所以检测结果的好坏会极大地影响最终结果。为了能够更好地定量分析跟踪算法的性能,消除检测性能带来的影响,许多数据集也提供了公开的检

测器。还有一些算法则使用自定义检测器,通过提高检测质量提高整体性能,如Faster R-CNN^[17]及其变体、SSD^[18]以及YOLO系列等。经典的SORT^[19]算法就是使用Faster R-CNN目标检测网络替换MOT数据集原本使用的通道聚合特征(Aggregate channel features, ACF)获得的检测,使得跟踪速度和准确度大幅提升。通过在Faster R-CNN基础上添加一个分支网络可以得到Mask R-CNN^[20],在实现目标检测的同时分割目标实例。Zhou等^[21]使用ACF方法检测和跟踪行人。Zhao等^[22]使用单阶段SSD作为检测器和多尺度增强方法训练数据。Zhang等^[23]提出的ByteTrack方法使用YOLOX^[24]作为其检测器。

3.2 跟踪阶段

跟踪阶段是MOT算法的核心,在这一阶段利用检测阶段输出的结果,使用运动、位置、外观线索或它们的组合进行检测框和已存在轨迹的跨帧关联。下面将按照基于外观模型、基于运动模型、基于外观和运动模型、基于图卷积网络分别进行算法分析与比较。

3.2.1 基于外观模型的跟踪算法

目标的外观特征是MOT中计算相似度矩阵的重要线索。在SOT算法中也有不少基于外观的模型,其主要区别在于SOT侧重于构建复杂的外观模型以区分目标和背景的不同,而MOT中则大多是利用目标的外观信息区分不同目标。

Sun等^[25]提出了一个深度亲和网络(Deep affinity network, DAN)用于对不一定相邻帧之间目标的外观进行建模并评估其相似性,以便后续进行可靠的轨迹匹配。行人重识别(Person Re-identification),也简称为Re-ID,是利用计算机视觉技术判断图像或者视频序列中是否存在特定行人的技术,受Re-ID任务的启发,许多MOT任务也采用了基于Re-ID的方法提取目标的外观特征。Yu等^[26]基于GoogLeNet设计了一个特征提取网络以提取目标外观特征,并嵌入一个简单的跟踪器当中,实验表明凭借高性能的检测和外观特征,可以达到与最先进的离线跟踪器相当的性能。Huang等^[27]使用了自监督外观模型进行Re-ID特征的提取,在不使用跟踪注释的情况下就能训练一个新的外观嵌入模型。而以上基于Re-ID模型的方法在实际应用中的缺点是:一旦存在衣着相似、背景或者光线干扰等因素,产生的外观特

征就有可能出错。为了解决上述问题, Kim等^[28]提出了一种多轨迹池模块,可以使用所有的轨迹联合更新外观模型从而提高目标在外观相似时的匹配可靠性,并且利用轨迹之间的依赖关系提出了一种新的池训练策略,由此产生基于双线性长短时记忆网络(Long-short term memory network, LSTM)架构的跟踪器,其精度与当时最先进的在线跟踪器相当但速度较慢。

基于外观模型的MOT算法可以很好地提取检测框内具有鉴别力的特征以便对目标进行更好的描述,从而提高算法性能,在目标运动模式复杂或相机运动场景下表现更好。缺点是外观模型容易受到光照、遮挡等情况的影响,进而造成误判。

3.2.2 基于运动模型的跟踪算法

运动特征是MOT跟踪阶段另一个常用的特征。运动模型通过捕捉目标的动态行为预估目标在未来帧的潜在位置,从而缩小搜索空间。其中最常用的是线性运动模型,它假设目标在每个时间段内的运动速度是恒定的,但是在复杂场景下如物体突然加速、变向、遮挡等情况下线性运动模型就不再适用。

SORT算法^[19]把目标帧间的位移视作线性匀速运动,通过卡尔曼滤波器基于前一帧的目标位置预测当前帧的目标位置,当视频帧率较高时,即使运动目标在整个运动过程中是非线性的,目标的运动依然可以在很短的时间间隔内视为线性运动,故算法也能很好地工作,但是在目标跟踪的过程中物体发生遮挡,导致目标在视频序列上的时间间隔变大,即使一个目标在丢失一段时间后仍可以通过SORT算法重新关联,但是由于时间误差放大,卡尔曼滤波参数已经远偏离真实值,因此,可能会再次丢失。此外, SORT算法在高帧率视频的连续帧中物体的位移和噪声可能有相同的量级,且噪声会累积到后续位置估计中。为了改善这一问题, Cao等^[29]提出了OC-SORT, 采用以观测为中心的更新策略减小累积误差, 即一旦一个跟踪轨迹在丢失跟踪一段时间后又重新与观测结果相关联, 便会根据丢失目标期间开始和结束的观测值生成虚拟轨迹, 用虚拟轨迹重新更新丢失期间的卡尔曼滤波器参数, 这样就避免了误差的积累。另外, 文章还提出用观测值代替估计值以减少运动方向计算的噪声, 并引入运动方向的一致性项帮助关联, 提出以观测为中心将轨迹最

后一次观测与新检测到的观测相关联以恢复轨迹,实验表明此方法相较于SORT对遮挡和非线性物体运动具有更强的鲁棒性。Han等^[30]提出了一个即插即用的运动感知跟踪器MAT,其中包括3个模块:①运动集成模块。用来平衡非刚性的目标和相机运动的兼容问题。②动态重连接上下文模块。当目标拥挤或遮挡时仍对遮挡轨迹进行预测,以便在将来的某一帧恢复轨迹,与OC-SORT不同的是,这里是基于相机运动和目标非线性运动的程度,动态决定目标的存活时间,以确保基于长距离运动的重连接鲁棒性。③3D积分图像模块。施加时空约束过滤无用的轨迹和检测关联。Qin等^[31]同样也是致力于解决长时跟踪的问题。该模型设计了一个交互模块,根据目标相邻两帧的偏移,计算出一个表征目标之间相互影响的矩阵,随后利用该矩阵经过图卷积网络和多层感知机得到目标的预测位置;设计了一个恢复模块拼接碎片化的轨迹,对于没有匹配成功的检测和轨迹,计算它们之间的速度-时间关系,保留匹配程度高的组合。上述方法大多采用检测器检测到目标固定次数就认为该目标是新出现的目标并将目标初始化为新的轨迹,忽略了检测的置信度,概率跟踪器使用概率模型估计目标的位置和运动状态。Larsen等^[32]发现考虑目标运动的距离,利用检测器的置信度以及建模非均匀杂波特性和对于概率跟踪器非常重要,进而提出了贝叶斯近似单假设跟踪器BASE。该方法使得模型能够感知目标的远近,缓解了显式建模目标距离在某些复杂环境下如目标被其他物体遮挡而导致距离测量不准确的问题。

基于运动模型的MOT算法可以很好地改善目标遮挡和检测器漏检导致的轨迹碎片化问题,相比于利用外观特征的跟踪器,速度更快,对检测器的性能依赖更低,在一些外观特征提取不可靠的复杂场景中更具鲁棒性。但这类算法非常依赖运动模型预测目标的位置,因此,对于运动轨迹复杂场景跟踪能力会有所下降。

3.2.3 基于外观和运动模型的跟踪算法

在一些复杂场景中,如存在遮挡或相机移动情况时,仅依靠外观或者运动特征很难对目标进行鲁棒跟踪。为了解决上述问题,有研究者将目标的外观信息和运动信息引入MOT算法,以提高对目标跟踪的准确性和鲁棒性。SORT算

法^[19]就是利用运动特征对目标的位置进行预测,而DeepSORT算法^[33]则是在其基础上借助了Re-ID领域模型引入外观信息,从而改善SORT算法ID Switch过大的问题。Karunasekera等^[34]提出了一种基于目标运动、外观、结构相异性度量的MOT算法。为了充分发掘目标内在的运动模式,也有一些研究者使用LSTM的方法。Seiden-schwarz等^[35]发现在Re-ID领域表现很好的模型如果直接引入MOT中用于提取目标的外观特征,并不一定会直接带来MOT性能的提升,因为在Re-ID模型提取跟踪目标的特征时,与当前帧相邻近的帧目标外观变化较小,相距较远的帧目标形变严重,因此,文中对不同时间跨度的视频帧采用不同的管理方式。此外,文中还分析了不同能见度、遮挡时间、相机运动条件下哪些情况仅使用Re-ID是不够的,进而引入线性运动模型,使用加权的方式将外观特征和运动特征结合起来从而获得最优性能。Yang等^[36]指出大多数MOT方法通过显式或隐式地利用强线索(即空间和外观信息),这些线索表现出强大的实例级别区分。然而当两个物体在当前帧中高度重叠时,检测和估计轨迹位置之间的交集会变得模糊,两个物体的外观特征都会被前景特征所主导,研究者通过结合弱线索(置信状态、高度状态和速度方向)补偿强线索,可以有效改善MOT中目标间的相互遮挡问题。

基于外观和运动模型的MOT方法可以弥补单独使用外观模型或运动模型的缺点,在复杂的场景下有更强的鲁棒性,但是其算法复杂度较高,实时性也有所下降。

3.2.4 基于图卷积网络的跟踪算法

图卷积神经网络(Graph convolutional network, GCN)在MOT中将检测对象视为节点,将检测对象与轨迹之间的关联关系视为边构建图,从而将MOT中检测对象和轨迹的匹配问题转化为最小成本网络流量优化问题。虽然目标的状态、位置、运动特征等发生变化,但算法仍使用之前的信息进行目标关联跟踪,这会导致跟踪准确性下降,这种情况在采用静态图进行目标匹配时尤为明显。针对上述问题,Li等^[37]设计了一个外观图网络和一个运动图网络分别捕捉目标间的外观相似性和运动相似性,另外,还精心设计了更新机制,使得图中的节点、边和全局变量都可以更

新,最后还提出了一种处理遗漏检测的策略,以弥补检测器的缺陷。Brasó 等^[38]利用 MOT 经典网络流公式定义了一个基于消息传递网络完全可微的框架,在该框架内外观和几何线索可以在整个检测集合中传播从而实现全局推理。Liu 等^[39]针对以往 MOT 算法多考虑单个对象的特征表示而对于对象之间的关系研究较少情况,提出了图相似度模型,该模型既考虑单个目标的特征,也考虑对象之间的关系,对遮挡和具有相似外观的目标具有更强的鲁棒性。一般跟踪器专注于短期关联的场景,缺乏长期身份保存的鲁棒性;而那些 MOT 性能好的算法是对长短期关联分别进行处理,又缺乏通用性。Cetintas 等^[40]提出 SUSHI 算法用于分层处理视频,该算法利用分层结构,每一层都是相同的图卷积网络,底层关注短期关联且高层关注长期关联。此方法和分治算法思想类似,通过底层合并可以逐步使得关联的视频长度增加,进而关联的范围可以推广到多个时间跨度,具有高度的通用性和良好的可扩展性。

传统跟踪算法大多只考虑目标的外观特征和运动特征,而基于图卷积网络的 MOT 算法可以很好地考虑目标与目标之间的空间关系对目标关联的影响,因此,基于图卷积网络的 MOT 算法在建模能力、鲁棒性方面表现出色,是处理复杂 MOT 问题的重要方法。

总的来讲,基于深度学习的 DBT 算法将 MOT 问题分为两个阶段:检测和跟踪,此时 MOT 系统中至少需要检测器和跟踪器两个计算密集型构件。这样整个算法执行时间大约是两个构件各自执行时间之和,并且会随着场景中目标数的增多而变大,不太适用于构建实时的 MOT 系统。为了降低时间复杂度,基于深度学习的 JDT 算法逐渐引起研究者的关注。

4 基于深度学习的 JDT 算法

JDT 是另一类常见的基于深度学习的 MOT 算法,是将检测阶段和跟踪阶段结合在一起,有的 JDT 算法是将检测和跟踪阶段设计成一个端到端的网络,联合优化;还有的 JDT 算法是将检测阶段的网络和跟踪阶段特征提取部分的网络设计成一个网络,或者通过特定设计直接用检测部分网络去生成跟踪阶段所需的运动特征,它们的区别是检测阶段和跟踪阶段融合程度的不同。下面

将分别介绍基于 Transformer、基于孪生网络和基于深度特征融合复用的 JDT 算法。

4.1 基于 Transformer 的 JDT 算法

Transformer^[41]是 Google 提出的一种基于注意力机制的深度学习模型,能够有效捕捉序列数据中的长距离依赖关系,最初被应用于自然语言处理领域,后由于其出色的特征表征能力被广泛应用于计算机视觉领域。Sun 等^[42]提出的 TransTrack 首次将 Transformer 应用到 MOT 任务中,设计了两个并行的解码器分别用于检测和跟踪,并利用 Query-Key 机制跟踪当前帧中已存在的目标并检测新的目标。Meinhardt 等^[43]提出了一种全新的 MOT 框架 TrackFormer,该框架参考 DETR^[44]思路,是一个端到端并且以 Tracking-by-attention 为全新思路实现了帧间的数据关联。注意力机制则确保了该模型同时考虑位置、遮挡和目标的识别特征。以上涉及的 MOT 算法大多是基于目标框表示方法,但在极度稠密场景下,一个目标框可能包含多个人的外观信息,这就导致训练的基于视觉外观表示的模型不准确,造成关联歧义。为了解决此问题,Xu 等^[45]提出了 TransCenter 方法,该方法抛弃了以往从稀疏查询输出稀疏目标框的方式,采用像素级多尺度密集查询预测目标中心点的方式,一定程度上解决了目标框重叠问题。此外,TransCenter 在训练过程中不需要烦琐的匹配算法,只需要简单的回归中心位置即可。为了更加有效地对多个目标之间进行时空建模,Chu 等^[46]提出了 TransMOT,它将一种新的时空图转换器用于 MOT,通过将检测的轨迹和候选框表述为稀疏带权图以显式建模它们之间的时空关系。由于图的稀疏性,其计算起来更加有效。另外,作者还建立了一个级联关联结构,用于低置信度检测从而解决长时间遮挡问题。MOT 研究关键在于目标轨迹的时序建模,现有方法多使用如外观相似度等简单的启发式方法,并不足以对复杂变化进行建模,即时间建模的能力不足。针对此问题,Zeng 等^[47]提出了一个完全的端到端跟踪框架 MOTR,并提出全新的 Track Query 概念:每个 Track Query 将为一个目标进行完整的追踪建模,在帧与帧之间传输、更新从而实现无缝检测和跟踪。此外,文中还提出了时域聚合网络融合多帧训练并为长程时域依赖关系建模,隐式地进行时间关联。Zhang 等^[48]在 MOTR 基础上提出

了MOTRv2,他们认为DETR网络同时进行学习检测和数据关联两个不同的任务是有冲突的,因此,在MOTR基础上加入了额外的目标检测器,而整体上依旧保持了模型端到端的特性。Gao等^[49]通过自定义的记忆-注意力层将检测特征长期记忆注入轨迹特征中,这显著提高了模型的目标关联能力,同时也加强了模型长时间信息建模的能力,使同一目标的轨迹特征更加稳定和可区分。此外,还提出了一种自适应聚合算法融合相邻帧的目标特征,以提高跟踪的鲁棒性。

总体而言,基于Transformer的JDT算法具有强大的建模能力和全局信息处理能力,但计算复杂度较高且实时性较差。在实际应用中,需要根据具体场景和需求选择是否采用基于Transformer的JDT算法。

4.2 基于孪生网络的JDT算法

孪生网络是基于两个人工神经网络构建的耦合架构,起初常用于单目标跟踪领域。基于孪生网络的跟踪算法是将跟踪问题转化为检测目标与目标模板的相关性匹配问题。

首次将孪生网络应用于单目标跟踪领域的是SiamFC^[50],图5为算法中孪生网络的结构示意图,该网络具有两个权值共享的分支:一个分支以目标模板图像作为输入,另一个分支利用目标模型在待搜索区域进行滑动密集搜索,将这些候选区域作为输入,然后计算目标模板与各个候选区域的相关性,相关性最大的候选区域就是待跟踪目标在当前帧的位置。把单目标跟踪迁移到多目标跟踪,最直接的方法就是将MOT中的每一个目标都当作一个独立的SOT,但是MOT场景目

标数量太多,同时维持大量的单目标跟踪器会严重影响跟踪的速度,且目标之间相互遮挡,会进一步影响目标外观建模,影响跟踪性能。Xu等^[51]基于双向循环神经网络构建了一种深度匈牙利网络,该网络可以近似看作可微形式的匈牙利算法,通过这种方式,MOT评价指标的梯度损失可以反向传播到整个MOT方法,同时这也使得该方法能够同时端到端地并行训练多个单目标跟踪器以进行多目标跟踪。Bergmann等^[52]通过前一帧轨迹的边界框作为当前帧轨迹初始的边界框进行回归,以得到当前帧轨迹的边界框,省略了匹配的过程,同时为了保持方法在线,引入孪生网络生成外观向量以进行短期的重试别。为了更充分地探索MOT中的外观相似性,Pang等^[53]选取一对临近的图像,通过区域建议网络(Region proposal network, RPN)产生大量的候选区域,采用一种简单的正负样本指定策略和非参数化的Softmax交叉熵损失函数以一种对抗学习的方式完成对嵌入特征的优化,同时使用基于双端Softmax的目标相似度计算保证了相互匹配的一致性,使用相似度阈值完成轨迹-检测之间的关联。Gao等^[54]将单目标跟踪SiameRPN作为MOT中的预测器替换了卡尔曼滤波算法,提高了算法对高速、非线性等场景的适应性。Shuai等^[55]探索了MOT中的运动建模,作者将基于区域的检测网络Faster R-CNN^[17]和两个思路源于孪生单目标跟踪的运动模型相结合,分别是隐式运动模型(Implicit motion model, IMM)和显式运动模型(Explicit motion model, EMM)。不同于CenterTrack^[56]基于点特征进行隐式的目标运动预测,SiamMOT使用区域特征并开发了显式模板匹配策略估计模板的运动,使得此方法在高速运动场景下更具鲁棒性。

孪生网络作为一种特殊的网络结构,将跟踪问题转化为相似度学习问题,起初被广泛应用于SOT中,但是在MOT中目标众多,如果直接维持多个SOT跟踪器会极大降低推理速度,且目标之间相互遮挡也会干扰相似度的计算,因此,在MOT算法中孪生网络大多用于匹配阶段的相似度计算。

4.3 基于深度特征融合复用的JDT算法

在基于DBT的算法中,有两个用于提取特征的网络,分别用于目标检测和数据关联的深度特

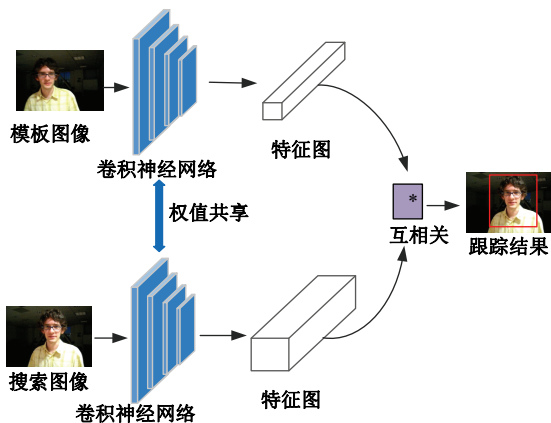


图5 孪生网络目标跟踪框架

Fig. 5 Schematic diagram of siamese network target tracking

征提取,二者提取的特征存在差异性,不能够直接使用。通过目标检测和 Re-ID 任务共享特征提取网络的方式可以提高模型的复用性。一般的做法是在目标检测网络里添加一个专门用于提取数据关联特征的分支。2019 年, Wang 等^[57]提出了 JDE 算法,创新性地目标检测环节和外观特征提取环节这两部分融合设计为一个网络,这极大提高了 MOT 算法的推理速度,但跟踪精度有所降低,同时也为 MOT 发展提供了一个新的思路。Lu 等^[58]在 RetinaNet^[59]基础上增加了提取实例级的 Re-ID 特征分支用于后续的数据关联。曲优等^[60]提出了一种基于锚框对齐卷积特征的目标检测-表观特征提取联合网络 AAC-JDAN,改善了联合检测和重识别类的算法提取表观特征与旋转目标之间关联性弱这一问题,使算法获得了检测旋转目标的能力。Liang 等^[61]分析了 JDE 算法性能下降的原因,认为这是目标表征存在竞争引起的,即检测阶段需要区分目标的类别(例如行人),具有相似的语义,而 Re-ID 则是倾向于区分两个不同的行人,这本质上与检测任务是相矛盾的,这意味着在一项任务中追求高性能可能会导致另一项任务性能下降。此外, JDE 算法在检测阶段引入特征金字塔网络,在该网络中,不同尺度会被分配给不同分辨率的特征,这种分配方式同样不适合跟踪器中的 Re-ID 任务,这会导致语义级别的错位。因此,文中提出了互惠关系网络和尺度感知注意力网络缓解上述提到的目标表征竞争和语义错位。Zhang 等^[62]认为,在人物密集场景中,一个锚框可能包含多个目标,一个目标也可能被多个锚框覆盖,因此, JDE 基于锚框的检测不利于 Re-ID 任务,故作者提出了 FairMOT,其在基于无锚框的检测网络 CenterNet^[63]中添加一个并行的分支提取像素级的 Re-ID 特征,实验结果表明其大幅优于之前提出的方法,为该 MOT 算法提供了一种不错的基线算法。Liang 等^[64]从标签分配和损失函数两个角度去探索如何更好地联合训练检测和 Re-ID,提出了身份感知的标签分配,该种分配方式联合考虑检测和 Re-ID 成本,自适应地为每个目标框选择合适的正样本,因此,更加公平地对待两个子任务,此外还设计了判别性焦点损失函数,监督训练重点关注具有强身份辨别力的正样本。

基于深度特征融合复用的 JDT 算法将跟踪

阶段的网络与检测阶段的网络融合,提高了模型的复用性,速度上也具有很大优势,但目标检测和重识别本质上是不同的两个任务,因此,平衡二者之间的差异性未来该类算法研究的重点。

5 评估指标和数据集

5.1 评估指标

为了能够定量地从不同角度评价 MOT 算法性能,研究者们提出了各种不同的评估指标。目前,在 MOT Challenge 线上评估系统中有一套公认的评估指标衡量 MOT 算法的性能,最常用的有 CLEAR MOT 指标^[65]、ID 指标^[66]和高阶跟踪准确率(Higher order tracking accuracy, HOTA)指标^[67]。此外,最能体现 MOT 算法综合性能的指标是多目标跟踪准确度(Multiple objection tracking accuracy, MOTA),它主要评估的是检测能力,计算公式如式(1)所示:

$$MOTA = 1 - \frac{FN + FP + IDSW}{GT} \in (-\infty, 1] \quad (1)$$

式中:FN 为未被正确检测的真实目标数目,即漏检数;FP 为错误的检测数目,即虚警数;IDSW 为跟踪过程中目标身份标识切换次数;GT 为真实框个数。由于算法中身份标识切换次数可能会比真实框还多,因此, MOTA 取值可能是负数。

多目标跟踪精度(Multiple objection tracking precision, MOTP)可以衡量跟踪算法在正确跟踪的情况下目标位置估计的精度,即跟踪算法在空间上的定位精度。MOTP 计算公式如下所示:

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_t C_t} \quad (2)$$

式中: C_t 为第 t 帧中目标和轨迹匹配数目; $d_{t,i}$ 为第 t 帧中第 i 对检测框中心点和真实框中心点之间的距离。

IDF₁ 表示在 ID 保持不变情况下跟踪准确率和召回率的 F_1 值,相比于考虑 IDSW 的 MOTA, IDF₁ 指标更注重跟踪能力。IDTP、IDFP 和 IDFN 分别为目标 ID 的正确匹配数、虚警数和漏检数。IDF₁ 计算公式为:

$$IDF_1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \quad (3)$$

MOTA 和 IDF₁ 分别强调检测与关联的性能,因此,文献[67]提出了 HOTA 系列指标,该指标可以更好地兼顾考虑检测和关联的性能。De-

tA 是检测的准确率,计算公式为:

$$\text{DetA}_\alpha = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}| + |\text{FP}|} \quad (4)$$

$$\text{DetA} = \int_0^1 \text{DetA}_\alpha d\alpha \quad (5)$$

式中:TP 为只考虑视频序列中目标的检测被正确检测的目标数目,即正样本数;FN 和 FP 与上述 MOTA 公式中的含义相同; α 为真实框和检测框是否匹配的阈值。最终的 DetA 是在不同的阈值下计算 DetA_α 然后求平均得到的。AssA 是关联的准确率,计算公式为:

$$A(c) = \frac{|\text{TPA}(c)|}{|\text{TPA}(c)| + |\text{FNA}(c)| + |\text{FPA}(c)|} \quad (6)$$

$$\text{AssA}_\alpha = \frac{1}{|\text{TP}|} \sum_{c \in \{\text{TP}\}} A(c) \quad (7)$$

$$\text{AssA} = \int_0^1 \text{AssA}_\alpha d\alpha \quad (8)$$

式中: c 为预测的 ID 编号; α 为真实框和检测框是否匹配的阈值;TP 为目标检测的正样本数;TPA(c)表示目标检测 TP 集合中真实框和检测框 ID 均为 c 的数目;FNA(c)表示 TP 集合中真实框 ID 为 c 但预测不为 c 和目标检测 FN 集合中真实框为 c 的数目;FPA(c)为 TP 集合中检测框 ID 为 c 但真实框 ID 不为 c 以及在 FP 集合中检测的 ID 为 c 的数目。最终,AssA 是在不同的阈值下计算 AssA_α 然后求平均得到。

HOTA 则是检测和关联两个准确率的几何平均值,如此可以确保该指标能够兼顾检测和关联,更全面地评估 MOT 算法的整体性能,最终,HOTA 是在不同的阈值 α 下计算 HOTA_α 然后求平均得到。计算公式为:

$$\text{HOTA}_\alpha = \sqrt{\text{DetA}_\alpha \times \text{AssA}_\alpha} \quad (9)$$

$$\text{HOTA} = \int_0^1 \text{HOTA}_\alpha d\alpha \quad (10)$$

5.2 评估数据集

一个典型的 MOT 数据集是由多个视频帧序

列组成。在这些序列中,对于每个视频帧数据集通常会提供目标的位置信息,为每个目标分配唯一的标识符,以便跟踪同一目标在不同帧之间的运动轨迹。KITTI 数据集^[68]包含用于自动驾驶和目标跟踪的大量现实世界驾驶场景的图像序列,可以提供包含车辆、行人和自行车等目标的标注边界框及关联轨迹。KITTI 数据集以其真实性和多样性而闻名,被广泛应用于评估 MOT 在复杂真实场景下的表现。MOTChallenge 系列数据集是常用 MOT 数据集,目前包括 4 个行人多目标跟踪数据集,即 MOT15^[69]、MOT16^[70]、MOT17^[70]和 MOT20^[71],这些数据集目标密集程度不断提高,涵盖了不同条件下的跟踪场景:有摄像机固定的跟踪场景,也有摄像机放在移动车辆的场景,此外还包含了俯视、平视和仰视不同的拍摄视角;同时这些视频是在不同的光照条件下拍摄的,包括晴天、阴天、夜晚等。DanceTrack 数据集^[72]搜集了 100 段视频,内容包括集体舞蹈、功夫、体操等,视频具有如下特点:①目标人物穿着相似甚至一致;②目标之间有大量的遮挡和位置交错;③目标的运动模式非常复杂多样,呈现明显的非线性,并且时常伴随多样的肢体动作。

以上用于评估的 MOT 数据集提供了丰富的视频序列,包括行人、车辆等类型目标,并提供准确的标注信息如目标边界框和关联轨迹,常用于评估多目标跟踪算法的准确性、鲁棒性和实时性。表 1 为 MOT 常用数据集概况。

5.3 方法性能分析

为了更好地对基于深度学习的 MOT 算法进行比较分析,本节对近些年的主流方法进行了整理,在 MOT17、MOT20、DanceTrack 3 个数据集上进行了对比,表 2 和表 3 分别为 MOT17、MOT20 算法性能评估结果,表 4 为 DanceTrack 算法性能评估结果。选取了 MOTA、IDF₁、HOTA、FN、FP、IDs、DetA、AssA 作为指标。“↑”表

表 1 MOT 典型数据集概况

Table 1 Overview of typical MOT datasets

数据集名称	年份	视频序列	大小/GB	特点及描述
KITTI Tracking ^[68]	2012	50	15	允许同时追踪人和车辆
MOT15 ^[69]	2015	22	1.3	场景丰富,密集程度低
MOT16 ^[70]	2016	14	1.9	场景丰富,规范标注,目标密集程度更高
MOT17 ^[70]	2016	14	5.5	增加更多场景,提供了更多的公开检测器
MOT20 ^[71]	2020	8	5.0	人群更加密集,相互遮挡情况更加严重
DanceTrack ^[72]	2022	100	16.5	目标外观相似且存在大量遮挡,运动模式复杂

表 2 MOT17数据集算法性能评估结果

Table 2 Performance evaluation results of MOT17 dataset algorithm

方法	类别	检测器	其他数据	MOTA/% $\uparrow$	IDF ₁ / % $\uparrow$	HOTA/% $\uparrow$	FN $\downarrow$	FP $\downarrow$	IDs $\downarrow$	速度/FPS
DeepSORT ^[33]	DBT	公共	否	60.3	57.4	—	—	36 111	5 529	8.1
DAN ^[25]	DBT	公共	否	53.9	49.5	—	234 592	25 423	8 431	6.3
GHOST ^[35]	DBT	私有	否	78.7	77.1	62.8	—	—	—	2 325
SHSHI ^[40]	DBT	公共	否	62.0	71.5	54.6	183 825	29 428	1 041	21.1
ByteTrack ^[23]	DBT	公共	是	67.4	70.0	56.1	172 636	9 939	1 331	27.4
ByteTrack* ^[23]	DBT	私有	是	80.3	77.3	63.1	83 721	25 491	2 196	29.6
MTA ^[30]	DBT	公共	否	67.1	69.2	—	161 547	22 756	1 279	18.5
MAT ^[30]	DBT	私有	否	69.5	63.1	—	138 741	30 660	2 844	18.5
OC-SORT* ^[29]	DBT	私有	是	78.0	77.5	63.2	108 000	15 100	1 950	—
MotionTrack ^[31]	DBT	私有	是	81.1	80.1	65.1	81 660	23 802	1 149	—
TransTrack ^[42]	JDT	私有	是	74.5	63.9	54.1	112 137	28 323	3 363	10.0
TransMOT ^[46]	JDT	私有	是	76.7	75.1	61.7	93 150	36 231	2 346	9.6
TransCenter ^[45]	JDT	私有	否	71.9	62.3	54.4	137 008	17 378	4 046	1.0
TrackFormer ^[43]	JDT	私有	是	74.1	68.0	57.3	108 777	34 602	2 892	7.4
MOTR ^[47]	JDT	私有	否	73.4	68.6	57.8	135 561	21 123	2 115	7.5
QDTrack ^[53]	JDT	私有	否	68.7	66.3	—	146 643	26 589	3 378	—
SiamMOT ^[55]	JDT	公共	是	65.9	63.3	—	170 955	18 098	3 040	—
MOTRv2 ^[48]	JDT	私有	否	78.6	75.0	62.0	—	—	—	—
JDE ^[57]	JDT	私有	是	63.0	59.5	—	162 927	39 888	6 171	18.8
CStrack ^[61]	JDT	私有	是	74.9	72.3	—	114 303	23 847	3 567	16.4
FairMOT ^[62]	JDT	私有	是	73.7	72.3	59.3	117 477	27 507	3 303	25.9
AdaMOT ^[64]	JDT	私有	是	74.9	75.5	—	112 134	26 883	2 613	26.0
AdaMOT* ^[64]	JDT	私有	是	75.7	75.5	—	95 385	39 777	2 226	26.0

表 3 MOT20数据集算法性能评估结果

Table 3 Performance evaluation results of MOT20 dataset algorithm

方法	类别	检测器	其他数据	MOTA/% $\uparrow$	IDF ₁ / % $\uparrow$	HOTA/% $\uparrow$	FN $\downarrow$	FP $\downarrow$	IDs $\downarrow$	速度/FPS
ByteTrack ^[23]	DBT	私有	是	77.8	75.2	61.3	87 594	26 249	1 223	17.5
GHOST ^[35]	DBT	公共	否	52.7	55.3	43.4	—	—	1 216	—
GHOST ^[35]	DBT	私有	否	73.7	75.2	61.2	—	—	—	—
SHSHI ^[40]	DBT	公共	否	61.6	71.6	55.4	168 098	29 429	1 053	5.5
OC-SORT* ^[29]	DBT	私有	否	75.5	75.9	62.1	108 000	18 000	913	—
MotionTrack ^[31]	DBT	私有	是	78.0	76.5	62.8	84 152	28 629	1 165	—
BASE ^[32]	DBT	私有	是	78.2	77.6	63.6	84 211	27 606	984	16.8
CStrack ^[61]	JDT	私有	是	66.6	68.6	—	144 358	25 404	3 196	4.5
FairMOT ^[62]	JDT	私有	是	61.8	67.3	54.6	88 901	103 440	1 331	13.2
AdaMOT ^[64]	JDT	私有	是	68.8	71.8	—	117 009	41 993	2 406	12.1
AdaMOT* ^[64]	JDT	私有	是	69.1	71.4	—	99 833	58 471	1 792	12.1
TrackFormer ^[43]	JDT	私有	是	68.6	65.7	54.1	140 373	20 384	1 532	5.7
TransTrack ^[42]	JDT	私有	是	65.0	59.4	48.9	150 197	27 191	3 068	14.9
TransCenter ^[45]	JDT	私有	是	61.0	49.8	43.5	147 890	49 189	4 493	1.0
MOTRv2 ^[48]	JDT	私有	是	76.2	72.2	60.3	—	—	—	—

示指标值越大越好,“ $\downarrow$ ”表示指标值越小越好。方法后加*表示该方法使用了轨迹插值,检测器分为数据集提供的公共检测器和算法自定义的私有检测器,其他数据一栏指的是除了该数据集之外,

算法是否还使用了其他数据集进行训练,速度的单位是帧/s(FPS)。

根据各个算法在 3 个测试集上的性能对比,可以得到以下结论:

表 4 DanceTrack数据集算法性能评估结果

Table 4 Performance evaluation results of DanceTrack dataset algorithm

方法	类别	检测器	其他数据	HOTA/% ↑	DetA/% ↑	AssA/% ↑	MOTA/% ↑	IDF ₁ /% ↑
ByteTrack*[²³]	DBT	私有	是	47.7	71.0	32.1	89.6	53.9
OC-SORT*[²⁹]	DBT	私有	是	55.1	80.4	40.4	92.2	54.9
SUSHI[⁴⁰]	DBT	私有	是	63.3	80.1	50.1	88.7	63.4
Hybrid-SORT[³⁶]	DBT	私有	否	62.2	63.0	47.4	81.9	91.6
FairMOT[⁶²]	JDT	私有	是	39.7	66.7	23.8	82.2	40.8
CenterTrack[⁵⁷]	JDT	私有	否	41.8	78.1	22.6	86.8	35.7
MOTR[⁴⁷]	JDT	私有	是	54.2	73.5	40.2	79.7	51.5
MOTRv2[⁴⁸]	JDT	私有	是	73.4	83.7	64.4	92.1	76.0
MeMOTR[⁴⁹]	JDT	私有	否	68.5	80.5	58.4	89.9	71.2

(1) 参与对比的这 3 个数据集侧重点各有不同,MOT17 的主要特点是场景丰富,MOT20 的主要特点是人群密集且目标间相互遮挡严重,而 DanceTrack 的主要特点则是目标运动模式复杂且外观相似。以 ByteTrack^[23]和 OC-SORT^[29]为例,这两种算法都是使用 YOLOX^[24]作为检测器,算法均是对跟踪器做了改进,因此,本文更加关注偏向于跟踪的 IDF₁ 指标。在 MOT17 上,ByteTrack 和 OC-SORT 的 IDF₁ 分别为 77.3% 和 77.5%,在人群密集且遮挡严重的 MOT20 上,二者的 IDF₁ 分别为 75.2% 和 75.9%,在目标运动轨迹更加复杂的 DanceTrack 上,二者的 IDF₁ 分别为 52.5% 和 54.9%,通过对比可以看出 OC-SORT 的 IDF₁ 指标在 DanceTrack 上提升最明显,而在 MOT20 上进步不明显是因为该数据集虽然目标间遮挡严重,但其中线性运动居多,实验的结果与 OC-SORT 提高目标遮挡且非线性运动跟踪的鲁棒性是相吻合的。由此可以得出结论:不同的算法专注改进的点各有不同,在实际应用场景中需要根据场景的特点更加灵活地选择算法。

(2) 基于深度特征融合复用的 JDT 算法通过把目标检测和匹配阶段提取目标外观特征这两个计算量比较大的部分组合在一起,减少了该类算法的整体计算量,但是早期该类算法在跟踪精度上没有太大竞争力。如 JDE^[57]在 MOT17 上速度达到了 18.8 FPS,但是 MOTA 和 IDF₁ 仅为 63.0% 和 59.5%。CSTrack^[61]从检测和重识别目标表征存在竞争的角度去改进 JDE,将 MOTA 和 IDF₁ 提高到了 74.9% 和 72.3%。FairMOT^[62]从锚框不适合重识别任务角度选用了不基于锚框的目标检测算法改进 JDE,将 MOTA 和 IDF₁ 提高到了 73.7% 和 72.3%。AdaMOT^[64]则是从标签分配和损失函数的角度改进此类 MOT

算法,并以 FairMOT 作为基线算法,将 MOTA 和 IDF₁ 提高到了 75.7% 和 75.5%。由此可以得出结论:联合检测和重识别的 JDT 算法改进的重点是平衡检测和重识别之间的差别,从而使这两种不同的任务能够更好地联合训练。

(3) 通过对比可以发现,大多数 MOT 算法漏检数量 FN 远大于虚警数量 FP。这是因为在 MOT 中,当目标被其他物体或者场景元素遮挡时就会导致目标漏检,或者小目标在图像中像素比较小,易受环境影响也容易发生漏检;而虚警通常是由于误检或者环境噪声所引起的,因而数量相对较少。即漏检产生的原因更复杂,减少小目标以及被遮挡目标的漏检数量是提高 MOT 算法性能的关键因素之一,使用高质量的检测器,可以直接减少 FN 和 FP 的数量,进而直接导致 MOTA 的提高和间接导致匹配阶段 IDF₁ 等指标的提高。以 ByteTrack^[23]为例,跟踪器不变,在使用 MOT Challenge 提供的公共检测器时,FN 和 FP 分别为 172 636 和 9 939,MOTA 和 IDF₁ 分别为 67.4% 和 70.0%,在换用 YOLOX 检测器后 FN 和 FP 变为 83 721 和 25 491,尤其是 FN 大幅度降低,MOTA 和 IDF₁ 变为 80.3% 和 77.3%。由此可以推断出 MOT 任务非常依赖于高质量的检测,一个好的检测器可以大幅提高 MOT 中的各个指标。

4) 在 MOT20 中,TransCenter^[45]的 MOTA 和 IDF₁ 分别为 61.0% 和 49.8%,但速度仅为 1 FPS,TrackFormer^[41]的 MOTA 和 IDF₁ 分别为 68.6% 和 65.7%,速度为 5.7 FPS。在目标运动模式复杂的 DanceTrack 上,MOTRv2^[48]的 HOTA、MOTA 和 IDF₁ 分别达到了 74.4%、92.1% 和 71.2%,是目前最好的结果。由于 Transformer 具有建模能力强、可扩展性好、上下文信息利用充

分等优点,基于 Transformer 的 MOT 算法在复杂的场景下依然可以有很高的跟踪准确性和鲁棒性,但是基于 Transformer 类的算法模型复杂,参数和计算量较大,通常推理较慢,难以达到实时性的要求,而且此类算法需要大量的数据进行训练,如何在不影响跟踪性能的同时降低此类算法的复杂度以及探索无监督训练可以很大程度提高此类算法在实际场景中的适用性。

6 待解决的问题和未来研究方向

目前,MOT 仍然是计算机视觉领域重要的研究方向,其应用广泛且具有重要的实际意义。然而,这个研究领域仍然存在着许多问题和挑战:

(1) 多机位的 MOT。在多摄像头存在的情况下,通常有两种情况,一种是多摄像头拍摄同一场景,此时同一个目标会被多个摄像头捕捉到不同图像,融合来自不同摄像机的信息是此场景下的关键问题;另一种情况是每个摄像机拍摄不同的场景,在此场景下如何实现在不同的拍摄角度下跨相机目标关联是一个关键问题,也可以看作跨摄像机的重识别问题。如何在以上场景中更加鲁棒、准确地进行多目标跟踪值得进一步探索。

(2) 多模态 MOT 问题。在现实场景中,MOT 通常需要同时利用多种传感器进行目标检测和跟踪。如自动驾驶领域,当目标距离较远摄像头难以提取有效的像素信息,需要激光雷达辅助获取周边目标的位置信息,当在夜间行驶时又需要红外传感器辅助获取周边的情况。然而,不同传感器的数据来源之间存在着差异,如何将多种模态的数据进行对齐融合处理,以实现模态的互补,有待进一步探索。

(3) 更加有效的长时 MOT 技术。在长时间的视频序列中目标遇到遮挡、消失、重新出现以及状态变化等复杂情况,如果想持续、准确地跟踪目标,所面临的挑战包括目标连续性维护、目标重识别、数据质量稳定性、多样的运动模式、计算复杂性等,这些问题都十分具有挑战性。

(4) 设计基于无监督学习的 MOT 算法。近年来,研究者们主要关注的是有监督学习 MOT 算法。然而监督学习需要大量的标注进行训练,耗费大量人力物力;且有监督 MOT 算法容易出现过拟合的现象,受到数据集质量和多样性限制,很难满足实际应用需求。设计基于无监督学习的

MOT 算法可以一定程度上避免上述缺点,但由于缺乏适合目标跟踪的高质量特征,设计无监督 MOT 算法仍具有很大的挑战性。

7 结束语

MOT 作为计算机视觉的研究热点,可以作为多种下游任务的基础,包括视频监控分析、自动驾驶、人机交互、军事航天等,为这些任务提供准确的目标位置、轨迹等信息,以用于更高级别的视觉理解和决策。然而,在实际应用中,MOT 面临着系列挑战,如目标的遮挡、外观变化、运动模式不确定等,如何在复杂场景下实现稳定跟踪是研究者们关注的重点。本文主要对近些年基于深度学习的 MOT 算法进行了总结,首先回顾了 MOT 中面临的挑战及传统方法的局限性,根据检测和跟踪是否独立将常见的基于深度学习的 MOT 算法分为两大类进行介绍,分别讨论了其优缺点,接下来介绍了 MOT 中常用的性能指标和数据集并对典型 MOT 方法进行了分析,最后展望了 MOT 未来的发展方向,为推动此研究领域发展提供有价值的参考。

参考文献:

- [1] 金沙沙,龙伟,胡灵犀,等.多目标检测与跟踪算法在智能交通监控系统中的研究进展[J].控制与决策,2023,38(4):890-901.
Jin Sha-sha, Long Wei, Hu Ling-xi, et al. Research progress of detection and multi-object tracking algorithm in intelligent traffic monitoring system[J]. Control and Decision, 2023, 38(4): 890-901.
- [2] Cui Y, Zeng C, Zhao X, et al. SportsMOT: a large multi-object tracking dataset in multiple sports scenes [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 9921-9931.
- [3] Peng J, Wang T, Lin W, et al. TPM: multiple object tracking with tracklet-plane matching[J]. Pattern Recognition, 2020, 107: No. 107480.
- [4] Ren W, Wang X, Tian J, et al. Tracking-by-counting: using network flows on crowd density maps for tracking multiple targets[J]. IEEE Transactions on Image Processing, 2020, 30: 1439-1452.
- [5] Shi J. Good features to track[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, USA, 1994: 593-600.

- [6] Broida T J, Chellappa R. Estimation of object motion parameters from noisy images[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1986(1): 90-99.
- [7] Isard M, Blake A. Condensation—conditional density propagation for visual tracking[J]. *International Journal of Computer Vision*, 1998, 29(1): 5-28.
- [8] Nummiaro K, Koller-meier E, Van Gool L. An adaptive color-based particle filter[J]. *Image and Vision Computing*, 2003, 21(1): 99-110.
- [9] 杨欣, 刘加, 周鹏宇, 等. 基于多特征融合的粒子滤波自适应目标跟踪算法[J]. *吉林大学学报: 工学版*, 2015, 45(2): 533-539.
Yang Xin, Liu Jia, Zhou Peng-yu, et al. Adaptive particle filter for object tracking based on fusing multiple features[J]. *Journal of Jilin University (Engineering and Technology Edition)*, 2015, 45(2): 533-539.
- [10] Comaniciu D, Ramesh V, Meer P. Real-time tracking of non-rigid objects using mean shift[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head, USA, 2000: 142-149.
- [11] Jeyakar J, Babu R V, Ramakrishnan K. Robust object tracking with background-weighted local kernels [J]. *Computer Vision and Image Understanding*, 2008, 112(3): 296-309.
- [12] Bolme D S, Beveridge J, Draper B A, et al. Visual object tracking using adaptive correlation filters[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, 2010: 2544-2550.
- [13] Henriques J F, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels[C]//*European Conference on Computer Vision*, Florence, Italy, 2012: 702-715.
- [14] Danelljan M, Shahbaz K F, Felsberg M, et al. Adaptive color attributes for real-time visual tracking[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, 2014: 1090-1097.
- [15] Zhang K, Zhang L, Liu Q, et al. Fast visual tracking via dense spatio-temporal context learning[C]//*European Conference on Computer Vision*, Zürich, Swiss-Confederation, 2014: 127-141.
- [16] Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks[J]. *Advances in Neural Information Processing Systems*, 2012, 60: 84-90.
- [17] 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*, 2016, 39(6): 1137-1149.
- [18] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C]//*European Conference on Computer Vision*, Amsterdam, Netherland, 2016: 21-37.
- [19] Bewley A, Ge Z, Ott L, et al. Simple online and realtime tracking[C]//*IEEE International Conference on Image Processing (ICIP)*. Phoenix, USA, 2016: 3464-3468.
- [20] He K, Gkioxari G, Dollár P, et al. Mask R-CNN [C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, Venice, Italy, 2017: 2961-2969.
- [21] Zhou Z, Xing J, Zhang M, et al. Online multi-target tracking with tensor-based high-order graph matching[C]//*24th International Conference on Pattern Recognition (ICPR)*, Beijing, China, 2018: 1809-1814.
- [22] Zhao D, Fu H, Xiao L, et al. Multi-object tracking with correlation filter for autonomous vehicle[J]. *Sensors*, 2018, 18(7): 2004.
- [23] Zhang Y, Sun P, Jiang Y, et al. Bytetrack: multi-object tracking by associating every detection box[C]//*European Conference on Computer Vision*, Tel Aviv, The State of Israel, 2022: 1-21.
- [24] Ge Z, Liu S, Wang F, et al. YOLOX: exceeding yolo series in 2021[DB/OL]. [2021-08-06]. <https://arxiv.org/pdf/2107.08430.pdf>.
- [25] Sun S J, Akhtar N, Song H S, et al. Deep affinity network for multiple object tracking[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 43(1): 104-119.
- [26] Yu F, Li W, Li Q, et al. POI: multiple object tracking with high performance detection and appearance feature[C]//*Computer Vision-ECCV 2016 Workshops*, Amsterdam, The Netherlands, 2016: 36-42.
- [27] Huang K, Sun B, Chen F, et al. Reidtrack: multi-object track and segmentation without motion[DB/OL]. [2023-08-03]. <https://arxiv.org/pdf/2308.01622.pdf>.
- [28] Kim C, Fu X L, Alotaibi M, et al. Discriminative appearance modeling with multi-track pooling for real-time multi-object tracking[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Virtual, 2021: 9553-9562.

- [29] Cao J, Pang J, Weng X, et al. Observation-centric sort: rethinking sort for robust multi-object tracking [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 9686-9696.
- [30] Han S, Huang P, Wang H, et al. MAT: motion-aware multi-object tracking[J]. *Neurocomputing*, 2022, 476: 75-86.
- [31] Qin Z, Zhou S, Wang L, et al. Motiontrack: learning robust short-term and long-term motions for multi-object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 17939-17948.
- [32] Larsen M, Rolfsjord S, Gusland D, et al. Base: probably a better approach to multi-object tracking [DB/OL]. [2023-09-21]. <https://arxiv.org/pdf/2309.12035.pdf>
- [33] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]//2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017: 3645-3649.
- [34] Karunasekera H, Wang H, Zhang H. Multiple object tracking with attention to appearance, structure, motion and size[J]. *IEEE Access*, 2019, 7: 104423-104434.
- [35] Seidenschwarz J, Brasó G, Serrano V, et al. Simple cues lead to a strong multi-object tracker[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 13813-13823.
- [36] Yang M, Han G, Yan B, et al. Hybrid-sort: weak cues matter for online multi-object tracking[C]//Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2024: 6504-6512.
- [37] Li J, Gao X, Jiang T. Graph networks for multiple object tracking[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, USA, 2020: 719-728.
- [38] Brasó G, Leal-taixé L. Learning a neural solver for multiple object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2020: 6247-6257.
- [39] Liu Q, Chu Q, Liu B, et al. GSM: graph similarity model for multi-object tracking[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Yokohama, Japan, 2020: 530-536.
- [40] Cetintas O, Brasó G, Leal-taixé L. Unifying short and long-term tracking with graph hierarchies[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 22877-22887.
- [41] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, Long Beach, USA, 2017: 5999-6009.
- [42] Sun P, Cao J, Jiang Y, et al. Transtrack: multiple object tracking with transformer[DB/OL]. [2021-05-04]. <https://arxiv.org/pdf/2012.15460.pdf>.
- [43] Meinhardt T, Kirillov A, Leal-taixé L, et al. Trackformer: multi-object tracking with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 8844-8854.
- [44] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European Conference on Computer Vision, Virtual, 2020: 213-229.
- [45] Xu Y, Ban Y, Delorme G, et al. TransCenter: transformers with dense representations for multiple-object tracking[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 45(6): 7820-7835.
- [46] Chu P, Wang J, You Q, et al. Transmot: spatial-temporal graph transformer for multiple object tracking[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, USA, 2023: 4870-4880.
- [47] Zeng F, Dong B, Zhang Y, et al. MOTR: end-to-end multiple-object tracking with transformer[C]//European Conference on Computer Vision, Tel Aviv, The State of Israel, 2022: 659-675.
- [48] Zhang Y, Wang T, Zhang X. MOTRv2: bootstrapping end-to-end multi-object tracking by pretrained object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 22056-22065.
- [49] Gao R, Wang L. MeMOTR: long-term memory-augmented transformer for multi-object tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 9901-9910.
- [50] Bertinetto L, Valmadre J, Henriques J, et al. Fully-convolutional siamese networks for object tracking [C]//European Conference on Computer Vision, Amsterdam, Netherlands, 2016: 850-865.

- [51] Xu Y, Osep A, Ban Y, et al. How to train your deep multi-object tracker[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2020: 6787–6796.
- [52] Bergmann P, Meinhardt T, Leal-Taixe L. Tracking without bells and whistles[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, USA, 2019: 941–951.
- [53] Pang J, Qiu L, Li X, et al. Quasi-dense similarity learning for multiple object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2021: 164–173.
- [54] Gao X, Shen Z, Yang Y. Multi-object tracking with siamese-RPN and adaptive matching strategy[J]. Signal, Image and Video Processing, 2022, 16(4): 965–973.
- [55] Shuai B, Berneshawi A, Li X, et al. SiamMOT: siamese multi-object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2021: 12372–12382.
- [56] Zhou X, Koltun V, Krähenbühl P. Tracking objects as points[C]//European Conference on Computer Vision, Virtual, 2020: 474–490.
- [57] Wang Z, Zheng L, Liu Y, et al. Towards real-time multi-object tracking[C]//European Conference on Computer Vision, Virtual, 2020: 107–122.
- [58] Lu Z, Rathod V, Votel R, et al. Retinatrack: online single stage joint detection and tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2020: 14668–14678.
- [59] Lin T, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 2017: 2980–2988.
- [60] 曲优, 李文辉. 基于多任务联合学习的多目标跟踪方法[J]. 吉林大学学报: 工学版, 2023, 53(10): 2932–2941.
Qu you, Li Wen-hui. Multiple object tracking method based on multi-task joint learning[J]. Journal of Jilin University (Engineering and Technology Edition), 2023, 53(10): 2932–2941.
- [61] Liang C, Zhang Z, Zhou X, et al. Rethinking the competition between detection and reid in multiobject tracking[J]. IEEE Transactions on Image Processing, 2022, 31: 3182–3196.
- [62] Zhang Y, Wang C, Wang X, et al. FairMOT: on the fairness of detection and re-identification in multiple object tracking[J]. International Journal of Computer Vision, 2021, 129: 3069–3087.
- [63] Duan K, Bai S, Xie L, et al. Centernet: keypoint triplets for object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 2019: 6569–6578.
- [64] Liang T, Li B, Wang M, et al. A closer look at the joint training of object detection and re-identification in multi-object tracking[J]. IEEE Transactions on Image Processing, 2022, 32: 267–280.
- [65] Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: the clear mot metrics[J]. EURASIP Journal on Image and Video Processing, 2008, 2008: 1–10.
- [66] Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]//European Conference on Computer Vision, Amsterdam, Netherlands, 2016: 17–35.
- [67] Luiten J, Osep A, Dendorfer P, et al. HOTA: a higher order metric for evaluating multi-object tracking[J]. International Journal of Computer Vision, 2021, 129: 548–578.
- [68] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012: 3354–3361.
- [69] Leal-taixé L, Milan A, Reid I, et al. MOTchallenge 2015: Towards a benchmark for multi-target tracking [DB/OL]. [2015-04-08]. <https://arxiv.org/pdf/1504.01942.pdf>.
- [70] Milan A, Leal-taixé L, Reid I, et al. MOT16: a benchmark for multi-object tracking[DB/OL]. [2016-05-03]. <https://arxiv.org/pdf/1603.00831.pdf>.
- [71] Dendorfer P, Rezatofighi H, Milan A, et al. MOT20: a benchmark for multi object tracking in crowded scenes[DB/OL]. [2020-03-19]. <https://arxiv.org/pdf/2003.09003.pdf>.
- [72] Sun P, Cao J, Jiang Y, et al. Dancetrack: multi-object tracking in uniform appearance and diverse motion [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 20993–21002.