跨摄像头多目标跟踪方法综述

张 鹏[1)] 雷为民[1)] 赵新蕾[2)] 董力嘉[2)] 林兆楠[1)] 景庆阳[1)]

[1)](东北大学计算机科学与工程学院 沈阳 110167)

[2)](沈阳二一三电子科技有限公司 沈阳 110027)

摘 要 单摄像头目标跟踪将目标跟踪范围限定在单一摄像头视野中,难以满足复杂应用场景需求,跨摄像头多目标跟踪融合多个摄像头的信息实现多个摄像头之间的特征传递和轨迹关联,可以将跨摄像头之间的多个目标在多个监控区域下联合跟踪,对现实复杂场景实时监控具有重要意义,成为目标跟踪领域研究热点.本文介绍了跨摄像头多目标跟踪的基本概念,结合实际应用需求将跟踪模型分为3类:包括重叠视角、非重叠视角以及混合视角的跨摄像头多目标跟踪.详细对比分析了重叠视角跨摄像头多目标跟踪相关的网络流优化方法、单应性约束方法、强化学习方法、超图方法和Transformer方法;以及基于双阶段轨迹关联、单阶段轨迹关联的非重叠视角的跨摄像头多目标跟踪方法;并总结了混合视角的跨摄像头多目标跟踪方法,混合视角方法可以在重叠视角数据集和非重叠视角数据集都能使用并且算法性能和精度都能达到良好的平衡.对比了各类方法的优缺点及其适用场景;分析了目前跨摄像头多目标跟踪常用的数据集和评估标准;总结了跨摄像头多目标跟踪存在的问题,并对相关技术的发展趋势进行了展望.

关键词 跨摄像头;多目标跟踪;摄像头关联模型;重叠视角;非重叠视角;混合视角

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2024.00287

A Survey on Multi-Target Multi-Camera Tracking Methods

ZHANG Peng[1)] LEI Wei-Min[1)] ZHAO Xin-Lei[2)] DONG Li-Jia[2)] LIN Zhao-Nan[1)] JING Qing-Yang[1)]

[1)](*Department of Computer Science and Engineering, Northeastern University, Shenyang* 110167)

[2)](*Shenyang* 213 *Electronic Technology Co., Ltd, Shenyang* 110027)

Abstract Multi-Target Single Camera Tracking limits the target tracking range to the field of view of a single camera, which is difficult to meet the needs of complex application scenarios. Multi-Target Multi-Camera Tracking (MTMCT) fuses the information from multiple cameras to realize the feature transfer and trajectory correlation between multiple cameras, which can jointly track multiple targets across multiple cameras in multiple monitoring areas, which is of great significance to the real-time monitoring of real-time complex scenarios. It has become a research hotspot in the field of Multi-Object Tracking. This paper introduces the basic concepts of MTMCT and classifies the tracking models into three categories, including overlapping view, non-overlapping view and mixed view, and the MTMCT model with an overlapping view is analyzed and introduced in detail in five subclasses: network flow optimization-based method,

收稿日期:2023-02-27;在线发布日期:2023-12-05.本课题得到2022年辽宁省"揭榜挂帅"科技重大专项(公共交通车辆火灾报警与灭火关键技术的智能化系统)(2022JH1/10400025)、中央高校基本科研业务费专项资金资助项目(N2216010)、国家重点研发计划基金资助项目(2018YFB1702000)资助.**张 鹏**,博士研究生,中国计算机学会(CCF)会员,主要研究领域为多目标跟踪和人体行为分析. E-mail:wind82465@yeah.net. **雷为民**(通信作者),博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为视频增强、传输优化、视频语义压缩编码. E-mail:leiweimin@ise.neu.edu.cn. **赵新蕾**,硕士,主要研究领域为多目标跟踪和人体行为分析. **董力嘉**,硕士,主要研究领域为人体行为分析和人体动作质量评估. **林兆楠**,博士研究生,高级工程师,主要研究领域为计算机视觉和目标检测. **景庆阳**,博士研究生,主要研究领域为目标跟踪和视频语义压缩编码.

homography constraints method, reinforcement learning method, hypergraph method and transformer method. subclasses are analyzed and introduced in detail. The network flow optimization method is a special Bayesian network model to solve the MTMCT task by calculating probabilities from binary images detected by multiple cameras. Methods based on network flow optimization can cause the phenomenon of false targets to affect the detection results, so the homography constraint method appears. The homography constraint defaults to all targets being of the same height, and the detection results of all views are projected onto the same ground plane after adding the constraint, and the intersection point formed is the position of the target. With the continuous development of deep learning methods, methods based on reinforcement learning appear, which mainly use hierarchical combination models to track the target. Most of the reinforcement learning methods are offline, and the algorithm parameters are large and difficult to train. In order to solve this problem, hypergraph-based methods have emerged, which introduce weighted hypernet work into the MTMCT task, and the weighted hypernet work can reduce the number of parameters and redundancy by sharing parameters across layers. The current transformer-based method can be realized to achieve online tracking with good performance at the same time. MTMCT with non-overlapping viewpoints is divided into two types of two-stage trajectory association and single-stage trajectory association for detailed analysis and introduction. MTMCT based on two-stage trajectory association refers to outputting multi-target tracking trajectories within a single camera first, and then carrying out multi-camera trajectory associations; MTMCT based on single-stage trajectory association refers to directly considering all trajectories globally to be associated. Different from the overlapping and non-overlapping views, the mixed view MTMCT method can be used on both overlapping and non-overlapping view datasets with a good balance of algorithmic performance and accuracy. In addition, we also compare the advantages and disadvantages of overlapping view, non-overlapping view, and mixed view approaches and their applicability scenarios. Finally, this paper analyzes the commonly used datasets and evaluation indexes for MTMCT, and summarizes the problems of MTMCT. We also look forward to the future trends of MTMCT technology, such as more MTMCT datasets, end-to-end models, richer evaluation metrics, higher crowd density, visual Transformer, and lightweight models.

Keywords Multi-Target Multi-Camera Tracking; Multi-Object Tracking; camera association model; overlapping perspective; non-overlapping perspective; mixed perspective

1 引 言

多目标跟踪(Multi-Object Tracking,MOT)[1-5]一直是学术界的研究热点.其主要任务是通过分析给定的视频或图像序列中的外观信息或运动信息,同时对多个目标进行定位标记和轨迹预测,并记录特定目标的轨迹路径.跨摄像头多目标跟踪(Multi-Target Multi-Camera Tracking,MTMCT)作为多目标跟踪的一个扩展,其目的是在特定范围内(如校园、火车站、广场)对目标进行跟踪和重新识别(通常是行人或车辆).跨摄像头多目标跟踪旨在从多个摄像头采集的视频中确定每个人的位置.近年来,跨摄像头多目标跟踪无论在运算效率和检测精度上都获得了明显提升.随着跟踪算法逐步成熟,跨摄像头多目标跟踪技术已经应用到不同的领域,包括监控系统[6-10]、行为分析[11-12]、姿态估计[13-19]、自动驾驶[20-27]、无人机[28]、体育运动[29-30]、机器人[31-32]、人脸跟踪[33]和海面运动目标的跟踪与检测[34]等.

跨摄像头多目标跟踪依然存在单摄像头目标跟踪中的问题,比如运动目标之间频繁的遮挡、相似的外观、多目标间的相互影响等都会影响跟踪结果.并且在具体实施过程中,如检测、特征提取、亲和力估计和数据关联等都可能会影响跟踪性能.所以,

跨摄像头多目标跟踪任务仍面临很多挑战，包括运动目标的在线跟踪、目标运动方向变化、形态变化、尺度变化、光照变化、复杂环境、低分辨率图像、不可靠的检测、摄像头距离变化等.

现阶段跨摄像头多目标跟踪识别的运动目标主要是行人与车辆，本文更关注的是基于行人目标的跟踪. 跨摄像头多目标跟踪数据集主要可分为两种：多摄像头之间有重叠区域和摄像头间无重叠区域. 其中，重叠视角指多个摄像头间拍摄视角重叠的场景，由于监控角度、摄像头景深以及像素差异等原因，同一个目标在不同摄像头拍摄的视频图像会出现不同的目标大小、视角、清晰度和运动速度等问题，所以有重叠区域的多目标跟踪任务需要进行图像特征融合. 重叠视角指多个摄像头间拍摄视角无重叠的场景，非重叠视角存在视野盲区导致摄像头之间的轨迹关联相较于重叠视角会更困难. 目标离开一个摄像头视角后进入另一个摄像头视角时多个摄像头间的协同关联是解决摄像头无重叠区域目标跟踪任务的关键.

现有研究综述大多针对单摄像头多目标跟踪算法(Multi-Target Single Camera Tracking，MTSCT)[35]，文献[36]对单摄像头多目标跟踪的数据关联任务进行了总结和回顾，并根据概率、层次、交互式多模型、卡尔曼滤波、模糊关联和基于新技术的方法进行了分类. 文献[37]介绍了跨摄像头非重叠视角的行人跟踪数据集与行人重识别数据集，并总结了基于时空特征、连续实体关联(Continuous Entity Association，CEA)、长短期记忆网络(Long Short Term Memory Network，LSTM)的时空跟踪器的行人跟踪方法. 目前，跨摄像头多目标跟踪任务集中在重叠视角环境下和非重叠视角，没有将混合视角的跟踪算法进行系统的分析和总结. 针对该问题，本文将跨摄像头多目标跟踪任务按照场景进行分类并分别进行详细的综述，如图1所示，分为重叠视角、非重叠视角和混合视角.

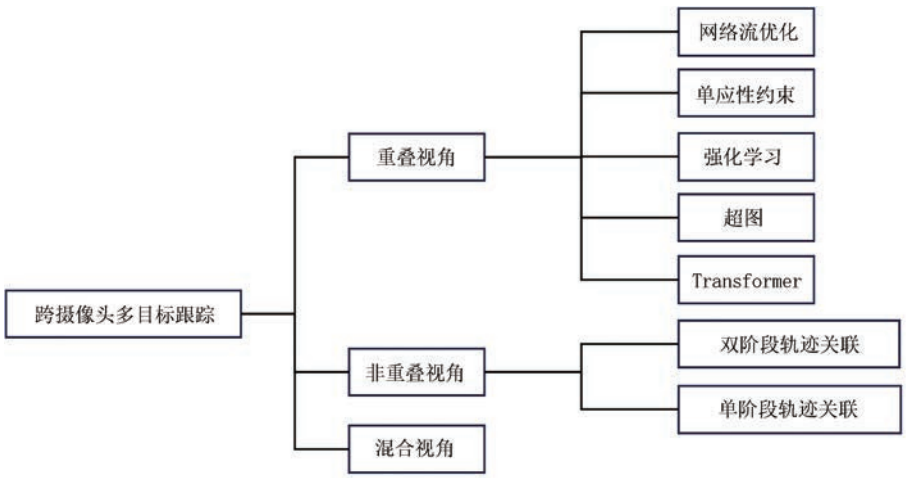


图1　跨摄像头多目标跟踪方法的分类

重叠视角的跨摄像头多目标跟踪方法主要适用场景为小区、校园、公园、交叉路口等公共场所，可以将行人高精度定位，常与人体姿态识别连用. 此类方法均为3D定位目标位置，实时性较强，已成为目前跨摄像头多目标跟踪的研究热点. 早期此类方法使用3D几何约束对2D轨道进行跨视图重建，简称概率图估计法，将重叠视野的多个摄像头利用视图之间的几何或外观一致性获得准确的3D定位[38]. 概率图估计法可以实时有效地跟踪目标状态，但在处理复杂跟踪场景时效果不佳. 早期算法采用网络流优化[39]、单应性约束[40]等方式解决数据关联问题. 近期算法为解决早期算法的不足，引用超图[41]、强化学习等多种方法解决跨摄像头的数据关联问题. 2017年随着Transformer[42]的提出，后续很多方法不再使用循环神经网络(Recurrent Neural Network，RNN)[43]或卷积神经网络，而是利用注意力机制获取全局信息解决跨摄像头多目标任务.

非重叠视角的跨摄像头多目标跟踪方法主要适用场景为智能视频监控[44]、智能安防、交通路口、人脸跟踪、现代安全、体育分析[45]和零售系统等复. 利用跨摄像头多目标跟踪算法对视频监控中的行人进行分析，获取监控中行人的运动轨迹. 跨摄像头多目标跟踪通常由目标检测、相似度估计以及数据关联组成. 通常重叠视角的跨摄像头多目标跟踪任务根据数据关联的方法分为两种类别：一类是基于双阶段轨迹关联的跨摄像头跟踪方法；另一类是基于

单阶段轨迹关联的跨摄像头跟踪方法.

混合视角的跨摄像头多目标跟踪(即非重叠与重叠均适用)主要应用在监控系统.混合视角指跨摄像头多目标跟踪算法既适用于重叠视角又适用于非重叠视角的情况,可使两者的跟踪性能达到平衡.此类方法为近期出现的研究热点,主要代表性算法为TRACTA[46]、DyGLIP[47]、LMGP[48]等.

本文从重叠、非重叠视角和混合视角讨论跨摄像头多目标跟踪的研究现状.按照数据集的类型将跨摄像头多目标跟踪分为三种类别进行阐述,主要介绍目前跨摄像头多目标检测常用的数据集和评估标准,并将不同方法在该数据集上的结果进行对比和分析.最后对跨摄像头多目标跟踪领域未来的发展方向进行总结和展望.

2 跨摄像头多目标跟踪

目前跨摄像头多目标跟踪研究主要有三个分支:第一个分支是重叠视角的跨摄像头多目标跟踪,主要适用于具有重叠视角的场景.多摄像头间具有重叠区域增加了行人的多角度外观和运动信息,为单摄像头下因单一视角出现的遮挡现象提供了解决方案.在现实生活中,具有重叠区域的摄像头会在同一时刻记录下同一目标的不同视角信息,但仍存在实时性差、遮挡、假目标等挑战.由于重叠视角的特殊性,多目标之间的特征信息互相干扰,极容易导致目标身份切换(ID Switch,ID)问题的出现.第二个分支是非重叠跨摄像头多目标跟踪,指多个摄像头记录的视频中不存在重叠的区域,也就是说同一个目标不会在同一时刻在不同的摄像头中出现,但目标在多摄像头之间移动的过程中存在视觉盲区.当目标在单帧错误的情况下后续的连续几帧内会一直出现误报,极大地增加了跨摄像头多目标跟踪的难度,所以区别目标之间相似性是解决该问题的关键.第三个分支是基于混合视角下的跨摄像头多目标跟踪,这种方式属于目前新兴模式,是指轨迹目标不仅在重叠视角的跨摄像头多目标跟踪方法中能有效跟踪,还能在非重叠视角的跨摄像头多目标跟踪方法中具有良好的性能.第三个分支能够平衡前两个分支的弊端.其特点是适用性强,运行速度快,能够基本满足实际工业领域的需求.

2.1 重叠视角的跨摄像头多目标跟踪

数据关联作为多目标跟踪中最重要的一环,其目的是在连续帧的多个目标中找到关联目标对的最优解,进一步形成多个目标的轨迹.因此数据关联可以看作是一种“约束优化”问题.本节在重叠视角的环境下,将跨摄像头多目标跟踪的数据关联方法分为5个子类,分别是网络流优化方法、单应性约束方法、强化学习方法、超图法和Transformer的方法.

2.1.1 网络流优化方法

网络流优化方法最早由Fleuret等人[38]提出.作者将跨摄像头多目标跟踪问题看作是一种对概率图的预测估计问题(Probabilistic Occupancy Map,POM).还提出一种轨迹优化算法应用于MTMCT任务中,可以实现在数千帧数的视频中准确跟踪.该方法的主要思想是通过不同视角运动检测生成的二值图像计算目标出现的概率,首先假定所有目标的高度是相同的,候选位置被投影回每个视角进行检测,将背景相减图像作为输入,并依靠平均场推断来计算候选位置在监控平面中的占用概率.为了实现对轨迹的全局优化,引入维特比算法动态匹配单条轨迹,并将多个轨迹的全局优化分解为最优单一轨迹.这种方式有效地降低了遮挡对跨摄像头多目标跟踪造成的干扰,使单个目标的轨迹在每个时间帧独立.但POM在重叠摄像头下处理速度较慢,很难满足实时性的要求.

2011年,Berclaz等人[39]在POM的基础上进行改进,通过在概率图中寻找K最短路径,并通过引入K最短路径算法[49](K-Shortest Paths,KSP)来解决凸问题.KSP[39]通过判断POM底层位置最可能的占位序列,将检测到的边界框按照时间顺序组织成有向的多条马尔可夫链,最后输出K最短路径(即找到了轨迹预测的最优解).该方法证明在不使用外观信息的情况下进行跟踪,也可以保证获得全局最优解,解决了跨摄像头多目标跟踪轨迹连接带来的优化困难的问题.KSP[39]提升了POM算法的速度,对实现在线跨摄像头多目标跟踪具有重要意义.

文献[38-39]均采用单纯的线性规划(Linear Programming,LP)[50]求解器求解的方式.此类方式会影响算法的速度同时也会导致结果出现非整数解.当摄像头的数量超过三个时,数据关联的轨迹之间会出现非确定性多项式时间困难(Nondeterministic Polynomially,NP)的问题.为了更好地解决线性分配导致的问题,Leal-Taixé等人[51]提出将数据关联建模为最小费用流问题,并使用二进制整数规划技术得到最优解.通过定义一个图结构,将跟踪目标看作图节点,构建网络流图.该方法通过寻找最小代价流来进行跟踪.图结构可以

同时捕捉帧间的时间相关性和空间相关性. 该方法通过使用多商品流(Multi-Commodity Flow)[52]保持全局外观约束. 通过使用丹齐格-沃尔夫分解(Dantzig-Wolfe decomposition)[53]减少计算时间并获得整数解.

2.1.2　单应性约束方法

单应性约束常被用于优化解决单摄像头多目标的跟踪结果. 原理是将所有视图的检测结果投影到同一地平面上以找到它们的交点,这些交点被视为行人的位置. 对于真实的监控场景中,目标之间的严重遮挡会导致检测出假目标.

为了解决假目标问题以及行人高度差异的影响,Huang等人[41]提出了一种基于多视角的贝叶斯网络模型(Multi-View Bayesian Network,MvBN),MvBN的网络结构如图2所示. MvBN模型首先利用现有的行人检测方法预定义行人的高度,得到一组初步的检测结果. 然后使用贝叶斯网络来建模每个摄像头视图中所有候选对象之间的遮挡关系,组合多个贝叶斯网络,通过从地平面到摄像头视图的单应性投影形成最终的MvBN. MvBN包括两类节点,分别代表行人候选点P-nodes和地平面上的位置点G-nodes,其中G-nodes用于组合来自不同贝叶斯网络的推断结果. 该方法在四个数据集上PETS09 S2L1[54], PETS09 City Center (CC), APIDIS[55]、EPFLTerrace[38]进行了实验. 因为MvBN初始版本[56]无法处理摄像头校准噪音和不同行人的身高,所以作者提出了一种高度自适应投影(Height Adaptive Projection,HAP)方法. HAP通过细化行人高度和位置邻域内的局部搜索过程优化检测结果.

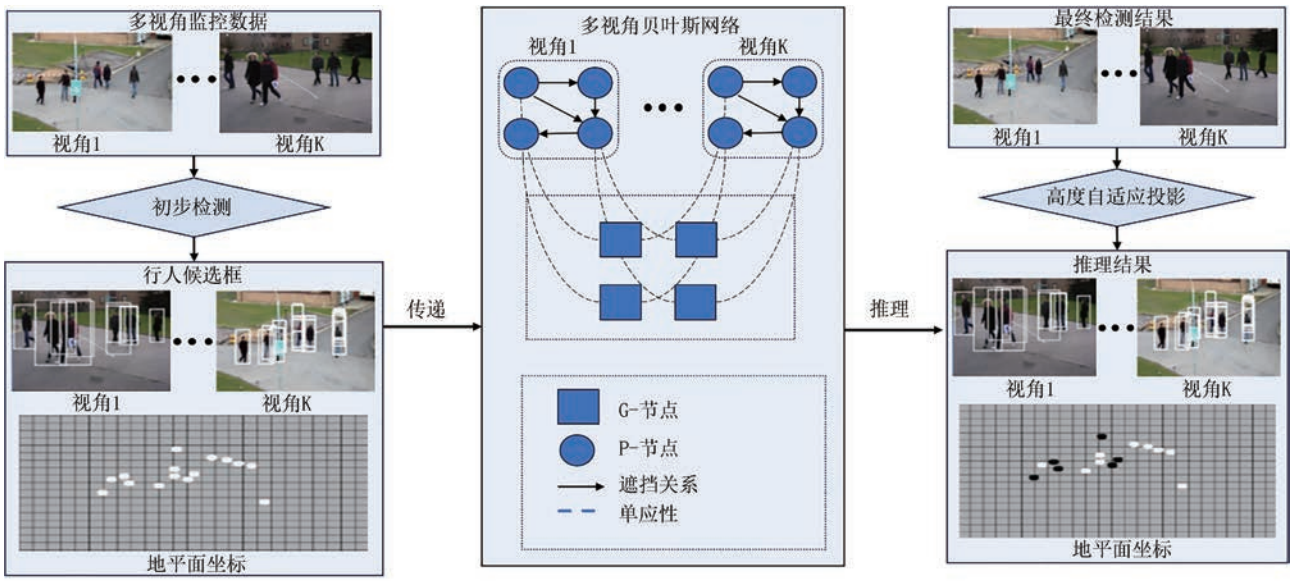


图2　MvBN[41]网络流程图

现有的跨摄像头多目标跟踪通常融合多个视角的视觉线索来优于单摄像头多目标跟踪,然而,跨摄像头多目标跟踪的主要缺点是产生幻影,因为在摄像头视图和地平面之间的单应性中,一些关键信息可能会丢失.

2.1.3　强化学习方法

早期的网络流优化法依赖于背景减法而不是深度学习导致网络性能不佳. 为提升整体精度,在网络流优化的基础上,Xu等人[57]提出了一种层次组合模型,本文将此类算法归类于强化学习方法. 强化学习方法首先将目标轨迹表示为组合层次结构,并设置了一套组合标准. 每个标准对应一个特定的线索信息,将外观相似度、运动一致性、稀疏度、3D定位重合度等不同的线索用适当的调度方法组合起来,这种方式可以增加一个图节点与子节点之间的约束. 另外,基于外观信息,用深度卷积神经网络(Deep Convolution Neural Network, DCNN)[58]建模外观变化,而不是使用传统的特征描述方式度量外观差异,其次,使用极大似然估计法在标注数据上学习组合标准,自动发现对象轨迹的最佳组合层次结构,以便处理更广泛的跟踪场景,并且用迭代贪心追踪算法[59]有效地构造分层图实现组合调度,通过渐进式组合过程来逼近层次结构. 基于强化学习的方法将跨摄像头多目标跟踪任务看作层次结构优化

问题容易产生漏检，因为二维检测是彼此独立执行的，并且映射到地平面上再组合会涉及重投影错误并忽略遮挡情况.

基于网络流优化的方法仅用背景减法作为输入，导致信息量越来越少，Baqué等人[60]使用多摄像头的几何结构来解决歧义问题. 作者设计了一个端到端可训练的卷积神经网络(Convolutional Neural Network，CNN)和条件随机场(Conditional Random Fields，CRF)[61]联合模型：通过衡量CNN预测与模型生成预测之间的差异进行损失函数的设计，通过提出联合CNN/CRF模型的组合方式省略了目标检测中的非极大值抑制(Non-Maximum Suppression，NMS)操作有效地解决遮挡问题. 此外，通过输出行人在地平面上的存在概率，可以用基于流的方法将其连接到完整的轨迹，降低了轨迹连接的难度.

Chavdarova等人[62]提出第一个完整的基于深度学习的多摄像头人物检测器方法，用来研究重叠视角下的概率图估计问题，检测部分可以使用GoogleNet[63]、AlexNet[64]以及ImageNet[65]上进行预训练，将最后一个全连接层替换为带有两个输出单元的随机初始化层，再对结果进行微调，使用点对点技术组合单摄像头检测器提取信息. 该工作的另一个贡献是提出一个大型高清重叠视角数据集WILDTRACK[66]，为以后其他学者研究重叠视角下的跨摄像头多目标跟踪提供了数据基准.

文献[38-39，41，57，60，62]等方法都是基于离线跟踪的方式，早期重叠视角下的跨摄像头多目标跟踪任务大部分停留在离线跟踪. 随着技术的更新，对于实时性的需求越来越大，实时跨摄像头多目标跟踪研究成为新的研究热点. Lan等人[67]提出一种半在线的跨摄像头多目标跟踪方法，其结构如图3所示. 该方法将跟踪问题看作重新识别问题，重新识别现有轨迹的检测假设，并建模为一个多标签马尔可夫随机场(Multi-Label Markov Random Field，MLMRF). MLMRF通过优化定义现有轨迹显式地处理目标的长期遮挡和外观变化. MLMRF考虑了目标的外观、运动、时间以及视图约束，引入扩展α-expansion[68]求解马尔可夫随机场，获得显著的加速比. 这种方法能够跟踪不同外观的行人，在光照变化、亮度变化和外观特征差别大的同一行人也能达到良好的追踪效果. 以前的跨摄像头多目标跟踪任务侧重于离线设计. MLMRF可以在半在线的情况下运行工作，为在线检测打下基础.

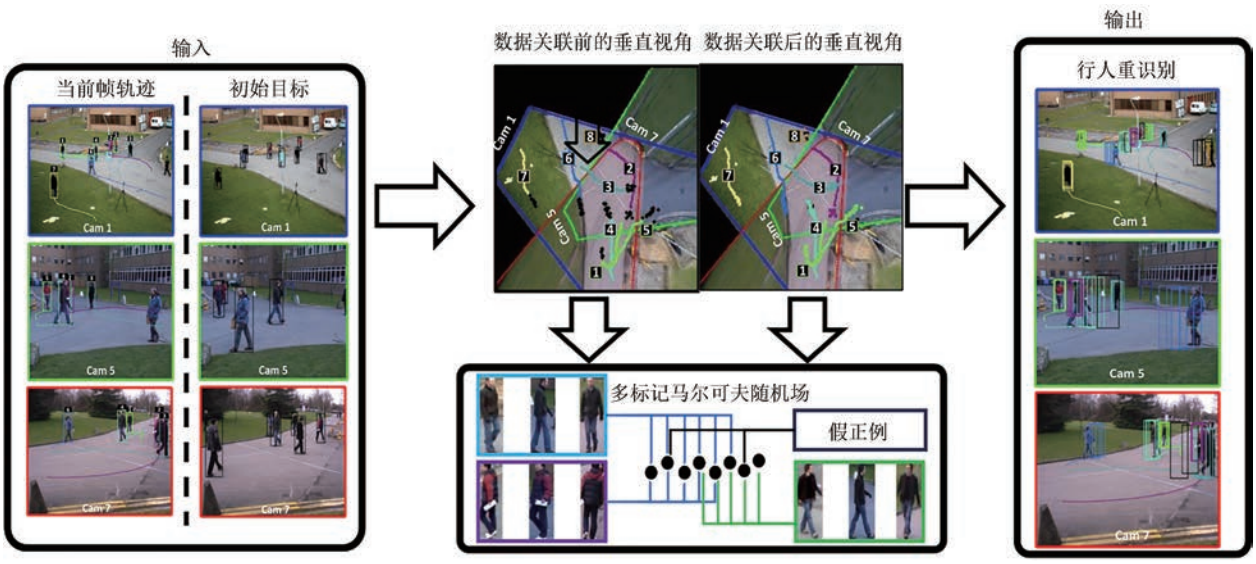


图3 MLMRF[67]工作流程图

2.1.4 超图方法

早期方法[38-39]单独解决跟踪和轨迹连接，没有利用这两个任务之间的相互引导关系. Wen等人[42]将加权超网络引入跨摄像头多目标跟踪，提出了时空视图超网络模型(Space-Time-View，STV). 首先在单摄像头下进行多目标跟踪，多个视图中的轨迹正确关联后，在3D环境中重建跟踪轨迹，最后利用采样的方法将跨摄像头多目标跟踪问题建模为在STV超图上搜索稠密的子超图，并进一步累加. 通过引入STV超图推断跨空间、时间和摄像头视角的轨迹点之间的高阶相关性增加了2D轨迹之间的高阶亲和性，同时验证了特征在3D几何、外观、运动连

续性和轨迹平滑方面的一致性.STV超网络的节点对应于2D轨道的潜在3D耦合，二维轨迹在形成耦合时的几何一致性用每个节点的权重进行编码.STV超图的超边及其相关权重反映了耦合之间的亲疏关系.然而，此类方法对跨摄像头视图的两个以上2D检测候选关联集之间的依赖关系无法有效地建模.

2.1.5 Transformer方法

2020年，Hou等人[69]提出了一种简单而有效的基于无锚框(anchor-free)的特征透视变换网络(Multi-View Detection，MVDet)用于跨摄像头多目标跟踪，MVDet可以进行端到端的训练.输入几张不同视角的3通道图像，通过ResNet[70]将特征提取后进行投影变换，将之前的3D视角的特征图投影到2D平面，并与坐标信息进行融合，得到俯视图(Bird's Eye View，BEV).BEV可以对整个场景的信息进行完整表述，将BEV特征叠加后再进行卷积得到行人位置预测以及检测结果.此类方法极大缓解了遮挡对跟踪系统的影响，但是MVDet在摄像头数量变少的情况下效果可能会变差，并且还是基于离线的，不能满足实时性要求.

针对跨摄像头多目标跟踪计算复杂度较高的问题，You等人[71]提出一种实时的跨摄像头3D多目标跟踪(Deep Multi-Camera Tracking，DMCT).DMCT设计了一个深度网络来估计每个目标在虚拟地平面上的投影——地面点，并将投影透视效应考虑进地面热图中，建立轻量化的深度掠影网络(Deep Glimpse Network，DGN)来捕获人体行为.DMCT可以同时处理视频中的多个帧，类似于人体关键点检测.DMCT可以在8个摄像头间实现每秒15帧的跟踪速度.Transformer的空间级联操作使不同的模态能够超越局部限制进行交互.为了处理各种图像失真问题，Hou等人[72]提出一种多尺度可变形的网络(Multi-View Detection Transformer，MVDeTr)来聚合多视角信息.MVDeTr[72]将多尺度可变形注意力扩展到跨摄像头多目标跟踪任务.首先将特征图的每个视图投影到地平面上，利用可变形注意力机制使模型在各个视角建立不同的关注点.但MVDeTr[72]没有解决由2D投影引起的空间结构断裂的问题.

为了解决空间断裂问题，Lee等人[73]提出一种新的多视图目标变换方法(Multi-View Target Transformation，MVTT).MVTT[73]通过限制行人特征空间大小使空间聚集在一个有限的感受野内.它首先利用单摄像头检测结果提取和编码每个行人的完整特征.其次，依靠提取的脚部特征来构建一个辅助特征图，对完整的目标特征进行编码并限制投影特征的感兴趣区域.最后定位行人的位置信息.MVTT[73]解决了多视图聚合中固有的失真问题.

2.2 非重叠视角的跨摄像头多目标跟踪

本文将非重叠视角的跨摄像头多目标跟踪任务分为2个子类：2.2.1节基于双阶段轨迹关联的跨摄像头跟踪方法，指先输出单摄像头内的多目标跟踪轨迹，再进行跨摄像头之间的轨迹关联；2.2.2节基于单阶段轨迹关联的跨摄像头跟踪方法，指直接全局考虑所有的轨迹进行关联.基于单阶段轨迹关联的跨摄像头跟踪法不需要得到全部的轨迹片段，就可以直接在各个帧进行关联检测.

2.2.1 基于双阶段轨迹关联的跨摄像头跟踪方法

由于单摄像头多目标跟踪任务已经非常成熟，很多方法可以实现在线跟踪.但是跨摄像头之间由于不存在重叠视角，导致同一目标不能同时出现在不同摄像头内.当单摄像头的轨迹预测错误时，后续的数据关联结果也会受到很大的影响.算法的精度表现不佳，需要借助行人重识别技术来区分不同行人之间的外观信息.如何将单摄像头预测的结果与新的检测轨迹之间进行关联并避免身份切换是解决跨摄像头多目标跟踪问题的关键.

2010年，Kuo等人[74]提出了一个在线学习的判别外观亲和力模型进行关联轨迹.该模型方法采用多示例学习(Multiple Instance Learning，MIL)适应模型学习过程中标记的模糊性，再将学习到的判别外观亲和力模型与时空信息相结合，最后计算轨迹关联的亲和度.模型架构图如图4所示，通过观察时间滑动窗口内的时空约束来收集在线训练样本，两个摄像头中在时间上重叠的轨迹对被认为是负样本.外观相似度计算是根据两个来自不同摄像头的轨迹的外观描述和相似性度量来判断它们是否属于同一个目标，其中，外观描述由颜色直方图、协方差矩阵和方向梯度直方图特征组成.通过MIL算法学习每两个轨迹对中的亲和性，最后分类器输出的预测置信度.此类方法的优势是能很好地区分在非重叠视角下不同的目标，虽然模型的精度较低，但是模型提出的框架经典且可借鉴性强，为后续研究奠定基础.

Cai等人[75]提出了一种利用时空上下文收集正负训练样本的跨摄像头多目标跟踪方法.模型架构

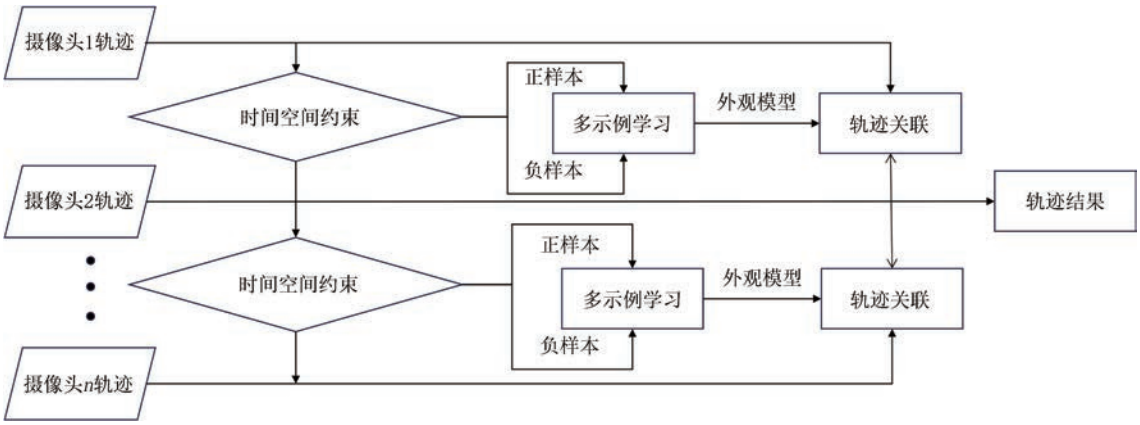


图4 基于多示例学习的跨摄像头多目标跟踪方法

图如图5所示。在重叠视角场景下，由于单一目标不能在同一时间出现在不同的摄像头下，利用这个时空上下文信息来收集正反例样本，进一步做区分性外观学习. 引入特定于目标的判别外观模型来区分不同的目标. 相关性外观上下文建模了行人间距较近时的外观相似性. 由于人群的聚集性，相同的一组人往往会在相邻的摄像头中重现，所以提出邻近集的概念. 将距离较近的人群按组划分，利用群体信息为个体的外观匹配提供了重要的视觉背景，查询两个摄像头下每个邻近集对象之间的外观相似性. 区分性外观学习可以对摄像头间匹配的个体外观进行消歧. 此类方法能从人群中区分视觉上非常相似的目标. 支持任何其他颜色、纹理和形状描述符插入到框架中，此类方法在NLPR_MCT[76]数据集上效果良好.

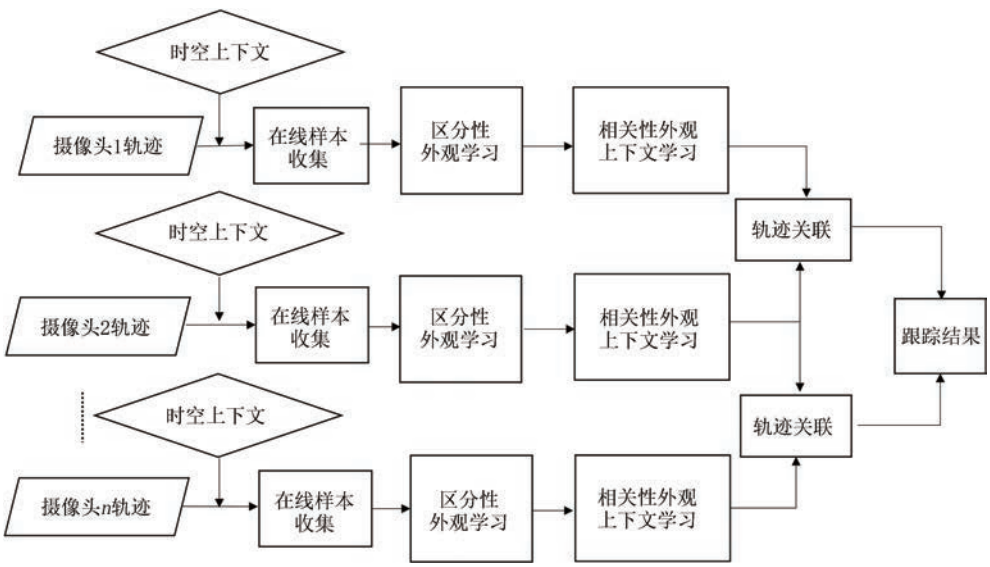


图5 基于时空上下文[75]的跨摄像头多目标跟踪方法

Zhang等人[77]在Duke MTMC数据集[78]上引入重排序和层次聚类实现跨摄像头多目标跟踪. 通过使用Faster R-CNN[79]作为检测器，使用行人重新识别模型来提取外观特征. 在单摄像头下，通过贪婪算法(Kuhn-Munkras，KM)[80]将相邻帧的边界框合并为小轨迹；在跨摄像头间，使用层次聚类将小轨迹合并为轨迹，无需更新距离矩阵. 层次聚类首先计算距离矩阵，轨迹与小轨迹之间的距离包括三个部分：外观相似性距离、分离部分的距离和重叠部分的距离. 将所有轨迹都放到所有摄像头中，对于每条轨迹进行重排序，根据从小到大的距离合并轨迹，直到最小距离达到阈值. 此类方法适用于长期跟踪，对重识别(Re-identification，ReID)特征进行平均得到了更鲁棒的特征.

Lee等人[81]提出了一种完全无监督的在线学习方法. 该方法将基于上下文的外观特征与外观线索结合起来，有效地集成了判别性视觉特征和上下文特征，并结合先进的检测算法和多核反馈来保持精

确的前景分割．在第一阶段使用耦合特征，在第二阶段使用整体颜色、区域颜色/纹理特征，由双向高斯混合模型拟合组成两阶段的特征提取器，可以有效且鲁棒地识别在不同摄像头下的同一个行人．上下文信息被表示为耦合特征，表示一对行人，将耦合特征与摄像头模型的融合特征权重进行融合，达到三个姿态不变颜色特征的有效融合．

2018年，Jiang等人[82]提出一种基于行人重识别和摄像头拓扑估计的跨摄像头多目标跟踪算法．该方法检测器使用Mask-RCNN[83]，首先生成单个摄像头中的轨迹，再利用行人重识别（Orientation-Driven Person Re-identification，ODPR）算法获得轨迹中每幅图像的外观特征．根据估计的摄像头拓扑结构，剔除不满足时空约束的冗余轨迹．对于满足传输时间约束的轨迹，计算轨迹相似性以实现摄像头间轨迹关联．通过设计方向驱动的损失函数和方向感知权重来缓解行人方向变化．这种方式可以缩小检索范围，提高了时间效率，为大规模监控环境下的智能摄像头间轨迹关联提供了可能，同时，结合全局特征和稳定躯干特征生成外观特征，有利于提高判别性特征表示．

Li等人[84]将基于状态感知的重识别功能引入跨摄像头多目标跟踪．单摄像头下，将遮挡状态、方位信息以及人体姿态信息都应用到ReID模型中．其中，只对稳定不被遮挡的目标进行ReID特征提取，遮挡的目标特征则会被舍弃．多摄像头下利用距离矩阵和贪婪算法实现跟踪，数据关联包括轨迹修正和轨迹聚类．采用同时出现的轨迹片段不能关联、目标的位移不能超过最大速度、目标不能消失很长时间等三种物理约束．轨迹修正可以让目标遮挡后回归跟踪．实时跟踪通常会造成轨迹的片段化，离线的跟踪会对片段化的轨迹做一个关联来生成最后的轨迹．轨迹聚类采用与文献[77]类似的策略，用来解决轨迹片段化的问题，让所有的轨迹片段之间建立联系．与文献[77]不同的是，当一个轨迹与其他轨迹相关联时，会更新距离相应的行和列．

2.2.2 基于单阶段轨迹关联的跨摄像头跟踪方法

Liu等人[85]提出基于全局信息关联的外观特征检测方法．该方法使用局部最大表示特征算法（Local Maximal Occurrence Representation，LOMO）[86]提取外观和动态运动的相似性．针对动态信息，建立每个轨迹汉克尔矩阵，并用迭代汉克尔矩阵最小方差算法[87]（Iterative Hankel Total Least Squares，IHTLS）对其进行秩估计，后将两个特征结合起来为图形提供边权重．在重叠视角的数据集EPFL Terrace[38]和Duke MTMC[78]，此方法仍然有改善空间．跟踪器上可以再包含更丰富的运动信息，数据关联部分在改进了文献[88]的基础上，提出全局最大优化算法（Global Maximum Clique Optimization，GMMCP），该算法是基于每两个低级别轨迹之间的边权重中找到最优解．边权重代表两个轨迹片段之间的相似性得分．

Tesfaye等人提出一个统一的三层框架算法MTMC_CDSC[89]来解决跨摄像头多目标跟踪任务．在前两层中，每个摄像头内均生成轨迹解决摄像头内跟踪问题，在第三层中，以同步的方式将同一个人在所有摄像头上的所有轨迹关联起来，以解决跨摄像头跟踪问题．引入约束支配集聚类技术与标准二次优化的参数化版本将多跨摄像头多目标跟踪建模为从一个图中找到受约束的支配集．通过减少搜索空间来进一步加快优化速度．通过提出的结合时空信息的表观模型改进了跨摄像头轨迹关联，解决了跨摄像头的目标交接问题．在数据集Duke MTMC[78]和MARS[90]上测试的结果表明行人重识别在非重叠摄像头中的多目标跟踪方法有效应用．

目前ReID好的效果往往依赖于深度学习、数据增强和特殊的损失函数等．但是ReID网络效果好不能说明跨摄像头多目标跟踪任务的精度高．基于此，Ristani等人提出一种提取特征的难例挖掘方法DeepCC[91]，将OpenPose[92]作为特征提取器同时提取运动特征和外观特征．这些特征将被轮流转换成相关性，并且会用相关聚类优化的方式打上标签，选择相关性聚类来进行数据关联，最后再对漏检插值和去除低置信度的轨迹做后处理．跨摄像头之间采取时间约束的方法，排除不可能的关联，通过联合外观相关性和运动相关性产生相关性矩阵，这种方法无需通过联合优化的方式测量轨迹质量，也可以让训练变得更加简单，减少模型代价．

跟踪的局部匹配过程与ReID外观特征的全局性质之间的不匹配可能会影响跨摄像头多目标跟踪的性能．Hou等人[93]提出了一种新的ReID训练方法，用计算相似度的三层感知机代替欧氏距离，设计一种基于局域感知的外观度量方式（Locality Aware Appearance Metric，LAAM），分为单摄像头内度量和跨摄像头间度量两种方式．前者是在同一台摄像头内某一时间段的轨迹学习，后者通过相邻摄像头间进行轨迹学习（目标可能连续出现）．LAAM[93]可以应用于全局学习的ReID中．单摄像

头内度量以及摄像头间度量用不同的指标，而且从各自采样的数据中进行训练，而不是在ReID透视图中进行全局采样．此类方法认为局部度量学习更适合跨摄像头多目标跟踪任务，具有较强的决策边界和敏感性．

Kwangjin等人[94]应用了多假设跟踪算法(Multiple Hypothesis Tracking，MHT)[95]来处理基于非重叠视角的跨摄像头多目标跟踪．首先设计一个跟踪假设树，分支代表一个目标在跨摄像头中的轨迹，目标是可以随时移动的．再通过跟踪摄像头内的多个目标来产生观测，轨迹假设树中的每个节点都指定了特定的观测，并且所有叶子节点都有一个状态，其中，节点描述观测与现有轨迹假设之间关联，最后将获得的观测值组成轨迹假设树．通过操纵每个跟踪假设的状态，同步产生轨迹假设树．此外，还提出了速度门控和时间门控，以处理不同的跟踪场景．对于观测的外观特征，为了捕捉姿态变化，将姿态分析应用于人的图像块后，采用简单平均的颜色直方图作为外观模型．此类MHT方法可实现在线和实时运行．

2.3 混合视角的跨摄像头多目标跟踪

目前跨摄像头多目标跟踪领域的算法数据集种类繁多，但是缺乏规律性，各类研究形式繁杂，缺乏系统的总结和归纳．近期对于摄像头多目标跟踪任务越来越多的研究不是偏向于一类重叠视角或者另一类非重叠视角，而是在重叠视角数据集和非重叠视角数据集都能使用并且算法性能和精度都能达到良好的平衡．因此本文将此类算法归类为基于混合视角的跨摄像头多目标跟踪．

He等人[46]将跨摄像头间的数据关联部分设计为局部轨迹与目标直接关联匹配(Tracklet-to-Target Assignment，TRACTA)的方式．这种关联方法使其既适用于重叠视角的视频场景又适用于非重叠视角的视频场景．设计了使用受限非负矩阵分解算法来建立最佳分配矩阵，用以整合来自所有局部轨迹集的信息．由于轨迹与目标直接关联需要依赖目标数目，又提出估计算法来估计整个网络中的目标数量．TRACTA以统一的方式跟踪多个目标，可以校正由局部轨迹集中存在的遮挡和漏检而引起的跟踪错误，并为跨摄像头中的每个目标生成完整、准确的全局轨迹．

Quach等人[47]首次将链路预测和动态图的概念引入跨摄像头多目标跟踪．将跨摄像头之间的全局轨迹看作动态图中的连接分配问题(Dynamic Graph Model with Link Prediction，DyGLIP)，并将动态图与注意力机制相结合．自注意力模块可以捕捉到跨多机位和多时间步骤的结构和时间变化，通过加入自注意力模块进一步提高了动态图的鲁棒性．DyGLIP框架如图6所示，通过图节点与轨迹关联．动态图表示允许利用每个对象的特征表示和移动模式来改进目标分配，随着时间的推移，节点被陆续添加到图中．它的转换特征通过关注图中的现有节点来计算，再使用从图中嵌入的结构和时间注意力特征来预测连接使得DyGLIP在具有重叠和非重叠视角的跨摄像头数据集中均有良好的运行效果．DyGLIP不仅适用于跨摄像头行人跟踪，还在车辆跟踪数据集Cityflow[96]上达到很好的效果．虽然DyGLIP在目前跨摄像头的行人数据集上均有良好

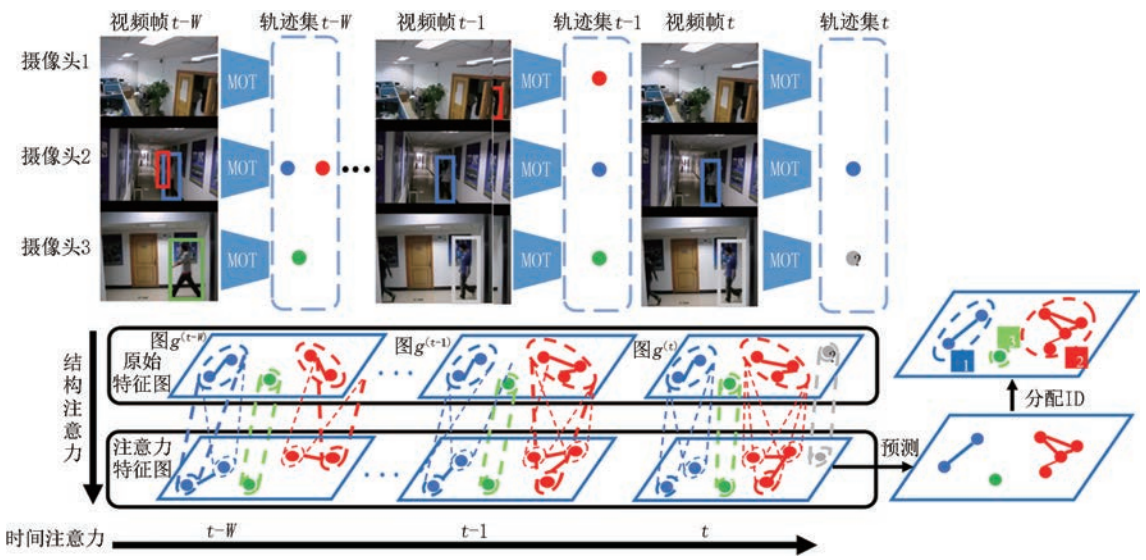


图6 DyGLIP[47]网络架构图

的表现，但在实际应用中，新轨迹生成还是会受到ID切换的影响，尤其在人群杂乱拥挤的场景下，所以基于混合视角的跨摄像头多目标跟踪还有很大的提升空间.

Nguyen等人[48]首次将基于时空提升的多切割公式(Lifted Multicut Meets Geometry Projections，LMGP)应用于跨摄像头多目标跟踪任务. LMGP[48]是一个提取全方位信息的跨摄像头多目标跟踪算法，基于单摄像头多目标跟踪任务生成检测目标轨迹，避免了大量的ID切换错误. 该方法通过引入了集中式表示范式中的概念，利用预聚类步骤来生成集中式表示思想. LMGP提出一种新的时空优化模型的数据关联方法，在预聚类步骤的占用图中出现ID切换错误时，将单摄像头多目标跟踪生成的初始轨迹分解，并为时间和空间亲和度建立精确的亲和度成本，在跨摄像头环境下有不俗的表现.

3　评估标准与数据集

本节将对不同评估任务的数据集以及不同方法在该数据集上的结果进行对比和分析.

3.1　评价指标

跨摄像头多目标跟踪数据集的评价指标与MOT任务的评价指标基本相同. 我们将常用的MOT指标应用于MOT Challenge[97]中的单摄像头多目标跟踪性能评估，包括文献[98]中提出的多目标跟踪精度(Multiple Object Tracking Precision，MOTP)和多目标跟踪精度(Multiple Object Tracking Accuracy，MOTA)；还有Ergys等人[99]提出的IDF1(Identification F-Score)、最大跟踪数(Mostly Tracked，MT)、最少丢失数(Mostly Lost，ML)等. 在跨摄像头多目标跟踪中选用文献[76]提出的跨摄像头跟踪准确度(Multi-cameraTracking Accuracy，MCTA)评估模型. 另外本文还将目标切换IDSW(ID Switch)[100]、多目标检测的准确性(Multiple Object Detection Accuracy，MODA)[101]、多目标检测精度(Multiple Object Detection Precision，MODP)[101]和高阶度量精度(Higher Order Tracking Accuracy，HOTA)[102]等评价指标进行介绍.

MOTA[98]：表示多目标跟踪的准确度. 通过计算跟踪所有帧中所有目标的误检、漏检和错误匹配得出. 其中，FN_t、FP_t和$IDSW_t$分别是t帧时漏检、误检和错误匹配的数量，g_t是地面真值目标矩形框的数量. 其计算公式：

$$\mathrm{MOTA}=1-\frac{\sum_t(FN_t+FP_t+IDSW_t)}{\sum_t g_t} \tag{1}$$

MOTP[98]：表示多目标跟踪的精度. 作用是量化检测器的定位精度. 其中，d_t^i代表第i个检测目标与给它分配的真值之间在所有帧中的平均度量距离，C_t代表在当前帧匹配成功的数目；其计算公式：

$$\mathrm{MOTP}=\frac{\sum_{i,t}d_t^i}{\sum_t C_t} \tag{2}$$

IDF1[99]指每个目标框中目标ID识别的F值. 其计算公式：

$$\mathrm{IDF1}=\frac{2}{\frac{1}{\mathrm{IDP}}+\frac{1}{\mathrm{IDR}}}=\frac{2\mathrm{IDTP}}{2\mathrm{IDTP}+\mathrm{IDFP}+\mathrm{IDFN}} \tag{3}$$

其中识别精确度(Identification Precision，IDP)是指每个行人框中行人ID识别的精确度，识别召回率(Identification Recall，IDR)是指每个行人框中行人目标ID识别的召回率. IDTP是真正ID数，IDFP是假正ID数，IDFN是假负ID数.

MT[99]：表示最多跟踪的目标数量(即跟踪目标在80%的时间以上都能够成功匹配的轨迹数量)；

ML[99]：表示最少丢失的目标数量(即跟踪目标在20%的时间以下能成功匹配的轨迹数量)；

IDSW[100]：目标身份切换的总次数(即ID改变的次数).

MCTA[76]：表示跨摄像头跟踪准确度，MCTA是衡量多个摄像头下跟踪的准确度，是目前少有的专门用来衡量多摄像头跟踪性能的评价指标. 其计算公式：

$$\mathrm{MCTA}=\mathrm{Detection}\times\mathrm{Tracking}^{\mathrm{SCT}}\times\mathrm{Tracking}^{\mathrm{ICT}} \tag{4}$$

其中Detection也就是$\mathrm{F}_{1\text{-score}}$；作用是统计漏检和错检，其计算公式：

$$F_{1\text{-score}}=\mathrm{Detection}=2\times\frac{P\times R}{P+R} \tag{5}$$

P和R分别代表精准率和召回率，其计算公式：

$$\mathrm{P}=1-\frac{\sum_t fp_t}{\sum_t r_t}\quad \mathrm{R}=1-\frac{\sum_t m_t}{\sum_t g_t} \tag{6}$$

其中，fp_t、r_t、m_t和g_t分别表示在第t帧时的错检数量、检测总数、漏检数量和标注总数，

$\mathrm{Tracking}^{\mathrm{SCT}}$表示单摄像头多目标跟踪能力，其计算公式：

$$\text{Tracking}^{\text{SCT}}=1-\frac{\sum_t mme_t^s}{\sum_t tp_t^s} \tag{7}$$

其中，mme_t^s表示单摄像头下所有的错误跟踪的总数，tp_t^s表示单摄像头下所有的匹配次数和；

$\text{Tracking}^{\text{ICT}}$表示跨摄像头多目标跟踪能力，其计算公式：

$$\text{Tracking}^{\text{ICT}}=1-\frac{\sum_t mme_t^c}{\sum_t tp_t^c} \tag{8}$$

其中，mme_t^c表示跨摄像头下所有的错误跟踪的总数，tp_t^c表示跨摄像头下所有的匹配次数和.

MODA[101]：MODA是指多目标检测的准确性，将漏检和误检的相对数纳入考虑范围. 其计算公式：

$$\text{MODA}(t)=1-\frac{C_m(m_t)+C_f(fp_t)}{N_G^t} \tag{9}$$

其中，m_t是第t帧里的漏检的数量，fp_t表示false positive，也就是假阳性的数量；$C_m()$、$C_f()$分别是自定义的漏检、FP的损失函数；N_G^t代表在第t帧里ground truth物体的数量.

MODP[101]：MODP是指多目标检测精度，其计算公式：

$$\text{MODP}(t)=1-\frac{\text{Overlap Ratio}}{N_{\text{mapped}}^t} \tag{10}$$

其中，OverlapRatio是目标检测中的交并比(Intersection over Union，IOU)，IOU是基于空间上衡量目标检测定位的准确性. IOU值反映了两个边框之间的重叠程度，它的计算方式是将两个边框的交集面积除以并集面积. 而N_{mapped}^t代表单帧检测出目标的数量，当N_{mapped}^t为0时，MODP也是0.

HOTA[102]：HOTA是指高阶度量，其计算公式：

$$\text{HOTA}_\alpha=\sqrt{\frac{\sum_{c\in\{\text{TP}\}}A(c)}{|\text{TP}|+|\text{FN}|+|\text{FP}|}} \tag{11}$$

其中α为定位阈值，c为某一正样本轨迹TP、TP、FN、FP分别代表只考虑检测框的正确预测样本、误检数量、漏检数量. 其中$A(c)$是指评价数据关联的准确率，其计算公式：

$$A(c)=\frac{|\text{TPA}(c)|}{|\text{TPA}(c)|+|\text{FNA}(c)|+|\text{FPA}(c)|} \tag{12}$$

对于TP集合中，TPA(c)是指预测的ID与检测框均为c的情况；$FNA(c)$是指在TP集合中，真实值为c但预测的ID不为c的情况，以及FN集合中，真实值为c的情况；FPA(c)是指在TP集合中，预测的ID为c但真实值不为c的情况，以及FP集合中，预测的ID为c的情况.

3.2 数据集

跨摄像头行人跟踪算法中常用的基准数据集可分为两类；一类为重叠视角的多摄像头行人数据集，包括PETS2009[54]、EPFL[38]、EPFL-RLC[57]、WILDTRACK[66]、MTA[103]、Multiview X[69]、MMPtrack[104]、GMVD[105]等；另一类非重叠视角的多摄像头行人数据集，此类方法多使用行人重识别相关数据集，包括Campus[58]、Duke MTMC[78]、NLPR_MCT[76]等. 此外还有其他属性的数据集，如跨摄像头车辆数据集CityFlowV1[96]、CityFlowV2[96]、足球场多视角跟踪数据集MVMPT[106]、虚拟跟踪数据集Waymo Open Dataset[107].

3.2.1 重叠视角的多摄像头行人数据集

PETS 2009 S2-L1[54]：PETS2009数据集包含S0、S1、S2、S3四个子集，S0为训练数据，S1为行人计数和密度估计，S2为行人跟踪，S3为流分析和事件识别；PETS2009通过3个监控摄像头和4个DV摄像头，摄像头的视野是部分重叠的，拍摄场景在英国雷丁大学校园内，总共记录10-50个行人. 视频分辨率为720×576，帧速率设置为7 fps. 在PETS 2009 S2子集中：低密度S2 L1序列记录19个行人，包含795帧图像，中等密度S2 L2序列记录43个行人，包含436帧图像，高密度S2 L3序列记录44个行人，包含240帧图像. 其中PETS 2009 S2 L1序列是跨摄像头多目标跟踪算法性能的广泛使用的基准数据集之一. PETS 2009 S2 L1包含7个室外摄像头的7个序列，每个序列由795帧组成. 数据主要来自于模拟数据，而非真实的监控场景.

EPFL[38]：EPFL数据集包含四个序列Terrace、Passageway、Campus和Basketball. 所有序列都有几个同步的视频流，以不同角度拍摄同一区域. 所以EPFL数据集的四个序列均为重叠视角. 所有摄像头都位于离地面约2 m的地方. 拍摄场景包括室内和室外. 每个序列由4个不同角度的DV摄像头拍摄6-11个行人步行或跑步，持续3.5分钟-6分钟. 每个视图以25 fps和较低的分辨率360×288进行拍摄. 对于每个视频序列，都提供了摄像头校准信息. 其中Terrace序列是在室外的露台上拍摄的. 多达7人在4个DV摄像头前散步，大约3.5分钟. EPFL数据集中最常被使用的就是Terrace序列. Passageway序列是在通往火车站的地下通道中拍

摄的．也由四个摄像头拍摄，由于视频光线不好，导致识别难度增大．Campus序列由三个摄像头在室外拍摄；Basketball由四个摄像头在室内篮球场拍摄．

EPFL-RLC Dataset：数据集在洛桑联邦理工大学内的公共图书馆使用三个经过校准的静态高清摄像头拍摄，摄像头的视野是重叠的．帧率为每秒60帧．每个摄像头的分辨率为1920×1080，相比EPFL数据集拍摄的视频更高清，引起近期算法广泛使用．测试集为真实标注的视频最后300帧数据．对于每个视角，无论它是否包含行人，都对负样本进行标注．这使得数据集可以用于单目行人训练．

WILDTRACK[66]：WILDTRACK是一个监控录像数据集，使用七个具有重叠视野的高科技静态定位摄像头拍摄苏黎世联邦理工大学主楼外的数千名学生．与EPFL、NLPR_MCT等数据集不同，这些视频是在“无剧本”、“非演员但真实的环境中”拍摄的．WILDTRACK数据集使用了三个GoProHero 4和四个GoPro Hero 3摄像头．覆盖12×36 m^2的区域．对于注释，地平面被量化为480×1440网格，其中每个网格单元是2.5 cm的正方形．7个摄像头以1080×1920的分辨率捕获图像，并以每秒2帧的速度进行注释．平均而言，WILDTRACK数据集中每帧有20个人，场景中的每个位置由3.74个摄像头覆盖．总共标记了313个行人，拍摄400帧．其中前360帧用于训练，其余49帧用于测试．WILDTRACK提供高清分辨率的数据并拍摄了室外场景．

MTA Dataset[103]：目前最大的重叠视角综合数据集．与其他重叠视角的行人数据集不同的是，MTA Dataset数据集中的数据并不是实景拍摄的，而是采用侠盗猎车手(Grand Theft Auto，GTA)视频游戏虚拟世界中城市场景．此数据集的数据更容易进行标注并消除了真实场景拍摄导致的隐私侵犯的风险；数据集包含了多样的天气状况、白天时间、室内和室外场景，MTA数据集记录了包含超过2800个行人目标，使用6个摄像头并且视频长度每个摄像头超过100分钟．视频总长度为10小时．并支持多种评估标准对其进行基线评估．

Multiview X[69]：是使用Unity引擎和PersonX的人体模型生成的合成大规模跨摄像头虚拟场景数据集，MultiviewX数据集覆盖的面积稍小，为16×25 m^2．使用2.5cm的网格将地平面量化为640×1000的网格．MultiviewX数据集中共有6个具有重叠视角的相机，每个相机图像分辨率为1080×1920．与WILDTRACK相同，MultiviewX中以每秒2帧的速度为400帧生成注释，但是行人的拥挤程度比WILDTRACK增加了一倍．平均有4.41台摄像头覆盖了同一地点．MultiviewX默认设置每帧有40个人，视频长度为60 min．

MMPtrack[104]：是大规模的密集标记的跨摄像头跟踪数据集，如图7(a)所示，MMPtrack录制模拟视频共计9.6小时，逐帧进行标注了超过50万个边界框；数据集包含零售、大堂、工业、咖啡厅和办公室等5个场景；其中，约5小时视频用作训练，1.5小时视频用作测试．帧率为15帧每秒．MMPtrack包含2D和3D的共同注释．假设一个大小为100 cm×100 cm×h的立方体，h代表人的高度，底部以每个跟踪的目标为中心．三维边界框投影到每个摄影机视图中．每个视图中的2D边界框是包围该视图中投影的3D边界框的最紧密的矩形．

GMVD[105]：是目前最大的跨摄像头行人合成数据集；使用Unity和侠盗猎车手(Grand Theft Auto，GTA)游戏环境来捕获WILDTRACK和MultiViewX均使用初始帧进行训练，后续帧用于测试，这种方法容易出现过拟合，因此，GMVD数据集不区分帧数而是选择区分场景；GMVD在7个不同场景中记录的约53个序列，包含一个室内场景(地铁)；还有六个室外场景，其中相机配置、天气、照明条件、行人外观等存在显著变化．GMVD数据集提供了白天变化(上午、下午、傍晚)和天气变化(晴天、多云、下雨、下雪)等多种类型．摄像头的数量也随场景而变化．使用6个场景进行训练，1个场景进行测试．

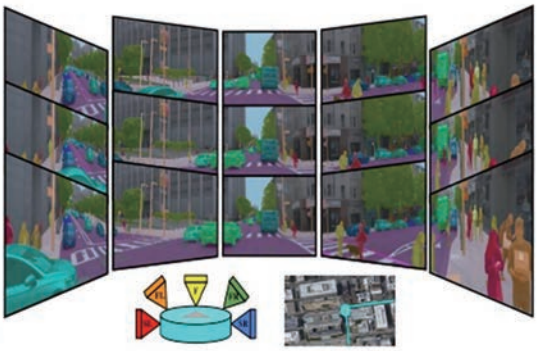
3.2.2　非重叠视角的多摄像头行人数据集

Duke MTMC[78]：该数据集是一个大规模标记的多目标多摄像头行人跟踪数据集，如图7(b)所示，数据集采集于杜克大学，由8个同步静态室外摄像头记录，所有目标均提供了人工标注和边框，图像分辨率可达到1080×1920；重叠视角的数据集指的是摄像头间严重重叠的情况，Duke MTMC只有两个摄像头稍微重叠，所以属于非重叠视角；视频长度85分钟，帧率为每秒60帧．这个数据集分为一个训练/验证集和两个测试集，简单测试集和困难测试集．训练集和验证集长50分钟，测试集总长35分钟．训练集含有702个人的16 522张样本，测试集则为17 661张样本，人均23.5张训练样本，是当前样本最丰富的数据集．

Campus[58]：该数据集中带有频繁的连接和遮

（b）Duke MTMC数据集[78]

（a）MMPtrack 数据集[104]



（c）WOD数据集[107]

图7 数据集

挡；融合复杂场景、动态背景以及各种对象尺度等多种特点而著名．包含四个序列Garden1、Garden2、Auditorium和ParkingLot，每个序列由安装在离地面1.5 m-2 m的3-4个高质量DV摄像头拍摄，每个摄像头覆盖重叠区域与非重叠区域．每个视频中有15-25个行人，由4个摄像头以30 fps的速度拍摄，视频长度约为3-4分钟，分辨率保持在1920×1080，以获得更好的精度和更丰富的信息．它有摄像头校准参数和地面平面三维投影．Garden1、Garden2拍摄的对象为室外公园内行人．

NLPR_MCT[76]：一共包含了四组多摄像头行人跟踪子数据库．四组数据图片尺寸均为320×240像素，每组子数据库包含了3到5个无重叠区域的摄像头，并且既包含了模拟数据，也包含了真实监控数据．行人总数最少有14个，最多有225个．该数据集既包含了室内场景也包含了室外场景．其中，Dateset1、Dateset2（第一组、第二组）包含3部非重叠的摄像头拍摄的20 min视频，帧率为20帧每秒；Dateset3、Dateset4（第三组、第四组）包含4-5部重叠的摄像头拍摄视频，帧率为25帧每秒．

跨摄像头行人跟踪常用的数据集及其属性如表1所列．

3.2.3 其他跨摄像头数据集

CityFlowV1[96]：发布于2019年人工智能城市挑战赛上，是一个城市规模的交通摄像头数据集，包含超过3小时的同步高清视频，来自10个路口的40个摄像头，同时两个摄像头之间的最长距离为2.5公里．CityFlow包含公路、住宅等多种真实场景．数据集包含将近23万个带注释的边界框，视频总长度大概3小时15分钟，标注了666辆车的跨摄像头轨迹．CityFlowV2是CityFlowV1的升级版．它记录了46个摄像头拍摄的880个目标．

MVMPT[106]：是一个大规模的球员跟踪数据集，视频来源于各种足球场景中的比赛实况．在几个足球场安装了多视角视频记录系统，以连续收集数据．该数据集采用1+N策略设计，由5×7个具有重叠视角的静态超高分辨率相机组成．总计17组2分钟的长视频，帧率为20帧每秒；总共1100分钟的多场景视频．

Waymo Open Dataset[107]：是首个将视频全景分割扩展到跨摄像头多目标跟踪的数据集，如图7(c)所示；Waymo Open Dataset（WOD）包含28个语义类别，2860个时间序列；还将车辆与人类进一步细分成多个子类，共包含1150个场景，增加了夜晚、雨天、密集城市等多种新场景；帧率为每秒10帧．数据集中的每个数据帧包括来自激光雷达设备的3D点云、来自五个相机（位于前、左前、右前、左侧面和右侧面）的图像，总共产生了10万张带标签的图像．每张图像都有全景

表1　跨摄像头多目标跟踪数据集

名称	场景	摄像头个数	年份	属性描述	目标	真实/虚拟	是否重叠	时长
Terrace [38]	Outdoor	4	2008	pedestrian	7	真实	Yes	3.5 min
Passageway [38]	Mixed	4	2011	pedestrian	4	真实	Yes	20 min
PETS09 S2-L1[54]	Outdoor	7	2009	pedestrian	19	真实	Yes	1 min
NLPR_MCT 1[76]	Mixed	3	2015	pedestrian	235	真实	No	20 min
NLPR_MCT 2[76]	Mixed	3	2015	pedestrian	255	真实	No	20 min
NLPR_MCT 3[76]	Indoor	4	2015	pedestrian	14	真实	Yes	4 min
NLPR_MCT 4[76]	Mixed	5	2015	pedestrian	49	真实	Yes	25 min
Duke MTMC[78]	Outdoor	8	2016	pedestrian	2834	真实	No	85 min
Campus[58]	Outdoor	4	2016	pedestrian	25	真实	Yes	4×4 min
EPFL-RLC [57]	Indoor	3	2017	pedestrian	-	真实	Yes	8000 frames
WILDTRACK [66]	Outdoor	7	2018	pedestrian	313	真实	Yes	~60 min
CityFlowV1[96]	Outdoor	40	2019	Car	666	真实	No	195 min
MTA [103]	Mixed	6	2020	pedestrian	2840	虚拟	Yes	102 min
MultiViewX [69]	Outdoor	6	2020	pedestrian	350	虚拟	Yes	~60 min
MMPtrack[104]	Indoor	4	2021	pedestrian	28	真实	Yes	576 min
CityFlowV2[96]	Outdoor	46	2021	Car	880	真实	No	215 min
WOD[107]	Outdoor	5	2022	Car，pedestrian	-	真实	No	50 s
GMVD[105]	Mixed	3,5,6,7,8	2023	pedestrian	2800	虚拟	Yes	-
MVMPT[106]	Outdoor	5,6,7	2023	Soccer player	316	真实	Yes	1100 min

分割标签，标注的真实标签包含3D和2D边界框.

3.3　实验效果评估对比

本文根据三种场景（重叠视角、非重叠视角、混合视角）将近年来跨摄像头多目标跟踪方法进行归纳分析，并在对应数据集上的结果进行比较，分析归纳跨摄像头多目标跟踪方法的实验效果.

表2总结了跨摄像头多目标跟踪算法在Terrace和PETS09 S2-L1数据集上的实验结果. 基于重叠视角下的数据集标注的行人目标一般都是3D定位的.

表2　基于重叠视角的跨摄像头多目标跟踪方法的指标评估结果

		数据集	Terrace[38]				PETS09 S2-L1[54]			
在线/离线	算法类别	评价指标	MOTA↑	MOTP↑	MODA↑	MODP↑	MOTA↑	MOTP↑	MODA↑	MODP↑
离线	网络流优化	POM[38]	58%	63%	19%	56%	-	-	65%	67%
离线	网络流优化	KSP[39]	67%	58%	80%		80%	57%	80%	61%
离线	单应性约束	MvBN[41]	-	-	82%	73%	-	-	87%	76%
离线	强化学习	HCT[58]	72%	71%	72%	69%	89%	73%	90%	72%
离线	强化学习	STP[108]	77%	79.6%	-	-	-	-	-	-
离线	超图	Wen, et al. [42]	-	-	-	-	95.08%	79.8%	-	-
半在线	强化学习	MLMRF[67]	-	-			96.8%	79.9%		

Terrace数据集为Fleuret等人于2008年提出，在Terrace数据集中，文献[38-39]属于早期传统的基于网络流优化的方式，MODA和MODP精度表现不佳. 基于单应性约束的方法对MODA和MODP两项指标做了很大的提升，例如文献[41]提出的遮挡关系有效去除了幻影（假目标）. 基于超图方法更集中于在PETS09 S2-L1数据集上进行测试，文献[42]提出的加权超网络将MOTA提升至95.08%. 表2中的重叠视角跨摄像头多目标跟踪方法[38-39,41,42,58,108]早期均为离线跟踪，只有文献[67]能够实现在线目标跟踪. 早期的方法都是在解决行人遮挡以及跨摄像头下的ID识别问题. 重叠视角下的行人的目标切换问题一直无法采取有效的措施进行解决. 文献[67]将跟踪任务建模为一个多标签马尔可夫随机场，这种方法使它在PETS09 S2-L1的精度达到最高，MOTA达到96.8%. MOTP为79.9%.

基于Transformer的方法因为研究较新，实验都

分布在WILDTRACK、Multiview X、MMPTrack数据集上. 所以本文将基于Transformer的重叠视角的跨摄像头多目标跟踪方法的指标评估结果进行对比分析，如表3所示. 其中，所有在线跟踪方法中MVDeTr[72]表现最佳，MODA可达到91.5%；MODP最高82.1%. 基于Transformer的方法为近期研究的主流，并且网络结构也更简单，具有更快的训练速度. 基于Transformer的方法DMCT[71]、MVDeTr[72]、UMMT[109]、UMPD[110]、MVFlow[111]都能够实现实时在线多目标跟踪. 但目前重叠视角下性能表现最好的方法MVTT[73]仍是基于离线跟踪的，在WILDTRACK数据集上，MODA可达到94.1%；比在线跟踪MVDeTr[72]的精度高2.6%. 在Multiview X数据集上MODP的精读达到95.0%.

表3 基于Transformer的重叠视角的跨摄像头多目标跟踪方法的指标评估结果

		数据集	WILDTRACK[66]				Multiview X[69]		MMPTrack[104]	
在线/离线	算法类别	评价指标	MOTA↑	MOTP↑	MODA↑	MODP↑	MODA↑	MODP↑	MOTA↑	IDF1↑
在线	Transformer	DMCT[71]	74.6%	78.9%-	-	-	-	-	88.8%	56.0%
离线	Transformer	MVDet[69]	-	-	88.2%	75.7%	83.9%	79.6%	-	-
在线	Transformer	MVDeTr[72]	-	-	91.5%	82.1%	93.7%	91.3%	-	-
在线	Transformer	UMMT[109]	95.2%	-	-	-	-	-	95.0%	84.3%
在线	Transformer	MVD[105]	-	-	80.1%	75.6%	70.7%	73.8%	-	-
在线	Transformer	UMPD[110]	-	-	76.2%	59.4%	24.2%	56.6%	-	-
在线	Transformer	MVFlow[111]	91.3%	57%	-	-	-	-	-	-
离线	Transformer	MVTT[73]	-	-	94.1%	81.3%	92.8%	95.0%	-	-

表4总结了近年内跨摄像头多目标跟踪算法在非重叠视角的数据集Duke MTMC的运行效果. 基于非重叠视角的跨摄像头多目标跟踪算法以Duke MTMC作为基准数据集. 其中，LAAM[93]的IDF1、IDP和IDR评估结果表现最优，IDF1可达到82.3%，IDP和IDR分别为87.4%和96.8%. Duke MTMC数据集的人物数量较多，测试视频平均有60个行人出现在4个非重叠的摄像头中. 目前大部分的非重叠视角的方法PT_BIPCC[112]、MTMC_CDSC[89]、Liu, et al.[85]、DeepCC[91]、Zhang, et al.[77]、Li, et al.[84]和LAAM[93]均为离线跟踪. 基于双阶段轨迹关联的MTMC_REID[77]使用了重新排序策略，MTMC_CDSC[89]则通过离线优化来恢复丢失的轨迹. 基于单阶段轨迹关联的DeepCC[91]将行人重识别(ReID)和全局运动相关性引入到跨摄像头多目标跟踪方法中. MTMC_REID[77]、BIPCC[113]、MyTracker[94]和Jiang et al.[82]等方法属于在线跟踪. 在线跟踪方法中表现最优的是Jiang, et al.[82]，IDF1可达到68.8%. 目前离线跟踪方法的精度值都普遍高于在线跟踪方法，并且离线跟踪大部分都属于单阶段轨迹关联.

表4 基于非重叠视角的跨摄像头多目标跟踪方法的指标评估结果

		数据集	Duke MTMC[78]		
在线/离线	算法类别	评价指标	IDF1↑	IDP↑	IDR↑
离线	单阶段轨迹关联	PT_BIPCC[112]	34.9%	41.6%	30.1%
在线	双阶段轨迹关联	BIPCC[113]	56.2%	67.0%	48.4%
离线	单阶段轨迹关联	MTMC_CDSC[89]	50.9%	63.2%	42.6%
离线	单阶段轨迹关联	Liu, et al.[85]	55.5%	78.9%	44.6%
离线	单阶段轨迹关联	DeepCC[91]	68.5%	75.9%	62.4%
在线	单阶段轨迹关联	MyTracker[94]	65.4%	71.1%	60.6%
在线	双阶段轨迹关联	Jiang, et al.[82]	68.8%	71.8%	66.0%
离线	双阶段轨迹关联	Zhang, et al.[77]	74.0%	81.4%	67.8%
离线	双阶段轨迹关联	Li, et al.[84]	81.3%	88.7%	75.1%
离线	单阶段轨迹关联	LAAM[93]	82.3%	87.4%	96.8%

表5总结了基于混合视角的跨摄像头多目标跟踪方法的指标评估结果；表5分别对比了重叠视角的PETS09 S2-L1数据集和非重叠视角的CAMPUS数据集中的Garden1序列和Parkinglot序列的评估结果；DyGLIP[47]和LMGP[48]在Parkinglot序列中ML甚至可以达到0，大大缩小了误差. 其中

LMGP[48]的精度最高，在重叠视角的PETS09 S2-L1数据集上MOTA达到97.8%，MOTP达到82.4%；在非重叠视角的CAMPUS数据集Garden1序列上MOTA达到76.9%，MOTP达到95.9%；远远超过表4中基于非重叠视角的跨摄像头多目标跟踪算法.

表5　基于混合视角的跨摄像头多目标跟踪方法的指标评估结果

	数据集	PETS09 S2-L1[54]		Garden1[58]				Parkinglot[58]			
在线/离线	评价指标	MOTA↑	MOTP↑	MOTA↑	MOTP↑	MT↑	ML↓	MOTA↑	MOTP↑	MT↑	ML↓
离线	TRACTA[46]	87.5%	79.2%	58.5%	74.3%	30.6%	1.6%	39.4%	74.9%	15.5%	10.3%
离线	DyGLIP[47]	93.5%	94.7%	71.2%	91.6%	31.3%	0	72.8%	98.6%	26.7%	0
离线	LMGP[48]	97.8%	82.4%	76.9%	95.9%	62.9%	0	78.1%	97.3%	62.1	0

基于混合视角的算法目前的研究较少，TRACTA[46]、DyGLIP[47]和LMGP[48]都属于离线跟踪.基于混合视角的算法TRACTA[46]和DyGLIP[47]的结果明显超越了基于重叠视角和基于非重叠视角的方法.基于离线的LMGP[48]方法虽然在精度上表现最佳但是无法满足实时运行的要求，还有进一步优化的空间.

表6总结了近年内跨摄像头多目标跟踪在NLPR_MCT数据集四个序列的运行效果.文献[114]采用匈牙利算法对轨迹进行数据关联.文献[115]属于单阶段轨迹关联方法.基于混合视角的跨摄像头多目标跟踪算法TRACTA[46]和DyGLIP[47]的精度较高，在四个序列均有着很高的精度.其中，Dataset3中的跨摄像头行人目标数量偏多，大大增加了识别难度，是NLPR_MCT数据集最难识别的序列.早期算法均无法解决跨摄像头中行人在来回迂回循环往复地行进这种特殊情况，常会出现大量的身份切换.DyGLIP[47]通过图节点与轨迹关联并加入自注意力模块解决了Dataset3中人物频繁切换导致的IDSW问题，MCTA达到89.4%，MOTA达到92.7%，后续基于混合视角的跨摄像头多目标跟踪将成为研究重点.

表6　基于NLPR_MCT行人数据集的跨摄像头多目标跟踪方法的指标评估结果

数据集	Dataset1		Dataset2		Dataset3		Dataset4	
评价指标	MCTA↑	MOTA↑	MCTA↑	MOTA↑	MCTA↑	MOTA↑	MCTA↑	MOTA↑
HFUTDSP[114]	28.1%	57.6%	28.2%	54.7%	3.6%	21.1%	6.1%	28.1%
CRIPAC_MCT[115]	12.5%	70.7%	10.8%	74.6%	1.1%	8.6%	2.1%	27.1%
NLPR[76]	41.2%	59.4%	47.9%	67.2%	18.6%	27%	28.4%	35.8%
USC_VISION[75]	59.5%	92.6%	62.6%	86.8%	5.6%	9.2%	34%	53.9%
ICLM[81]	61.2%	87.3%	67.7%	88.3%	37.2%	53.2%	54.3%	62.5%
TRACTA[46]	70.8%	94.9%	83.7%	93.4%	53.8%	58.5%	71.5%	79.6%
DyGLIP[47]	76.2%	86.7%	91.9%	95.7%	89.4%	92.7%	84.7%	92.5%

4　跨摄像头多目标跟踪发展趋势

跨摄像头多目标跟踪算法是多目标跟踪的一个新的分支，经过多年研究技术日趋成熟，其性能在重叠摄像头与非重叠摄像头数据集上效果日趋增加，但跨摄像头多目标跟踪任务仍然面临诸多挑战，未来的研究可能有以下趋势.

(1) 更多的跨摄像头跟踪数据集

在单摄像头跟踪任务，基准数据集MOT16[116]起着重要作用.虽然MOT16在跨摄像头跟踪领域也有测试数据集，但目前阶段针对跨摄像头领域数据集的规模还不够.主要表现在：标注的数据集很多都是基于模拟场景，不能够模拟真实场景中的天气、温度、白天、夜晚等差异；同一个人在一段时间内同时出现在多个摄像头下、同一摄像头下相似的外观、严重的遮挡等条件限制了行人数据集的构建.与此同时车辆数据集的发展较快，因此，未来应开发出多类别、多属性、多角度的真实场景拍摄的监控数据集.并且数据集应同时包括2D与3D定位，有利于支持各种类型的算法应用.

(2) 普适性更强的端到端的模型

目前，研究人员较少设计端到端模型进行跨摄像头跟踪.已有的端到端的跨摄像头多目标跟踪模型均为3D定位的方式，仅适用于重叠视角的摄像头，对于非重叠、混合视角的跨摄像头目标跟踪的算

法步骤繁琐.通常完整的跟踪模型包括目标检测、相似度估计以及数据关联3个阶段.此类模式过度依赖上游目标检测任务的结果,在实际监控场景中,各种错检、漏检情况都会存在.因此设计一个端到端的通用模型来降低对目标检测的依赖性将是未来的研究重点,也是不小的挑战.

(3)更丰富的评价指标

多目标跟踪任务已经有很成熟的模型评估指标,包括多目标跟踪精 度和多目标跟踪精度以及IDF1、MT、ML等.其中IDF1也常用在非重叠视场的跨摄像头多目标跟踪、行人重识别任务中,文献[76]提出的跨摄像头评价指标MCTA是目前针对多摄像头间的有效的衡量基准,但只支持在NLPR_MCT[76]数据集.同一摄像头内与不同摄像头之间的目标定位精度与检测精度都应该做好更有效的区分,HOTA虽然能够实现更好的定位精度,但是仍然没有做正负样本的区分.后续应当更加细化评价指标,使其能在多种算法中灵活应用.

(4)更高的人群密度

较高的人群密度对于目标检测、特征提取和数据关联与跟踪都是更高的挑战.目前跨摄像头跟踪领域研究的目标数量最多的也只有2800个目标,人群分布都比较稀疏,然而实际应用场景,例如火车站、机场、商场等场所的监控系统中人群分布广并且非常密集,行人活动存在大量不确定性.未来应建立更多关联实际场景下密集人群的多目标跟踪与定位.

(5)视觉Transformer

随着ViT(Vision Transformer,ViT)[117]的提出,Transformer被引入到了计算机视觉中.Transformer与CNN相比,可以对上下文有更好的理解,有着更强的全局感知能力.处理时序关系较强的数据时,Transformer可以处理更长的时间维度的数据,增强了模型的时序关系.Transformer可以构建更具描述性的视觉表达,或者用Transformer来代替CNN来提取特征,并结合混合视角开发出满足实际需求的网络模型.目前基于Transformer的跨摄像头多目标跟踪任务已经可以实现在线跟踪,但是模型都需要较多的预训练过程,对数据要求较为苛刻,后续可以针对该问题进行更深入的研究.

(6)轻量化模型

近些年的目标跟踪算法的研究主要集中在模型的精度上,但是部分模型会增加模型的资源占用,实用性较差.因此,研究轻量化的目标跟踪模型与算法非常必要,后续可以采用轻量化模型、剪枝、知识蒸馏等方式降低网络复杂度,从而减少模型计算量,提升推理速度,进一步提升模型的泛化能力.

5 总 结

本文对跨摄像头多目标跟踪任务提出了新的分类方式,将基于重叠视角、基于非重叠视角和混合视角的三种方法进行了系统的分析.对于重叠视角,本文将网络流优化、单应性约束、超图、强化学习等数据关联的方式实现重叠视角的跨摄像头多目标跟踪任务,并总结其实现方式.对于非重叠视角环境下,按照数据关联的步骤进行分类,包括双阶段轨迹关联和单阶段轨迹关联的两种方式.对于提出的混合视角,本文将适用于该场景的算法也进行了详细的介绍与分析,并进一步介绍了相关数据集以及评价指标,将跨摄像头多目标跟踪在不同数据集的效果进行比较并总结;最后,进一步分析了跨摄像头多目标跟踪任务的现状以及存在的问题,并讨论了该领域未来的发展方向,希望能给后续的研究工作提供有价值的帮助.

参 考 文 献

[1] Weng X, Wang Y, Man Y, et al. Gnn3dmot: graph neural network for 3d multi-object tracking with 2d-3d multi-feature learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 6499-6508

[2] Zhang Y, Wang C, Wang X, 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

[3] Wan X, Wang J, Kong Z, et al. Multi-object tracking using online metric learning with long short-term memory// Proceedings of the 25th IEEE Int. Conf. Image Process. Athens, Greece, 2018: 788-792

[4] Wu J, Cao J, Song J, et al. Track to detect and segment: an online multi-object tracker//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 12352-12361

[5] Zhou X, Koltun V, PhilippKrhenbühl. Tracking objects as points//Proceedings of the European Conference on Computer Vision. Virtual, 2020: 474-490

[6] Fuentes L M, Velastin S A. People tracking insurveillance applications. Image and Vision Computing, 2006, 24(11): 1165-1171

[7] Huang C M, Fu L C. Multitarget visual tracking based effective

surveillance with cooperation of multiple active cameras. IEEE Transactions on Systems Man and Cybernetics Part B，2011，41(1)：234-247

[8] Nascimento J ，Marques J. Performance evaluation of object detection algorithms for video surveillance. IEEE Transactions on Multimedia，2006，8(4)：761-774

[9] Lee C Y，Lin S J，Lee C W，et al. An efficient continuous tracking system in real-time surveillance application. Journal of Network and Computer Applications，2012，35(3)：1067-1073

[10] Wang X. Intelligent multi-camera video surveillance：a review. Pattern Recognition Letters，2013，34(1)：3-19

[11] Hu W M，Tan T N，Wang L，et al. A survey on visual surveillance of object motion and behaviors. IEEE Transactions on Systems，Man，and Cybernetics，Part C (Applications and Reviews)，2004，34(3)：334-352

[12] Khairunissa J，Wahjuni S，Soesanto I R H，et al. Detecting poultry movement for poultry behavioral analysis using the multi-object tracking algorithm//Proceedings of the 2021 8th International Conference on Computer and Communication Engineering. Piscataway，USA，2021：265-268

[13] Gu R，Wang G，Jiang Z. Multi-person hierarchical 3d pose estimation in natural videos. IEEE Transactions on Circuits and Systems for Video Technology，2019，30(11)：4245-4257

[14] Chen L，Ai H，Chen R，et al. Cross-view tracking for multi-human 3d pose estimation at over 100 fps//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA，2020：327-3288

[15] Chu P，Fan H，Tan C，et al. Online multi-object tracking with instance-aware tracker and dynamic model refreshment//Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision. Waikoloa Village，USA. 2019：161-170

[16] Xiu Y，Li J，Wang H，et al. Pose Flow：efficient online pose tracking. arXiv preprint arXiv：1802.00977，2018.

[17] Alejandro P Y，Antonio A. Matching and recovering 3D people from multiple views//Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa，USA，2022：1184-1193

[18] Andriluka M，Roth S，Schiele B. Monocular 3D pose estimation and tracking by detection//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco，USA，2010：623-630

[19] Ku J，Pon A D，Walsh S，et al. Improving 3D object detection for pedestrians with virtual multi-view synthesis orientation estimation//Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Macao，China，2019：3459-3466

[20] Tang Z，Wang G，Xiao H，et al. Single-camera and inter-camera vehicle tracking and 3d speed estimation based on fusion of visual and semantic features//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops. Salt Lake City，USA，2018：108-115

[21] Li F，Wang Z，Nie D，et al. Multi-camera vehicle tracking system for AI city challenge 2022//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops. New Orleans，USA，2022：3264-3272

[22] Liu C，Zhang Y，Luo H，et al. City-scale multi-camera vehicle tracking guided by crossroad zones//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual，2021：4129-4137

[23] Specker，A，Lucas F，CoMickael，et al. Improving multi-target multi-camera tracking by track refinement and completion//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans，USA，2022：3198-3208

[24] Nikodem M，Slabicki M，Surmacz T，et al. Multi-camera vehicle tracking using edge computing and low-power communication. Sensors，2020，20(11)：3334

[25] Herzog F，Chen J，Teepe T，et al. Synthehicle：multi-vehicle multi-camera tracking in virtual cities//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa，USA，2023：1-11

[26] Chung N M，Le H，Nguyen V，et al. Multi-camera multi-vehicle tracking with domain generalization and contextual constraints//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans，USA，2022：3326-3336

[27] He Y，Han J，Yu W，et al. City-scale multi-camera vehicle tracking by semantic attribute parsing and cross-camera tracklet matching//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Virtual，2020：576-577

[28] Du Y，Wan J，Zhao Y，et al. Giaotracker：a comprehensive framework for mcmot with global information and optimizing strategies in visdrone //Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual，2021：2809-2819

[29] Naik B，Hashmi M，Geem Z，et al. DeepPlayer-Track：player and referee tracking with Jersey color recognition in soccer. IEEE Access，2022，10：32494-32509

[30] Khan S D，Ullah H，Ullah M，et al. Person head detection based deep model for people counting in sports videos//Proceedings of the 16th IEEE International Conference on Advanced Video and Signal Based Surveillance. Taipei，China，2019：1-8

[31] Michela Z，Mikhail G，Riccardo M，et al. Multi-robot multiple camera people detection and tracking in automated warehouses//Proceedings of the IEEE 19th International Conference on Industrial Informatics.Palma de Mallorca ，Spain，2021：1-6

[32] Tsokas N A，Kyriakopoulos K J. Multi-robot multiple hypothesis tracking for pedestrian tracking with detection uncertainty//Proceedings of the 19th Mediterranean Conference on Control Automation. Corfu，Greece，2011：315-320

[33] Zhang S，Gong Y，Huang J B，et al. Tracking persons-of-interest via adaptive discriminative features//Proceedings of the European Conference on Computer Vision. Amsterdam，Netherland ，2016：415-433

[34] Rao J，Xu K，Chen J，et al. Sea-surface target visual tracking

with a multi-camera cooperation approach. Sensors (Basel), 2022, 22(2): 693

[35] Mangi S N. Multi-target tracking for video surveillance using deep affinity network: a brief review. arXiv preprint arXiv: 2110.15674, 2021

[36] Rakai L, Song H, Sun S J, et al. Data association in multiple object tracking: a survey of recent techniques. Expert Systems with Application, 2022,192(4):1-19

[37] Narayan N, Sankaran N, Setlur S, et al. Learning deep features for online person tracking using non-overlapping cameras: a survey. Image and Vision Computing, 2020,89(9):222-235

[38] Fleuret F, Berclaz J, Lengagne R, et al. Multi-camera people tracking with a probabilistic occupancy map. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(2): 267-282

[39] Berclaz J, Fleuret F, Turetken E, et al. Multiple object tracking using K-shortest paths optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011,33(9):1806-1819

[40] Sankaranarayanan A, Veeraraghavan A, Chellappa R. Object detection, tracking and recognition for multiple smart cameras. Proceedings of the IEEE, 2008, 96(10):1606-1624

[41] Peng P, Tian Y, Wang Y, et al. Robust multiple cameras pedestrian detection with multi-view Bayesian network. Pattern Recognition: The Journal of the Pattern Recognition Society, 2015, 48(5): 1760-1772

[42] Wen L, Lei Z, Chang M C, et al. Multi-camera multi-target tracking with space-time-view hyper-graph. International Journal of Computer Vision. 2017, 122(2): 313-333

[43] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA, 2006: 369-376

[44] Balasundaram A, Chellappan C. An intelligent video analytics model for abnormal event detection in online surveillance video. Journal of Real-Time Image Processing, 2020, 17(4): 915-930

[45] Zhang R, Wu L, Yang Y, et al. Multi-camera multi-player tracking with deep player identification in sports video. Pattern Recognition. 2020,102(C): 107260-107260

[46] He Y, Wei X , Hong X , et al. Multi-target multi-camera tracking by tracklet-to-target assignment. IEEE Transactions on Image Processing, 2020, 29: 5191-5205

[47] Quach K G, Nguyen P Le H, et al. DyGLIP: a dynamic graph model with link prediction for accurate multi-camera multiple object tracking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 13784-13793

[48] Nguyen D, Henschel R, Rosenhahn B, et al. LMGP: Lifted multicut meets geometry projections for multi-camera multi-object tracking. //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022:8866-8875

[49] Suurballe J W. Disjoint Paths in a Network. Networks, 1974, 4 (2):125-145

[50] Jiang H, Fels S, Little J. A linear programming approach for multiple object tracking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-8

[51] Leal-Taixé L, Pons-Moll G, Rosenhahn B. Branch-and-price global optimization for multi-view multi-target tracking// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012:1987-1994

[52] Shitrit H B, Berclaz J, Fleuret F, et al. Tracking multiple people under global appearance constraints//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011:137-144

[53] Rios J, Ross K. Massively parallel dantzig-wolfe decomposition applied to traffic flow scheduling. Journal of Aerospace Computing, Information, and Communication, 2010, 7(1): 32-45

[54] Ferryman J, Shahrokni A. Pets2009: dataset and challenge// Proceedings of the 20th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. SnowBird, USA, 2009: 1-6

[55] Vleeschouwer D, Chen F, Delannay D, et al. Distributed video acquisition and annotation for sport-event summarization. NEM Summit. 2008,8(10):1016-1023

[56] Peng P, Tian Y, Wang Y, et al. Multi-camera pedestrian detection with multi-view Bayesian network model//Proceedings of the British Machine Vision Conference. Surrey, UK, 2012: 1-12.

[57] Xu Y, Liu X, Liu Y, et al. Multi-view people tracking via hierarchical trajectory composition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4256-4265

[58] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modeling sentences. arXiv preprint arXiv: 1404.2188, 2014.

[59] Pirsiavash H, Ramanan D, Fowlkes C. Globally-optimal greedy algorithms for tracking a variable number of objects// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 1201-1208

[60] Baqué P, Fleuret F, Fua P. Deep occlusion reasoning for multi-camera multi-target detection//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 271-279

[61] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. // Proceedings of the International Conference on Machine Learning, Williamstown, USA, 2001: 282-289

[62] Chavdarova T, Fleuret F. Deep multi-camera people detection//Proceedings of the IEEE International Conference on Machine Learning and Applications. Cancun, Mexico, 2017: 848-853

[63] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1-9

[64] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84-90

[65] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3):211-252

[66] Chavdarova T, Baqué P, Bouquet S, et al. Wild-track: A multi-camera hd dataset for dense unscripted pedestrian detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 5030-5039

[67] Lan L, Wang X, Hua G, et al. Semi-online multi-people tracking by re-identification. International Journal of Computer Vision, 2020,128(7): 1937-1955

[68] Boykov Y, Veksler O, Zabih R. Fast approximate energy minimization via graph cuts. IEEE Transactions on pattern analysis and machine intelligence, 2001, 23(11):1222-1239

[69] Hou Y, Zheng L, Gould S. Multiview detection with feature perspective transformation//Proceedings of the European Conference on Computer Vision. Virtual, 2020:1-18

[70] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[71] You Q, Jiang H. Real-time 3d deep multi-camera tracking. arXiv preprint arXiv:2003.11753, 2020

[72] Hou Y, Zheng L. Multiview detection with shadow Transformer (and view-coherent data augmentation)//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China, 2021: 1673-1682

[73] Lee W Y, Jovanov L, Philips W. Multi-view target transformation for pedestrian detection//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2023: 90-99

[74] Kuo C, Huang C, Nevatia R. Inter-camera association of multi-target tracks by on-line learned appearance affinity models//Proceedings of the European Conference on Computer Vision. Heraklion, Greece, 2010: 5-11

[75] Cai Y, Medioni G. Exploring context information for inter-camera multiple target tracking//IEEE Winter Conference on Applications of Computer Vision. Steamboat Springs, USA, 2014: 761-768

[76] Chen K, Cao L, Chen X, et al. An equalized global graph model-based approach for multicamera object tracking. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(11): 2367-2381

[77] Zhang Z, Wu J, Zhang X, et al. Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on DukeMTMC project. arXiv preprint arXiv:1712.09531, 2017

[78] Ristani E, Solera F, Zou R, et al. Performance measures and a dataset for multi-target, multi-camera tracking//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherland, 2016: 17-35

[79] Girshick R. Faster r-cnn//Proceedings of the Conference on Computer Vision. Santiago, Chile, 2015:1440-1448

[80] Kuhn H W. The Hungarian method for the assignment problem. Naval Research Logistics (NRL), 2005, 52(1): 7-21.

[81] Lee Y, Zheng T, Hwang J. Online-learning-based human tracking across non-overlapping cameras. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(10): 2870-2883

[82] Jiang N, Bai S, Xu Y, et al. Online inter-camera trajectory association exploiting person re-identification and camera topology//Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Republic of Korea, 2018: 1457-1465

[83] Kaiming H, Georgia G, Piotr D, et al. Mask r-cnn//Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy,2017:2961-2969

[84] Li P, Zhang J, Zhu Z, et al. State-aware re-identification feature for multi-target multi-camera tracking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA, 2019:0-0

[85] Liu W, Camps O, Sznaier M. Multi-camera multi-object tracking. arXiv preprint arXiv:1709.07065, 2017

[86] Liao S, Yang H, Zhu X, et al. Person re-identification by Local Maximal Occurrence representation and metric learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2197-2206

[87] Dicle C, Camps O I, Sznaier M. The way they move: tracking multiple targets with similar appearance//Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, Australia ,2013: 2304-2311

[88] Roshan A, Dehghan A, Shah M. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs//Proceedings of the European Conference on Computer Vision. Florence, Italy ,2012: 343-356

[89] Tesfaye Y, Zemene E, Prati A, et al. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. International Journal of Computer Vision, 2019, 127: 1303-1320

[90] Zheng Z, Bie Z, Sun Y, et al. MARS: A video benchmark for large-scale person re-identification//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherland,2016: 868-884

[91] Ristani E, Tomasi C. Features for multi-target multi-camera tracking and re-identification//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6036-6046

[92] Cao Z, Gines H, Tomas S, et al. Open Pose: realtime multi-person 2D pose estimation using Part Affinity Fields. arXiv preprint arXiv:1812.08008, 2018

[93] Hou Y, Zheng L, Wang Z, et al. Locality aware appearance metric for multi-target multi-camera tracking. arXiv preprint arXiv:1911.12037, 2019

[94] Kwangjin Y, Young-Min S, Moongu J. Multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views. IET Image Processing, 2018, 12(7): 1175-1184

[95] Reid D. An algorithm for tracking multiple targets. IEEE

transactions on Automatic Control, 1979, 24(6): 843-854
[96] Tang Z, Naphade M, Liu M, et al. CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 8797-8806
[97] Leal-Taixé L, Milan A, Reid I, et al. Motchallenge 2015: Towards a benchmark for multi-target tracking. arXiv preprint arXiv:1504.01942, 2015
[98] Keni B, Rainer S. Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing, 2008, 2008: 1-10
[99] Ristani E, Tomasi C. Tracking multiple people online and in real time//Proceedings of the 12th Asian Conference on Computer Vision. Singapore, Singapore, 2014: 444-459
[100] Li Y, Huang C, Ram N. Learning to associate: Hybrid boosted multi-target tracker for crowded scene//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Miami, USA,2009: 2953-2960
[101] Kasturi R, Goldgof D, Soundararajan P, et al. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 31(2): 319-336.
[102] Luiten J, Osep A, Dendorfer P, et al. Hota: A higher order metric for evaluating multi-object tracking. International Journal of Computer Vision, 2021,129(2):548-578
[103] Philipp K, Andreas S, Arne S, et al. The mta dataset for multi-target multi-camera pedestrian tracking by weighted distance aggregation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Virtual, 2020: 1042-1043
[104] Han X, You Q, Wang C, et al. Mmptrack: Large-scale densely annotated multi-camera multiple people tracking benchmark. arXiv preprint arXiv: Arxiv-2111.15157, 2021.
[105] Vora J, Dutta S, Jain K, et al. Bringing generalization to deep multi-view pedestrian detection//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW).Waikoloa, USA, 2023: 110-119
[106] Fu X, Huang W, Sun Y, et al. A Novel Dataset for Multi-View Multi-Player Tracking in Soccer Scenarios. Applied Sciences, 2023,13(9): 5361
[107] Mei J, Zhu A Z, Yan X, et al. Waymo open dataset: Panoramic video panoptic segmentation//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 53-72
[108] Zhu S, Lei Q, Liu X, et al. Cross-view people tracking by scene-centered spatio-temporal parsing//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, USA,2017: 4299-4305
[109] Yang F, Odashima, S, Yamao, et al. A unified multi-view multi-person tracking framework. arXiv preprint arXiv: 2302.03820,2023
[110] Liu M, Zhu C, Ren S, et al. Unsupervised multi-view pedestrian detection. arXiv preprint arXiv:2305.12457, 2023.
[111] Engilberge M , Liu W , Fua P. Multi-view tracking using weakly supervised human motion prediction//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA,2023:1582-1592
[112] Maksai A, Wang X, Fleuret F , et al. Non-markovian globally consistent multi-object tracking//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2544-2554
[113] Liang Y, Zhou Y. Multi-camera tracking exploiting person re-id technique//Proceedings of the International Conference on Neural Information Processing. Guangzhou, China, 2017: 397-404
[114] Chen W, Cao L, Chen X, et al. Multi-camera object tracking challenge//Proceedings of the ECCV Workshop on Visual Surveillance and Re-identification. Zürich, Switzerland, 2014: 32-37
[115] Chen W, Cao L, Chen X, et al. A novel solution for multi-camera object tracking//Proceedings of the IEEE International Conference on Image Processing. Paris, France, 2014: 2329-2333
[116] Milan A, Leal-Taixé L, Reid I, et al. MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831, 2016
[117] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020

ZHANG Peng, Ph.D. candidate. His main research interests include multi-target tracking and human behavior analysis.

LEI Wei-Min, Ph. D., professor, his main research interests are video enhancement, real-time transmission optimization, and video semantic compression coding.

ZHAO Xin-Lei, M.S. Her main research interests include multi-target tracking and human behavior analysis.

DONG Li-Jia, M.S. Her research interests include human behavior analysis and human movement quality assessment.

LIN Zhao-Nan, Ph. D. candidate, senior engineer. His research interests include computer vision and target detection.

JING Qing-Yang, Ph. D. candidate. Her research interests include target tracking and video semantic compression coding.

Background

The research of the Multi-Target Single Camera Tracking method is becoming more and more mature, but the single-camera multi-target tracking task can only recognize a single region, and with the advancement of video surveillance technology, the method can no longer meet the needs of the surveillance system. With the advancement of video surveillance technology, this method can no longer meet the needs of the surveillance system. It has become the main demand of the intelligent surveillance system to open up the recognition area between cameras and to extend the single camera area to the whole area recognition.

In recent years, in order to solve the problem of full-area recognition and tracking has prompted many researchers to address Multi-Target Multi-Camera Tracking, and some results have been achieved. However, there is no comprehensive article on the Multi-Target Multi-Camera Tracking task to summarize the field, and most of them are used as a complementary description of Multi-Target Single Camera Tracking, which does not have a very clear and more detailed classification and introduction of the task, and there is also a clear collection and organization of related datasets and metrics.

In view of the current situation of Multi-Target Multi-Camera Tracking, this paper classifies the task in detail by combining with the practical application requirements, and divides the Multi-Target Multi-Camera Tracking task into overlapping view, non-overlapping view and mixed view. The advantages and disadvantages of these methods and their applicable scenarios are compared; the current commonly used datasets and evaluation criteria for Multi-Target Multi-Camera Tracking are analyzed; and finally, the problems of Multi-Target Multi-Camera Tracking are summarized and outlook is given.

This research was jointly funded by the 2022 Key Science and Technology Project of Liaoning Province (Intelligent System of Key Technologies for Public Transport Vehicle Fire Alarm and Extinguishment) (2022JH1/10400025), the Special Fund for Basic Research Business Fees of Central Universities (Grant Number: N2216010), and the National Key R&D Program of China (Grant Number: 2018YFB1702000).