

面向足球比赛视频的跨镜头全局运动跟踪方法

宋子恺 马林涛 唐 颖 杨 卫 于俊清

(华中科技大学计算机科学与技术学院 武汉 430074)

摘 要 足球视频中的跨镜头目标跟踪技术对于实现整场比赛的全局连续分析具有重要意义,然而,当前算法大多局限于单镜头跟踪,难以应对足球比赛中频繁的镜头切换,导致跟踪中断和对象失配,无法满足全局数据分析的需求。针对这一关键问题,本文提出了一种跨镜头全局跟踪算法,通过单镜头跟踪增强和跨镜头关联匹配两方面的创新优化,显著提升了足球视频分析的完整性和准确性。在方法层面,首先设计了基于球员状态的关系交互模型,有效解决了运动模糊、外观相似和遮挡等问题,设计了基于重叠窗口的层级关联策略,解决逐帧关联容易轨迹碎片化的足球跟踪难题;其次创新构建了融合球员位置和姿态的长跨度关联匹配方法,提出边界轨迹衔接后处理方法,实现跨镜头精准匹配。实验方面,构建了全新的足球视频跨镜头跟踪数据集,所提方法在跨镜头关联上取得86.7%的准确率和97.0%的召回率,全局跟踪方面在HOTA指标上提升14.0%,此外,所提方法在SoccerNet-Tracking和SportsMOT公开数据集上达到最优性能。研究成果为体育智能化发展提供了重要的技术支撑。

关键词 计算机视觉;智能体育分析;足球运动分析;多目标跟踪;跨镜头跟踪

中图法分类号 TP391

DOI号 10.11897/SP.J.1016.2026.02181

Global Cross-shot Object Tracking for Soccer Video Analysis

SONG Zi-Kai MA Lin-Tao TANG Ying YANG Wei YU Jun-Qing

(School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074)

Abstract Cross-shot object tracking is essential for enabling comprehensive match analysis in soccer video sequences. While current tracking methods predominantly operate within individual camera shots, they exhibit significant limitations when confronted with the dynamic multi-shot nature of broadcast soccer footage. Frequent shot transitions often result in tracking discontinuities and identity mismatches, ultimately compromising the integrity of global match analytics. To address this key challenge, this paper proposes a cross-shot global tracking algorithm, featuring innovative optimizations in both single-shot tracking enhancement and cross-shot object association, significantly improving the completeness and accuracy of soccer video analysis. Methodologically, the study first introduces a player-state-based relational interaction model to effectively handle motion blur, appearance similarity, and occlusion, and introduces a hierarchical association strategy with overlapping time windows to mitigate trajectory fragmentation in frame-by-frame tracking. Second, we develop a long-range cross-shot association method that fuses player positional and pose features, along with a boundary trajectory stitching post-processing step, to achieve precise cross-shot matching. For experimentation, a novel cross-shot soccer video tracking dataset was constructed. The proposed method achieves 86.7% accuracy and 97.0% recall in

收稿日期:2025-09-08;在线发布日期:2026-03-26。本课题得到国家自然科学基金青年科学基金项目(No. 62272184)、国家自然科学基金面上项目(No. 62402189)、中国博士后科学基金资助项目(2024M751012、GZC20230894、2025T180429)、湖北省博士后创新人才培养项目(2024HBBHCXB014)资助。宋子恺,博士,中国计算机学会(CCF)会员,主要研究领域为视觉运动分析、社交网络分析。E-mail: skyesong@hust.edu.cn。马林涛,硕士,主要研究领域为视觉目标跟踪。唐 颖,博士研究生,主要研究领域为多任务学习和特征分析与融合。杨 卫,博士,副研究员,主要研究领域为三维视觉。于俊清(通信作者),博士,教授,主要研究领域为智能媒体计算、网络空间安全。E-mail: yjqing@hust.edu.cn。

the cross-shot association, while improving the HOTA metric by 14.0% in global tracking. It also demonstrates state-of-the-art performance on public datasets such as SoccerNet-Tracking and SportsMOT. These research outcomes provide critical technical support for the advancement of intelligent sports analytics.

Keywords computer vision; intelligent sports analysis; soccer analysis; multi-object tracking; cross-shot tracking

1 引 言

作为全球最具影响力的体育项目,足球不仅拥有广泛的群众基础,还蕴含巨大的商业潜力。随着计算机视觉技术的快速发展,目标跟踪^[1-5]在足球视频分析中的应用日益广泛,已成为推动体育智能化进程的关键技术之一。该技术在足球视频分析中的多个层面展现出显著价值^[6-7]:对于普通观众,通过对球员的精准跟踪,可实现精彩镜头的自动提取与聚焦展示,增强观赛的沉浸感与互动性;对于教练团队,可借助目标跟踪技术获取球员的运动轨迹、跑动距离与速度等关键参数,从而优化训练方案与战术部署;对于裁判系统,目标跟踪有助于精确定位球员位置,辅助识别犯规行为,从而提高判罚的准确性与公正性。

目标跟踪在足球视频分析中具有显著的理论意义与实用价值,然而,现有的目标跟踪算法在足球领域中的应用存在两大问题:(1)跟踪连续性不足。足球比赛视频中,镜头通常呈现间断性切换的特点,如图1所示,但现有算法通常只能在单一镜头内跟踪,镜头切换时需重新初始化,导致跟踪中断和对对象失配,这种间断性处理方式易造成跟踪中断与对象失配,限制了球员行为数据的全局整合与分析深度;(2)复杂场景干扰严重。比赛过程中球员的高速运动导致预测算法难以准确估计其位置,球员穿着相似的队服使得外观特征区分度降低,比赛过程中频繁发生的遮挡现象进一步削弱了跟踪的稳定性与鲁棒性,这些复杂的干扰因素叠加,使得现有方法难以在足球场场景下取得理想效果。

针对上述难点,本文设计并实现一种面向足球比赛视频的跨镜头全局跟踪算法(Cross-shot soccer player tracking approach, CSTracker),如图1所示,通过单镜头内跟踪增强与跨镜头间匹配关联的双重优化,来提升复杂足球环境下跟踪的准确性与鲁棒性。这项工作首次在足球领域构建了全局跨镜

头跟踪框架,极大增强了算法的实用性,此外,针对跨镜头足球跟踪数据集稀缺的现状,构建了全新的足球视频跨镜头跟踪数据集(Cross-shot soccer player tracking benchmark, CSoccerBench),为后续研究提供了坚实的数据基础和新的评估基准。



图1 全局运动跟踪中的单镜头跟踪与跨镜头匹配示意图

本文创新点可以归纳如下:

(1)在单镜头跟踪方面,设计了球员状态的关系交互模型和基于重叠窗口的层级关联策略:关系交互模型通过消息传递网络让球员在时空维度上交换信息,不仅考虑单个球员的特征,还能利用球员之间的相对关系,提升球员遮挡或运动模糊场景下的鲁棒性;层级关联策略首先在窗口内做短期匹配,然后在重叠时间窗口之间层级关联,逐步拼接成长期轨迹,解决逐帧关联容易轨迹碎片化的问题;

(2)在跨镜头关联匹配方面,构建了融合球员位置和姿态的长跨度关联匹配方法:位置关联利用预测位置和实际位置的偏差来建模球员轨迹的连贯性,姿态关联利用人体关键点估计获得外观难以区分时的补充信息,将位置和姿态特征自适应地融合,实现跨镜头球员精准匹配和轨迹修正,最终达到对整场比赛的全局连续跟踪;

(3)构建了全新的足球视频跨镜头跟踪数据集 CSoccerBench,在此数据集上,本文方法在跨镜头关联上达到 86.7% 的准确率和 97.0% 的召回率,全局跟踪准确度相比于基准方法在主要指标 HOTA 上提升 14.0%。除此以外,本文提出的方法在公开的足球视频跟踪数据集 SoccerNet-Tracking^[8] 和 SportsMOT^[9] 上面也达到了当前最优异的跟踪性能。

2 相关工作

本章节主要讨论现有多目标跟踪方法、跨镜头跟踪方法以及面向体育场景的跟踪方法,为读者对本文的理解打下基础。

2.1 单镜头多目标跟踪

多目标跟踪方法的核心可以归纳为两个步骤,分别是当前帧的目标定位和前后帧的目标匹配。依据这两个核心步骤中目标定位和目标匹配处理流程的不同,多目标跟踪算法可以分为基于检测的两阶段跟踪算法(Tracking-by-detection Two-stage Methods)和端到端的单阶段跟踪算法(End-to-end One-stage Methods)。

两阶段多目标跟踪算法是指采用两个独立的模块,即目标检测模块用于当前帧目标定位和位移估计模块用于前后帧的目标匹配。两阶段跟踪算法^[10-19]遵循先检测后跟踪(Tracking-by-Detection)的范式,定位模块通常使用较为成熟的目标检测算法^[20-21],匹配模块一般使用位移估计方法(如卡尔曼滤波^[22])计算跟踪对象的位移分布,匹配模块会利用位移估计与目标检测结果,利用目标框的交并比重叠程度,计算出前后帧目标的对应关系。两阶段跟踪算法通常难以应对复杂运动的场景,其运动估计以卡尔曼滤波为基础,以物体的线性运动为假设前提进行估计运动轨迹,在人群监控等固定镜头、位移单一的场景下表现较好,但是在足球运动等目标运动复杂的场景下很难实现准确的运动估计。此外,基于图网络的跟踪方法^[23-25]将多目标跟踪问题建模为图网络,将不同视频帧中的目标表示为图节点,关联可能性表示为边,并通过融合外观、运动等多线索为每条边计算关联代价。

单阶段多目标跟踪算法将目标检测和目标匹配集成到一个算法框架中,以端到端的方式完成多目标跟踪任务。主流的单阶段多目标跟踪算法可以根据其算法结构分为两类,查询嵌入算法和同步回归

算法。查询向量(Query)嵌入的跟踪方法^[26-33]将每个跟踪对象视为一个 Query,跟踪器对所有 Query 在图像帧中回归出其位置信息;同步回归的跟踪方法^[34-38]则是在网络结构中设计多输出分支,包含用于目标定位的检测分支和用于位移估计的匹配分支,同步回归目标的位置信息与位移信息从而完成跟踪任务。单阶段跟踪算法没有显式地进行目标的位移估计,使之不能继承先前位移估计方法(如级联匹配^[11],低置信度再匹配^[13]等)的优势,查询嵌入方法 Query 数组中目标对象的数量设定为一个固定值,不能动态调整,在应对拥挤场景或者目标数量变化较大场景时存在一定的不足。

本文提出的全局跟踪方法,包括单镜头跟踪与跨镜头匹配两个部分,其中,单镜头跟踪算法采用两阶段框架。与现有单镜头跟踪方法仅依赖单目标特征提取不同,本方法针对足球场景中常见的球员遮挡与运动模糊问题,创新性地设计了球员关系交互模型,通过消息传递网络实现球员间的信息交换,不仅充分利用单个球员的外观与运动特征,还融合了球员之间的相对关系,从而显著提升在复杂场景下的鲁棒性;另一方面,现有方法通常在逐帧基础上进行匹配,容易导致轨迹碎片化,为此,本文提出了层级关联策略,先在重叠时间窗口内进行短期匹配,再逐步拼接为长期轨迹,有效缓解了逐帧关联造成的轨迹碎片化问题。

2.2 跨镜头跟踪

跨镜头多目标跟踪是指从多个镜头采集的视频中确定每个跟踪目标的位置。根据镜头之间的视角是否重叠,跨镜头多目标跟踪方法主要可以分为重叠视角、非重叠视角和混合视角。

重叠视角的跟踪算法中镜头能记录下同一个目标在同一时刻不同视角的信息,为单一视角导致的遮挡现象提供了解决方案,但多目标之间的特征信息在重叠视角下容易互相干扰,导致目标身份切换(ID Switch)问题的出现。根据目标关联的方法,重叠视角下的跨镜头多目标跟踪可以分为网络流优化方法^[39-42]、单源性约束方法^[43-44]、强化学习方法^[45-47]、超图法^[48]和基于 Transformer 的方法^[49-52]。非重叠视角的算法中同一个跟踪目标在不同镜头之间移动的过程中存在视觉盲区,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。非重叠视角下的跨镜头多目标跟踪算法主要可以分为双阶段关联算法^[53-60]和单阶段关联算法。双阶段关联算法采用分步策略,先生

成每个独立的镜头内的多目标跟踪轨迹,然后以这些轨迹作为单元,进行跨摄像头的匹配与连接;单阶段轨迹关联算法^[61-65],从全局出发,综合考虑所有镜头的全部目标,直接进行统一的身份分配与轨迹生成。混合视角算法旨在解决跨镜头多目标跟踪算法场景普适性问题,使其能同时适用于重叠与非重叠的摄像机视角。TRACTA^[66]通过非负矩阵分解将局部轨迹与全局目标直接匹配,并能估计目标总数以纠正局部错误;DyGLIP^[67]则首次引入动态图与链路预测,利用注意力机制捕捉时空变化;LMGP^[68]应用时空多切割公式主动分解单摄像头阶段产生的错误轨迹,并建立精确的亲和度成本来重组,实现更准确的关联。Fan 等人^[69]通过生成目标运动的时空一致性,在图注意力机制中融合外观特征,构建具有高区分度的关联特征。

与现有的跨镜头跟踪方法不同,本文提出的跨镜头匹配并非利用多个摄像头在同一时刻、不同视角下对目标进行联合跟踪,而是针对足球场景中同一摄像头在镜头切换前后、间隔较长时间跨度后的匹配问题。由于时间跨度过长,球员的运动距离往往较大,因此本文提出利用球场区域先验知识进行镜头对齐,从而实现基于位置的匹配。同时,不同于传统跨镜头方法主要依赖外观特征,足球比赛中同队球员队服高度相似,外观特征区分度不足,为此,本文引入球员姿态特征,利用其在运动过程中的差异性与独特性提升区分能力。将位置与姿态特征进行自适应融合,实现跨镜头下球员的精确匹配与轨迹修正,保证整场比赛的全局连续跟踪。

2.3 面向体育场景的跟踪方法

足球场景的目标跟踪算法经历了从早期模板匹配到如今视觉大模型的转变。初期,算法主要围绕如何利用颜色和形状等基本特征展开,Seo 等人^[70]和 Yoon 等人^[71]利用模板直方图反向投影(Histogram Backprojection)技术,初步解决了球员在遮挡情况下的跟踪难题。Ekin 等人^[72]在此基础上,针对足球比赛中的不同情景(如射门、无遮挡和遮挡等),分别设计了场地线匹配、区域检测和自适应融合的跟踪策略。为应对更复杂的动态和个体差异,基于概率模型的跟踪方法逐渐成为主流,Filters 等人^[73]将粒子滤波器(Particle Filter)引入球员跟踪,利用颜色信息来区分个体。Liu 等人^[74]使用马尔科夫链蒙特卡洛方法对目标进行分析,通过更优的数据驱动模型提高了整体的跟踪效率。Mazzeo 等人^[75]使用诸如质心、速度、外接矩形的状态向量增

强跟踪中球员的特征表达能力。Song 等人^[76]提出的干扰感知算法通过动态编码干扰球员语义信息,增强了追踪鲁棒性。图跟踪方法^[77-78]使用图表示相邻帧中球员的位置关系,并通过形态学分割和图的前后推理实现球员跟踪。近年来随着大模型的兴起,Cui 等人^[79]在大型体育运动多目标跟踪数据集(SportsMOT)的基础上,提出了面向足球、篮球和排球运动场景的基于混合注意力(MixFormer)^[80]的跟踪算法,有效融合了外观和运动信息,在球员关联上取得了优异表现。

足球场景具有遮挡频繁、轨迹复杂以及球员外观高度相似等特点,这些因素给目标跟踪带来了显著挑战,导致现有算法在鲁棒性与跟踪稳定性方面表现不足。更为关键的是,现有方法通常仅能在单镜头内完成局部跟踪,一旦发生镜头切换,跟踪便被迫中断。而针对跨镜头的全局跟踪仍缺乏有效解决方案,这一技术空白严重制约了足球视频分析的实际应用价值。本文首次提出了面向足球比赛视频的跨镜头全局运动跟踪方法,为实现长时间、跨镜头的球员连续跟踪提供了新的解决思路。

3 全局运动跟踪方法

全局运动跟踪方法由单镜头跟踪增强和跨镜头关联匹配两部分构成,单镜头跟踪增强是指在单一镜头内保持球员轨迹的连续与稳定,跨镜头关联匹配是指在镜头切换前后实现球员身份的准确关联,从而保证整场比赛中的全局连续跟踪。

3.1 单镜头跟踪

单镜头跟踪的流程如图 2 所示,包括状态建模、时空交互以及层级关联三个部分:状态建模将检测到的球员表示为包含外观、运动的球员状态,建立初始的关系图模型,为后续信息交互与匹配提供可比较、可传播的基础表示;关系交互利用消息传递网络传递与融合球员的状态,并在单个时间窗内实现目标匹配,增强球员表示的鲁棒性,更好地应对遮挡与运动模糊;层级关联在多个时间窗口之间做层级匹配,逐步拼接为长期轨迹,缓解逐帧匹配带来的轨迹碎片化问题。

3.1.1 状态建模

状态建模作为单镜头跟踪的初始化模块,包含两个部分:(1)状态表征构建对每个球员从外观特征和时空特征两方面来表示球员状态,增强球员的辨识度;(2)建立关系图模型将球员的检测结果转化

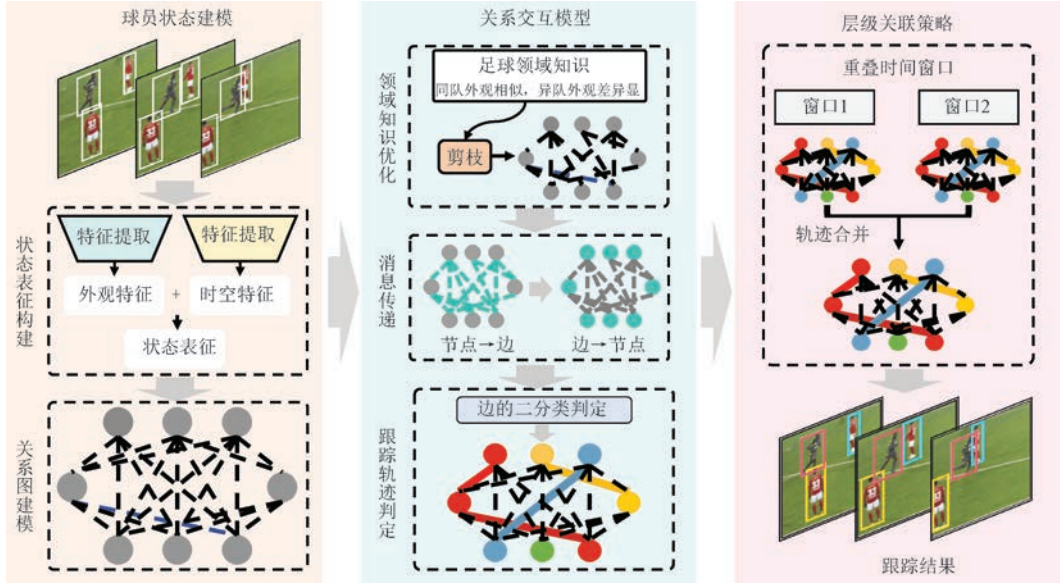


图2 单镜头头内跟踪的整体算法结构示意图

为二分图匹配的形式,对图进行计算优化,得到初步的关联关系。

状态表征构建。对于足球比赛视频中的每一帧图像,利用目标检测器识别出视频帧中的球员。本文采用多目标跟踪中流行的YOLOX^[20]作为检测器,设当前为时间 t 的视频帧,当前帧的目标数量为 n ,目标检测的结果集合为 $D^t = \{d_1^t, d_2^t, \dots, d_n^t\}$,其中 d_i^t 表示的单个目标。

得到球员检测结果后,对每个球员从外观特征和时空特征两个方面构建球员的状态。外观特征方面,采用FastReID^[81]提取特征,为适应模型的尺寸,利用多层感知机对外观特征降维映射,目标 d_i 的外观特征表示为 r_i 。时空特征方面,由于足球场场景中队友球员身着相似队服导致外观相似度极高,为了增加特征的区别性,使用球员的边框特征作为确定目标轨迹的重要线索,对于目标 d_i ,其包围盒矩形框表示为 $b_i = (x_i, y_i, w_i, h_i)$ 分别对应其左上角坐标和矩形框的宽高,构建融合时间信息的五元组特征 $st_i = (x_i/W, y_i/H, w_i/W, h_i/H, t_i/T)$,其中 W 和 H 表示图像帧的宽和高, t_i 和 T 表示当前帧和总帧数,将该时空特征通过感知机网络 φ 映射到高维空间后,与外观特征 r_i 相连接,得到目标的最终状态表示 $f_i = \varphi_1(\text{cat}\{\varphi_2(st_i), r_i\})$ 。

关系图建模。对于窗口内相邻帧 t 和 $t+1$ 的检测结果 D^t 和 D^{t+1} ,建立不同时刻样本对两两之间的全局联系,依此类推,建立窗口内所有帧的初始全局关系,表示为无向图 $G=(V, E)$,图中节点表示目标,边表示目标是否匹配,关系交互模型的目标就

是将具有关联关系的节点状态进行传递融合,然后确定每条边的二分类情况,将所有活跃的边组成跟踪轨迹。初始关系图建立之后,对于相邻的两个节点 v_i 和 v_j 组成的边 $e_{(i,j)}$,使用二进制变量 $y_{(i,j)}$ 表示该边是否代表活跃轨迹,如下所示:

$$y_{(i,j)} = \begin{cases} 1, & \exists e_{(i,j)} \in \text{Traj s.t. } e_{(i,j)} \in E \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

其中, Traj 表示ground-truth的轨迹集合,在关系图构建中,1表示前后相邻两帧中的目标存在匹配关系(即这两个目标在ground-truth中被标注为同一目标),0表示没有匹配关系(即这两个目标在ground-truth中被标注为不同目标)。这种标注方式确保了关系图与ground-truth的匹配一致性。

3.1.2 关系交互模型

关系交互模型是在单个窗口内进行球员状态的传递融合、确定球员匹配关系从而完成窗口内的球员跟踪,一共包含三个步骤:(1)领域知识优化是利用足球领域的特定知识,对初始化的关系图模型进行剪枝优化,增强关系图的模型效率和准确率;(2)消息传递机制在关系图模型基础上,利用足球领域知识对球员状态进行传递与融合,增强状态表达;(3)跟踪轨迹判定则是利用优化后的球员状态,在关系图模型上确定球员的对应关系,得到单个窗口内的球员跟踪轨迹。

领域知识优化。在足球场场景中,同一球队的球员身着相同或相似颜色的队服,这为节点关联提供了天然的视觉区分依据,使得球队标签成为一种强

先验信息,有效约束了节点之间的关联范围。为降低消息传递网络的计算与内存开销,并缓解边标签的严重不平衡问题,引入了基于 K 近邻(KNN)剪枝的图优化策略。该策略的核心是定义一个融合了多种特征的关联贡献度 $d(v_i, v_j)$ 作为边的重要性度量,仅保留对每个节点贡献度最高的多条边。关联贡献度 $d(v_i, v_j)$ 是外观特征距离 d_{app} 、运动特征距离 d_{motion} 和边框距离 d_{pos} 的加权和, λ_1 和 λ_2 分别表示外观特征和运动特征的权重。 λ_1 为外观特征距离的权重,值设为 0.5, λ_2 为运动特征距离权重,设为 0.3, $(1-\lambda_1-\lambda_2)$ 表示位置特征的权重。这一加权和方式融合不同特征的信息,在 KNN 剪枝中根据每条边的贡献度,选择最重要的边进行保留,从而有效降低计算复杂度,缓解边标签的不平衡问题,确保节点关联更加准确和高效。具体如下:

$$d(v_i, v_j) = \lambda_1 \cdot d_{\text{app}}(v_i, v_j) + \lambda_2 \cdot d_{\text{motion}}(v_i, v_j) + (1 - \lambda_1 - \lambda_2) \cdot d_{\text{pos}}(v_i, v_j) \quad (2)$$

其中,运动特征距离 d_{motion} 由广义交并比(GIOU)计算得出:

$$d_{\text{motion}}(v_i, v_j) = 1 - \text{GIOU}(b_i, b_j) \quad (3)$$

在剪枝前,算法利用通过 K -means 聚类得到的球队标签提供了强有力的先验知识:我们通过聚类排除了外观特征差异较大的节点对(例如球衣颜色不同,属于不同队伍的球员)。这种方法确保了在图结构简化的同时,能够有效避免因外观相似而可能导致的错误匹配。值得注意的是, K -means 聚类使用的外观特征是通过前一步骤状态表征构建的 FastReID 提取的,因此在此步骤中并没有增加额外的计算负担,直接使用已提取的外观特征进行处理。最终,对于每个节点,仅保留与其关联贡献度高的近邻节点所连接的边,以此构建一个更稀疏、更准确的图结构用于后续的消息传递更新。

消息传递机制。消息传递更新模块通过图神经网络中的迭代式信息传递机制,优化节点与边的状态表示。该过程在固定的迭代次数 N_s 内进行,每一轮迭代 $s \in \{1, 2, \dots, N_s\}$ 包含两个关键步骤:(1)节点到边的更新($v \rightarrow e$),边的特征 $f_{(i,j)}$ 通过聚合其两端节点的状态 f_i 和 f_j 及自身上一轮的状态进行更新:

$$f_{(i,j)}^{(s)} = \varphi_e([f_i^{(s-1)}, f_j^{(s-1)}, f_{(i,j)}^{(s-1)}]) \quad (4)$$

其中, φ_e 是一个多层感知机(MLP), $[]$ 表示向量拼接操作。在边特征的初始化阶段,边的特征被设置为统一常数值,这些占位符特征通常

初始化为零向量或统一常数值。在后续的迭代过程中,边特征将通过节点状态特征的聚合进行更新,确保边的表示能够有效反映节点之间的关联关系。(2)边到节点的更新($e \rightarrow v$),为防止违反跟踪匹配的约束(即一个节点最多连接一个前继节点和一个后继节点),节点的更新被分解为对过去帧邻居和未来帧邻居的两个独立过程。首先计算来自两个方向的消息:

$$m_{(i,j)}^{(s)} = \begin{cases} \varphi_v^p([f_i^{(s-1)}, f_{(i,j)}^{(s)}, f_{(i,j)}^{(0)}]), & \text{if } e_{(i,j)} \in E \quad t_j < t_i \\ \varphi_v^f([f_i^{(s-1)}, f_{(i,j)}^{(s)}, f_{(i,j)}^{(0)}]), & \text{if } e_{(i,j)} \in E \quad t_j > t_i \end{cases} \quad (5)$$

其中, φ_v^p 和 φ_v^f 为 MLP,分别处理过去和未来方向的邻居信息, t_i 表示节点 v_i 所处的时刻。随后聚合两个方向的消息并更新节点状态:

$$f_i^{(s)} = \varphi_v([\sum_{t_j < t_i} m_{(i,j)}^{(s)}, \sum_{t_j > t_i} m_{(i,j)}^{(s)}]) \quad (6)$$

经过 N_s 轮迭代,每个节点的状态表示 f_i 都融合了其在时空图中的上下文信息,为后续的边分类和轨迹生成提供了判别性更强的特征表示。

跟踪轨迹判定。将跟踪任务建模为图中的边二分类问题,判断任意两个节点(即球员检测框)是否属于同一轨迹。该过程通过一个可学习的网络获取匹配置信度,并根据阈值进行二值化实现匹配结果,同时该过程采用贪心舍入策略确保输出解满足实际物理约束。在推理阶段,网络为每条边 $e_{(i,j)} \in E$ 输出一个分数解 $\hat{y}_{(i,j)}$,通过阈值(0.5)进行二值化,生成初始的二进制边标签 $y_{(i,j)}$ 。随后,算法迭代地检查每个检测节点是否满足跟踪任务的一对一匹配原则(即一个球员在某一帧中只能对应一条轨迹,不能同时匹配到多个目标)。对于违反该原则的节点,算法仅保留其过去帧和未来帧邻居中分数最高的边,而将其余边置为不活跃($y_{(i,j)} = 0$)。该贪心过程最终产生一个合理的二进制解 y_i ,其中所有活跃的边($y_{(i,j)} = 1$)即构成最终的跟踪轨迹。在训练阶段,通过最小化所有边上预测值与真实标签之间的二元交叉熵损失来优化网络参数。损失函数定义如下:

$$\text{Loss} = \frac{1}{|E|} \sum_s \sum_{e_{(i,j)}} [\omega \cdot y_{(i,j)} \log(\hat{y}_{(i,j)}^{(s)}) + (1 - y_{(i,j)}) \log(1 - \hat{y}_{(i,j)}^{(s)})] \quad (7)$$

其中, s 表示迭代步数, ω 是权重参数,用于平衡正负样本(活跃边与非活跃边)的贡献,以缓解类别不平衡问题。该损失函数监督网络在消息传递的多次迭代中逐步优化边的分类置信度。

3.1.3 层级关联策略

得到单个窗口内的跟踪轨迹之后,将相邻窗口的轨迹进行合并,如图3所示,以层次化的方式逐层两两归并,进而实现单个镜头内的完整跟踪。在足球比赛视频中,球员运动具有高度的非线性和突发性,相邻窗口间的轨迹可能因球员位置剧烈变化或相机运动而缺乏连续性,非重叠时间窗口划分难以适应这种动态变化,导致关联失败。为此,本章节提出了重叠窗口分层关联策略,通过在相邻窗口间引入重叠区域,更好地捕捉球员运动的连续性。

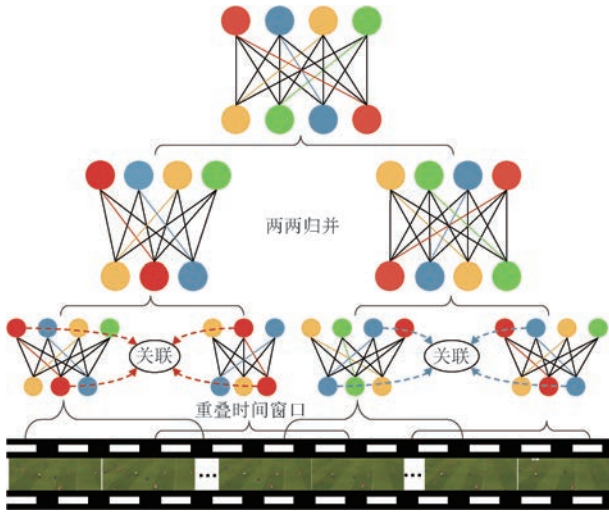


图3 层级关联策略示意图

本部分采用滑动窗口机制处理相邻时间窗口间的轨迹关联。设时间窗口长度为 L_w , 视频总长度为 L , 则窗口数量 K 由重叠因子决定:

$$K = \frac{2L}{L_w} - 1 \quad (8)$$

每个窗口与前一个窗口保持 $L_w/2$ 帧的重叠。对于相邻窗口中的轨迹 $Traj_l$ 和 $Traj_{l+1}$, 其关联代价由运动相似性和外观相似性共同决定:

$$c(Traj_l, Traj_{l+1}) = \lambda \cdot \cos(F_l, F_{l+1}) + (1 - \lambda) \cdot \text{IoU}(Traj_l, Traj_{l+1}) \quad (9)$$

其中, F_l 表示窗口 l 内每个球员的状态, 状态取窗口内轨迹中最后一个时刻, $\cos(\cdot)$ 表示计算球员状态的余弦相似度, $\text{IoU}(\cdot)$ 计算两个窗口内重叠轨迹的边界框的平均交并比, λ 为平衡权重(取 0.8)。得到相邻窗口的所有轨迹的关联代价 c 之后, 通过匈牙利算法求解最优匹配, 实现重叠窗口间的轨迹关联。

3.2 跨镜头匹配

跨镜头匹配旨在确保对于发生镜头切换前后的相邻远镜头间的跟踪连续性, 从而实现整场比赛全

局跟踪。如图4所示, 镜头切换导致球员轨迹出现了断层, 针对此情况, 跨镜头匹配流程的整体设计包含以下三个关键环节: (1) 关联信息构建, 提取位置和姿态两种特征, 球员在球场上具有明显的站位区分, 位置特征提供了稳定和可靠的几何约束, 姿态特征能帮助区分穿着相同队服的外观相似的球员, 提供更丰富的细节来识别和关联目标; (2) 特征融合匹配, 构建自适应的特征融合矩阵实现两种关键特征的有效融合, 实现跨镜头目标的精准匹配; (3) 轨迹优化, 在跟踪完成后, 针对跟踪轨迹在跨镜头复杂场景中可能会出现断裂或碎片化的问题, 将断裂与碎片化的轨迹进行关联, 进行轨迹平滑处理, 减少跟踪轨迹中的抖动和漂移误差。

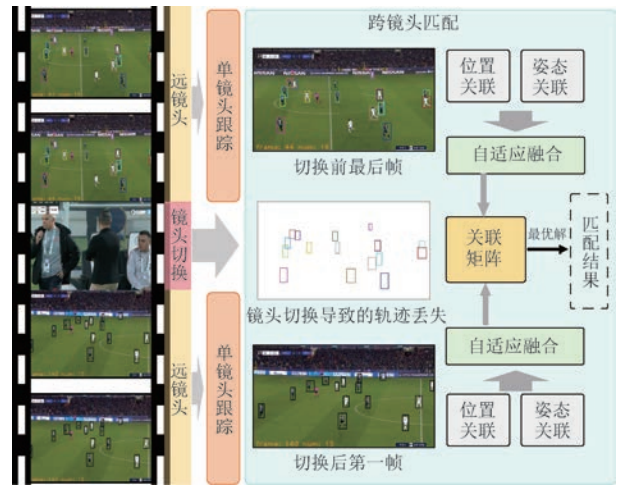


图4 跨镜头匹配的整体算法结构示意图

3.2.1 关联信息构建

针对足球比赛球员站位具备明显的区分性以及外观特征在同队球员中难以区分的特性, 我们构建了两种互补的特征, 基于球场映射的位置特征和基于动作识别的姿态特征。

位置关联构建。镜头切换前后的远镜头中, 球员坐标可能基于不同的参考系, 无法直接通过视频帧中的图像坐标进行关联。因此, 我们采用了基于球场区域的单应矩阵^[82]实现球员从图像坐标到真实球场坐标的映射, 如图5所示。通过球场分割、直线检测与分组、中圈点集识别等方法, 从足球视频帧中提取出球场上的关键点(如直线交点、中圈上的点等)。为了保证在远镜头中获得足够的关键点(大于4对), 我们结合了球场的几何特征, 通过水平与垂直直线的关系、直线之间的距离关系以及球场中圈与直线的几何关系来确定多组交点作为关键点(具体请参考文献[82])。这些关键点随后与标准二维球场模型中预先定义的关键点进行匹配。最后, 利

用至少 4 对匹配点(且任意 3 点不共线),通过最小二乘法或类似算法求解一个 3×3 的单应矩阵 $\mathbf{H}$,使得视频帧中的点 (u, v) 通过 $\mathbf{H}$ 映射到标准球场上的点 (u', v') ,完成从图像坐标系到标准球场坐标系的投影变换。对于在镜头切换后临界帧中消失的球员,采用卡尔曼滤波方法通过其历史运动信息预测其在当前帧的可能位置。对于镜头切换前后的两批球员轨迹,将其全部映射到真实球场坐标系下,计算两两球员之间的欧氏距离 d_{pos} ,并以此构建位置关联矩阵 $\mathbf{D} \in \mathbb{R}^{m \times n}$,其中 m 和 n 分别为切换前后镜头中的球员数量,矩阵元素 $D_{i,j}$ 的值越小,表示球员 i 与球员 j 的位置越接近。

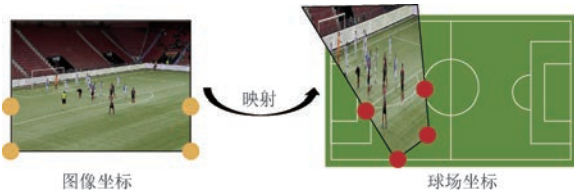


图 5 图像坐标到球场坐标映射示意图

姿态关联构建。球员姿态特征因其对运动模式和动作习惯的良好表征能力,具有比外观特征更高的稳定性和区分度。采用姿态特征进行跨镜头关联

的整体流程如图 6 所示,具体步骤如下:(1)使用姿态引导的对齐方法 PGFA^[83] 姿态估计器提取球员的关键点特征信息,形成初步的姿态特征表示 f_{pos} ; (2)为增强姿态特征的鲁棒性,引入基于 Vision Transformer^[84] 的特征提取模块,将球员图像分割为多个图像块,提取所有分块表示的上下文局部特征 $F_p = \{f_p^1, f_p^2, \dots\}$,然后再以球员整体图像作为输入提取球员的全局特征 f_g ; (3)考虑到球员姿态中不同的部位对区分球员有独特贡献,利用球员的局部特征来增强其姿态特征,即将姿态特征 f_{pos} 与全局特征 f_g 以对应元素相乘的方式得到融合特征 $f'_{pos} = f_{pos} \cdot f_g$,然后计算融合特征与所有局部特征的相似度,筛选出对球员动作贡献最大的局部特征:

$$ind = \operatorname{argmax} \left(\frac{\langle f_p^i, f'_{pos} \rangle}{|f_p^i| \cdot |f'_{pos}|} \right), f_p^i \in F_p \quad (10)$$

将该局部特征 f_p^{ind} 与姿态特征 f'_{pos} 相连接得到最终的姿态表示; (4)计算镜头切换前后两两球员之间的姿态特征的余弦距离,以此构建姿态关联矩阵 $\mathbf{A} \in \mathbb{R}^{m \times n}$ 。

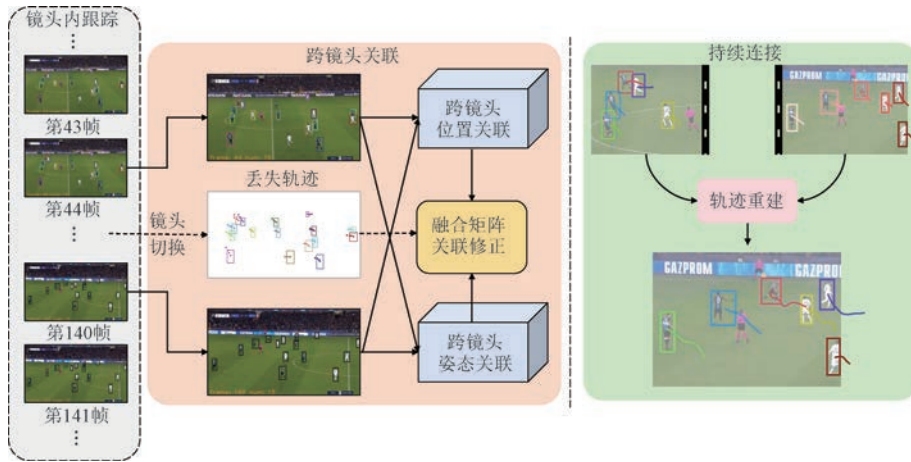


图 6 跨镜头关联整体流程

3.2.2 信息融合匹配

位置和姿态特征从不同维度提供了关联信息,本节设计一种自适应融合策略,将二者结合为一个更鲁棒的匹配代价矩阵。简单的线性加权求和无法应对球场上的复杂场景。我们观察到,在球员分散站位时,位置信息区分度更高;而在球员集中站位(如争抢球权)时,遮挡严重,姿态信息更为可靠。因此,我们根据场景类型自适应调整融合权重。定义融合后的关联矩阵 $\mathbf{C}$ 为

$$\mathbf{C} = \lambda \overline{\mathbf{D}'} + (1 - \lambda) \overline{\mathbf{A}} \quad (11)$$

$\mathbf{D}$ 为位置关联矩阵,值越小表示越可能匹配,对其每个元素求倒数为 $\mathbf{D}'$, 然后进行归一化得到 $\overline{\mathbf{D}'}$, $\mathbf{A}$ 为归一化的姿态关联矩阵,值越大表示越可能匹配, λ 表示自适应权重,在球员分散站位时设置较大的值,在集中站位时设置较小的值。在初始阶段,权重 λ 的值设置为 0.5,但随着算法训练的进行,这些权重会通过反向传播不断优化和调整。通过训练过程中的反向传播,我们能够根据数据的实际分布和需求动态地调整这些权重,使得模型能够自动学习最合适的权重配置。在融合前,我们对两个关联矩阵进

行预处理,剔除明显不可能的匹配对,最终通过使用匈牙利算法在融合矩阵 C 上求解最优匹配,实现跨镜头间球员身份的关联匹配。

3.2.3 轨迹优化

跨镜头关联完成后,整个视频中的轨迹可能仍存在因长期遮挡或出镜造成的碎片化问题。我们提出一种基于时空约束的持续连接后处理算法,将跨镜头关联后可能存在的碎片化轨迹进行衔接和优化:(1)对于一条结束的轨迹 $Traj_{lost}$,在其结束时间点 t_{end} 之后的一个时间窗口 Δt 内,寻找所有新出现的轨迹段 $Traj_{new}$;(2)计算消失轨迹 $Traj_{lost}$ 的末段运动方向与候选轨迹 $Traj_{new}$ 的起始段运动方向之间的夹角,若夹角小于阈值 θ_c (60°),则认为运动方向一致,保留为候选;(3)对通过运动方向约束的候选轨迹,计算其起始部分与消失轨迹末尾部分的外观特征与姿态特征余弦相似度,如果其相似度高于阈值 θ_{reid} ,则判定为同一球员,将两条轨迹连接为一条完整的轨迹。

通过对全局轨迹的优化衔接,算法能够有效减少身份切换(ID Switch)和轨迹碎片化,最终输出贯穿整场比赛的、身份一致的全局球员轨迹。

4 实验结果与分析

4.1 跨镜头球员跟踪数据集

数据集构建。为填补足球视频跨镜头跟踪研究的数据空白,构建了一个名为 CSoccerBench 的跨镜头足球跟踪数据集。CSoccerBench 数据集选取了数十场高质量职业足球比赛视频,从中提取了 50 个视频序列,所有视频均以 1080p 分辨率、25 帧/秒的标准录制。每个序列均包含典型的球员运动模式和多种镜头切换类型,确保了数据的多样性和代表性。数据集共包含 39906 帧图像、2494 条轨迹和 676751 个标注边界框,覆盖了远镜头、中镜头、近镜头等多种拍摄角度,并针对遮挡、模糊、消失等复杂场景进行了精细化标注。

数据集特点。CSoccerBench 数据集在场景设置和标注粒度上具有显著特点:根据球员站位密度将跨镜头场景分为球员分散站位和球员集中站位两类,每类各 50 个场景;标注中细化了四类遮挡情况(同队遮挡、异队遮挡、多人遮挡、完全遮挡)以及运动模糊和消失属性,并采用实时单目标跟踪器辅助标注,提升标注质量与一致性。与 MOT17^[85]、DanceTrack^[86]、SportsMOT^[9] 等主流多目标跟踪数据

集相比,CSoccerBench 在运动复杂度、外观相似度和遮挡频率方面更具挑战性,尤其适合验证算法在足球视频中的鲁棒性和跨镜头关联能力。如图 7 所示,对 MOT17 和我们的 CSoccerBench 的样本用 t-SNE 将外观特征降维可视化,MOT17 中目标特征聚类明显、区分度高,而 CSoccerMOT 中特征分布混杂、难以区分,体现了 CSoccerBench 数据库在外观相似度方面更具有挑战性。

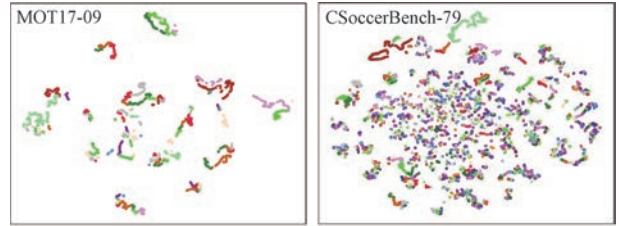


图7 数据集的外观特征可视化对比

评价指标分别针对单镜头跟踪和跨镜头跟踪设定了评价指标,单镜头跟踪采用多目标跟踪指标,包括高阶跟踪准确率(Higher Order Tracking Accuracy, HOTA)^[87]、身份 F1 分数(Identification F1 Score, IDF1)^[88]、关联准确率(Association Accuracy, AssA)^[88]、多目标跟踪准确率(Multiple Object Tracking Accuracy, MOTA)^[88]等,其中 HOTA、IDF1 和 MOTA 作为核心指标以突出算法在身份维持与关联方面的能力。跨镜头跟踪则额外定义了匹配精确率(Match Precision, MP)、匹配召回率(Match Recall, MR)和匹配 F1 值(Match F1-Measure, MF1),用于衡量镜头切换前后球员身份的连续性和一致性。所有指标共同构成了对算法在复杂足球场景下综合性能的系统评估体系。

4.2 实验设置

模型训练与参数设置如下:目标检测任务采用 YOLOX 模型^[20],输入图像尺寸设置为 1440×800 ,使用在 COCO^[89]数据集上预训练的 YOLOX-X 权重进行初始化。训练采用 AdamW 优化器,初始学习率设置为 1×10^{-3} ,通过余弦退火策略进行衰减,衰减因子为 0.1,在训练集上共训练 50 个 epoch。外观特征提取采用 FastReID 模型^[81],输入尺寸为 128×384 ,使用 sbs_R50-ibn 配置文件和 Market1501 数据集预训练权重,初始学习率设为 3.5×10^{-4} ,采用余弦退火学习率调度,批量大小为 64,共训练 58 个 epoch。时空分层模型以 3×10^{-4} 的学习率进行训练,AdamW 优化器的 β_1 和 β_2 参数分别设置为 0.9 和 0.999,权重衰减为 1×10^{-4} ,交叉熵损失权重设为 1,共训练 200 个 epoch,单个 GPU 批

量大小为 4,使用两个 GPU 并行训练,总批量大小为 8。实验环境采用 NVIDIA Tesla V100S PCIe 的 GPU 型号,深度学习框架为 PyTorch 1.9.1,并基于 CUDA 11.1 进行 GPU 的加速计算。

4.3 综合实验对比

本节从单镜头对比和跨镜头对比两个方面对本文构建的 CSTracker 方法和目前最优的跟踪算法进行了对比。单镜头跟踪对比中,首先在本文构建的 CSoccerBench 中,按照每个镜头一个序列的形式,以单镜头跟踪的方式进行了对比,接着在主流的足球跟踪数据集 SportsMOT^[9]和 SoccerNet-Tracking^[8]中进行了对比。跨镜头匹配对比中,由于目前还没有专门针对足球跨镜头跟踪的数据集和相关算法,在 SOTA 算法中构建了基于 IoU 和基于 ReID 的跨镜头匹配进行对比。

单镜头跟踪对比。在本文构建的 CSoccerBench 数据集上的实验结果如表 1 所示,我们的 CSTracker 方法在多项关键指标上表现优异,尤其在身份一致性方面优势显著,其 HOTA、IDF1 和 AssA 分别达到 69.9%、82.9% 和 73.1%,显示出

在球员频繁遮挡与再现的足球场景中出色的身份维持能力。虽然在 MOTA(75.3%)和 DetA(67.0%)等偏重检测性能的指标上略低于 OC-SORT(MOTA 77.7% 和 DetA 67.7%),以及在 SportsMOT 中的 MOTA(93.2%)指标低于其他方法,但这源于 CSTracker 为提升关联稳定性而主动过滤低置信检测框的策略。该策略旨在减少错误的匹配,确保更高的身份一致性,尤其是在目标密集且存在大量干扰的场景下。因此,CSTracker 的低 MOTA 值并不意味着跟踪效果差,反而更符合实际足球分析中对目标身份一致性的高要求,确保了在复杂场景下的鲁棒性和稳定性。

涵盖足球、篮球等多种运动的 SportsMOT^[9]数据集如表 2 所示,CSTracker 展现出优秀的跨场景泛化能力。其 HOTA(74.3%)、IDF1(78.3%)和 AssA(65.9%)全面领先于其他算法,值得注意的是,CSTracker 的 IDF1 指标较第二名高出 3.9 个百分点,表明其不同运动、不同队服外观及运动模式背景下,能充分利用时空上下文信息实现可靠的身份关联,适用于复杂的多体育项目场景。

表 1 单镜头跟踪对比中,各个算法在本文构建的 CSoccerBench 数据集上的表现

方法	HOTA	IDF1	MOTA	AssA	DetA
单阶段:					
CenterTrack ^[37]	62.6	71.1	74.5	60.1	65.4
FairMOT ^[90]	56.8	70.6	71.7	54.4	59.3
TransTrack ^[26]	65.2	75.7	75.8	64.2	66.3
Trackformer ^[27]	64.8	76.5	76.8	63.7	66.1
MOTR ^[28]	66.8	77.6	76.1	67.1	66.7
MOTRv2 ^[29]	67.1	77.6	77.1	66.8	67.5
MOTIP ^[31]	68.5	78.0	77.4	67.9	68.0
SAMBA ^[91]	66.6	76.9	76.2	66.3	67.3
DiffusionTrack ^[17]	64.2	75.6	73.7	64.1	63.4
DiffMOT ^[92]	66.0	75.8	77.3	66.5	65.6
两阶段:					
SORT ^[10]	61.6	73.0	74.9	59.8	63.6
DeepSORT ^[11]	60.7	74.9	73.7	60.2	61.3
TrackMPNN ^[23]	61.8	72.4	73.8	61.6	62.3
GTR ^[24]	64.7	75.9	74.4	64.7	64.7
ByteTrack ^[13]	66.7	76.5	75.6	66.4	67.0
Bot-SORT ^[14]	67.2	77.5	75.4	66.7	67.7
OC-SORT ^[16]	69.0	79.8	77.7	70.3	67.7
MixSORT ^[9]	68.2	79.0	76.0	69.5	67.1
SUSHI ^[25]	69.0	81.2	74.2	72.0	66.2
HybridSORT ^[18]	67.5	77.7	75.1	66.3	67.4
OFTrack ^[19]	69.3	81.5	75.9	72.2	67.5
TrackTrack ^[93]	68.9	80.1	77.0	71.7	67.7
CSTracker	69.9	82.9	75.3	73.1	67.0

注:加粗的结果表示排名第一,下同。

表 2 单镜头跟踪对比中,各个算法在 SportsMOT^[9]数据集上的表现

方法	HOTA	IDF1	MOTA	AssA	DetA
单阶段:					
CenterTrack ^[37]	62.7	60.0	90.8	48.0	82.1
FairMOT ^[90]	49.3	53.5	86.4	34.7	70.2
QDTrack ^[94]	60.4	62.3	90.1	47.2	77.5
TransTrack ^[26]	68.9	71.5	92.6	57.5	82.7
MeMOTR ^[30]	70.0	—	—	57.8	82.0
两阶段:					
GTR ^[24]	54.5	55.8	67.9	45.9	64.8
ByteTrack ^[13]	62.8	69.8	94.1	51.2	77.1
ByteTrack* ^[13]	64.1	71.4	95.9	52.3	78.5
Bot-SORT ^[14]	68.7	70.0	94.5	55.9	84.4
OC-SORT ^[16]	71.9	72.2	94.5	59.8	86.4
OC-SORT* ^[16]	73.7	74.0	96.5	61.5	88.5
MambaTrack ^[95]	72.6	72.8	95.3	60.3	88.8
MambaMOT ^[96]	71.3	71.1	94.9	58.6	86.7
MixSORT ^[9]	74.1	74.4	96.5	62.0	88.5
CSTracker	74.3	78.3	93.2	65.9	83.8

注:方法后的星号(*)表示额外引入了训练集。

在 SoccerNet-Tracking 数据集^[8]上,如表 3 所示,CSTracker 取得了 HOTA(80.14%)、AssA(66.48%)和 DetA(94.72%)的最佳成绩,相比 OC-SORT 分别提升了 2.05%、1.79% 和 0.45%。该数据集场景相对简单、遮挡较少,检测器性能普遍较高,CSTracker 全面提升各项指标,说明其不仅在复杂场景中稳

健,在优质视频条件下也能进一步发挥检测与关联协同优化的优势,具备良好的综合性能和稳定性。

表 3 单镜头跟踪对比中,各个算法在 SoccerNet-Tracking^[8]数据集上的表现

方法	HOTA	AssA	DetA
DeepSORT ^[11]	69.6	58.7	82.6
FairMOT ^[90]	57.9	50.5	66.6
ByteTrack ^[13]	71.5	60.7	84.3
Bot-SORT ^[14]	77.0	63.5	93.5
OC-SORT ^[16]	78.1	64.7	94.3
CSTracker	80.1	66.5	94.7

跨镜头匹配完整结果对比。选用双阶段跟踪算法 Bot-SORT、SUSHI 与 OFTrack 作为参照,测试其在不处理跨镜头关联,及增加基于 IoU 和 ReID (采用 FastReID 模型)的跨镜头关联模块时的表现,结果如表 4 所示。CSTracker 取得了 HOTA 63.7%、IDF1 70.5%、AssA 59.9%的成绩,显著优于其他对比算法。与未处理跨镜头关联的基线模型(如 Bot-SORT 的 HOTA 50.5%)相比,CSTracker 凭借其融合位置与姿态特征的自适应关联机制,在维持身份一致性方面表现出明显优势。而与采用传统 IoU 或 ReID 进行跨镜头关联的改进模型(如 SUSHI+ReID 的 HOTA 得分为 59.3%,OFTrack+ReID 为 59.5%)相比,CSTracker 的领先优势进一

表 4 跨镜头匹配对比中,各个算法在 CSoccerBench数据集上的表现

方法	HOTA	IDF1	MOTA	AssA	DetA
不处理跨镜头的切换:					
SORT ^[10]	48.7	48.5	74.6	39.6	63.3
DeepSORT ^[11]	50.1	53.1	74.0	42.3	62.8
ByteTrack ^[13]	47.7	48.7	75.0	42.4	64.5
Bot-SORT ^[14]	50.5	54.6	76.1	44.4	64.8
OC-SORT ^[16]	52.2	56.5	76.5	45.4	66.0
TrackMPNN ^[23]	49.7	51.3	74.5	43.3	64.1
SUSHI ^[25]	54.3	60.1	77.0	50.9	67.5
SAMBA ^[91]	52.5	57.6	75.8	44.2	65.1
DiffusionTrack ^[17]	49.7	50.3	74.6	44.7	63.2
DiffMOT ^[92]	51.8	56.1	76.0	44.9	65.5
HybridSORT ^[18]	51.1	55.5	74.7	45.6	64.8
OFTrack ^[19]	54.0	59.6	77.1	50.5	67.1
TrackTrack ^[93]	55.1	62.0	77.5	51.6	68.2
处理跨镜头的切换:					
Bot-SORT+IoU	53.9	57.8	76.3	48.1	65.0
Bot-SORT+ReID	54.4	58.4	76.8	49.0	65.4
SUSHI+IoU	58.3	63.9	77.7	54.8	67.8
SUSHI+ReID	59.3	65.1	77.8	57.0	68.3
OFTrack+IoU	58.0	63.2	77.5	54.8	67.3
OFTrack+ReID	59.5	65.3	78.0	57.5	68.5
CSTracker	63.7	70.5	79.1	59.9	70.1

步验证了其专门设计的特征融合框架在处理镜头切换导致的表观与运动突变方面的有效性和必要性,为足球视频全局跟踪提供了可靠的解决方案。

4.4 消融实验分析

4.4.1 单镜头跟踪的消融实验

单镜头跟踪的消融实验在 CSoccerBench 数据集上进行,将每个镜头视为一条独立的跟踪序列,不考虑镜头切换以及跨镜头匹配的情况,用于评估单镜头跟踪设计中各个模块的性能。

各模块评估。为评估单目标跟踪算法设计中各个模块的性能,将状态建模、关系交互和层级关联三个模块依次集成于基准跟踪器进行对比实验,基准跟踪器采用图网络跟踪算法 TrackMPNN^[23]。实验结果如表 5 所示,可以发现各模块的引入均带来性能的持续提升。引入外观特征和时空特征相融合的状态建模,相比于基准算法仅采用外观模型而言性能提升显著,使 HOTA 和 IDF1 分别提升 1.2% 和 1.3%。值得关注的是,引入基于重叠时间窗口的层级关联模块相对于基准算法提升最为显著,在 HOTA 和 IDF1 上提升 3.1% 和 4.9%,表明层级关联策略增强了长时身份一致性。完整的 CSTracker 算法集成所有模块后,在所有评价指标上达到最优性能。结果表明,各模块间存在显著协同效应,状态建模、关系交互与层级关联策略共同作用,显著提升了足球视频多目标跟踪的综合性能。

表 5 单镜头跟踪中,各个模块的评估对比

方法	HOTA	IDF1	MOTA	AssA
基准方法	64.3	74.8	73.0	65.3
基准+状态建模	65.5	76.1	74.2	65.3
基准+关系交互	66.0	76.8	74.6	65.9
基准+层级关联	67.4	79.7	74.9	71.5
CSTracker	69.9	82.9	75.3	73.1

球员的状态表征构建。表 6 评估了球员状态表征中外观特征与运动特征的贡献,其中运动特征由时间信息和空间信息组成。实验表明,在外观特征中单独引入位置信息可使 HOTA、IDF1 和 MOTA 分别提升 0.7%、1.7% 和 1.0%,有效增强了空间定位能力。而单独使用时序运动约束虽提高了关联指标,但导致 MOTA 下降 0.4%,反映出其在抑制误匹配的同时也带来一定的漏检风险。融合两者后,模型在各项指标上均获得稳定提升,证明时空联合建模对复杂足球场景中轨迹连贯性的关键作用。

重叠窗口的层级关联。为评估重叠窗口层级关联模块的性能,本文设计了层级深度与窗口规模的

对比实验。模型采用 1~7 层分层结构,对应的窗口数量为 $[2^1, 2^2, \dots, 2^7]$ 。特征有效性实验通过逐层消融不同特征(位置、外观、运动),以 IDF1 损失评估各特征在不同层级中的贡献度。

表 6 单镜头跟踪中,球员状态表征消融实验

球员状态表征			HOTA	IDF1	MOTA
外观	位置	时间			
✓	—	—	68.7	81.6	74.1
✓	✓	—	69.4	83.3	75.1
✓	—	✓	69.0	82.5	73.7
✓	✓	✓	69.9	84.3	75.3

层数与窗口大小对关联性能的影响如图 8 所示,实验发现单一时间窗口下扩大窗口尺寸会导致性能下降:128 帧窗口的 IDF1 较 32 帧窗口降低 3.3%,表明单纯延长时序范围会引入噪声并破坏局部关联一致性。而采用 7 层重叠窗口结构后,IDF1 显著提升 18.1%,证明多尺度分层机制能有效协调短时匹配精度与长时关联鲁棒性。这表明重叠窗口设计通过保持局部关联连续性,缓解了长时跨度下的身份漂移问题。

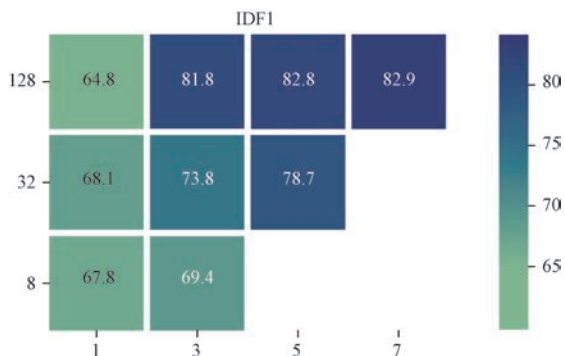


图 8 层数与窗口大小对关联性能的影响

特征在不同层级中的贡献度如图 9 所示,外观特征在所有层级中影响最为显著,其在高层级(4~7 层)的失效会导致 IDF1 下降 7.2%,远大于添加低层级(第二列到第一列)的 1.4%损失,说明高层级关联时更依赖外观特征的身份判别能力。运动特征在各层级均保持稳定贡献(IDF1 损失 0.8%~2.9%),体现出其跨尺度的时序一致性建模作用。位置特征在高层级的影响较弱(损失仅 0.1%),表明其在长时期关联中对几何约束的依赖性降低。

4.4.2 跨镜头匹配的消融实验

跨镜头匹配的消融实验在本文构建的 CSoccer-Bench 数据集上进行实验,其中各模块评估的消融,由于需要体现各个模块在整体算法中的作用,选择在完整的跨镜头跟踪序列中进行实验对比,评估指

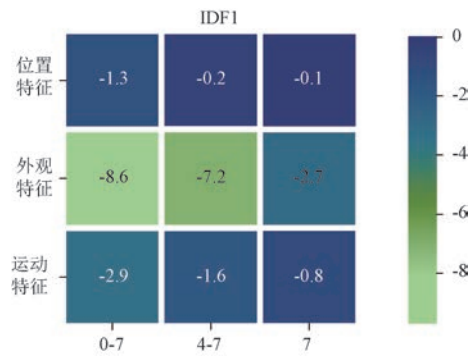


图 9 特征在不同层级中的贡献度

标也采用多目标跟踪的评估。其他各个子模块的消融实验,则是在跨镜头前后最近邻帧间作目标匹配,以此剔除掉长序列跟踪的影响,突出比较跨镜头目标匹配的性能。

各模块评估。将目标跟踪的基准算法 TrackMP-NN 与跨镜头匹配的各模块集成后进行评估,实验结果如表 7 所示,从实验结果中可以看出轨迹优化的后处理单独使用可以提升 HOTA 的 2.8%的性能,体现了该策略在短时轨迹碎片恢复方面的优势。基准方法与融合匹配模块集成后,系统在长时身份保持方面展现出显著优势,AssA 关联性能提升 6.7%。此外,就关联线索而言,位置关联对 AssA 的提升(+8.3%)比姿态关联(+6.3%)更为显著,表明两类特征分别在解决跟踪准确性和身份一致性的不同维度问题上,位置关联更具重要性。最终系统在跨镜头场景中实现 HOTA 63.7%、IDF1 70.5%的综合性能,证明了多模块协同在解决足球视频跨镜头匹配难题中的有效性。

表 7 跨镜头匹配中,各个模块的评估对比

方法	HOTA	IDF1	MOTA	AssA
基准方法	50.3	53.6	71.3	45.2
基准+轨迹优化	53.1	55.2	71.4	46.2
基准+位置关联	59.8	62.3	77.1	53.5
基准+姿态关联	58.8	60.4	76.2	51.5
基准+融合匹配	61.1	68.6	78.0	57.7
CSTracker	63.7	70.5	79.1	59.9

姿态关联。姿态关联模块的实验结果揭示了外观与姿态特征在跨镜头场景中的互补特性。如表 8 所示,基础 ViT 特征(全局特征)仅达到 50.5%的匹配精确率,反映出全局外观特征对视角变化的敏感性。单独使用姿态特征表现出色(MP=54.1%),特别是在处理同队球员外观相似情况时展现出独特优势。将局部外观特征与姿态特征结合后,MP 进一步提升至 60.4%,且召回率(MR)达到 86.7%,

表明两种特征分别解决了匹配精度和覆盖度的关键问题。这一发现为跨镜头关联中的特征设计提供了重要方向:需同时兼顾局部细节感知和运动姿态表征。

表 8 跨镜头匹配中,姿态关联构建的消融分析

方法	MP	MR	MF1
姿态特征	54.1	79.5	64.0
全局特征	50.5	77.1	61.9
姿态+局部特征	60.4	86.7	71.3

信息融合匹配。信息融合策略的比较实验如表 9 所示,揭示了跨镜头关联中的决策机制本质。在仅仅测试跨镜头匹配的准确率上,位置关联与姿态关联与全局跟踪的表现一致,位置关联在决策中仍占据更重要的位置。然而,实验结果表明,尽管步态特征为球员身份区分提供了补充信息,但我们的主要依赖仍是位置特征,特别是在球员站位具备明显区分性的情况下(如球场中的位置特征)。在跨镜头身份关联时,正如正文中的表 9 所示,单独使用姿态特征(这里的姿态特征融合了外观与姿态)时,球员身份匹配的准确率(MP)为 60.4%;而单独使用位置特征时,准确率则为 79.6%。通过融合姿态特征,准确率提升至 86.7%。这一结果凸显了位置特征作为主要区分依据的重要性,尤其是在动态变化的场景中。在融合方式上,固定加权方法虽然能够整合多种特征,但自适应融合策略在 MP 上的表现进一步领先了 4.3%,这一差距凸显了静态加权融合方法的局限性。深入分析发现,自适应融合策略在球员集中站位场景中更多依赖姿态特征,而在分散站位场景中则更侧重于位置特征。此动态调整能力使得模型在复杂场景中保持了稳定的性能,特别是在处理部分遮挡和目标交叉的情景下,表现出更强的鲁棒性。

表 9 跨镜头匹配中,信息融合匹配的消融分析

方法	MP	MR	MF1
位置关联	79.6	91.0	81.4
姿态关联	60.4	86.7	71.3
固定加权	82.4	94.1	85.9
自适应融合	86.7	97.0	91.5

4.5 讨论分析

跟踪速度分析。为了探究 CSTracker 算法引入的跨镜头全局跟踪设计对跟踪速度的影响,对基准算法以及相关图网络匹配算法在模型效率和跟踪速度上做了对比,实验结果如表 10 所示。其中 Params 表示模型参数量,以百万(M)为单位,FLOPs

表示模型的浮点运算次数,以十亿次(G)为单位,FPS 为速度的单位,表示每秒计算的帧数。从实验结果中可以看出,尽管本文方法 CSTracker 的模型复杂度有所增加,但其参数量(80.1M)和计算量(247.6G FLOPs)略高于对比方法,但其跟踪性能取得了显著突破,HOTA 指标达到 63.7,较 SUSHI 和 TrackMPNN 分别大幅提升 9.4 和 14.0 个百分点,相对提升幅度超过 17%。尤为重要的是,尽管计算开销增加,CSTracker 仍保持了 22.1 FPS 的实时处理速度,达到了实际应用中视频分析算法的帧率要求。这表明本文提出的跨镜头全局关联机制以可接受的效率代价换来了性能的质的飞跃,在足球视频全局跟踪这一复杂场景中实现了精度与速度的最佳平衡。

表 10 图网络匹配跟踪算法中的模型效率对比

方法	HOTA	Params	FLOPs	FPS
TrackMPNN ^[23]	49.7	69.2M	197.7G	30.4
SUSHI ^[25]	54.3	77.3M	220.5G	24.5
CSTracker	63.7	80.1M	247.6G	22.1

可视化结果。图 10 对 CSoccerBench-21 序列的跨镜头跟踪结果进行可视化分析,揭示了不同算法在长时身份保持能力上的本质差异。该序列包含 96 帧的镜头切换空窗期,对跟踪器的跨镜头关联能力构成极端挑战。



图 10 跨镜头跟踪结果可视化对比

实验表明,Bot-SORT^[14]在产生镜头切换后仅能重新关联 5 个正确目标,且这些目标在切换前后位移较小,说明其依赖的外观特征和线性运动模型难以应对跨镜头场景的几何与外观突变。SUSHI^[25]则表现出更强的遮挡鲁棒性,成功关联 8 个目标,但

其性能在球员密集区域显著下降,反映出基于外观的图模型对空间拥挤和尺度变化的敏感性。相比之下,CSTracker在镜头切换后保持了所有球员的身份连续性。深入分析发现,该算法在目标位移剧烈、外观变化显著的场景中仍能正确关联,证明了其融合位置映射与姿态特征的多模态策略的有效性。特别值得注意的是,该算法在球员高度聚集区域(如禁区争抢场景)依然保持稳定性能,说明其设计的自适应特征融合机制能够有效应对复杂交互场景下的身份混淆问题。这些可视化结果从实证角度验证了:在跨镜头跟踪中,单一特征源(如外观或运动)具有固有局限性,而结合几何位置约束与姿态语义的表征学习是实现鲁棒全局关联的关键突破点。

5 结束语

足球比赛的视频分析中,跨镜头全局跟踪长期缺乏有效研究,而这一任务在实际应用中却至关重要。为应对镜头切换频繁、球员高速运动及外观相似等复杂干扰,本文设计了CSTracker算法,从两个方面优化跟踪性能:在单镜头内,通过球员关系交互模型与层级关联策略增强轨迹连续性与鲁棒性;在跨镜头匹配中,结合位置预测与姿态特征实现精准匹配与轨迹修正。实验结果表明,CSTracker不仅在新构建的CSoccerBench数据集上显著优于现有方法,而且在公开的SoccerNet-Tracking与SportsMOT数据集上也取得了最优性能,验证了方法的优异性能。这一研究不仅填补了跨镜头足球跟踪领域的空白,而且在实际应用中具有重要价值,可以辅助教练进行战术分析、提升裁判判罚准确性,并增强观众的观赛体验。未来的工作将进一步探索多模态信息融合与大规模数据训练,以提升跨镜头跟踪的泛化能力和实战应用水平,推动足球智能分析向更高层次发展。

致 谢 感谢国家自然科学基金青年基金项目(No. 62272184)、国家自然科学基金面上项目(No. 62402189)、中国博士后科学基金资助项目(2024M751012、GZC20230894、2025T180429)、湖北省博士后创新人才培养项目(2024HBBHCXB014)资助。

参 考 文 献

[1] Zhang P, Lei W M, Zhao X L, et al. A survey on multi-target

multi-camera tracking methods. Chinese Journal of Computers, 2024, 47(2): 287-309 (in Chinese)

(张鹏, 雷为民, 赵新蕾等. 跨摄像头多目标跟踪方法综述. 计算机学报, 2024, 47(2): 287-309)

[2] Han R Z, Feng W, Guo Q, et al. Single object tracking research: A survey. Chinese Journal of Computers, 2022, 45(9): 1877-1907 (in Chinese)

(韩瑞泽, 冯伟, 郭青等. 视频单目标跟踪研究进展综述. 计算机学报, 2022, 45(9): 1877-1907)

[3] Song Z, Yu J, Chen Y P P, et al. Transformer tracking with cyclic shifting window attention//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 8791-8800

[4] Song Z, Tang Y, Luo R, et al. Autogenic language embedding for coherent point tracking//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia, 2024: 2021-2030

[5] Song Z, Luo R, Yu J, et al. Compact transformer tracker with correlative masked modeling//Proceedings of the AAAI Conference on Artificial Intelligence; Vol: 37. Washington, USA, 2023: 2321-2329

[6] Yu J, Lei A, Song Z, et al. Comprehensive Dataset of Broadcast Soccer Videos//2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). Miami, USA, 2018: 418-423

[7] Feng N, Song Z, Yu J, et al. SSET: a dataset for shot segmentation, event detection, player tracking in soccer videos. Multimedia Tools and Applications, 2020, 79(1): 28971-28992

[8] Giancola S, Amine M, Dghaily T, et al. SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos//Proceedings of the CVPRW. Salt Lake City, USA, 2018

[9] Cui Y, Zeng C, Zhao X, et al. Sportsmot: A large multi-object tracking dataset in multiple sports scenes//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France, 2023: 9921-9931

[10] Bewley A, Ge Z, Ott L, et al. Simple online and realtime tracking//2016 IEEE International Conference on Image Processing (ICIP). Phoenix, USA, 2016: 3464-3468

[11] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric//2017 IEEE International Conference on Image Processing (ICIP). Beijing, China, 2017: 3645-3649

[12] Du Y, Song Y, Yang B, et al. Strongsort: Make deepsort great again. arXiv preprint arXiv:2202.13514, 2022

[13] Zhang Y, Sun P, Jiang Y, et al. Bytetrack: Multi-object tracking by associating every detection box//Proceedings of the ECCV. Tel Aviv, Israel, 2022: 1-21

[14] Aharon N, Orfaig R, Bobrovsky B Z. BoT-SORT: Robust associations multi-pedestrian tracking. arXiv preprint arXiv: 2206.14651, 2022

[15] Zhao Z, Wu Z, Zhuang Y, et al. Tracking objects as pixel-wise distributions//Proceedings of the ECCV. Tel Aviv, Is-

- rael,2022;76-94
- [16] Cao J, Weng X, Khirodkar R, et al. Observation-centric sort: Rethinking sort for robust multi-object tracking. *arXiv preprint arXiv:2203.14360*, 2022
- [17] Luo R, Song Z, Ma L, et al. Diffusion track: Diffusion model for multi-object tracking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024,38(5):3991-3999
- [18] Yang M, Han G, Yan B, et al. Hybrid-sort: Weak cues matter for online multi-object tracking//*Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, Canada,2024,38:6504-6512
- [19] Song Z, Luo R, Ma L, et al. Temporal coherent object flow for multi-object tracking//*Proceedings of the AAAI Conference on Artificial Intelligence*. Philadelphia, USA,2025,39:6978-6986
- [20] Ge Z, Liu S, Wang F, et al. YOLOx: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021
- [21] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors//*Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022;7464-7475
- [22] Bishop G, Welch G, et al. An introduction to the kalman filter. *Proc of SIGGRAPH*, Course, 2001,8(27599-23175):41
- [23] Rangesh A, Maheshwari P, Gebre M, et al. Trackmpnn: A message passing graph neural architecture for multi-object tracking. *arXiv preprint arXiv:2101.04206*, 2021
- [24] Zhou X, Yin T, Koltun V, et al. Global tracking transformers//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, USA,2022;8771-8780
- [25] Cetintas O, Brasó G, Leal-taixé L. Unifying short and long-term tracking with graph hierarchies//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada,2023;22877-22887
- [26] Sun P, Cao J, Jiang Y, et al. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020
- [27] Meinhardt T, Kirillov A, Leal-taixé L, et al. Trackformer: Multi-object tracking with transformers//*Proceedings of the CVPR*. New Orleans, USA,2022;8844-8854
- [28] Zeng F, Dong B, Zhang Y, et al. Motr: End-to-end multiple-object tracking with transformer//*Proceedings of the ECCV*. Tel Aviv, Israel,2022;659-675
- [29] Zhang Y, Wang T, Zhang X. MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, Canada, 2023;22056-22065
- [30] Gao R, Wang L. MeMOTR: Long-term memory-augmented transformer for multi-object tracking//*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France,2023;9901-9910
- [31] Gao R, Qi J, Wang L. Multiple object tracking as ID prediction//*Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. Nashville, USA,2025;27883-27893
- [32] Jiang J, Wang Z, Zhao M, et al. SAM2MOT: A novel paradigm of multi-object tracking by segmentation. *arXiv preprint arXiv:2504.04519*, 2025
- [33] Yu E, Wang T, Li Z, et al. Motrv3: Release-fetch supervision for end-to-end multi-object tracking. *arXiv preprint arXiv:2305.14298*, 2023
- [34] Bergmann P, Meinhardt T, Leal-taixé L. Tracking without bells and whistles//*Proceedings of the ICCV*. Seoul, Republic of Korea,2019;941-951
- [35] Xu Y, Osep A, Ban Y, et al. How to train your deep multi-object tracker//*Proceedings of the CVPR*. Seattle, USA, 2020;6787-6796
- [36] Guo S, Wang J, Wang X, et al. Online multiple object tracking with cross-task synergy//*Proceedings of the CVPR*. Nashville, USA,2021;8136-8145
- [37] Zhou X, Koltun V, Krähenbühl P. Tracking objects as points//*Proceedings of the ECCV*. Glasgow, UK, 2020;474-490
- [38] Xu Y, Ban Y, Delorme G, et al. TransCenter: Transformers with dense representations for multiple-object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022,45(6):7820-7835
- [39] Fleuret F, Berclaz J, Lengagne R, et al. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008,30(2):267-282
- [40] 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
- [41] Leal-taixé L, Pons-moll G, Rosenhahn B. Branch-and-price global optimization for multi-view multi-target tracking//*2012 IEEE Conference on Computer Vision and Pattern Recognition*. Providence, USA,2012;1987-1994
- [42] Cheng C C, Qiu M, Chiang C K, et al. ReST: A reconfigurable spatial-temporal graph model for multi-camera multi-object tracking//*2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France,2023;10017-10026
- [43] Peng P, Tian Y, Wang Y, et al. Robust multiple cameras pedestrian detection with multi-view Bayesian network. *Pattern Recognit.*, 2015,48:1760-1772
- [44] Teepe T, Wolters P, Gilg J, et al. Lifting multi-view detection and tracking to the bird's eye view//*2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, USA,2024;667-676
- [45] Xu Y, Liu X, Liu Y, et al. Multi-view people tracking via hierarchical trajectory composition//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA, 2016;2145-2154

- tion (CVPR). Las Vegas, USA, 2016:4256-4265
- [46] Baqué P, Fleuret F, Fua P V. Deep occlusion reasoning for multi-camera multi-target detection//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017:271-279
- [47] Chavdarova T, Fleuret F. Deep multi-camera people detection//Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). Cancun, Mexico, 2017:848-853
- [48] 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*, 2016, 122:313-333
- [49] Hou Y, Zheng L, Gould S. Multiview detection with feature perspective transformation. *arXiv preprint arXiv:2007.07247*, 2020
- [50] You Q, Jiang H. Real-time 3D deep multi-camera tracking. *arXiv preprint arXiv:2003.11753*, 2020
- [51] 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
- [52] Lee W Y, Jovanov L, Philips W. Multi-view target transformation for pedestrian detection//Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). Waikoloa, USA, 2023:1-10
- [53] Kuo C H, 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 (ECCV). Heraklion, Greece, 2010
- [54] Cai Y, Medioni G G. Exploring context information for inter-camera multiple target tracking//Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). Steamboat Springs, USA, 2014:761-768
- [55] 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
- [56] Lee Y G, Tang Z, Hwang J N. Online-learning-based human tracking across non-overlapping cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018, 28: 2870-2883
- [57] Jiang N, Bai S Y, 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
- [58] Li P, Zhang J, Zhu Z, et al. State-aware re-identification feature for multi-target multi-camera tracking//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, USA, 2019:1506-1516
- [59] Jeon Y, Tran D Q, Park M, et al. Leveraging future trajectory prediction for multi-camera people tracking//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver, Canada, 2023:5399-5408
- [60] Sakaguchi S, Amagasaki M, Kiyama M, et al. Multi-camera people tracking with spatio-temporal and group considerations. *IEEE Access*, 2024, 12:36066-36073
- [61] Liu W, Camps O I, Sznaiier M. Multi-camera multi-object tracking. *arXiv preprint arXiv:1709.07065*, 2017
- [62] Tesfaye Y T, Mequanint E Z, Prati A, et al. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. *arXiv preprint arXiv:1706.06196*, 2017
- [63] Ristani E, Tomasi C. Features for multi-target multi-camera tracking and re-identification//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018:6036-6046
- [64] Hou Y, Zheng L, Wang Z, et al. Locality aware appearance metric for multi-target multi-camera tracking. *arXiv preprint arXiv:1911.12037*, 2019
- [65] Yoon K, Song Y, Jeon M. Multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views. *IET Image Process*, 2018, 12:1175-1184
- [66] 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
- [67] 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 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA, 2021:13779-13788
- [68] Nguyen D M H, Henschel R, Rosenhahn B, et al. LMGP: Lifted multicut meets geometry projections for multi-camera multi-object tracking//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA, 2021:8856-8865
- [69] Fan X, Zheng H. An Instance-level motion-aware graph model for multi-target multi-camera tracking//Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN). Yokohama, Japan, 2024:1-8
- [70] Seo Y, Choi S, Kim H, et al. Where are the ball and players? Soccer game analysis with color based tracking and image mosaick//Proceedings of the International Conference on Image Analysis and Processing. Florence, Italy, 1997
- [71] Yoon H S, Bae Y J, Yang Y K. A soccer image sequence mosaicking and analysis method using line and advertisement board detection. *ETRI Journal*, 2002, 24(6):443-454
- [72] Ekin A, Tekalp A M. Robust dominant color region detection and color-based applications for sports video//Proceedings 2003 International Conference on Image Processing (ICIP). Barcelona, Spain, 2003, 1:1-21
- [73] Filters U P. Tracking football player movement from a single moving camera//Proceedings of the 3rd European Conference on Visual Media Production (CVMP). London, UK, 2006: 29-37
- [74] Liu J, Tong X, Li W, et al. Automatic player detection, labeling and tracking in broadcast soccer video. *Pattern Recog-*

- dition Letters, 2009,30:103-113
- [75] Mazzeo P L, Spagnolo P, Leo M, et al. Visual players detection and tracking in soccer matches//Proceedings of the 2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance. Santa Fe, USA,2008;326-333
- [76] Song Z, Wan Z, Yuan W, et al. Distractor-aware tracker with a domain-special optimized benchmark for soccer player tracking//Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR). Taipei, China,2021
- [77] Nillius P, Sullivan J, Carlsson S. Multi-target tracking-linking identities using bayesian network inference//Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). New York, USA, 2006(2):2187-2194
- [78] Sullivan J, Carlsson S. Tracking and labelling of interacting multiple targets//Proceedings of the European Conference on Computer Vision (ECCV). Graz, Austria,2006
- [79] Cui Y, Zeng C, Zhao X, et al. SportsMOT: A large multi-object tracking dataset in multiple sports scenes//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France,2023;9887-9897
- [80] Cui Y, Jiang C, Wang L, et al. MixFormer: End-to-end tracking with iterative mixed attention//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA,2022;13608-13618
- [81] He L, Liao X, Liu W, et al. Fastreid: A pytorch toolbox for general instance re-identification//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada,2023;9664-9667
- [82] Ji X, Song Z, Yu J. Player movement data analysis on soccer field reconstruction. Journal of Beijing University of Aeronautics and Astronautics, 2022,48(8):1543-1552
- [83] Miao J, Wu Y, Liu P, et al. Pose-guided feature alignment for occluded person re-identification//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Republic of Korea,2019
- [84] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale//Proceedings of the International Conference on Learning Representations (ICLR). Virtual, Vienna, Austria,2021
- [85] Milan A, Leal-taixé L, Reid I, et al. MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831,2016
- [86] Sun P, Cao J, Jiang Y, et al. Dancetrack: Multi-object tracking in uniform appearance and diverse motion//Proceedings of the CVPR. New Orleans, USA,2022;20993-21002
- [87] Luiten J, Osep A, Dendorfer P, et al. Hota: A higher order metric for evaluating multi-object tracking. International Journal of Computer Vision, 2021,129:548-578
- [88] Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing, 2008,2008:1-10
- [89] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context//Proceedings of the European Conference on Computer Vision (ECCV). Zurich, Switzerland, 2014;740-755
- [90] 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, 2021,129:3069-3087
- [91] Segu M, Piccinelli L, Li S, et al. Samba: Synchronized set-of-sequences modeling for multiple object tracking. arXiv preprint arXiv:2410.01806, 2024
- [92] Lv W, Huang Y, Zhang N, et al. DiffMOT: A Real-time Diffusion-based Multiple Object Tracker with Non-linear Prediction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA,2024: 19321-19330
- [93] Shim K, Ko K, Yang Y, et al. Focusing on tracks for online multi-object tracking//Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). Nashville, USA,2025;11687-11696
- [94] Pang J, Qiu L, Li X, et al. Quasi-dense similarity learning for multiple object tracking//Proceedings of the CVPR. Virtual, Nashville, USA,2021;164-173
- [95] Xiao C, Cao Q, Luo Z, et al. Mambatrack: a simple baseline for multiple object tracking with state space model//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia,2024;4082-4091
- [96] Huang H W, Yang C Y, Chai W, et al. Exploring learning-based motion models in multi-object tracking. arXiv preprint arXiv:2403.10826, 2024



SONG Zi-Kai, Ph. D. His research interests include visual motion analysis and social network analysis.

MA Lin-Tao, M. S. His research interests include visual object tracking.

TANG Ying, Ph. D. candidate. Her research interests include multi-task

learning and feature analysis and fusion.

YANG Wei, Ph. D. , associate research fellow. His research interests include 3D vision.

YU Jun-Qing, Ph. D. , professor. His research interests include intelligent media computing and cyberspace security.

Background

Soccer video analysis is a prominent research area in computer vision, particularly in multi-object tracking (MOT) within dynamic, real-world environments. Continuous tracking of players throughout a match is essential for tactical evaluation, performance analysis, and automated highlight generation. While traditional MOT methods perform well in single-camera scenarios, broadcast soccer footage poses significant challenges: frequent camera switches, motion blur, occlusion, and high visual similarity among players often lead to identity switches, fragmented trajectories, and incomplete data, limiting the reliability of global match analytics.

Internationally, deep learning-based single-shot trackers have achieved impressive results on standard datasets by leveraging spatiotemporal interaction modeling, appearance embedding, and motion prediction. However, cross-shot tracking remains largely underexplored. Existing approaches often fail to maintain identity consistency across camera transitions or rely on post-hoc corrections, resulting in limited applicability to real-world soccer broadcasts. Public datasets such as SoccerNet-Tracking and SportsMOT provide benchmarks for single-shot tracking but are insufficient for evalua-

ting large-scale, multi-shot scenarios.

To address these limitations, We propose CSTracker, a cross-shot global tracking framework that preserves continuous player trajectories across multiple camera views. CSTracker introduces two key innovations: enhancing single-shot tracking through a hierarchical model with overlapping time windows and relational message passing, and achieving accurate cross-shot association via fused positional and pose-based features with boundary trajectory reconstruction. This integrated approach effectively mitigates occlusion, appearance ambiguity, and motion variations, achieving high accuracy and recall in cross-shot association and substantial improvements in global tracking metrics.

This research was supported by the National Natural Science Foundation of China(No.62272184,62402189), China Postdoctoral Science Foundation under Grant Nos. 2024M751012, GZC20230894, 2025T180429, and the Postdoctor Project of Hubei Province under Grant Number 2024HBBHCXB014. This work presents the first global cross-shot tracking framework for soccer, enabling continuous player trajectories across multiple shot views.