

分类号\_\_\_\_\_

学号 M201971982

学校代码 10487

密级\_\_\_\_\_

华中科技大学  
硕士学位论文  
(学术型■ 专业型□)

复杂场景高性能多目标跟踪技术研究

学位申请人：张一夫

学科专业：信息与通信工程

指导教师：刘文予 教授

答辩日期：2022年5月18日

**A Dissertation Submitted in Partial Fulfillment of the Requirements  
for the Master Degree in Engineering**

**High-Performance Multi-Object Tracking Techniques in  
Complex Environments**

**Candidate : ZHANG Yifu**

**Major : Information and Communication  
Engineering**

**Supervisor : Prof. Wenyu Liu**

**Huazhong University of Science and Technology**

**Wuhan 430074, P. R. China**

**May, 2021**

## 独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期： 年 月 日

## 学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保 密口，在\_\_\_\_\_年解密后适用本授权书。

本论文属于 不保密口。

(请在以上方框内打“√”)

学位论文作者签名：

指导教师签名：

日期： 年 月 日

日期： 年 月 日

# 华中科技大学硕士学位论文

---

## 摘要

在计算机视觉领域中，多目标跟踪是一个非常重要的研究课题，其目的是得到视频中感兴趣物体的运动轨迹，包括每帧中代表物体位置信息的包围框和代表物体身份信息的编号。多目标跟踪有广泛的应用价值，是安防监控、自动驾驶、智慧城市等应用中的关键技术。多目标跟踪中包含目标检测、行人重识别等其他研究方向的技术，其中，目标检测常用来获取目标位置，行人重识别常用来得到目标的身份。当前主流多目标跟踪方法大多遵循“先检测后跟踪”的范式，即先使用一个目标检测网络得到视频帧中物体的位置，再根据物体的外观、运动、位置等信息使用数据关联的方法将当前帧的检测结果和之前帧的跟踪结果匹配起来，得到当前帧目标的身份。

多目标跟踪大多对行人进行跟踪，场景较为复杂，包括遮挡、拥挤、运动模糊、相机运动快、场景多样化等难点。本文从网络模型、数据关联和多视角信息三个不同的方面来解决这些难点问题。在网络模型方面，本文提出了一个基于中心点特征的联合检测和重识别网络，速度达到实时，能够公平并很好地同时优化两个任务，并且使用一个特征点代表一个物体，在拥挤和遮挡场景下效果相较于框的物体表示性能提升明显。在数据关联方面，本文提出一种高低分检测框层次关联方法，从低分检测框中利用跟踪轨迹和检测框的相似度，从低分框中找到遮挡、模糊等困难样本，大幅减少漏检并且保持跟踪轨迹的连贯性。在多视角信息方面，本文提出一个基于体素的三维网络，利用多视角信息得到行人在三维空间中的位置和姿态，并融合同一个人在不同视角下的重识别特征，利用二维的外观信息和三维的位置信息完成跟踪，解决目标在单个视角下的遮挡问题。

本文提出的高性能多目标跟踪技术在多个主流多目标跟踪评测集上均达到业内领先的水平，大幅领先之前方法，在不同的数据集上精度比目前最好方法提升 3 到 21 个百分点不等，速度首次在单块 Nvidia V100 显卡上达到每秒 30 帧，在复杂场景下可以得到鲁棒的结果。大量实验也证明了本文方法的有效性。

# 华 中 科 技 大 学 硕 士 学 位 论 文

---

---

**关键词：** 多目标跟踪； 目标检测； 行人重识别； 人体姿态估计； 数据关联

## Abstract

Multi-object tracking is a long-standing yet active goal in computer vision, which aims at estimating the trajectories of the interested objects in videos. The trajectory includes the bounding box and identity of the object. Multi-object tracking has a wide range of application, including surveillance, autonomous driving and intelligent city. Some techniques from other research areas such as object detection and person re-identification are applied in multi-object tracking. Object detection acquires the object location and person re-identification obtains the object identity. Most methods follow the “tracking-by-detection” paradigm, which first detects the bounding boxes of objects and then associate them according to appearance, motion and location cues in time to obtain the identities.

Multi-object tracking is mostly aimed at pedestrians. The scene is very complex, including some challenges such as occlusion, crowded objects, motion blur, camera motion and diversified scenes. This dissertation deals with the challenges from three different perspectives including the network structure, data association and multi-view information. From the perspective of network structure, this dissertation proposes a point-based joint detection and re-identification which runs at real-time speed. It can fairly and effectively optimize both tasks. Moreover, the point-based representation of an object shows superior advantages over box-based representation in crowded scenes. From the perspective of data association, this dissertation proposes a hierarchical data association based on both high score and low score detection boxes, which seeks the occluded or blurred objects from the low score detection boxes using motion similarity. The proposed approach significantly reduces missing detections and keeps the consistency of trajectories. From the perspective of multi-view information, this dissertation proposes a voxel-based 3D network to estimate the location and pose of people in the 3D space. The re-identification features of the same person from different views are fused and then utilized for tracking. The proposed approach is able to solve the occlusion cases in each single view.

The proposed high-performance multi-object tracking techniques achieve

# 华中科技大学硕士学位论文

---

state-of-the-art performance on several multi-object tracking benchmarks and remarkably outperforms previous methods by 3 to 21 points. For the first time, it achieves 30 fps running speed on a Nvidia V100 GPU. It can also obtain very robust results in complex environments. The comprehensive experiments also prove the effectiveness of the techniques.

**Key words:** Multi-Object Tracking, Object Detection, Person Re-Identification, Human Pose Estimation, Data Association

# 华中科技大学硕士学位论文

## 英文缩写对照表

英文名称	英文缩写	中文名称
Ambiguity Resolution Network	ARN	歧义解决网络
Convolution Neural Network	CNN	卷积神经网络
Deep Layer Aggregation	DLA	深层特征融合
False Accept Rate	FAR	错误接受率
Fully Connected	FC	全连接
False Negative	FN	假阴性
False Positive	FP	假阳性
Feature Pyramid Network	FPN	特征金字塔网络
Identity F1 score	IDF1	身份的 F1 值
Identity Switches	IDs	身份跳变
Intersection over Union	IoU	交并比
Joint Estimation Network	JEN	关节点估计网络
Multi-Object Tracking	MOT	多目标跟踪
Multi-Object Tracking Accuracy	MOTA	多目标跟踪精度
Mean Per Joint Position Error	MPJPE	所有节点平均位置误差

# 华中科技大学硕士学位论文

---

Non-Maximum Suppression	NMS	非极大值抑制
Person Re-Identification	Re-ID	行人重识别
state-of-the-art	SOTA	最先进的
True Positive Rate	TPR	真阳性率

# 华中科技大学硕士学位论文

---

## 目 录

<b>摘 要</b> .....	I
<b>ABSTRACT</b> .....	III
<b>英文缩写对照表</b> .....	V
<b>1 绪论</b>	
1.1 研究背景与意义 .....	(1)
1.2 多目标跟踪国内外研究现状 .....	(2)
1.3 多目标跟踪目前存在的问题 .....	(13)
1.4 本文主要内容和贡献 .....	(16)
<b>2 公平对待目标检测和行人重识别的单阶段多目标跟踪网络</b>	
2.1 研究动机 .....	(19)
2.2 主要思路 .....	(19)
2.3 模型结构 .....	(20)
2.4 模型训练和推理 .....	(23)
2.5 数据关联 .....	(24)
2.6 实验结果与分析 .....	(25)
2.7 本章小结 .....	(34)
<b>3 高低分检测框层次关联的多目标跟踪方法</b>	
3.1 研究动机 .....	(36)
3.2 主要思路 .....	(37)
3.3 目标检测模型 .....	(38)
3.4 重识别模型 .....	(38)

# 华 中 科 技 大 学 硕 士 学 位 论 文

---

3.5	数据关联方法 .....	(40)
3.6	实验结果 .....	(43)
3.7	本章小结 .....	(54)
<b>4 基于体素特征的多视角三维人体姿态估计和跟踪</b>		
4.1	研究动机 .....	(55)
4.2	主要思路 .....	(55)
4.3	三维姿态估计 .....	(57)
4.4	遮挡感知的人体跟踪 .....	(61)
4.5	实验结果 .....	(63)
4.6	本章小结 .....	(74)
<b>5 总结与展望</b>		
5.1	全文总结 .....	(76)
5.2	展望 .....	(76)
<b>致谢 .....</b> (78)		
<b>参考文献 .....</b> (79)		
<b>附录 1 攻读硕士学位期间取得的研究成果 .....</b> (92)		

## 1 绪论

### 1.1 研究背景与意义

#### 1.1.1 计算机视觉

计算机视觉（Computer Vision）是当下非常火热的一个研究领域，它的目的是赋予计算机识别图像或者视频中的内容的能力。随着深度学习（Deep Learning）和卷积神经网络（Convolution Neural Network, CNN）<sup>[1-3]</sup>出现，计算机视觉发展迅速，各种算法的性能也得到了显著的提升。计算机视觉在人们的生活中已经得到了广泛的应用，例如人脸识别、客流统计、图像搜索等。近年来，研究者们正尝试利用更先进的算法和运算平台让计算机进行越来越复杂的任务，例如智慧城市、自动驾驶、虚拟现实等等。然而计算机距离人类的水平还有很大一段差距，在一些复杂的场景下，计算机视觉算法的准确度还不能令人满意。同时，在将算法落地应用时，运行效率也是非常值得关注的一个问题。

#### 1.1.2 多目标跟踪

多目标跟踪（Multi-Object Tracking, MOT）<sup>[4-6]</sup>是计算机视觉一个重要分支，其目的是得到一段视频中感兴趣物体的运动轨迹，包括代表物体位置信息的包围框（bounding box）和代表物体身份信息的编号（identity），同一个物体的编号在视频中应当保持不变，如图 1-1 所示，框代表人的位置，框上面的数字代表人的身份，为了更好的视觉效果，不同颜色的框代表不同身份的人。目前实际中比较有应用价值的是对行人和车辆进行跟踪。多目标跟踪在计算机视觉中具有广泛应用<sup>[7,8]</sup>，例如视频内容分析、人体行为识别需要先得到视频中目标的编号，再分别对每个目标进行动作识别和分析。在智慧城市中也需要应用多目标跟踪技术，比如得到经过红绿灯路口的行人和车辆的轨迹并进行计数，再动态的调节红绿灯时长，让城市高效率运转。近年来自动驾驶中同样也要用到多目标跟踪的技术，自动驾驶车辆的感知技术需要对周围行人、车辆等物体进行跟踪并预测行进轨迹，再做决策。本文主要研究复杂场景下的高效率多目标跟踪技术。

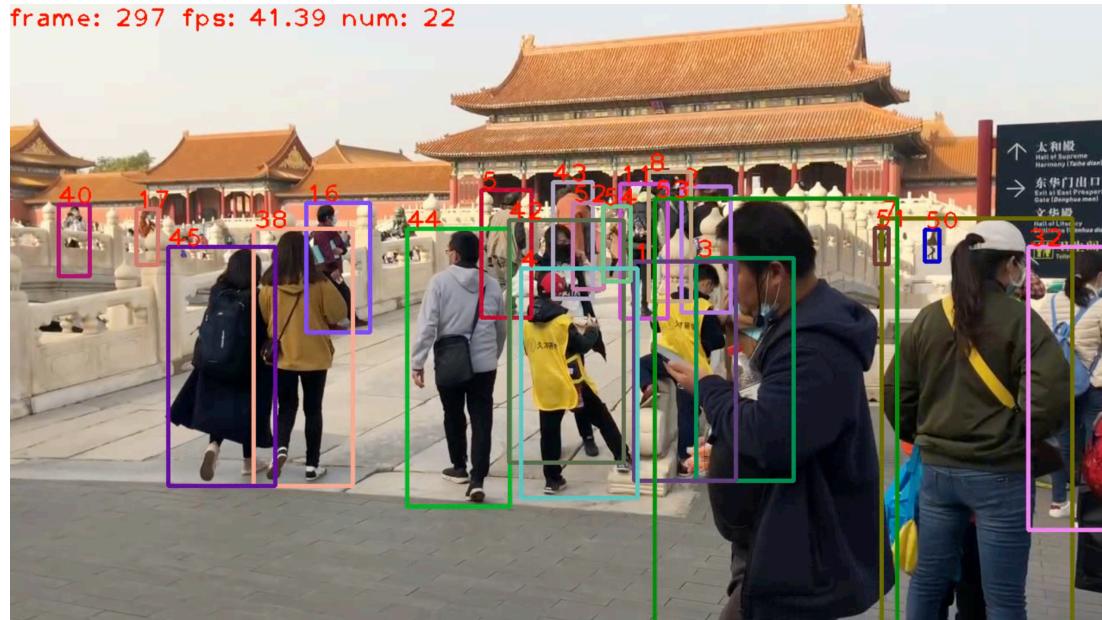


图 1-1 多目标跟踪示意图

## 1.2 多目标跟踪国内外研究现状

多目标跟踪是一个较为复杂的研究方向，包含了许多其他研究方向的技术，例如目标检测（Object Detection）、行人重识别（Person Re-Identification）、视频目标检测（Video Object Detection）、人体姿态估计（Human Pose Estimation）等。本节将首先介绍主流多目标跟踪框架以及各个组件，再介绍其他一些与多目标跟踪有联系的方向的技术及它们之间的关联。

### 1.2.1 多目标跟踪主流框架

当前主流多目标跟踪方法遵循“先检测后跟踪（tracking-by-detection）”的范式，即先使用一个目标检测模型得到当前帧所有目标的检测框，如图 1-2 所示，再将当前帧的检测框（detection bounding box）与之前帧的跟踪框（tracklet）进行数据关联（data association），得到当前帧检测框的编号（identity），完成跟踪。数据关联是多目标跟踪中的独有且核心的问题，如图 1-3 所示。



图 1-2 实际采集视频目标检测示意图

DeepSORT<sup>[9]</sup>是多目标跟踪中非常经典的一个方法，在数据关联时，它同时考虑了物体的外观（appearance）、运动（motion）和位置（location）信息。在得到检测框之后，DeepSORT 用行人重识别网络来提取每个行人的外观特征，并用卡尔曼滤波（Kalman filter）<sup>[10]</sup>作为运动模型来预测每个行人在下一帧可能出现的位置。在计算检测框和跟踪框之间的相似度时，DeepSORT 同时考虑了外观相似度和运动相似度。外观相似度由检测框和跟踪框之间的外观特征的余弦距离（Cosine Distance）计算得到，运动相似度由检测框和跟踪框经过卡尔曼滤波得到的预测框的马氏距离（Mahalanobis Distance）或者交并比（Intersection over Union, IoU）得到。最终的相似度通过外观相似度和运动相似度加权求和得到，在得到相似度矩阵之后，通过一个简单的匈牙利算法（Hungarian Algorithm）<sup>[11]</sup>便能得到当前帧检测框的编号，完成跟踪。

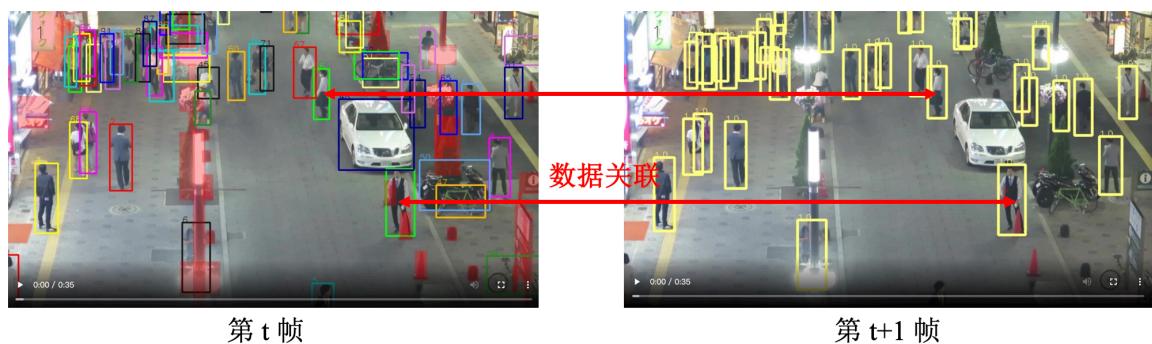


图 1-3 多目标跟踪中数据关联示意图

## 1.2.2 多目标跟踪中的目标检测模型

目标检测<sup>[12-14]</sup>是计算机视觉领域一个经典的研究方向，它也是多目标跟踪的基础。主流多目标跟踪数据集 MOT17<sup>[15]</sup>提供了三种不同检测模型得到的检测结果，包

括 DPM<sup>[16]</sup>, Faster R-CNN<sup>[17]</sup>和 SDP<sup>[18]</sup>。许多多目标跟踪方法<sup>[19-21]</sup>都在这些检测结果的基础上进行，提供这些公开的检测结果可以公平地比较不同跟踪方法的数据关联能力。

随着目标检测的发展，越来越多的多目标跟踪方法开始使用更加高效的目标检测模型来得到更高的跟踪精度。单阶段的目标检测模型 RetinaNet<sup>[23]</sup>开始被一些多目标跟踪方法使用，该方法提出了 Focal Loss 来平衡单阶段模型中正负样本不均衡的问题，提高了单阶段检测模型的精度。CenterNet<sup>[24]</sup>是一个非常简单且高效的单阶目标段检测模型，使用热力图（heatmap）的形式学习每个物体的中心点坐标，再用一个额外的网络分支回归出物体的宽和高，得到物体最终的包围框，如图 1-4 所示。YOLO 系列的检测模型<sup>[25-27]</sup>在经过网络模型、数据增强、训练策略等方面的高度优化之后精度和速度都达到了非常高的水平，因此也被越来越多的跟踪方法使用。目前大多数方法都是对单帧图像进行目标检测，再进行数据关联。

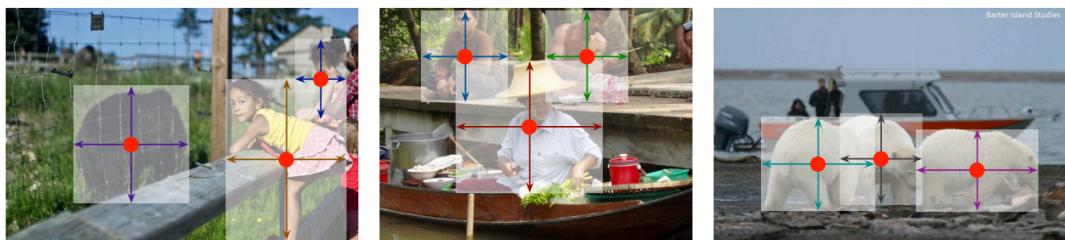


图 1-4 CenterNet 检测物体方法图<sup>[24]</sup>

### 1.2.3 数据关联中的相似度计算方法

在数据关联中，首先要计算当前帧的检测结果和之前帧的跟踪轨迹两两之间的相似度，再根据相似度对检测结果和跟踪轨迹进行匹配。相似度计算是数据关联中的核心模块，位置、运动、外观是相似度计算中的三个有力的参考。SORT<sup>[28]</sup>使用了一个非常简单的方法将运动信息和位置信息结合起来计算相似度，它首先使用卡尔曼滤波<sup>[10]</sup>预测跟踪轨迹在当前帧中的位置，接着计算当前帧的检测框和预测框之间的交并比 IoU 作为相似度。IoU-Tracker<sup>[29]</sup>只使用了两帧间物体的 IoU 进行关联，速度非常快。但是由于很难做到跨帧的长距离关联，将短暂消失的物体找回，所以会带来比较多的身份跳变（identity switches）。近年来，有一部分方法设计神经网络<sup>[30-32]</sup>

来预测物体的运动，这些方法在相机运动剧烈或者低帧率视频上能够取得比卡尔曼滤波更好的效果。总体来说，基于运动的相似度在短距离内的关联结果是非常可靠的，但是在处理物体被遮挡又出现时的这种长距离的关联上性能还不够好。

外观特征对于长距离关联有比较大的帮助，利用物体的外观信息可以在其被遮挡一段时间又出现的时候找回其身份，外观相似度可以使用物体的重识别特征间的余弦距离计算得到。DeepSORT<sup>[9]</sup>使用了一个单独的重识别模型，在大量行人重识别数据集上训练好了之后，在进行数据关联时，将当前帧的检测框从图上裁剪下来之后送入重识别模型中，提取每个物体的重识别特征，最后计算检测框特征和跟踪框特征之间的余弦距离得到外观相似度。Bae 等人提出了一个在线更新外观特征的方法<sup>[33]</sup>，可以用来适应视频中行人外观的变化。随着多任务学习的兴起<sup>[34-36]</sup>，近年来有部分方法把目标检测和行人重识别放在同一个网络中，通过共享大量特征的方法减少计算量，使基于外观相似度的多目标跟踪方法首次达到了实时的可能。JDE<sup>[37]</sup>在目标检测模型 YOLOv3<sup>[25]</sup>的基础上加入了一个重识别分支，如图 1-5 所示，用一个网络同时进行目标检测和行人重识别，并且整理了能够联合训练检测和重识别的六个数据集。Track R-CNN<sup>[38]</sup>在 Mask R-CNN<sup>[39]</sup>的基础上加入了一个重识别分支，一个网络能同时完成目标检测、实例分割和行人重识别。这类方法因为其简单和高效的特性，越来越受到大家的欢迎。

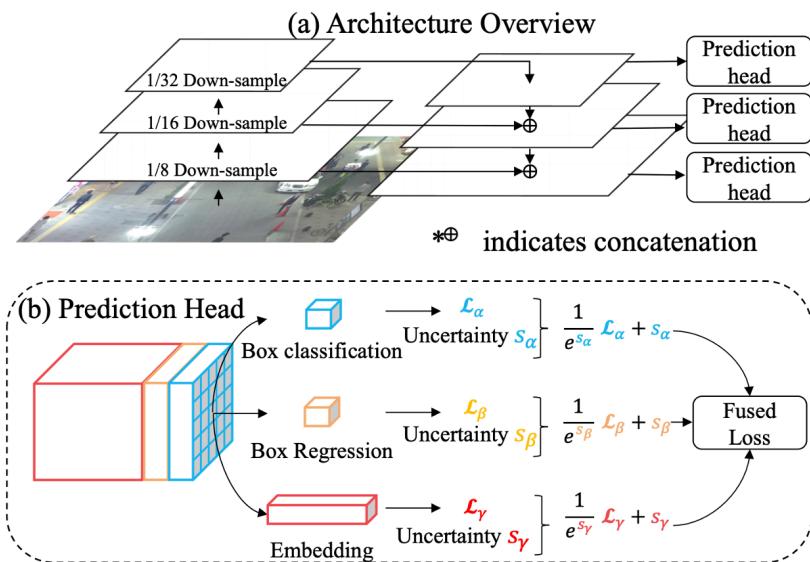


图 1-5 JDE 网络模型结构图<sup>[37]</sup>

## 1.2.4 数据关联中的匹配策略

数据关联中，得到当前帧的检测结果和先前帧的跟踪轨迹之间两两的相似度之后，需要经过一些匹配（关联）策略得到每个目标的身份。最简单的匹配方法是匈牙利算法<sup>[11]</sup>和贪心算法，都是为了找到总距离最小的匹配结果。在相似度计算较为准确的情况下，简单的匹配方法便足以取得很好的效果。SORT<sup>[28]</sup>只对检测框和跟踪轨迹进行了一次匹配。DeepSORT<sup>[9]</sup>提出了一个级联的匹配策略，首先将检测结果与上一帧中激活（非丢失）的跟踪轨迹进行匹配，再和丢失 30 帧以内的跟踪轨迹进行匹配。MOTDT<sup>[40]</sup>首先使用重识别特征在所有检测结果和跟踪轨迹之间进行一次匹配，再使用 IoU 作为相似度在第一次匹配中没有匹配上的检测结果和跟踪轨迹之间再进行一次匹配，第二次匹配中的对象大多都是遮挡严重的物体，重识别特征不再可靠，所以用 IoU 作为相似度。近年来，随着注意力机制<sup>[41]</sup>（Transformer）的发展，可以直接把检测框从上一帧传播到当前帧，从而隐式地完成数据关联。TrackFormer<sup>[42]</sup>和 MOTR<sup>[43]</sup>提出了跟踪查询（track query）的概念，可以在当前帧中直接找到之前帧跟踪轨迹的位置，如图 1-6 所示，在  $t=0$  帧时使用 Transformer 检测出所有物体，接着在后续帧中将上一帧物体的特征作为当前帧的 Transformer 的查询（query）输入，通过解码器直接得到上一帧物体的包围框，只检测新出现的物体，这样便能够保持上一帧物体的身份与当前帧物体的直接对应。该方法在注意力交互的过程中隐式地完成了匹配，不需要匈牙利算法。

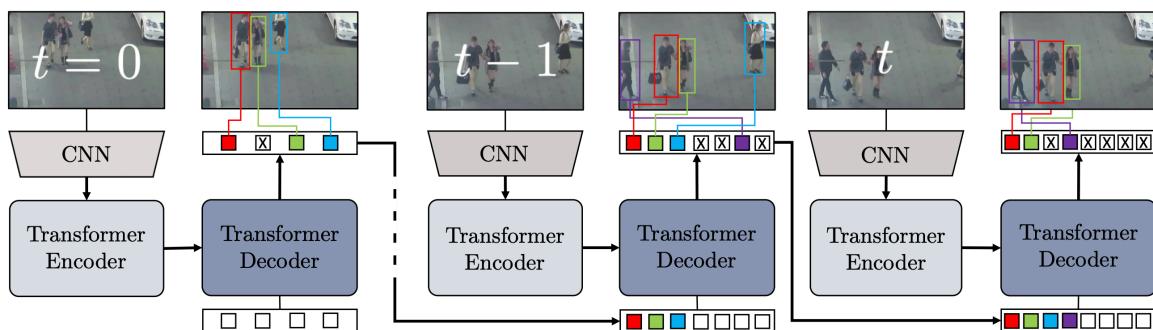


图 1-6 TrackFormer 跟踪示意图<sup>[42]</sup>

## 1.2.5 离线多目标跟踪方法

以上讨论的都是在线多目标跟踪方法，即只使用视频当前帧和过去帧的信息，

在实际应用中有着较为广泛的用途。离线多目标跟踪方法<sup>[44-46]</sup>可以利用未来帧的信息，也叫批量多目标跟踪方法，离线方法的精度往往高于在线方法因为可以在整个视频上做整体优化，但是在应用场景上不如在线方法，只能对视频进行离线处理。

Zhang 等人将多目标跟踪等效成一个图模型<sup>[44]</sup>，每个节点代表检测结果，并使用最小成本流算法搜索最优分配，该算法利用图的特定结构比线性规划更快达到最优。Berclaz 等人同样将数据关联视作一个流优化问题<sup>[45]</sup>，使用 K-最短路径算法来解决这个问题，显著加快计算速度并减少需要调整的参数。MPNTrack<sup>[22]</sup>中提出了一个可以学习的图神经网络(Graph Neural Network)对所有帧中的检测结果进行全局关联，让多目标跟踪这个任务变得完全可导，如图 1-7 所示。Lif\_T<sup>[47]</sup>将多目标跟踪定义为一个提升的不相交路径问题，并提出了提升的边缘来进行远距离的时间交互，显著减少了身份跳变，还能够将丢失的目标的身份重新识别正确。

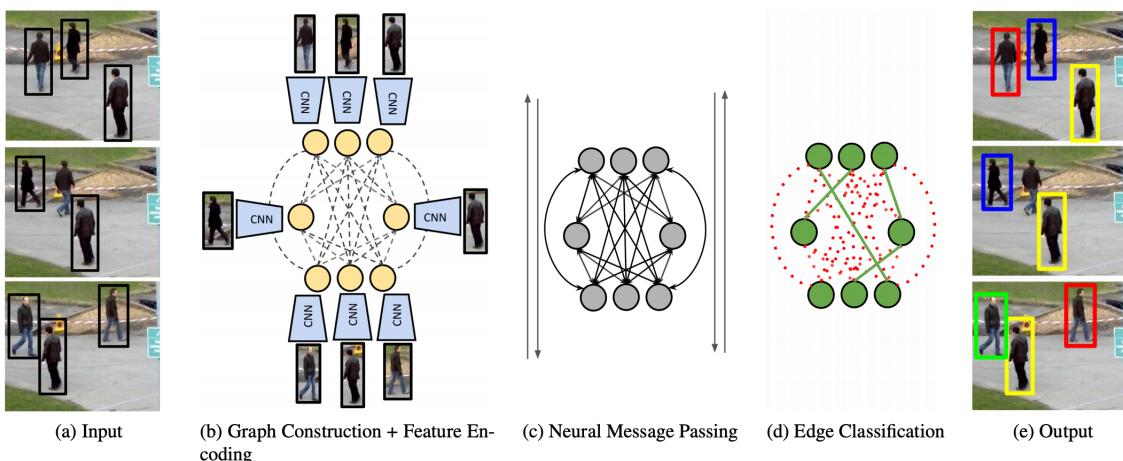


图 1-7 MPNTrack 离线跟踪示意图<sup>[22]</sup>

## 1.2.6 视频目标检测方法

视频目标检测<sup>[48-50]</sup>和多目标跟踪也比较相关，它使用跟踪的方法来提高在有挑战性的场景下目标检测的表现，其中的一些想法在多目标跟踪领域非常有价值。Tang 等人<sup>[50]</sup>将多帧的物体当成管道检测出来，目的是使用前后帧的信息增强当前帧物体的分类得分，该方法对于小物体的检测精度提升明显。相似的想法在其他视频目标检测的工作中也有探索过<sup>[51]</sup>。这类基于管道的方法的一个局限性是当视频中物体数量很多的时候，检测的速度很慢。多目标跟踪中也出现了基于管道的方法 TubeTK<sup>[52]</sup>，

利用多帧的信息对每个目标形成一个管道，解决一些单帧检测无法解决的遮挡问题，如图 1-8 所示。

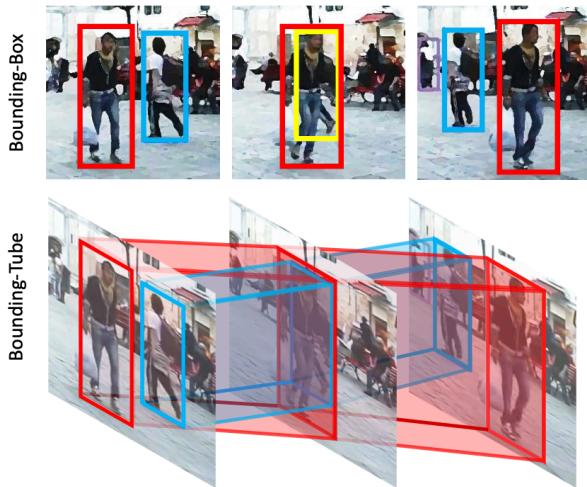


图 1-8 TubeTK 用管道代替框进行跟踪示意图<sup>[52]</sup>

### 1.2.7 三维人体姿态估计方法

在二维世界中，有很多行人遮挡问题是很难解决的，利用多视角信息可以得到行人在三维空间中的位置，解决单个视角下的遮挡问题，这个位置可以通过人体的关节点坐标表示，在得到行人位置的同时还能够得到人体姿态。在多人三维人体姿态估计中有两大难点，第一个难点是需要将同一个人的关节点关联起来，第二个难点是需要使用外观特征<sup>[54]</sup>或几何特征<sup>[55]</sup>将同一个人在不同视角下的二维关节点关联起来，但是，在人体遮挡较为严重时这些特征都会变得非常不稳定。一些基于模型的方法通过最大化模型投影和图像观察之间的一致性来解决第二个难点，比如将图结构模型（Pictorial Structure Model）拓展到多人三维人体姿态估计上<sup>[56,57]</sup>。但是，人与人之间的交互会在图中引入循环，这会使优化变得特别复杂。这些挑战限制了这些方法的三维姿态估计精度。Dong 等人<sup>[54]</sup>提出了一个多路匹配算法，同时使用外观线索和几何线索找到不同视角下的二维人体关节点之间的一致性，能够减少错误检测并处理视角之间的部分重叠，之后通过基于三角化（triangulation）的方法将不同视角下同一个人的二维关节点恢复成三维关节点。如图 1-9 所示。Chen 等人<sup>[55]</sup>利用视频中的时间一致性将检测到的二维关节点与估计的三维关节点直接在三维空间

中匹配，并通过跨视角多人跟踪的方法迭代更新三维人体关节点，精度和速度较先前方法都有提升。

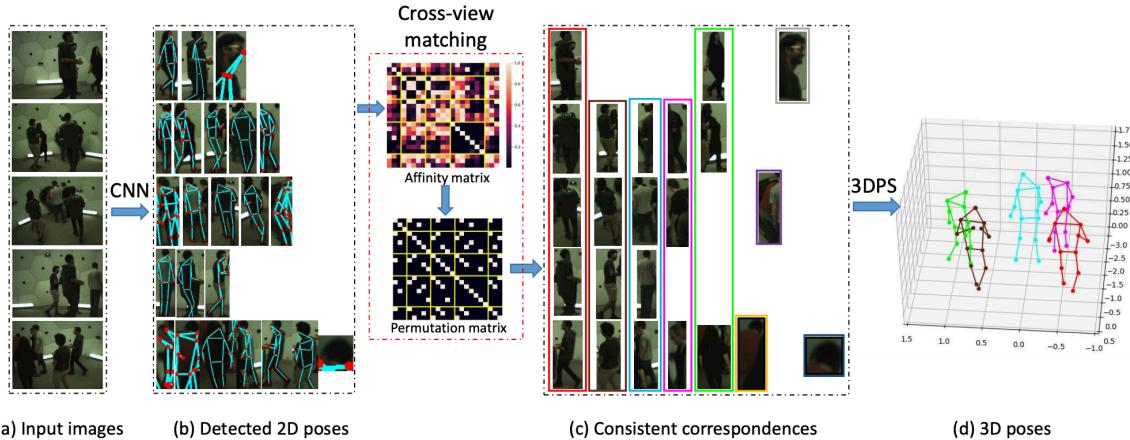


图 1-9 Dong 等人提出的三维人体姿态估计方法<sup>[54]</sup>

### 1.2.8 多目标跟踪数据集和评价指标

MOTChallenge<sup>1</sup>是全球最权威的多目标跟踪挑战赛，由慕尼黑工业大学（TUM）在 2015 年提出，后续也在不断地更新数据集，它包含了四个行人多目标跟踪数据集：MOT15<sup>[58]</sup>、MOT16<sup>[15]</sup>、MOT17<sup>[15]</sup>、MOT20<sup>[59]</sup>。这四个数据集都只需要跟踪行人一个类别的物体，其中 MOT15 包含了 11 个训练视频和 11 个测试视频，视频帧率在 2.5 FPS 到 30 FPS 不等，平均人群密度为每幅图 9 个人，较为稀疏，如图 1-10 所示。MOT16 和 MOT17 使用的是同样的视频，包含了 7 个训练视频和 7 个测试视频，视频帧率相较于 MOT15 有所提高，在 14 FPS 到 30 FPS 之间；人群密度有所提高，平均一幅图 25 个人；场景也更为丰富，有一半的视频都是相机运动较为剧烈的视频，如图 1-11 所示。MOT17 的标注比 MOT16 更加精细，标注了很多半遮挡和全遮挡的人，而且提供了来自 DPM、Faster R-CNN 和 SDP 三种检测器得到的检测结果，供大家在相同的检测结果上比较各自算法的跟踪性能。MOT20 中的人群更为密集，平均一幅图有 160 个人，包含 4 个训练视频和 4 个测试视频，每个视频的时长更长，在 400 到 3300 帧之间，帧率也有所提高，所有视频都是 25 FPS，如图 1-12 所示。

<sup>1</sup> <https://motchallenge.net/>

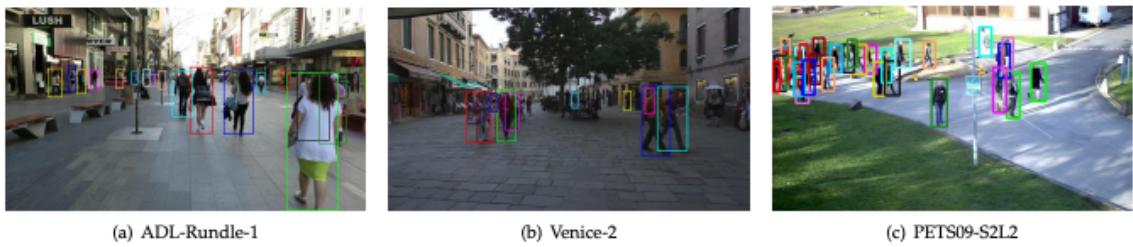


图 1-10 MOT15 数据集部分场景和标注<sup>[58]</sup>



图 1-11 MOT17 数据集场景图<sup>[15]</sup>



图 1-12 MOT20 数据集场景图<sup>[59]</sup>

HiEve (Human in Events)<sup>[60]</sup>是一个面向复杂场景的多目标跟踪数据集，由上海交通大学在 2020 年发布，也是只跟踪人，和 MOTChallenge 稍有不同，HiEve 需要跟踪各种姿态的人，不仅只跟踪行人。HiEve 的场景非常丰富，包括餐厅、广场、商场、地铁、公交等公众场合，其中还有一些非常罕见的场景比如地震、斗殴、抢劫等，如图 1-13 所示。它包含 19 个训练视频和 13 个测试视频，帧率在 14 FPS 到 30 FPS 之间。HiEve 除了提供人的包围框和身份的标注之外，还提供了人体关节点的标注，可以用来完成人体姿态估计任务。



图 1-13 HiEve 数据集场景图<sup>[60]</sup>

单类别多目标跟踪的性能评估采用 CLEAR 指标<sup>[79]</sup>, 其中包括多目标跟踪精度 (MOTA)、假阳性个数 (FP), 假阴性个数 (FN)、身份跳变次数 (IDs) 等和身份的 F1 值 (IDF1)<sup>[80]</sup>, 其中 MOTA 指标更偏向于模型的检测性能, 计算公式如下:

$$MOTA = 1 - \frac{FN + FP + IDs}{GT} \quad (1-1)$$

其中 FN 为总的假阴性数量, FP 为总的假阳性数量, IDs 为总的身份跳变次数, 如果同一个目标的身份发生了跳变, 那么 IDs 次数加一, GT 为总的标签数量。IDs 和 IDF1 指标更偏向于模型的关联性能, 其中 IDF1 的计算公式如下:

$$IDF1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \quad (1-2)$$

其中 IDTP 为目标身份识别正确的数量, IDFP 为目标身份识别为假阳性的数量, IDFN 为目标身份识别为假阴性的数量。

BDD100K<sup>[61]</sup>是面向自动驾驶场景的多目标跟踪数据集, 由美国加州大学伯克利分校 (UC Berkeley) 在 2020 年发布, 总共需要跟踪 8 个类别的物体, 包括行人、汽车、卡车、三轮车等, 数据量也大很多, 包含 1400 个训练视频, 200 个验证视频, 400 个测试视频, 每个视频每 5 帧标注一帧, 帧率较低, 每个视频总共标注 200 张图, 除了跟踪之外, 也提供了许多其他任务的标注, 如分割、车道线检测等, 如图 1-14 所示。因为 BDD100K 是多类别多目标跟踪数据集, 所以评价指标中加入了两个多类别的评价指标 mMOTA 和 mIDF1, 计算方式为对每个类别目标的 MOTA 和 IDF1 取平均值, 得到 mMOTA 和 mIDF1。

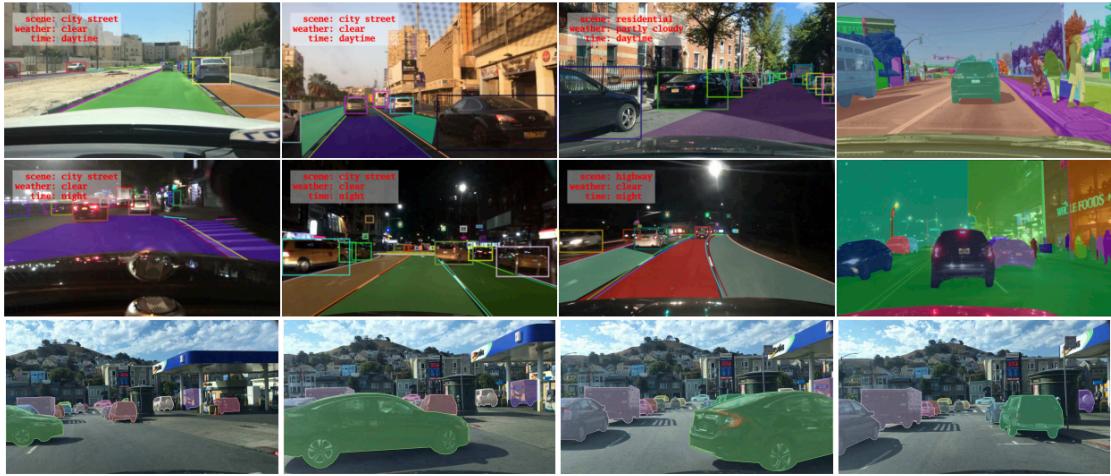


图 1-14 BDD100K 数据集场景图及标注<sup>[61]</sup>

Campus 和 Shelf<sup>[62]</sup>是用于三维人体姿态估计和跟踪的小规模的多相机数据集，它由慕尼黑工业大学（TUM）在 2014 年的时候提出。Campus 包含三个相机视角的视频，场景为学校室外场景，活动为 3 个人走路和谈话。Shelf 包含五个相机视角的视频，场景是室内场景，活动为 4 个人围绕一个书架在交谈和行走，如图 1-15 所示。其中 Shelf 中包含更多的人和书架间的遮挡。

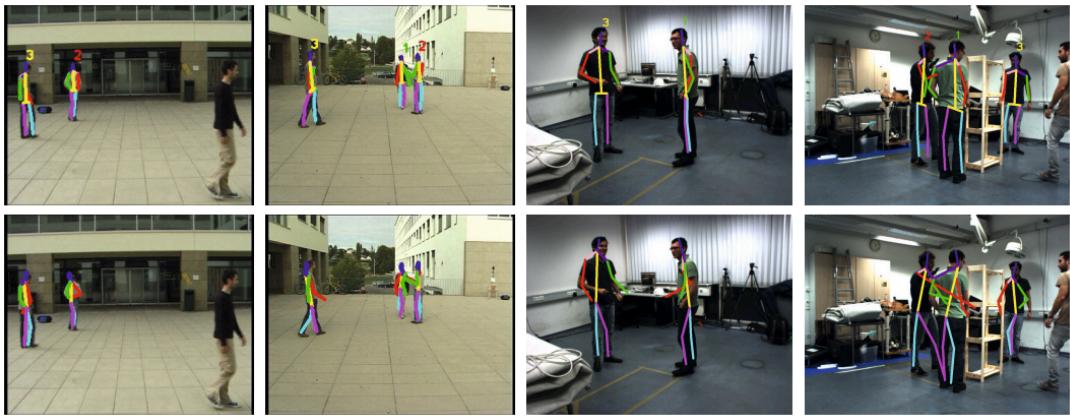


图 1-15 Campus 和 Shelf 数据集场景图及标注<sup>[62]</sup>

CMU Panoptic<sup>[63]</sup>是用于三维人体姿态估计和跟踪的大规模的多相机数据集，它由卡耐基梅隆大学（CMU）在 2015 年的时候提出，通过数十个相机捕获人们的日常活动。在一个实验室内，会有多人进行吃饭、闲逛、玩耍等活动。每段视频都有上万帧，分为 10 个训练视频和 4 个测试视频。本文从每段视频挑取 5 个相机捕获的画面作为数据。数据集提供相机参数和每个视频的三维人体关节点坐标标注、人的

身份标注，如图 1-16 所示。



图 1-16 CMU Panoptic 数据集场景图及标注<sup>[63]</sup>

## 1.3 多目标跟踪目前存在的问题

本部分从网络模型、数据关联和视角信息三个不同的方面分析目前多目标跟踪方法存在的问题以及多目标跟踪领域存在的一些挑战。

### 1.3.1 网络模型中的多任务不公平问题

目前多目标跟踪中把目标检测和行人重识别放在同一个网络中越来越受到欢迎，因为在骨干网络中共享了两个任务所需的大量特征，相较于用两个不同的网络分别完成这两个任务带来很大的速度提升，这类方法可称之为单阶段多目标跟踪网络。JDE<sup>[37]</sup>是一个代表方法，它在目标检测器 YOLOv3<sup>[25]</sup>的基础上加入了一个行人重识别分支，同时完成目标检测和行人重识别。但是，单阶段方法相较于双阶段方法<sup>[64]</sup>的跟踪精度有明显下降，身份跳变的次数大幅增加，这说明单阶段方法中的行人重识别特征没有学好。本文经过研究，发现问题在于目标检测和行人重识别这两个任务在同一个网络中没有被公平地对待。第一是体现在目标检测中的锚框上，与标签框的 IoU 大于 0.5 的锚框便被视为正样本进行训练，如图 1-17 所示，红色框为锚框，绿色框为标签框。这种训练方法对目标检测较为友好，但是会损伤行人重识别的学习效果。

习，比如在目标较为拥挤时，一个锚框可能会包含多个目标，同时一个目标可能会被多个锚框覆盖，给重识别的训练带来很大的模糊性，如图 1-18 所示。

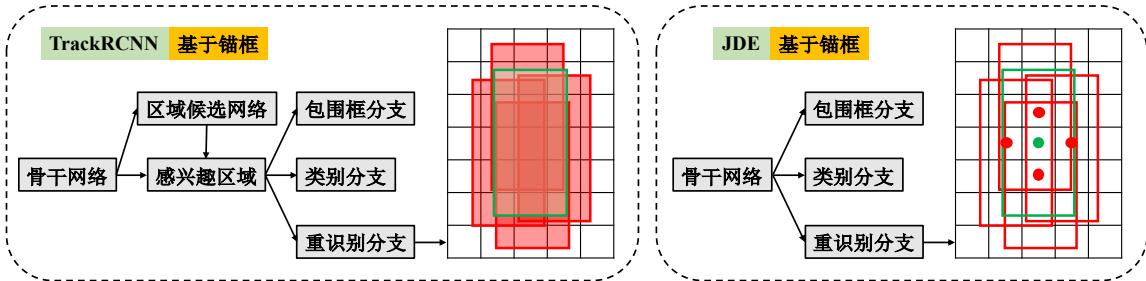


图 1-17 单阶段多目标跟踪网络中的锚框



图 1-18 使用锚框提取重识别特征带来的问题

第二处不公平体现在网络特征上，目标检测和行人重识别是两个完全不同的任务，目标检测需要网络学习目标之间的共性特征，要把所有的人都检测出来；而行人重识别需要学习目标之间的差异性特征，要把不同的人分开，所以两个任务的优化目标是冲突的，优化一个任务会加剧另一个任务的损失。从特征本身来说，目标检测需要高层的语义特征，而行人重识别需要低层的外观特征，这也表明了两个任务的差异性。第三处不公平体现在两个任务的特征维度上，目标检测通常需要较低的特征维度，目标的类别加坐标，一般在 10 以下。行人重识别通常使用较高的特征维度，比如 512<sup>[37]</sup>或 1024<sup>[65]</sup>，特征维度的差异也导致了两个任务无法得到很好的优化。

### 1.3.2 数据关联中的目标漏跟踪问题

目前的遵循“先检测后跟踪（tracking-by-detection）”范式的多目标跟踪方法都是在得到检测结果之后，只保留高分检测框做跟踪，为了去除背景便把低分检测框直接扔掉。这样的做法会带来大量的目标漏跟踪和轨迹中断的情况。如图 1-19 所示，第一行标出了连续三帧的检测结果和得分，第二行标出了跟踪结果，不同的颜色代

表不同的身份。可以看出只保留高分检测框的做法在  $t_2$  和  $t_3$  帧会漏掉第  $t_1$  帧中红色框的跟踪轨迹，带来不可逆转的目标漏跟踪和轨迹中断的问题。



图 1-19 只保留高分检测框做跟踪的结果

### 1.3.3 单视角视频中的遮挡问题

如果只使用单个视角的视频进行跟踪，当行人完全被遮挡时是非常难跟踪到的，当前最先进的检测器<sup>[27]</sup>也检测不出来，但是在实际应用比如自动驾驶中要求对这种完全被遮挡的目标进行跟踪，以防其没被遮挡后突然出现，如图 1-20 所示，在第 15 帧的时候目标完全被车遮挡，没有被跟踪上，为了突出被遮挡的目标，图中其余部分使用淡色表示。



图 1-20 单视角视频中的遮挡问题<sup>[67]</sup>

# 华中科技大学硕士学位论文

---

## 1.4 本文主要内容和贡献

### 1.4.1 拟解决的问题和研究目标

针对目前多目标跟踪领域存在的问题，本文主要拟解决 3 个关键问题：

1) 网络模型中的目标检测和行人重识别的不公平问题，这个问题导致单阶段多目标跟踪网络提取的行人重识别特征区分性很差。本文的研究目标是尽可能降低这两个任务在同一个网络模型中的相互影响，在保证目标检测精度的同时，提取具有很强区分性的行人重识别特征来进行多目标跟踪，使得单阶段多目标跟踪网络的性能可以超过或者和双阶段方法持平。

2) 目标检测结果的不稳定带来的目标漏跟踪和轨迹中断问题，许多遮挡、模糊的物体由于目标检测得分很低导致跟踪时被舍弃，带来大量的漏跟踪和轨迹中断情况。本文的研究目标是将得分低的遮挡、模糊的物体跟踪上，保持轨迹的连贯性，在相同目标检测的性能下，提升多目标跟踪精度 MOTA 和轨迹连贯性指标 IDF1。

3) 单个视角下目标的遮挡问题，在目标被完全遮挡时，几乎没办法被检测到。本文的研究目标是使用多视角的信息，得到目标在场景中的三维位置并进行跟踪，能够跟踪到单个视角下完全被遮挡的目标，降低由于遮挡带来的身份跳变次数。

### 1.4.2 主要研究内容和创新点

本文围绕复杂场景下的多目标跟踪问题，从网络模型、数据关联和视角信息三个不同的方面进行了三项研究，实现高性能的多目标跟踪，主要研究内容和创新如下：

1) **公平对待目标检测和行人重识别的单阶段多目标跟踪网络** 提出一个基于中心点特征的联合检测和重识别网络，用物体中心点代替目标检测中的锚框，使用两个平行的分支来提取检测和重识别的特征，并在骨干网络中加入多层特征融合模块降低特征冲突，使得目标检测和行人重识别两个任务在同一个网络中被公平地对待。在多目标跟踪领域达到 25 FPS 的运行速度，跟踪精度领先同期最优方法 5 个百分点。

2) **基于高低分检测框层次关联的多目标跟踪方法** 提出了一个简单有效，普适性很强的数据关联方法，将目标检测器得到的检测框分成高分和低分两部分。首先

# 华中科技大学硕士学位论文

对跟踪轨迹使用卡尔曼滤波预测其在当前帧的位置，再用高分检测框和跟踪轨迹进行第一次匹配，对没有匹配上的跟踪轨迹，再和低分检测框进行第二次匹配，从而找到低分检测框中遮挡严重、模糊的目标并去除背景，从而解决漏跟踪、轨迹中断的问题。该方法应用到当前大部分多目标跟踪模型上能够将跟踪精度提升 1-10 个百分点。实验表明该方法在 MOTChallenge 上达到目前最好的跟踪性能。

3) 基于体素特征的多视角三维人体姿态估计和跟踪框架 提出一个基于体素特征的三维网络，将整个三维场景建模成一个体素特征，利用多视角信息得到行人在三维空间中的位置和姿态，并融合同一个人在不同视角下的重识别特征进行跟踪，解决单个视角下的遮挡问题。该方法提供了一个三维人体姿态估计和跟踪的基线框架 (baseline)，利用多视角信息得到行人的三维位置，即使在单个视角完全遮挡的情况下也能成功精准跟踪目标，不会发生身份跳变。

研究内容 1 为多目标跟踪中的网络结构模型，主要用于得到视频帧中目标的包围框和重识别特征；研究内容 2 为多目标跟踪中的数据关联模块，主要根据帧之间目标的包围框和重识别特征将相同的目标匹配到一起，得到目标身份；研究内容 3 为一个完整的使用多视角信息作为输入的三维多目标跟踪框架，其中包含了研究内容 1 中的多目标跟踪网络模型和研究内容 2 中的数据关联技术，如图 1-21 所示。研究内容 1 和 2 的场景为单视角视频场景，研究内容 3 的场景为多视角视频场景。

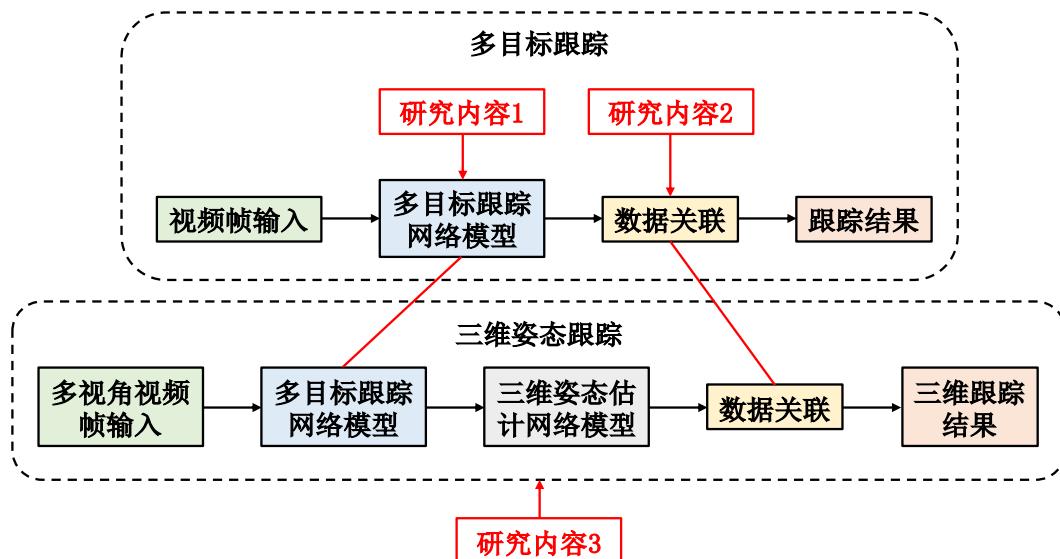


图 1-21 本文三个研究内容之间的关系

# 华 中 科 技 大 学 硕 士 学 位 论 文

---

---

## 1.4.3 本文主要内容安排

第一章绪论首先讲述了多目标跟踪的背景和研究意义，接着介绍了国内外研究现状，包括研究方法和常用数据集，最后简要地描述了本文的研究内容和创新点。

第二章介绍了本文提出的公平对待目标检测和行人重识别的单阶段多目标跟踪网络。

第三章介绍了本文提出的基于高低分检测框层次关联的多目标跟踪方法。

第四章介绍了本文提出的基于体素特征的多视角多人三维人体姿态估计和跟踪框架。

第五章是硕士期间工作总结和对多目标跟踪领域未来的展望。

## 2 公平对待目标检测和行人重识别的单阶段多目标跟踪网络

### 2.1 研究动机

本章将介绍本文提出的一种公平对待目标检测和行人重识别的单阶段多目标跟踪网络 FairMOT (FairMOT: On the Fairness of Detection and Re-Identification in Multi-Object Tracking)。针对之前基于锚框的单阶段多目标跟踪网络 JDE<sup>[37]</sup>和 Track R-CNN<sup>[38]</sup>无法提取到有区分力的行人重识别特征，跟踪中身份跳变次数多的问题。本文发现之前单阶段方法在处理目标检测和行人重识别着两个任务上的三点不公平的问题：(1) 目标检测网络中的锚框是针对目标检测设计的，不适合行人重识别，在人群密集和遮挡的情况下给网络的训练带来很大的歧义性，导致行人重识别特征无法分清靠得近的目标；(2) 目标检测和行人重识别两个任务的优化目标有冲突；(3) 目标检测和行人重识别所需的特征维度相差较大，导致网络训练的不一致性。本文主要解决单阶段多目标跟踪网络中的上述三点不公平问题，使得网络在得到精准的目标检测结果的同时，能够提取出区分力强的行人重识别特征，降低跟踪中身份跳变次数。本章的研究内容对应的场景为单视角场景，包含街道、商场等场景，本章的方法只对行人进行跟踪，人群较为密集。本章将从主要思路、模型结构、模型训练、数据关联和实验结果五个方面来介绍 FairMOT。

### 2.2 主要思路

针对单阶段多目标跟踪网络中的锚框问题，本文提出了一个基于中心点的特征提取网络，在训练阶段时，只使用目标的标注框中心点处对应的位置提取检测特征和重识别特征，保持标注和训练样本一对一，避免了基于锚框的特征提取方法中一个锚框可能包含多个目标、一个目标可能有多个锚框对应带来的歧义性。

针对目标检测和行人重识别两个任务的特征冲突问题，本文在骨干网络中加入了高低层特征融合结构，使得两个任务能够分别从不同的网络层中提取各自需要的特征，降低两个任务之间的冲突，实验证明多种不同的高低层特征融合结构都能够

起到降低冲突的作用。

针对目标检测和行人重识别特征维度相差大的问题，本文提出使用低维度（128 或 64）的行人重识别特征代替高维度特征。首先，低维度的行人重识别特征在训练时可以减少对目标检测任务的影响，得到更高的目标检测精度。其次，低维度的重识别特征具有更快的跟踪速度，计算量少。最后，由于多目标跟踪任务和行人重识别任务的差别在于跟踪任务只需要在相邻两帧中使用重识别特征间的相似度将当前帧的检测目标和之前帧的跟踪目标关联，其中目标中有一一对应的关系，所以不需要行人重识别任务中的高维度特征，低维度特征也能将目标很好地区分开。

## 2.3 模型结构

整个网络模型包括骨干网络、目标检测分支和行人重识别分支，其中目标检测分支和行人重识别分支平行地接在骨干网络之后，骨干网络输出的特征作为目标检测分支和行人重识别分支的输入，整个模型结构如图 2-1 所示。该模型结构主要有两个优点：一是简洁、推理速度快，不需要 Faster R-CNN<sup>[17]</sup>中的区域候选网络；二是平行的目标检测分支和重识别分支能够降低两个任务之间的相互影响，Track R-CNN<sup>[38]</sup>中重识别分支的输入依赖于区域候选网络的输出结果，也就是说重识别特征依赖目标检测的结果，本章提出的模型结构避免了这种依赖性，更加公平地对待这两个任务。

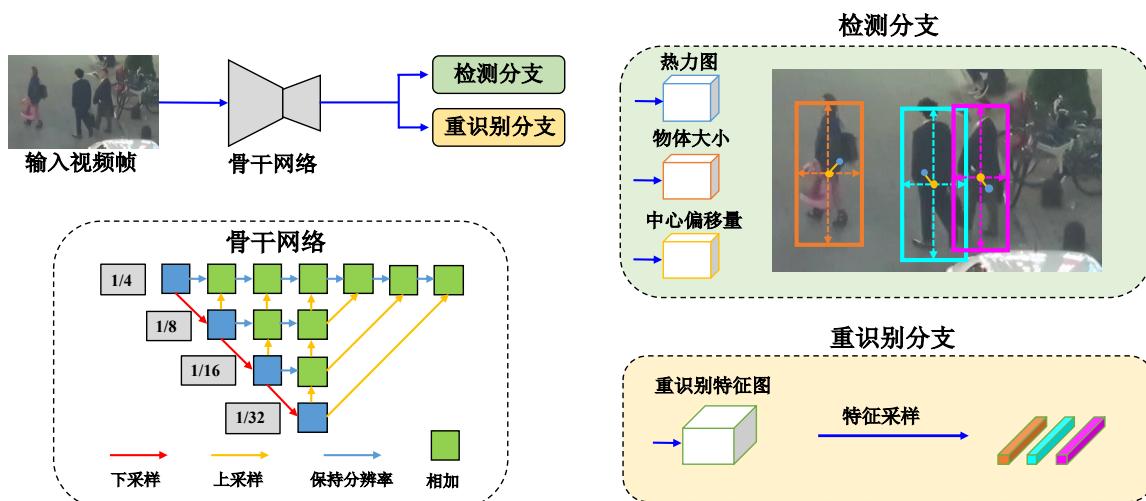


图 2-1 FairMOT 模型结构图

### 2.3.1 多层特征融合骨干网络

ResNet-34<sup>[2]</sup>可以作为最原始的骨干网络达到一个比较好的精度和速度平衡，但是 ResNet 系列的骨干网络缺少高低层特征融合，导致目标检测和行人重识别两个任务之间发生特征冲突，无法寻找各自所需的特征，例如目标检测需要高层语义特征，行人重识别需要低层外观特征。DLA (Deep Layer Aggregation) 网络<sup>[68]</sup>中包含高低层特征融合结构，本文在其基础上，在高低层之间加入了更多的跨层连接 (skip connections)，类似 FPN (Feature Pyramid Network)<sup>[69]</sup>的做法，如图 2-1 左下角的骨干网络所示。加入了跨层连接之后可以使目标检测和行人重识别这两个任务能够更加方便地从不同的层中学习各自需要的特征，比如检测可以从高层中学到语义特征，重识别可以从低层中学到外观特征。高低层特征融合结构减少了两个任务之间的冲突。本文将所有上采样卷积层（图 2-1 左下角黄色箭头）都替换成了可变形卷积 (deformable convolution)<sup>[70]</sup>，使网络根据物体的形状和姿态动态地调整感受野，能更好地覆盖到物体全身，这样一来网络能够更好地学习物体的大小和外观。修改过后的骨干网络称为 DLA-34，包含 34 个卷积层。输入图像在经过骨干网络之后输出特征图的宽和高为输入图像的四分之一。除了 DLA 网络，其他包含多尺度特征的网络也都可以用作 FairMOT 的骨干网络，例如 HRNet<sup>[71]</sup>、FPN<sup>[69]</sup>和 HarDNet<sup>[72]</sup>。

### 2.3.2 目标检测分支

FairMOT 的目标检测分支是在 CenterNet<sup>[24]</sup>的基础上设计的，包含三个平行的小分支：热力图分支 (heatmap head)、物体大小分支 (box size head) 和中心偏移量分支 (box offset head)。热力图分支负责以热力图的形式来学习物体中心点的位置，物体大小分支负责学习物体包围框的宽和高的值，中心偏移量分支负责学习中心点在经过骨干网络的四倍下采样之后造成的小数偏差值。每个小分支都包含一个通道数为 256，卷积核大小为的 $3 \times 3$ 的卷积和一个通道数为各自输出维度，卷积核大小为 $1 \times 1$ 的卷积。

热力图分支负责学习物体中心点位置。热力图是姿态估计中的标准做法，其通道数为物体类别数，在行人多目标跟踪中，只有行人一个类别，故其通道数为 1。热

力图上的值分布在 0 到 1 之间，如果某处与物体的中心位置的真实值重合，那么它的值就为 1，随着位置远离物体中心点，热力图上的响应值呈指数衰减。假设第*i*个物体包围框在图像上的真实值为  $\mathbf{b}^i = (x_1^i, y_1^i, x_2^i, y_2^i)$ ，物体中心点坐标  $(c_x^i, c_y^i)$  可以通过  $c_x^i = \frac{x_1^i + x_2^i}{2}$  和  $c_y^i = \frac{y_1^i + y_2^i}{2}$  得到，那么它在热力图上的位置可以通过图像的位置除以下采样的倍数并向下取整得到： $(\tilde{c}_x^i, \tilde{c}_y^i) = (\left\lfloor \frac{c_x^i}{4} \right\rfloor, \left\lfloor \frac{c_y^i}{4} \right\rfloor)$ ，最后可以得到热力图在  $(x, y)$  位置处的响应值  $M_{xy} = \sum_{i=1}^n e^{-\frac{(x-\tilde{c}_x^i)^2+(y-\tilde{c}_y^i)^2}{2\sigma_c^2}}$ ，其中  $N$  表示图中总的物体个数， $\sigma_c$  表示标准差。加入 Focal 损失<sup>[23]</sup>之后的热力图分支的损失函数计算方式如下：

$$L_{heat} = -\frac{1}{N} \sum_{xy} \begin{cases} (1 - \hat{M}_{xy})^\alpha \log(\hat{M}_{xy}), & M_{xy} = 1 \\ (1 - M_{xy})^\beta (\hat{M}_{xy})^\alpha \log(1 - \hat{M}_{xy}), & M_{xy} \neq 1 \end{cases} \quad (2-1)$$

其中  $\hat{M}$  表示由热力图分支估计出来的热力图， $\alpha$  和  $\beta$  是 Focal Loss 中预定义的超参数。

物体大小分支输出物体包围框的宽和高，物体中心偏移量分支能够更加精准的定位物体，因为最终的输出特征图的输入图像的四分之一大小，会带来 4 个像素之内的量化误差，偏移量分支会估计出中心点真实值和 4 倍下采样取整之后的值之间的偏移量。假设第*i*个物体包围框在图像上的真实值为  $\mathbf{b}^i = (x_1^i, y_1^i, x_2^i, y_2^i)$ ，它的大小真实值为  $\mathbf{s}^i = (x_2^i - x_1^i, y_2^i - y_1^i)$ ，类似地，中心点偏移量的真实值为  $\mathbf{o}^i = \left( \frac{c_x^i}{4}, \frac{c_y^i}{4} \right) - \left( \left\lfloor \frac{c_x^i}{4} \right\rfloor, \left\lfloor \frac{c_y^i}{4} \right\rfloor \right)$ ， $\hat{\mathbf{s}}^i$  和  $\hat{\mathbf{o}}^i$  分别表示包围框大小的预测值和中心偏移量的预测值，由物体中心点真实值在这两个分支输出的特征图上的位置得到，物体大小分支和中心偏移量分支的输出特征图的通道数都为 2。最终物体包围框的损失函数由向这两个分支加 L1 loss 得到：

$$L_{box} = \sum_{i=1}^N \|\mathbf{o}^i - \hat{\mathbf{o}}^i\|_1 + \lambda_s \|\mathbf{s}^i - \hat{\mathbf{s}}^i\|_1 \quad (2-2)$$

其中  $\lambda_s$  是一个权重参数，设置为 0.1。

### 2.3.3 行人重识别分支

行人重识别分支的目的是提取能区分不同物体的特征，理想化来说，视频中不同物体之间的相似度应该远小于相同的物体之间的相似度。为了达到这个目的，

FairMOT 中使用一个 128 通道的卷积核来为特征图上每个点提取重识别特征, 用  $\mathbf{E} \in \mathbb{R}^{128 \times H \times W}$  表示。单个物体的重识别特征  $\mathbf{E}_{x,y} \in \mathbb{R}^{128}$  能够通过位置  $(x, y)$  从特征图  $\mathbf{E}$  上提取, 其中  $(x, y)$  为物体中心点的坐标。

FairMOT 中使用分类任务的形式来学习重识别特征, 整个训练集中身份相同的物体视作同一个类别。对于图上每一个真实的物体包围框  $\mathbf{b}^i = (x_1^i, y_1^i, x_2^i, y_2^i)$ , 可以从热力图上得到物体的中心点坐标  $(\tilde{c}_x^i, \tilde{c}_y^i)$ , 接着提取该物体的重识别特征  $\mathbf{E}_{\tilde{c}_x^i, \tilde{c}_y^i}$ , 再用一个全连接层和一个柔性最大 (softmax) 操作将它映射为一个类别分布向量  $\mathbf{P} = \{\mathbf{p}(k), k \in [1, K]\}$ , 将类别标签的 one-hot 表征记作  $\mathbf{L}^i(k)$ , 最终重识别损失为一个标准的交叉熵损失函数, 由以下式子得到:

$$L_{identity} = - \sum_{i=1}^N \sum_{k=1}^K \mathbf{L}^i(k) \log (\mathbf{p}(k)) \quad (2-3)$$

其中  $K$  是整个训练集中的物体身份的个数,  $N$  表示图中总的物体个数, 只有物体中心点处的重识别特征会用来训练和测试。

## 2.4 模型训练和推理

训练 FairMOT 时, 目标检测分支和行人重识别分支一起训练, 总的损失函数为公式 (2-1), (2-2) 和 (2-3) 相加, 在相加时使用不确定度 (uncertainty)<sup>[35]</sup> 来自动平衡每个任务的损失的权重, 总的损失由下式计算得到:

$$L_{detection} = L_{heat} + L_{box} \quad (2-4)$$

$$L_{total} = \frac{1}{2} \left( \frac{1}{e^{\omega_1}} L_{detection} + \frac{1}{e^{\omega_2}} L_{identity} + \omega_1 + \omega_2 \right) \quad (2-5)$$

其中  $\omega_1$  和  $\omega_2$  是两个可学习的参数, 来自动平衡目标检测和行人重识别两个任务。

对于有身份标注的数据集比如 MOT17<sup>[15]</sup>, 公式 (2-5) 用作最终的优化目标, 对于只有物体包围框标注, 没有身份标注的数据集如 COCO<sup>[73]</sup> 和 CrowdHuman<sup>[74]</sup>, 可以只训练目标检测分支, 也可以用弱监督的方法同时训练目标检测分支和行人重识别分支。使用弱监督训练方法时, 需要人为生成物体的身份标注, 每个包围框都会被赋予一个单独的身份, 数据集中的每个物体实例会被视作一个单独的类别。输入图像会被添加 HSV 色彩增强、旋转、缩放、仿射变换等数据增强来赋予每个身份的

物体更多的训练样本。如此一来，更多的训练数据便能够被 FairMOT 利用，提高模型性能。

在模型推理时，首先使用目标检测分支得到代表物体中心点响应的热力图，接着对热力图使用基于响应值的非极大值抑制（Non-Maximum Suppression, NMS）操作来提取峰值点，非极大值抑制可以通过一个简单的最大池化操作完成，在非极大值抑制之后，保留响应值大于一个阈值的热力图点作为物体的中心点。在得到物体中心点之后，根据中心点的位置从物体大小分支和物体中心偏移量分支提取出对应位置的特征，由中心点、长宽、中心偏移量得到物体包围框的坐标。类似地，每个物体的重识别特征也根据中心点处的位置提取得到。

## 2.5 数据关联

在得到物体的包围框和重识别特征后，再使用数据关联得到每个物体的身份，FairMOT 中的数据关联同样遵循“先检测后跟踪”的范式。在第一帧时，将所有检测框都初始化为跟踪轨迹（tracklets），在接下来的视频帧中，新检测到的物体和跟踪轨迹进行匹配。整个匹配分为两个阶段，第一个阶段中共同使用卡尔曼滤波<sup>[10]</sup>和行人重识别特征，卡尔曼滤波用来预测每个跟踪轨迹在当前帧的位置，通过计算预测位置和实际检测位置之间的马氏距离（Mahalanobis distance）作为运动距离 $D_m$ 。新检测的物体和跟踪轨迹之间行人重识别特征的余弦距离作为外观距离 $D_r$ 。最终的距离 $D$ 为运动距离和外观距离的加权求和，由式（2-6）得到：

$$D = \lambda D_r + (1 - \lambda) D_m \quad (2-6)$$

其中 $\lambda$ 是一个权重参数，用来调整外观距离和运动距离之间的权重，因为外观距离在 0-1 之间，而运动距离没有做归一化，数量级往往在 1 以上，所以外观距离的权重 $\lambda$ 通常会设置的较大，该方法在实验中将其设置成 0.98，最后通过匈牙利算法<sup>[11]</sup>得到第一阶段的匹配结果，其中距离小于 $\tau_1=0.4$ 的能够匹配，实验中发现 0.4 的效果最好。

第二阶段匹配针对的是第一阶段匹配中未能匹配上的检测框和跟踪轨迹，通常是因为目标遮挡带来的行人重识别特征不可靠导致没有匹配上，如图 2-2 所示，第 t 帧中红色框的跟踪轨迹和第 t+1 帧中黄色框的检测结果就是这种情况。此时使用未能匹

配上的检测框和跟踪轨迹之间的交并比（Intersection over Union, IoU）作为相似度进行第二阶段的匹配，同样使用匈牙利算法，距离小于 $\tau_1=0.5$ 的能够匹配。两阶段的匹配都完成后，每个跟踪轨迹的行人重识别特征会使用当前帧的特征进行更新，使其更为平滑，对物体的外观变化更为鲁棒，更新后的重识别特征为跟踪轨迹的特征和当前帧物体的特征的加权求和。最后，没有匹配上跟踪轨迹的检测结果会被初始化成新的跟踪轨迹，表示新物体的出现；没有匹配上检测结果的跟踪轨迹会被保留 30 帧，视为丢失，如果其在后面的帧中再出现那么就有可能再匹配上，保持身份不变。

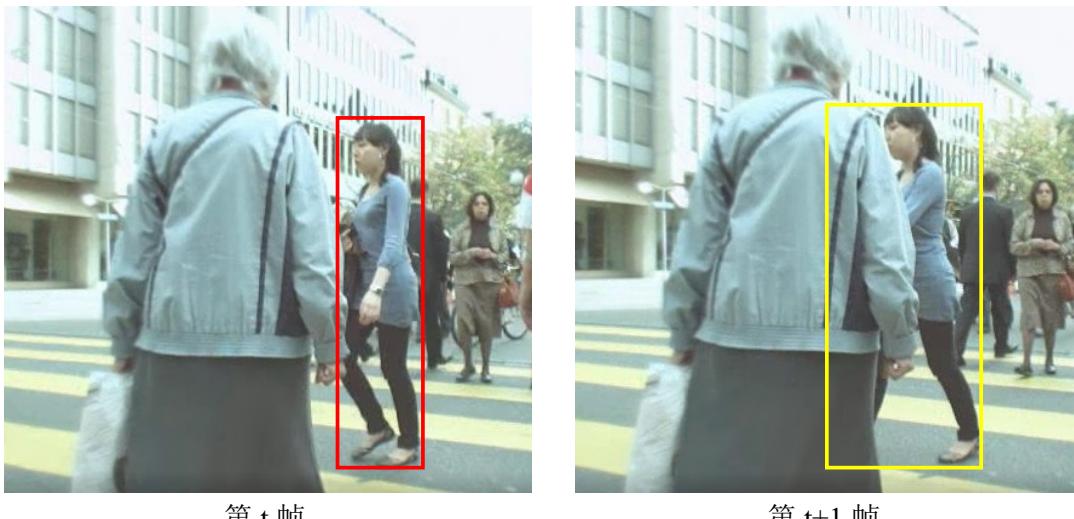


图 2-2 FairMOT 第二阶段数据关联示意图

## 2.6 实验结果与分析

### 2.6.1 数据集和评价指标

FairMOT 在训练时使用多目标跟踪、行人检测和行人搜索相关的数据集，ETH<sup>[75]</sup>、CityPerson<sup>[76]</sup>和 CrowdHuman<sup>[74]</sup>属于行人检测数据集，只提供物体包围框的标注，只用来训练目标检测分支。CalTech<sup>[77]</sup>、MOT17<sup>[15]</sup>属于多目标跟踪数据集，CUHK-SYSU<sup>[78]</sup>和 PRW<sup>[65]</sup>属于行人搜索数据集，这四个数据集都提供物体包围框和身份的标注，可以同时训练目标检测分支和行人重识别分支，详细训练策略在 2.2 节中。

FairMOT 在 MOT15<sup>[58]</sup>、MOT16<sup>[15]</sup>、MOT17<sup>[15]</sup>和 MOT20<sup>[59]</sup>四个数据集上评估

# 华中科技大学硕士学位论文

---

其跟踪性能。除 1.2.8 节中介绍的单类别多目标跟踪评价指标 MOTA、IDF1 和 IDs 之外，平均精度（Average Precision, AP）被用作评估目标检测性能的指标，错误接受率（False Accept Rate, FAR）为 0.1 时的真阳性率（True Positive Rate, TPR）被用作评估行人重识别特征区分力的指标，在提取行人重识别特征时使用物体包围框的真实值来提取，这样使得行人重识别的评估不会受到目标检测精度的影响，接着用每个特征去检索  $N$  个最相似的候选特征，最终得到错误接受率为 0.1 时的真阳性率。

## 2.6.2 实验环境和实现

本章方法使用的主要软件环境为 Python 3.8 和 Pytorch 1.7.0，其他相关的 pip 安装包包括 yacs、opencv-python、PyYAML、cython-bbox、scipy、progress、motmetrics、matplotlib、Pillow 等。硬件环境为两块 Nvidia RTX 2080 Ti 显卡，整个训练过程在 2 张 RTX 2080 Ti 显卡上花费 30 个小时，测试过程在单块显卡上的速度为每秒 25 帧，达到实时。

骨干网络默认使用改进过后的 DLA-34 网络<sup>[24]</sup>，模型参数先在 COCO 数据集<sup>[73]</sup>上进行预训练，再在 CrowdHuman<sup>[74]</sup>上使用弱监督的方法训练 60 轮迭代（epoch），最后在 2.6.1 节中提到的 6 个混合数据集上训练 30 轮迭代。优化器使用 Adam<sup>[81]</sup>，初始学习率设置为  $10^{-4}$ ，在 20 轮迭代后降为  $10^{-5}$ ，批大小为 12，训练时的数据增强包括旋转、缩放和色彩抖动等。输入图像的大小为  $1088 \times 608$ ，输出特征图的分辨率为  $272 \times 152$ 。

## 2.6.3 性能比较

FairMOT 和当前最先进的（state-of-the-art, SOTA）多目标跟踪方法在 MOT15、MOT16、MOT17 和 MOT20 四个数据集的测试集上进行比较，其中包括单阶段的方法和双阶段的方法，单阶段方法的速度较快，双阶段的方法精度较高。由于 FairMOT 使用的自身模型得到的检测结果，所以属于“私有检测器”赛道，即不使用 MOTChallenge 数据集官方提供的检测结果，使用自身模型得到的检测结果。结果见

# 华中科技大学硕士学位论文

---

表 2-1，所有结果均来自 MOTChallenge 竞赛官网<sup>2</sup>，MOTChallenge 是全球最权威的多目标行人跟踪挑战赛，↑代表该指标越高越好，↓代表越低越好，每个数据集上最高的指标加粗显示。

表 2-1 FairMOT 与多目标跟踪最先进方法比较

数据集	方法	MOTA↑	IDF1↑	MT↑	ML↓	IDs↓	FPS↑
MOT15	MDP_SubCNN <sup>[82]</sup>	47.5	55.7	30.0%	18.6%	628	<1.7
	CDA_DDAL <sup>[83]</sup>	51.3	54.1	36.3%	22.2%	544	<1.2
	EAMTT <sup>[84]</sup>	53.0	54.0	35.9%	19.6%	753	<4.0
	AP_HWDPL <sup>[85]</sup>	53.0	52.2	29.1%	20.2%	708	6.7
	RAR15 <sup>[86]</sup>	56.5	61.3	45.1%	14.6%	<b>428</b>	<3.4
	TubeTK*[52]	58.4	53.1	39.3%	18.0%	854	5.8
	<b>FairMOT(本文方法)*</b>	<b>60.6</b>	<b>64.7</b>	<b>47.6%</b>	<b>11.0%</b>	591	<b>30.5</b>
MOT16	EAMTT <sup>[84]</sup>	52.5	53.5	19.9%	34.9%	910	<5.5
	SORTwHPD16 <sup>[28]</sup>	59.8	53.8	25.4%	22.7%	1423	<8.6
	DeepSORT_2 <sup>[9]</sup>	61.4	62.2	32.8%	18.2%	781	<6.4
	RAR16wVGG <sup>[86]</sup>	63.0	63.8	39.9%	22.1%	<b>482</b>	<1.4
	TubeTK*[52]	64.0	59.4	33.5%	19.4%	1117	1.0
	JDE*[37]	64.4	55.8	35.4%	20.0%	1544	18.5
	CNNMTT <sup>[87]</sup>	65.2	62.2	32.4%	21.3%	946	<5.3
	POI <sup>[64]</sup>	66.1	65.1	34.0%	20.8%	805	<5.0
	CTrackerV1*[88]	67.7	57.2	32.9%	23.1%	1897	6.8
	<b>FairMOT(本文方法)*</b>	<b>74.9</b>	<b>72.8</b>	<b>44.7%</b>	<b>15.9%</b>	1074	<b>25.9</b>
MOT17	SST <sup>[89]</sup>	52.4	49.5	21.4%	30.7%	8431	<3.9
	TubeTK*[52]	63.0	58.6	31.2%	19.9%	4137	3.0
	CTrackerV1*[88]	66.6	57.4	32.3%	24.2%	5529	6.8
	CenterTrack*[30]	67.8	64.7	34.6%	24.6%	<b>2583</b>	17.5
	<b>FairMOT(本文方法)*</b>	<b>73.7</b>	<b>72.3</b>	<b>43.2%</b>	<b>17.3%</b>	3303	<b>25.9</b>
MOT20	<b>FairMOT(本文方法)*</b>	<b>61.8</b>	<b>67.3</b>	<b>68.8%</b>	<b>7.6%</b>	<b>5243</b>	<b>13.2</b>

\*注：带\*的代表单阶段多目标跟踪方法，不带\*的代表双阶段方法。

FairMOT 在四个数据集上都取得了第一名的结果，并且在 MOTA 和 IDF1 等主流指标上大幅领先第二名的方法，值得注意的是 FairMOT 的速度也是最快的，达到实

<sup>2</sup> <https://motchallenge.net/>

# 华中科技大学硕士学位论文

---

时 (25 FPS)，相比之下，许多性能还不错的两阶段方法<sup>[64,87]</sup>速度却要慢很多。其中，JDE 也是单阶段联合检测和重识别的方法，从表 2-1 中可以看出，MOT16 数据集上 FairMOT 比 JDE 在 MOTA 指标上高 10.5 个百分点，IDF1 指标上高 17.0 个百分点，并且降低 30% 的 IDs。这个结果说明基于中心点特征的 FairMOT 方法比基于锚框特征的 JDE 方法能够取得更好的跟踪性能，使目标检测和行人重识别被公平对待。图 2-3 中展示了一些 FairMOT 在 MOT17 数据集上的可视化结果，不同颜色的包围框代表不同的行人身份，图中每行为 MOT17 测试集中的一个视频中一定间隔的三帧。从 MOT17-01 视频结果可以看出，本方法对行人交错的情况处理得很好，黄色框和紫色框的行人交错经过之后各自的身份仍然能保持不变，可以看出本方法提取的重识别特征具有很强的区分能力。从 MOT17-03 的结果可以看出本方法在人群拥挤的场景下仍然能够取得很好的效果，验证了基于中心点特征的模型的有效性。



图 2-3 FairMOT 在 MOT17 测试集上的可视化结果图

## 2.6.4 消融对比实验

特征点比锚框更适合提取行人重识别特征 本部分采用 5 种不同的方式来从特征图上提取重识别特征，如表 2-2 所示，其中 ROI-Align、POS-Anchor 和 Two-Stage 这 3 种方法都是使用锚框来提取重识别特征，Center 和 Center-BI 这 2 种方法是使用中心点来提取特征。ROI-Align 和 Two-Stage 都来自于 Faster R-CNN<sup>[17]</sup>，即使用特征图上整个锚框内的特征作为重识别特征，Two-Stage 的做法是先做检测，再根据检测结果提取重识别特征。POS-Anchor 是 JDE<sup>[37]</sup>中的做法，即真实框附近的所有正样本锚框都会用来提取重识别特征。Center 是 FairMOT 中的标准做法，即只提取物体中心点处的特征作为重识别特征，Center-BI 是对带有小数点的中心点坐标进行双线性差值<sup>[39]</sup>，得到更加精确的中心点位置。表 2-2 中的结果显示，基于特征点的方法 Center 和 Center-BI 能够取得比基于锚框的方法更好的跟踪性能，尤其是 IDF1、IDs 和 TPR 这三个更偏向关联性能的指标，以上结果表明基于中心点的行人重识别特征提取方法能够得到更有区分力的特征。

表 2-2 基于锚框和基于中心点的重识别特征提取方法比较

特征提取方法	锚框	MOTA $\uparrow$	IDF1 $\uparrow$	IDs $\downarrow$	TPR $\uparrow$
ROI-Align <sup>[38]</sup>	✓	68.7	71.0	331	93.1
Two-Stage <sup>[38]</sup>	✓	69.0	68.2	388	90.5
POS-Anchor <sup>[37]</sup>	✓	69.0	70.3	434	93.9
<b>Center (本文方法)</b>		<b>69.1</b>	72.8	<b>299</b>	94.4
<b>Center-BI (本文方法)</b>		68.8	<b>74.3</b>	303	<b>94.9</b>

平衡多个任务的损失函数 本部分尝试几种多任务学习中平衡多个损失函数的方法，包括 Uncertainty<sup>[35]</sup>，GradNorm<sup>[34]</sup>和 MGDA-UB<sup>[36]</sup>，还有一个通过网格搜索得到的固定的损失函数权重的基线方法 Fixed，本文实现了两种不同的 Uncertainty 方法，第一种 Uncertainty-task 是只在目标检测的损失函数  $L_{detection}$  和  $L_{identity}$  之间使用 2 个不确定度估计权重，第二种 Uncertainty-branch 是在四个不同的 head 之间使用 4 个不确定度。最终结果如表 2-3 所示，可以看到 Fixed 方法取得了最高的 MOTA 和 AP，但是与 ID 相关的指标不行，说明模型更偏向目标检测，MGDA-UB 取得了最高的

# 华中科技大学硕士学位论文

TPR，但是 MOTA 非常低，说明模型更偏向行人重识别，Uncertainty 和 GradNorm 都能够取得不错的效果，由于 GradNorm 需要计算梯度，花费更长的训练时间，最终选择 Uncertainty-task 作为最后的损失函数，如公式（2-5）所示。

表 2-3 不同平衡多任务损失函数的方法比较

权重平衡方法	MOTA↑	IDF1↑	IDs↓	AP↑	TPR↑
<b>Fixed</b> (本文方法)	<b>69.6</b>	71.6	387	<b>81.9</b>	93.8
<b>Uncertainty-task</b> (本文方法)	69.1	72.8	<b>299</b>	81.2	94.4
Uncertainty-branch <sup>[35]</sup>	68.5	73.3	319	81.0	96.0
MGDA-UB <sup>[36]</sup>	63.6	67.9	355	78.5	<b>97.0</b>
GradNorm <sup>[34]</sup>	69.5	<b>73.8</b>	311	81.3	95.1

**多层特征融合对缓解任务冲突有帮助** 本文比较了多种不同的骨干网络对最终跟踪性能的影响，包括 ResNet<sup>[2]</sup>、DLA<sup>[24]</sup>、HRNet<sup>[71]</sup>、FPN<sup>[69]</sup>和 HarDNet<sup>[72]</sup>，结果如表 2-4 所示。对比 ResNet-34 和 ResNet-50 的结果可以发现，加深网络带来的收益很小，仅有 0.1 个百分点的 MOTA 提升。对比 ResNet-34 和 ResNet-34-FPN 可以发现，加入多层特征融合之后，目标检测和行人重识别的性能均有不错的提升，比如 AP 提高 2.6 个百分点，TPR 提高 3.3 个百分点，整体的跟踪指标也有提升。HRNet、DLA 和 HarDNet 中包含更多的多层特征融合结构，所以其整体结果也有大幅提升，因为 DLA-34 的速度是所有骨干网络中最快的，所以在 FairMOT 中默认使用 DLA-34 作为骨干网络。

表 2-4 不同骨干网络的性能比较

骨干网路	多层特征融合	MOTA↑	IDF1↑	IDs↓	AP↑	TPR↑
ResNet-34		63.6	67.2	435	75.1	90.9
ResNet-50		63.7	67.7	501	75.5	91.9
ResNet-34-FPN	√	64.4	69.6	369	77.7	94.2
HRNet-W18	√	67.4	74.3	315	80.5	94.6
DLA-34	√	69.1	72.8	299	81.2	94.4
HarDNet-85	√	<b>71.2</b>	<b>74.5</b>	<b>198</b>	<b>82.6</b>	<b>95.8</b>

图 2-4 中更加直观地展示了基于中心点的特征提取方法和多层特征融合结构这两个重要部分对行人外观相似度的影响，该图是查询图像（query image）中红色框的行人在目标图像（target image）全图中的重识别特征相似度响应，其中查询图像中红色框的行人和目标图像中绿色框的行人身份相同，图 2-4 的目的是对查询图像中红色框的人，找到其在目标图像中最相似的地方，理想结果是目标图像中绿色框行人处的相似度最高，颜色越红代表该处的相似度越高。可以看出基于中心点的特征提取方法和多层特征融合结构对重识别的帮助非常大，尤其在行人较为密集的情况下，相似度最高处基本都在绿色框行人处，验证了这两个部分的重要性。

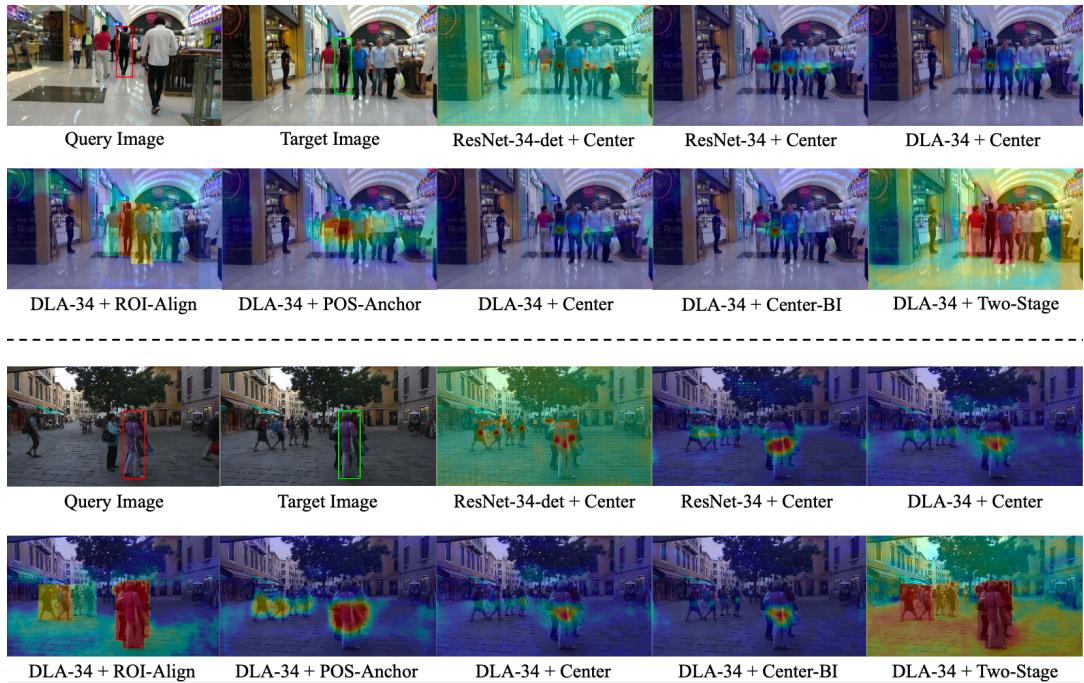


图 2-4 外观特征相似度响应图

**低维度的行人重识别特征更适合多目标跟踪** 在行人重识别任务中，通常使用 $1024^{[65]}$ 维度的特征，之前的单阶段多目标跟踪网络例如 JDE<sup>[37]</sup>通常使用 512 维度的行人重识别特征。本文在实验中发现特征维度在平衡目标检测和行人重识别上起到非常重要的作用，使用低维的重识别特征对检测的精度造成的损伤更小，并且能提高整体的跟踪速度。本文在 JDE 和 FairMOT 框架下比较不同的重识别特征维度对最后跟踪性能的影响，如表 2-5 所示，从表中可以看出 64 维度的重识别特征比 512 维度取得更好的 MOTA 和 AP 指标和更快的 FPS。

# 华中科技大学硕士学位论文

表 2-5 JDE 和 FairMOT 上不同重识别特征维度的性能比较

方法	维度	MOTA↑	IDF1↑	IDs↓	AP↑	TPR↑	FPS↑
JDE <sup>[37]</sup>	512	59.9	64.1	536	73.3	76.8	22.2
JDE <sup>[37]</sup>	64	<b>60.3</b>	<b>65.0</b>	<b>474</b>	<b>73.6</b>	<b>82.0</b>	<b>24.4</b>
FairMOT (本文方法)	512	68.5	<b>73.7</b>	312	80.9	<b>94.6</b>	24.1
FairMOT (本文方法)	64	<b>69.2</b>	73.3	<b>283</b>	<b>81.3</b>	94.3	<b>26.8</b>

**数据关联方法中各个组件的作用** 本部分评估数据关联中三个组件对最终跟踪性能的影响，包括包围框 IoU、重识别特征和卡尔曼滤波，结果如表 2-6 所示。如果只使用 IoU 的话会带来非常多的身份跳变，尤其在拥挤或者相机运动的情况下。使用重识别特征能够大幅提高 IDF1 的值，使得跟踪轨迹更加连贯，加上卡尔曼滤波之后又能再次提高结果，获得更加平滑的轨迹，最后加上 IoU 能够弥补一些重识别特征不够准确的情况，例如行人部分遮挡，三个组件在整个数据关联的过程中能起到互相补充的作用。

表 2-6 JDE 和 FairMOT 上不同重识别特征维度的性能比较

包围框 IoU	重识别特征	卡尔曼滤波	MOTA↑	IDF1↑	IDs↓
√			67.8	67.2	648
	√		68.1	70.3	435
	√	√	68.9	71.8	342
√	√	√	<b>69.1</b>	<b>72.8</b>	<b>299</b>

**整个系统运行时间** 本部分分模块地展示 FairMOT 整个系统的运行时间，包括目标检测、重识别匹配、卡尔曼滤波预测和 IoU 匹配。整个运行时间选择不同人群密度的视频进行测试，能够反映人数多少对各个模块的影响。整个运行时间如图 2-5 所示。其中提取目标检测和行人重识别特征的网络模块几乎不受到人数的影响，耗时稳定在 40 毫秒以下。卡尔曼滤波预测和 IoU 匹配的耗时在 1 毫秒到 2 毫秒左右，几乎可以忽略。重识别特征匹配的耗时随人数增长而线性增长，这是由于大量时间都花在每个跟踪轨迹的重识别特征更新上。

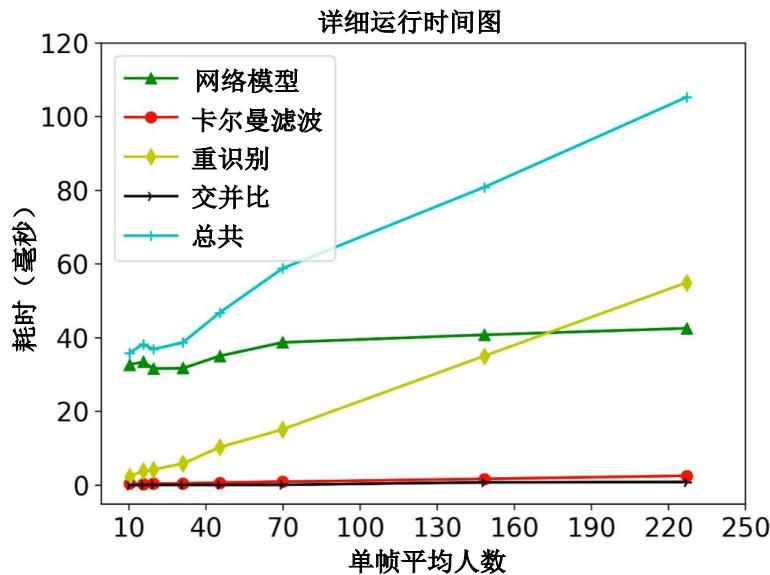


图 2-5 FairMOT 整个系统各个模块随人数变化的运行时间图

**不同数量的训练数据带来的性能提升** 本部分评估不同训练数据对最后跟踪结果的影响，结果如表 2-7 所示，其中 MIX 代表 2.4.1 节中提到的 ETH、CityPerson、Caltech、CUHK-SYSU 和 PRW 这五个数据集，CH 代表 CrowdHuman 数据集，所有结果都是在 MOT17 的测试集上得到的。从表 2-7 中可以看出只使用 MOT17 训练也能够得到不错的结果，随着数据量的增加，整体跟踪性能也会跟着提升，这在实际应用中是一个非常大的优势。

表 2-7 不同训练数据对跟踪性能的影响

训练数据	图像数	包围框数	身份数	MOTA↑	IDF1↑	IDs↓
MOT17	5K	112K	0.5K	69.8	69.9	3996
MOT17+MIX	54K	270K	8.7K	72.9	<b>73.2</b>	3345
MOT17+MIX+CH	73K	740K	8.7K	<b>73.7</b>	72.3	<b>3303</b>

## 2.7 本章小结

从探究单阶段多目标跟踪网络结果较双阶段方法相差较大的原因出发，本章发现其中最主要的原因是锚框的使用，多个锚框对应同一个物体和一个锚框包含多个物体会为网络中行人重识别部分的训练带来很大的模糊性，同时还存在目标检测和

# 华中科技大学硕士学位论文

---

行人重识别特征冲突和维度不一致的问题。为了解决这些问题，本章提出一个基于中心点特征的单阶段多目标跟踪网络 FairMOT，使用基于中心点的特征代替基于锚框的特征，避免重识别任务训练时的模糊性；在骨干网络中加入高低层特征融合结构，降低目标检测和行人重识别之间的特征冲突；降低重识别特征维度使其和目标检测特征维度更好地平衡。本方法在精度和速度上都大幅领先之前的方法，在最主流的多目标跟踪数据集 MOT17 上，MOTA 指标领先 6 个百分点，IDF1 指标领先 8 个百分点，速度首次达到实时（25 FPS）。本章的研究内容为多目标跟踪领域提供了一个简单有效的基线方法，并且在落地应用上也有很大的优势。

本章研究内容相关论文 “FairMOT: On the Fairness of Detection and Re-Identification in Multi-Object Tracking” 于 2021 年发表在国际计算机视觉顶级期刊 International Journal of Computer Vision 上，其 Github 开源代码获得 3.4k star 并且在工业界已经得到广泛应用，例如安防监控、智慧城市、智能超市等。

### 3 高低分检测框层次关联的多目标跟踪方法

#### 3.1 研究动机

本章将介绍一种高低分检测框层次关联的数据关联方法 ByteTrack，不像之前的多目标跟踪中的数据关联都只使用高分检测框进行关联，ByteTrack 考虑每一个检测框，从低分框中找出被遮挡、模糊的物体，大幅减少漏跟踪和轨迹中断的情况。因为每个检测框都是跟踪轨迹的基本单元，正如字节（Byte）是计算机程序的基本单元一样，本章提出的方法珍惜每一个检测框，所以命名为 ByteTrack。如图 3-1 所示，第一行是检测器得到的所有检测框，第二行是之前方法的结果，第三行是 ByteTrack 的跟踪结果。从第二行中可以看出之前方法只使用高分物体做数据关联会造成很多漏跟踪的情况，比如在第  $t_2$  帧和第  $t_3$  帧中便漏掉一些遮挡较为严重的目标。第三行中，加入高低分检测框层次关联的方法之后，可以跟踪到一些遮挡严重的行人，红色虚线框代表前一帧的跟踪轨迹经过卡尔曼滤波预测得到的在当前帧的位置，黄色框代表与其成功匹配的低分检测框，最终在第  $t_1$ 、 $t_2$  和  $t_3$  帧上得到完整的跟踪轨迹，并成功去除背景框。ByteTrack 是一种泛化性很强的数据关联方法，在“先检测后跟踪”的范式下，适用于任何检测器和不同的相似度计算方法。本章的研究内容对应的场景与第二章类似，在其基础上加入了驾驶场景，并对行人、车辆等多个类别的目标进行跟踪。本章将从主要思路、目标检测模型、重识别模型、数据关联方法和实验结果五个方面介绍 ByteTrack，其中重点是数据关联方法。

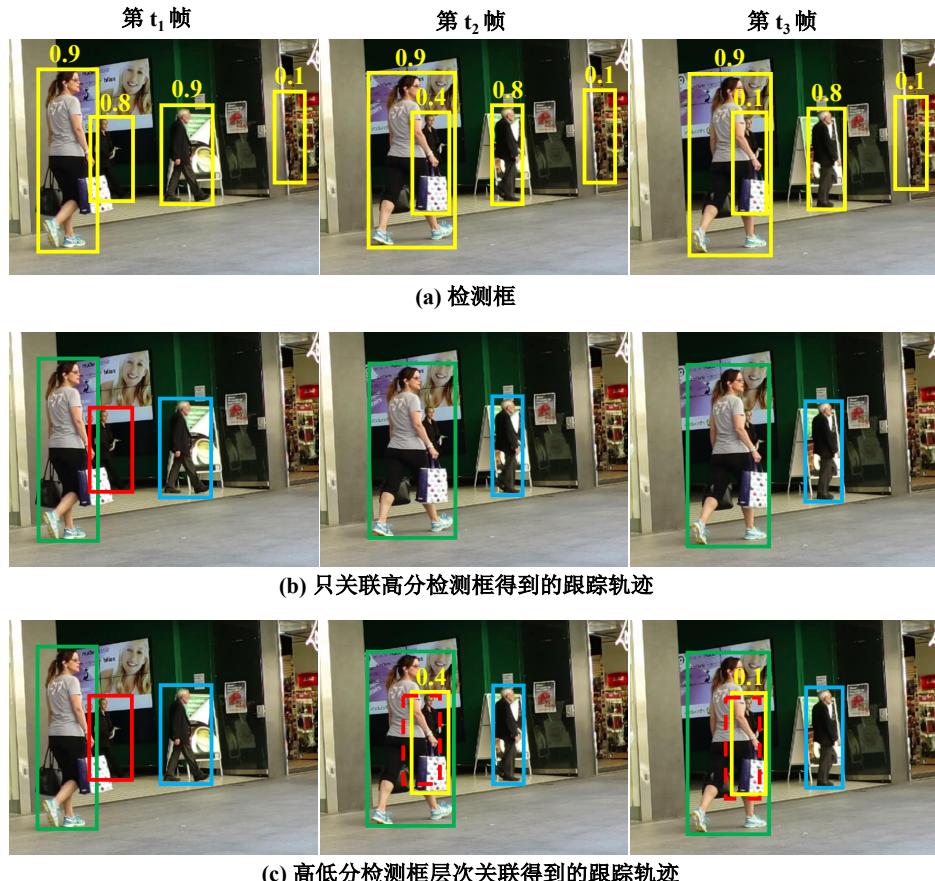


图 3-1 ByteTrack 和其他多目标跟踪方法的区别示意图

## 3.2 主要思路

针对之前多目标跟踪方法只使用高分检测框做跟踪带来的漏跟踪和轨迹中断的问题，本章提出使用高低分检测框层次关联的方法，先使用当前帧的高分检测框和先前帧的跟踪框进行匹配，形成跟踪轨迹，再使用低分检测框和没有匹配上高分检测框的跟踪框进行匹配，找出低分框中遮挡、模糊的物体，大幅减少漏跟踪和轨迹中断的情况。其中第一次匹配可以采用多样化的相似度，例如外观相似度和位置相似度，第二次匹配中相似度只能使用位置相似度。

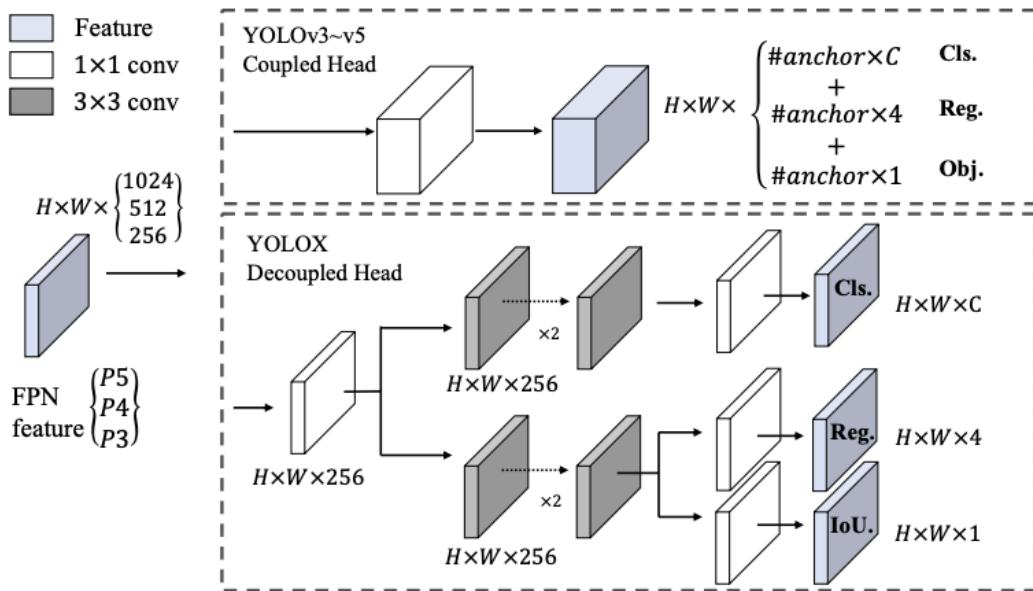
在多目标跟踪数据集场景中，行人之间的遮挡是一种非常普遍现象。在行人逐渐被遮挡的过程中，目标的检测得分往往伴随着遮挡的增多而降低，在遮挡逐渐消除时，检测得分也会随着遮挡的减少而升高，所以对遮挡物体的跟踪是保证跟踪轨

迹连贯性的关键。遮挡时物体的检测得分由高到低变化是高低分检测框层次关联方法有效找到遮挡物体的关键，如果该物体在第  $t_1$  帧得分高，在第  $t_2$  帧得分变低，那么在第  $t_2$  帧时便能使用低分框关联方法将该物体继续跟踪。虽然在物体被完全遮挡时本方法无法将其跟踪，但是在完全遮挡的这段时间中运动模型仍然会预测该物体的运动轨迹，在该物体又重新出现时将其和预测的运动轨迹进行匹配，便能够将身份找回。为了减少预测的模糊性，完全遮挡的时间应设置限定，超出时间设定之后便不会再进行运动预测，所以在有限的时间内使用低分框关联的方法将部分遮挡的物体跟踪上对保持整个跟踪轨迹的连贯性至关重要。

### 3.3 目标检测模型

本章使用的目标检测模型为 YOLOX<sup>[27]</sup>，是一个单阶段的目标检测模型，速度和精度能够达到一个很好的均衡。YOLOX 将之前 YOLO 系列的检测器<sup>[25,27]</sup>从基于锚框的形式转换成无需锚框的形式，还集成了许多先进的目标检测技术，包括将回归分支和分类分支分解，更强的数据增强（Mosaic<sup>[26]</sup>和 Mixup<sup>[113]</sup>）和高效的标签分配策略 SimOTA<sup>[114]</sup>等，最终在目标检测任务上达到最先进的（state-of-the-art，SOTA）水平。

YOLOX 采用和 YOLOv5<sup>[26]</sup>一样的骨干网络，都是改进后的 CSPNet<sup>[115]</sup>和一个额外的 PAN<sup>[116]</sup>输出分支提供多个分辨率的特征图。在骨干网络之后有两个分解的分支，一个负责回归，另一个负责分类。在回归分支中还有一个额外的交并比（Intersection over Union，IOU）感知分支，用来预测标注框和模型输出框之间的 IoU。回归分支对于每个像素点都会预测表示其位置的包围框的四个坐标值，例如左上角点的坐标和包围框的宽和高。回归分支使用 GIoU<sup>[117]</sup>损失进行监督，分类分支和 IoU 分支采用二元交叉熵损失进行监督，整个模型结构如图 3-2 所示。


 图 3-2 YOLOX 目标检测模型结构图<sup>[15]</sup>

YOLOX 中 SimOTA 标签分配策略会根据每个样本和标注的相似性自动地选择正样本，该相似性通过类别相似性和位置相似性的加权求和得到。它从物体中心点标注周围一定范围大小的区域内动态地选择相似度最高的 k 个 (top-k) 正样本。该标签分配策略有效地提升了目标检测的性能。

本章对 YOLOX 检测模型进行了改进使其更适用于多目标跟踪数据集 MOT17<sup>[15]</sup>。MOT17 需要得到行人的全身框，即使是被遮挡或者有部分在图像外面也需要被框住。YOLOX 的默认实现中将所有检测框的坐标都固定到图像区域内来，为避免在图像边界处的坐标错误，本方法在数据预处理和标签分配阶段不将物体原始包围框坐标固定到图像区域内，只删除在数据增强后全部在图像外的包围框。在 SimOTA 标签分配策略中，正样本需要分布在中心点周围，而有些目标的中心点在图像外，这时需要将中心点固定到图像区域内，但是四个顶点的坐标保持不变。MOT20<sup>[59]</sup>、HiEve<sup>[60]</sup> 和 BDD100K<sup>[61]</sup>这三个数据集的包围框坐标标注没有超过图像范围，所以不需要以上处理。

### 3.4 重识别模型

高低分检测框层次匹配方法具有高度的灵活性，可以不使用重识别模型，也可

以配合使用多种重识别模型。MOT17、MOT20 和 HiEve 数据集中视频帧率较高，相机运动不是很剧烈，只使用卡尔曼滤波运动模型就可以得到较好的跟踪结果，所以不使用重识别模型。BDD100K 数据集是面向自动驾驶场景的数据集，视频的标注帧率较低（5 FPS 左右），而且具有大量相机剧烈运动的情况（车辆拐弯），运动模型往往不能很好地预测物体运动，所以需要重识别模型提取物体的外观特征，利用外观相似度完成关联。

本章采用一个简洁、泛化性能好的方法 UniTrack<sup>[118]</sup>作为重识别模型，该方法最大的优点在于首次提出直接使用一个在 ImageNet 上进行预训练的分类模型作为跟踪任务中的重识别模型，不需要在某特定任务的数据集训练重识别模型。在 BDD100K 数据集的实验中，本方法使用在 ImageNet 上预训练的 ResNet-50<sup>[2]</sup>模型作为重识别模型，与常规行人重识别模型输入为每个物体的图像不同，本方法中重识别模型的输入是全图，在经过 ResNet-50 的重识别模型之后输出分辨率为输入图像大小 1/8 的特征图。接着根据每个物体的包围框，使用 ROI-Align<sup>[39]</sup>从输出特征图上提取每个物体的重识别特征。最后计算重识别特征之间的余弦距离作为最终的外观相似度。

## 3.5 数据关联方法

整个 ByteTrack 数据关联方法分为 5 部分：初始化跟踪轨迹、高分框匹配、低分框匹配、新建/删除轨迹和跟踪轨迹插值，其中跟踪轨迹插值是一个离线的后处理操作。除跟踪轨迹插值外，整个在线算法伪码描述如算法 3-1 所示。

### 3.5.1 初始化跟踪轨迹

ByteTrack 数据关联方法的输入是视频序列、检测器和超参数  $\tau$ ，代表低分检测框的和高分检测框的划分界限。该方法的输出是视频中所有跟踪轨迹，每个跟踪轨迹包含每帧中的包围框和身份。首先，对视频中的每一帧，检测器会预测出检测框和得分，根据阈值  $\tau$  可以将所有的检测框分成两部分  $D_{high}$  和  $D_{low}$ ，其中得分大于  $\tau$  的放入  $D_{high}$  中，得分小于  $\tau$  的放入  $D_{low}$  中。在视频第一帧中，将  $D_{high}$  中所有的物体都新建一个跟踪轨迹，得到所有跟踪轨迹  $T$ 。该部分位于算法 3-1 的第 1-13 行。

## 3.5.2 高分框匹配

从第二帧开始,先用卡尔曼滤波预测每个跟踪轨迹在当前帧的位置,然后将 $D_{high}$ 中所有检测框和 $T$ 中所有跟踪轨迹做第一次匹配,匹配时用到的相似度是通过 $D_{high}$ 中的检测框和 $T$ 中跟踪轨迹的预测框之间的交并比(Intersection over Union, IoU)计算得到,得到相似度之后使用匈牙利算法完成身份的分配。ByteTrack的灵活性非常高,在高分框匹配时可以使用多种相似度,比如加入外观相似度,可以使用重识别模型提取外观特征计算余弦距离,得到外观相似度,完成第一次匹配,整个第一次匹配过程位于算法3-1的第14-19行。

## 3.5.3 低分框匹配

在第一次匹配之后,对没有在 $D_{high}$ 中找到匹配对的跟踪轨迹 $T_{remain}$ ,将其与 $D_{low}$ 中的所有低分检测框做第二次匹配。非常重要的一点是在第二次匹配中相似度只能使用位置相似度IoU,因为低分框中的物体大部分都是遮挡严重的,外观特征非常不可靠,同样使用匈牙利算法得到物体的身份。本部分位于算法3-1的第20-21行。

## 3.5.4 新建轨迹和删除轨迹

在两次匹配都完成后,对于 $D_{high}$ 中没有匹配上跟踪轨迹的检测框,新建跟踪轨迹,对于 $D_{low}$ 中没有匹配上跟踪轨迹的检测框,把它们视为背景,全部舍弃。对于 $T$ 中在高分框匹配和低分框匹配中都没有匹配上检测框的跟踪轨迹,标记为丢失,保留特定的时间 $t$ ,比如30帧,输出时不输出丢失的轨迹,但是再之后做匹配时会考虑这些丢失的轨迹,防止其再次出现,这对保持轨迹的连贯性非常重要。如果一个跟踪轨迹丢失超过时间 $t$ ,那么就删除这个跟踪轨迹。本部分位于算法3-1的第22-24行。

## 3.5.5 跟踪轨迹插值

在MOT17数据集中有一些完全被遮挡(可见百分比为0)的物体也做了标注,要求被跟踪到。即使使用当前最先进的检测器,这种物体也几乎无法被检测到,所以本部分提出一种根据跟踪轨迹进行插值的方法来跟踪这类完全被遮挡的物体。

# 华中科技大学硕士学位论文

假设有一个跟踪轨迹 $T$ , 在第 $t_1$ 帧到 $t_2$ 帧之间由于遮挡, 没有跟踪到该物体, 该跟踪轨迹在第 $t_1$ 帧时候的包围框为 $B_{t_1} \in \mathbb{R}^4$ , 包含该包围框左上角和右下角的坐标,  $B_{t_2} \in \mathbb{R}^4$ 为 $T$ 在第 $t_2$ 帧时的包围框。接着设置一个超参数 $\sigma$ 代表进行跟踪轨迹插值的最大时间间隔帧数, 意味着当 $t_2 - t_1 \leq \sigma$ 时, 才会进行插值。跟踪轨迹 $T$ 在第 $t$ 帧时由插值得到的包围框能够通过下式计算得到:

$$B_t = B_{t_1} + (B_{t_2} - B_{t_1}) \cdot \frac{t - t_1}{t_2 - t_1} \quad (3-1)$$

其中 $t_1 < t < t_2$ 。在跟踪轨迹插值之后, 完成整个数据关联部分。

算法 3-1 高低分检测框层次关联算法

**输入:** 视频序列 $V$ ; 目标检测器 $Det$ ; 检测得分阈值 $\tau$

**输出:** 视频中的跟踪轨迹 $T$

```
/* 初始化跟踪轨迹 */
1: 初始化跟踪轨迹 $T \leftarrow \emptyset$ 
2: for 视频序列 $V$ 中的第 $f_k$ 帧 do
    /* 图 3-1 (a) */
    /* 预测检测框和得分 */
    3:  $D_k \leftarrow Det(f_k)$ 
    4:  $D_{high} \leftarrow \emptyset$ 
    5:  $D_{low} \leftarrow \emptyset$ 
    6: for  $D_k$ 中的每个检测框 $d$  do
        7:     if 检测框 $d$ 的得分大于 $\tau$  then
            8:          $D_{high} \leftarrow D_{high} \cup \{d\}$ 
        9:     end
        10:    else
            11:         $D_{low} \leftarrow D_{low} \cup \{d\}$ 
        12:    end
    13: end
    /* 预测每个跟踪轨迹的新位置 */
```

```
14: for  $T$  中的每个跟踪轨迹  $t$  do
15:      $t \leftarrow KalmanFilter(t)$ 
16: end
    /* 图 3-1 (b) */
    /* 高分框匹配 */
17: 使用相似度#1 关联跟踪轨迹  $T$  和检测框  $D_{high}$ 
18:  $D_{remain} \leftarrow D_{high}$  中没有匹配上跟踪轨迹的检测框
19:  $T_{remain} \leftarrow T$  中没有匹配上检测框的跟踪轨迹
    /* 图 3-1 (c) */
    /* 低分框匹配 */
20: 使用相似度#2 关联跟踪轨迹  $T_{remain}$  和低分检测框  $D_{low}$ 
21:  $T_{re-remain} \leftarrow T_{remain}$  中没有匹配上检测框的跟踪轨迹
    /* 删除没有匹配上检测框的跟踪轨迹 */
22:  $T \leftarrow T \setminus T_{re-remain}$ 
    /* 新建跟踪轨迹 */
23: for 高分框匹配中剩余的检测框  $D_{remain}$  中的每个检测框  $d$  do
24:      $T \leftarrow T \cup \{d\}$ 
25: end
26: end
27: 返回  $T$ 
```

## 3.6 实验结果

### 3.6.1 实验环境和实现

本章方法使用的主要软件环境为 Python 3.8 和 Pytorch 1.9.0，其他相关的 pip 安装包包括 loguru、opencv-python、scikit-image、cython-bbox、tqdm、thop、motmetrics、ninja、Pillow、lap 等。硬件环境为八块 Nvidia Tesla V100 显卡，测试过程在单块显卡上的速度为每秒 30 帧，达到实时。

# 华中科技大学硕士学位论文

---

本文将在 MOT17<sup>[15]</sup>、MOT20<sup>[59]</sup>、HiEve<sup>[60]</sup>和 BDD100K<sup>[61]</sup>四个数据集上对 ByteTrack 进行测试，额外的训练数据使用 CrowdHuman<sup>[74]</sup>、ETH<sup>[75]</sup>和 CityPerson<sup>[76]</sup>。评价指标除使用 1.2.8 节中提到的 MOTA、IDF1 和 IDs 之外，在测试集上还使用近期新出现的对跟踪性能评价更加合理的评价指标 HOTA<sup>[90]</sup>，显式的平衡了多目标跟踪中检测、关联和定位的表现。BDD100K 中还加入了多类别的跟踪指标 mMOTA 和 mIDF1，分别为每个类别的 MOTA 指标和 IDF1 指标的平均。

ByteTrack 的检测器使用的是 YOLOX<sup>[27]</sup>，模型参数使用在目标检测数据集 COCO 上预训练过的模型参数，训练数据混合了 MOT17、CrowdHuman、ETH 和 CityPerson 四个数据集，总共训练 80 轮迭代（epoch），优化器采用随机梯度下降优化器，权重衰减设置为  $5 \times 10^{-4}$ ，动量为 0.9，初始学习率为是，最开始 1 个迭代使用热启动（warm-up），采用余弦退火（cosine annealing）的学习率策略，批大小（batch size）为 48，整个训练过程在 8 张 NVIDIA Tesla V100 显卡上需要训练 12 个小时。

ByteTrack 的数据关联方法中，默认的高低分检测框阈值  $\tau$  为 0.6，在匹配阶段，如果相似度 IoU 小于 0.2，那么检测框和跟踪轨迹不能够进行匹配。对于标记为丢失的跟踪轨迹，会保留  $t$  为 30 帧。

## 3.6.2 性能比较

ByteTrack 和当前最先进的(state-of-the-art, SOTA)多目标跟踪方法在 MOT17、MOT20、HiEve 和 BDD100K 四个数据集的私有检测器赛道（能够使用自己得到的检测结果）上进行比较，结果分别在表 3-1、3-2、3-3 和 3-4 中，所有结果都通过官网测试得到<sup>3,4,5</sup>。

**1) MOT17 数据集** ByteTrack 在 MOT17 排行榜上位列第一名，如表 3-1 所示，它不仅取得最高的精度（80.3 MOTA，77.3 IDF1，63.1 HOTA），还取得最快的速度（30 FPS），性能大幅领先第二名的方法 ReMOT<sup>[91]</sup> (+3.3 MOTA, +5.3 IDF1, +3.4

---

<sup>3</sup> <https://motchallenge.net/>

<sup>4</sup> <http://humaninevents.org/>

<sup>5</sup> <https://eval.ai/web/challenges/challenge-page/1259/overview>

# 华中科技大学硕士学位论文

HOTA)。并且, ByteTrack 所需的训练数据(29K 张图像)比许多高性能方法<sup>[93,94]</sup>(73K 张图)都要少。值得注意的是 ByteTrack 只使用基于卡尔曼滤波的最简单的相似度计算方法, 不像其他方法<sup>[94,31]</sup>加入行人重识别特征或者注意力机制。所有这些表明 ByteTrack 是一个简单又强大的多目标跟踪方法。

表 3-1 ByteTrack 在 MOT17 测试集上与 SOTA 方法的比较

方法	MOTA↑	IDF1↑	HOTA↑	FP↓	FN↓	IDs↓	FPS↑
Tube_TK <sup>[52]</sup>	63.0	58.6	48.0	27060	177483	4137	3.0
MOTR <sup>[43]</sup>	65.1	66.4	-	45486	149307	2049	-
CTracker <sup>[88]</sup>	66.6	57.4	49.0	22284	160491	5529	6.8
CenterTrack <sup>[30]</sup>	67.8	64.7	52.2	<b>18498</b>	160332	3039	17.5
QuasiDense <sup>[95]</sup>	68.7	66.3	53.9	26589	146643	3378	20.3
TraDes <sup>[32]</sup>	69.1	63.9	52.7	20892	150060	3555	17.5
SOTMOT <sup>[96]</sup>	71.0	71.9	-	39537	118983	5184	16.0
TransCenter <sup>[97]</sup>	73.2	62.2	54.5	23112	123738	4614	1.0
GSDT <sup>[98]</sup>	73.2	66.5	55.2	26397	120666	3891	4.9
FairMOT(第二章方法)	73.7	72.3	59.3	27507	117477	3303	25.9
RelationTrack <sup>[99]</sup>	73.8	74.7	61.0	27999	118623	<b>1374</b>	8.5
PermaTrackPr <sup>[100]</sup>	73.8	68.9	55.5	28998	115104	2699	11.9
CSTrack <sup>[92]</sup>	74.9	72.6	59.3	23847	114303	3567	15.8
TransTrack <sup>[31]</sup>	75.2	63.5	54.1	50157	86442	3603	10.0
SiamMOT <sup>[93]</sup>	76.3	72.3	-	-	-	-	12.8
CorrTracker <sup>[94]</sup>	76.5	73.6	60.7	29808	99510	3369	15.6
ReMOT <sup>[91]</sup>	77.0	72.0	59.7	33204	93612	2853	1.8
<b>ByteTrack(本文方法)</b>	<b>80.3</b>	<b>77.3</b>	<b>63.1</b>	25491	<b>83721</b>	2196	<b>29.6</b>

**2) MOT20 数据集** 和 MOT17 相比, MOT20 中行人拥挤和遮挡的情况多很多, 平均一帧中有 170 个人。ByteTrack 同样在 MOT20 的排行榜上位列第一, 在所有指标上都大幅领先其余 SOTA 方法。如表 3-2 所示, 将之前的最高 MOTA 从 68.6 提升到 77.8, IDF1 从 71.4 提升到 75.2, 把 IDs 从 4209 降低到 1223, 降低足足 71%, 这个极低的身份跳变次数表明加入低分检测框关联在遮挡场景下非常有效, 能够有效跟踪到一些部分遮挡/严重遮挡的行人。

表 3-2 ByteTrack 在 MOT20 测试集上与 SOTA 方法的比较

方法	MOTA↑	IDF1↑	HOTA↑	FP↓	FN↓	IDs↓	FPS↑
MLT <sup>[101]</sup>	48.9	54.6	43.2	45660	216803	2187	3.7
FairMOT(第二章方法)	61.8	67.3	54.6	103440	88901	5243	13.2
TransCenter <sup>[97]</sup>	61.9	50.4	-	45895	146347	4653	1.0
TransTrack <sup>[31]</sup>	65.0	59.4	48.5	27197	150197	3608	7.2
CorrTracker <sup>[94]</sup>	65.2	69.1	-	79429	95855	5183	8.5
CSTrack <sup>[92]</sup>	66.6	68.6	54.0	<b>25404</b>	144358	3196	4.5
GSDT <sup>[98]</sup>	67.1	67.5	53.6	31913	135409	3131	0.9
SiamMOT <sup>[93]</sup>	67.1	69.1	-	-	-	-	4.3
RelationTrack <sup>[99]</sup>	67.2	70.5	56.5	61134	104597	4243	2.7
SOTMOT <sup>[96]</sup>	68.6	71.4	-	57064	101154	4209	8.5
<b>ByteTrack(本文方法)</b>	<b>77.8</b>	<b>75.2</b>	<b>61.3</b>	26249	<b>87594</b>	<b>1223</b>	<b>17.5</b>

3) **HiEve 数据集** 和 MOT17 和 MOT20 相比, HiEve 数据集包含更多更复杂的场景和更广阔的相机视角。本文使用 CrowdHuman 和 HiEve 的训练集来训练 ByteTrack, 也达到 HiEve 数据集排行榜第一, 并且大幅领先之前 SOTA 方法, 如表 3-3 所示, 将 MOTA 从 40.9 提升到 61.3, IDF1 从 45.1 提升到 62.9, 这也表明 ByteTrack 对复杂的场景非常鲁棒。

表 3-3 ByteTrack 在 HiEve 测试集上与 SOTA 方法的比较

方法	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
DeepSORT <sup>[9]</sup>	27.1	28.6	8.5%	41.5%	5894	42668	2220
MOTDT <sup>[40]</sup>	26.1	32.9	8.7%	54.6%	6381	43577	1599
IoUTracker <sup>[29]</sup>	38.6	38.6	28.3%	27.6%	9640	28993	4153
JDE <sup>[37]</sup>	33.1	36.0	15.1%	24.1%	9526	33327	3747
FairMOT(第二章方法)	35.0	46.7	16.3%	44.2%	6523	37750	<b>995</b>
CenterTrack <sup>[30]</sup>	40.9	45.1	10.8%	32.2%	3208	36414	1568
<b>ByteTrack(本文方法)</b>	<b>61.7</b>	<b>63.1</b>	<b>38.3%</b>	<b>21.6%</b>	<b>2822</b>	<b>22852</b>	1031

**4) BDD100K 数据集** 该数据集是一个多类别多目标跟踪数据集，面向自动驾驶场景，需要跟踪行人、汽车、卡车、自行车等 8 个类别的物体。主要的挑战在于低帧率和相机剧烈运动（车辆拐弯时）。在 BDD100K 上的实验中，高分框匹配时只使用外观相似度，不使用运动相似度。卡尔曼滤波在低帧率和相机运动剧烈时的预测结果很差。ByteTrack 在 BDD100K 排行榜上也排名第一，如表 3-4 所示，在验证集上将 mMOTA 从 36.6 提高到 45.5，在测试集上从 35.5 提高到 40.1。实验结果表明本方法能够有效解决自动驾驶场景中的多目标跟踪挑战。

表 3-4 ByteTrack 在 BDD100K 验证集和测试集上与 SOTA 方法的比较

方法	划分	mMOTA↑	mIDF1↑	FN↓	FP↓	IDs↓	MT↑	ML↓
Yu et al. <sup>[61]</sup>	验证集	25.9	44.5	122406	52372	8315	8396	3795
QDTrack <sup>[95]</sup>	验证集	36.6	50.8	108614	46621	<b>6262</b>	9481	3034
<b>ByteTrack(本文方法)</b>	验证集	<b>45.5</b>	<b>54.8</b>	<b>92805</b>	<b>34998</b>	9140	<b>9626</b>	<b>3005</b>
Yu et al. <sup>[61]</sup>	测试集	26.3	44.7	213220	100230	14674	16299	6017
QDTrack <sup>[95]</sup>	测试集	35.5	52.3	201041	80054	<b>10790</b>	17353	5167
<b>ByteTrack(本文方法)</b>	测试集	<b>40.1</b>	<b>55.8</b>	<b>169073</b>	<b>63869</b>	15466	<b>18057</b>	<b>5107</b>

### 3.6.3 消融对比实验

在所有的消融实验中，ByteTrack 都在 CrowdHuman 和 MOT17 half train 数据集上训练，最后在 MOT17 half val 数据集上测试得到结果。

**和其他数据关联方法的比较** 高低分层次关联方法 ByteTrack 本质上是一种数据关联的方法，所以本节会将 ByteTrack 和其他一些比较流行的数据关联方法做比较，包括 SORT<sup>[28]</sup>，DeepSORT<sup>[9]</sup>和 MOTDT<sup>[40]</sup>。为了只比较数据关联的性能，所有方法都使用相同的检测结果，都来自于 YOLOX<sup>[27]</sup>，结果如表 3-5 所示。

表 3-5 ByteTrack 和其余数据关联方法的比较

方法	重识别模型	MOTA↑	IDF1↑	IDs↓	FPS↑
SORT <sup>[28]</sup>		74.6	76.9	291	<b>30.1</b>
DeepSORT <sup>[9]</sup>	√	75.4	77.2	239	13.5
MOTDT <sup>[40]</sup>	√	75.8	77.6	273	11.1
<b>ByteTrack(ours)</b>		<b>76.6</b>	<b>79.3</b>	<b>159</b>	29.6

# 华中科技大学硕士学位论文

---

SORT 可以视作一个基线方法，因为和 ByteTrack 一样都只使用卡尔曼滤波预测物体在下一帧的运动。从表 3-5 可以看出 ByteTrack 将 SORT 的 MOTA 指标从 74.6 提升到 76.6, IDF1 指标从 76.9 提升到 79.3, IDs 指标从 291 降低到 159。因为 SORT 只使用高分检测框做关联，这个明显的提升表明加入低分检测框关联之后是有能力从低分框中找到真正的物体，去除背景，提高跟踪性能。

DeepSORT 使用额外的行人重识别模型增强行人被遮挡又出现这种情况的长距离的关联。和 DeepSORT 相比，ByteTrack 同样有很大的优势，这表明简单的卡尔曼滤波有能力进行长距离关联，在检测结果不错的情况下能够比行人重识别模型取得更好的 IDF1 和更少的 IDs。这是因为在严重的遮挡情况下，重识别特征非常不可靠，往往导致更多的身份跳变，卡尔曼滤波这种运动模型反而效果更好。

MOTDT 加入运动指导的包围框传播结果，即将卡尔曼滤波预测的框和实际检测到的框做了择优筛选，但效果还是不如本文提出的加入低分框匹配的方法，原因是 MOTDT 使用部分卡尔曼滤波预测的框作为最后的跟踪轨迹输出结果，在相机移动过大时会带来一些定位的漂移情况，而在 ByteTrack 中使用的低分检测框能够更加精确地找到被遮挡或者模糊的物体，如图 3-3 所示，中间一列红色三角形标注的就是找到的遮挡严重或者模糊的物体。

**轻量检测模型下跟踪性能的比较** 本部分使用轻量的检测骨干网络，在没那么精准的检测结果的情况下比较 ByteTrack 和 DeepSORT 的跟踪性能。检测模型使用 YOLOX 中的不同的骨干网络，输入图像大小为  $608 \times 1088$ , 608 是图像的高，1088 是宽。结果显示在表 3-6 中，可以看出和 DeepSORT 相比，ByteTrack 的关联方法能够带来稳定的 MOTA 和 IDF1 的提升，证明 ByteTrack 对检测的性能非常鲁棒，建议改为 不管是使用高性能的检测器还是低性能的检测器，都能够比 DeepSORT 有更大的提升。值得注意的是使用 YOLOX-Nano 作为骨干网络时，ByteTrack 比 DeepSORT 的 MOTA 高 3 个百分点，这让其在实际应用中拥有更大的优势。

# 华中科技大学硕士学位论文

表 3-6 轻量骨干网络下 ByteTrack 与 DeepSORT 的跟踪性能比较

骨干网络	参数量	计算量	方法	MOTA↑	IDF1↑	IDs↓
YOLOX-S	8.9 M	43.0 G	DeepSORT	69.6	71.5	<b>205</b>
YOLOX-S	8.9 M	43.0 G	ByteTrack	<b>71.1</b>	<b>73.6</b>	224
YOLOX-Tiny	5.0 M	24.5 G	DeepSORT	68.6	72.0	224
YOLOX-Tiny	5.0 M	24.5 G	ByteTrack	<b>70.5</b>	<b>72.1</b>	<b>222</b>
YOLOX-Nano	0.9 M	4.0 G	DeepSORT	61.4	66.8	212
YOLOX-Nano	0.9 M	4.0 G	ByteTrack	<b>64.4</b>	<b>68.4</b>	<b>161</b>

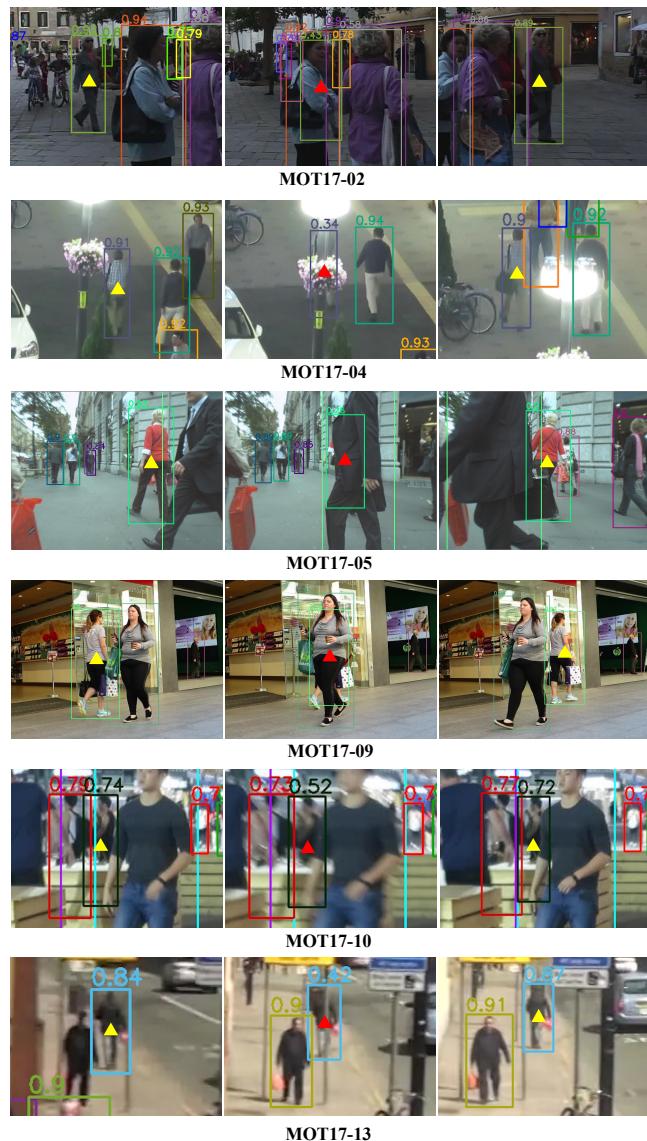


图 3-3 ByteTrack 在 MOT17 验证集上的可视化跟踪结果

**对检测得分阈值的鲁棒性** 检测得分阈值 $\tau$ 在多目标跟踪中是一个非常敏感的超参数，需要仔细地调试出合适的值。本节比较了 $\tau$ 在 0.2 到 0.8 之间时 SORT 和 ByteTrack 的性能，结果如图 3-4 所示，红线代表 ByteTrack，缩写为 BYTE，蓝线代表 SORT。可以看出 ByteTrack 对超参数的鲁棒性比 SORT 要好很多，原因是低分检测框匹配中考虑很多得分在 $\tau$ 以下的框，所以不管 $\tau$ 如何变化，都能考虑到所有的检测框。

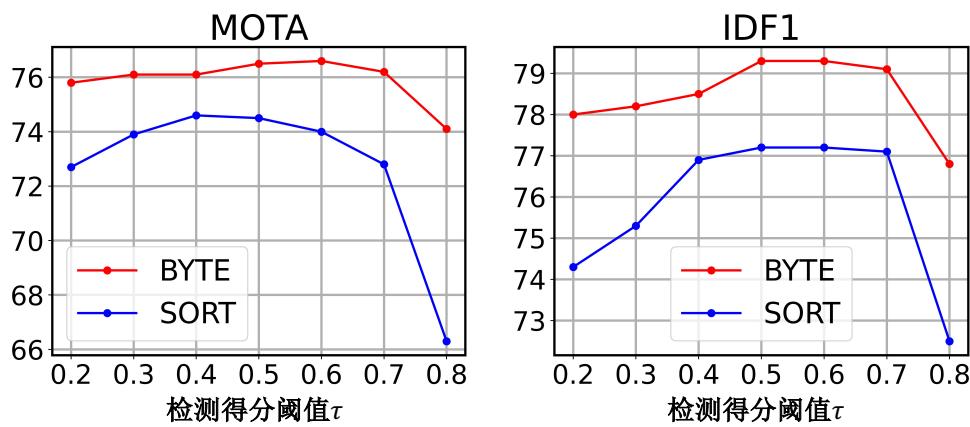


图 3-4 ByteTrack 和 SORT 在不同检测得分阈值下的 MOTA 和 IDF1 比较图

**对低分检测框的分析** 为证明高低分检测框层次关联比只使用高分框关联更有效，本节收集由 ByteTrack 得到的低分框中真阳性(True Positives, TPs)和假阳性(False Positives, FPs)的个数。首先，把所有检测框中得分在 $\tau$ 以下的都筛选出来，和标签做比对，划分出 TPs 和 FPs。接着从 ByteTrack 的跟踪结果中选出得分小于 $\tau$ 的跟踪框，也和标签做对比，得到 TPs 和 FPs，和总的低分检测框的 TPs 和 FPs 一起画在图 3-5 中，淡颜色的表示整个低分检测框中的 TP 和 FP 的数量，深颜色的表示 ByteTrack 的低分跟踪框中的 TP 和 FP 的数量。从图中可以看出 ByteTrack 能够从低分检测框中找到比 FP 多很多的 TP，即使很多视频（比如 MOT17-02）的低分检测框中本身含有更多的 FP。从低分检测框中获取的 TPs 很明显的将 MOTA 指标从 74.6 提升到 76.6，如图 3-5 所示。

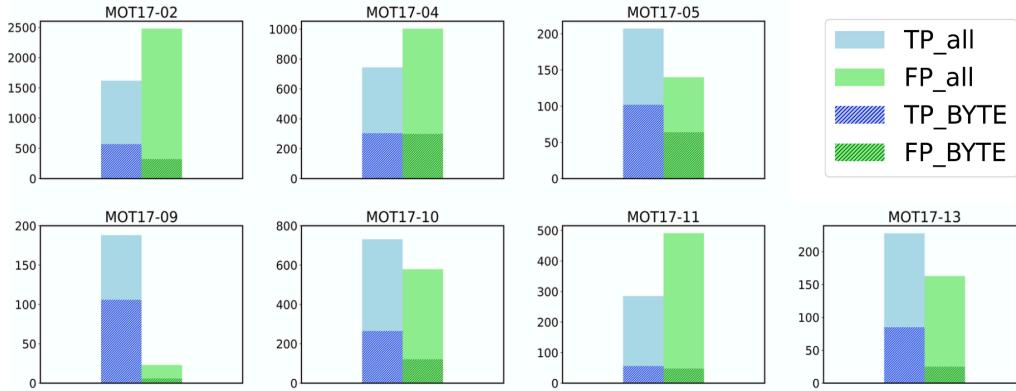


图 3-5 ByteTrack 在 MOT17 数据集不同视频上找回物体的 FP 和 TP 比较

将高低分检测框层次关联方法应用到其他多目标跟踪方法上 在本节中，该数据关联方法会被应用到其他 7 个当前最先进的 (state-of-the-arts, SOTA) 多目标跟踪方法上，包括 JDE<sup>[37]</sup>、CSTrack<sup>[92]</sup>、FairMOT (本文第二章方法)、QuasiDense<sup>[95]</sup>、CenterTrack<sup>[30]</sup>、Chained-Tracker<sup>[88]</sup>和 TransTrack<sup>[31]</sup>。在所有这些方法中，JDE、CSTrack、FairMOT 和 TraDes 使用运动相似度和外观相似度的结合，QuasiDense 只使用外观相似度，CenterTrack 和 TraDes 使用网络预测运动相似度，Chained-Tracker 使用链式结构同时输出连续两帧的结果，在同一帧内使用包围框之间的 IoU 做关联。TransTrack 和 MOTR 使用注意力机制<sup>[41]</sup>在帧间传播包围框。总的结果如表 3-7 所示，为验证 ByteTrack 的有效性，本节设置两种不同的模式将 ByteTrack 应用到其他的多目标跟踪方法上。

第一种模式是将低分检测框匹配插入到其他方法原始的关联过程中，结果在表 3-7 的每种方法的第二行。用 FairMOT 举例，在 FairMOT 完成本身的关联过程之后，将所有没有匹配上检测框的跟踪轨迹收集起来，再和低分检测框使用 IoU 进行匹配，和 3.4.3 节中的过程一样。这里值得注意的一点是对于得分较低的物体，重识别特征往往不太可靠，所以只用 IoU 进行匹配。在 Chained-Tracker 中，本节不使用第一种模式，因为在链式结构中无法进行。

第二种模式是直接使用每个多目标跟踪方法得到的检测框，然后使用 3.4 节中完整的高低分检测框层次关联方法，得到跟踪结果，结果在表 3-7 的每种方法的第三行。

在两个模式下，ByteTrack 都能带来稳定的提升，包括几乎所有的指标例如 MOTA，

# 华中科技大学硕士学位论文

IDF1 和 IDs。例如在 CenterTrack 中使用 ByteTrack 的关联方法，提升 1.3 个百分点的 MOTA 和 9.8 个百分点的 IDF1，在 Chained-Tracker 中使用 ByteTrack 能提升 1.9 个百分点的 MOTA 和 5.8 个百分点的 IDF1，TransTrack 上能提高 1.2 个百分点的 MOTA 和 4.1 个百分点的 IDF1。最终结果表明 ByteTrack 具有非常好的泛化能力，在现有的跟踪器上都能得到稳定的性能提升。

表 3-7 将高低分检测框层次关联应用到其他多目标跟踪方法上的性能比较

方法	相似度	加入本章方法	MOTA↑	IDF1↑	IDs↓
JDE <sup>[37]</sup>	运动+重识别	不加入	60.0	63.6	473
	运动+重识别	模式一	60.3 (+0.3)	64.1 (+0.5)	418
	运动	模式二	60.6 (+0.6)	66.0 (+2.4)	360
CSTrack <sup>[92]</sup>	运动+重识别	不加入	68.0	72.3	325
	运动+重识别	模式一	69.2 (+1.2)	73.9 (+1.6)	285
	运动	模式二	69.3 (+1.3)	71.7 (-0.6)	279
FairMOT (第二章方法)	运动+重识别	不加入	69.1	72.8	299
	运动+重识别	模式一	70.4 (+1.3)	74.2 (+1.4)	232
	运动	模式二	70.3 (+1.2)	73.2 (+0.4)	236
QuasiDense <sup>[95]</sup>	重识别	不加入	67.3	67.8	377
	运动+重识别	模式一	67.7 (+0.4)	72.0 (+4.2)	281
	运动	模式二	67.9 (+0.6)	70.9 (+3.1)	258
CenterTrack <sup>[30]</sup>	运动	不加入	66.1	64.2	528
	运动	模式一	66.3 (+0.2)	64.8 (+0.6)	334
	运动	模式二	67.4 (+1.3)	74.0 (+9.8)	144
Chained-Tracker <sup>[88]</sup>	链式	不加入	63.1	60.9	755
	运动	模式二	65.0 (+1.9)	66.7 (+5.8)	346
TransTrack <sup>[31]</sup>	注意力机制	不加入	67.1	68.3	254
	注意力机制	模式一	68.6 (+1.5)	69.0 (+0.7)	232
	运动	模式二	68.3 (+1.2)	72.4 (+4.1)	181

**速度与精度的权衡** 本部分使用不同的输入图像大小，比较 ByteTrack 在测试时的速度和精度。所有实验在训练时都是用多尺度训练的方法，在测试时使用不同的输入图像大小。表 3-8 展示了结果，输入图像的大小从  $512 \times 928$  提升至  $800 \times 1440$ ，其中第一位数字代表输入图像的高，第二位数字代表宽。运行时间从 17.9 毫秒提升至 30.0 毫秒，关联时间大约都是 4 毫秒。ByteTrack 能够以 45.7 FPS 的运行速度达到 75.0 的 MOTA，以 29.6 FPS 的运行速度达到 76.6 的 MOTA，在实际应用中有非常大的优势。

表 3-8 不同输入图像大小下 ByteTrack 的速度和精度比较

输入图像大小	MOTA $\uparrow$	IDF1 $\uparrow$	IDs $\downarrow$	耗时 (毫秒)
$512 \times 928$	75.0	77.6	200	<b>17.9+4.0</b>
$608 \times 1088$	75.6	76.4	212	21.8+4.0
$736 \times 1280$	76.2	77.4	188	26.2+4.2
$800 \times 1440$	<b>76.6</b>	<b>79.3</b>	<b>159</b>	29.6+4.2

**两次关联中相似度的选择** 在低分框关联时，使用交并比 IoU 作为相似度非常重要，因为低分检测框中通常包含很多严重遮挡或者运动模糊的物体，这些物体的重识别特征往往非常不可靠。表 3-9 中展示了在两次关联中使用不同类型的相似度对跟踪结果的影响，其中相似度#1 表示高分框关联中使用的相似度，相似度#2 表示低分框关联中使用的相似度。从结果可以看到相似度#1 使用交并比和重识别都可以得到不错的跟踪结果，相似度#2 只能使用交并比，使用重识别时三个指标都有明显下降，这表明低分框中物体的重识别特征不可靠。

表 3-9 ByteTrack 两次关联中不同相似度对跟踪结果的影响

相似度#1	相似度#2	MOTA $\uparrow$	IDF1 $\uparrow$	IDs $\downarrow$
交并比	重识别	75.8	77.5	231
交并比	交并比	<b>76.6</b>	79.3	<b>159</b>
重识别	重识别	75.2	78.7	276
重识别	交并比	76.3	<b>80.5</b>	216

**跟踪轨迹插值** 跟踪轨迹插值的结果如表 3-10 所示，在最大间隔帧数 $\sigma$ 为 20 时

# 华中科技大学硕士学位论文

---

能够将 MOTA 从 76.6 提升到 78.3, IDF1 从 79.3 提升到 80.2。跟踪轨迹插值是一种非常有效地跟踪完全遮挡物体后处理方式。本章中在 MOT17, MOT20 和 HiEve 的测试集结果中加入了跟踪轨迹插值的方法。

表 3-10 不同最大时间间隔下跟踪轨迹插值的性能比较

时间间隔	MOTA↑	IDF1↑	FP↓	FN↓	IDs↓
无	76.6	79.3	<b>3358</b>	9081	159
10 帧	77.4	79.7	3638	8403	150
20 帧	<b>78.3</b>	<b>80.2</b>	3941	7606	<b>146</b>
30 帧	<b>78.3</b>	<b>80.2</b>	4237	<b>7337</b>	147

## 3.7 本章小结

针对多目标跟踪之前方法只保留高分检测框做关联带来的漏跟踪和轨迹中断问题，本章提出了一个简单有效的数据关联方法高低分检测框层次关联，通过先匹配高分框，再匹配低分框的方法，能够根据位置相似度从低分框中找出遮挡、模糊的物体，大幅减少漏跟踪和轨迹中断的情况，而且能够被简单地应用到现有的多目标跟踪方法中并取得稳定的性能提升。该方法首次在全球最权威的多目标跟踪挑战赛 MOT17 数据集的测试集上取得 80.3 MOTA, 77.3 IDF1 和 63.1 HOTA，并且以 30 FPS 的速度运行。因为精准的检测结果和加入低分检测框关联，该方法能够跟踪许多被严重遮挡的目标，保持跟踪轨迹的连贯性。同时也为如何在多目标跟踪任务中最大程度利用检测结果提供了一个最简单初始的思路，为多目标跟踪的实际落地提供了可能。

## 4 基于体素特征的多视角三维人体姿态估计和跟踪

### 4.1 研究动机

第二章和第三章从多目标跟踪的网络模型结构和数据关联两方面出发进行一些研究，解决了一些目标拥挤、部分遮挡、模糊的问题。然而，对于完全遮挡的目标，修改模型结构和数据关联策略很难对其进行跟踪。在三维空间中，如果目标在某个视角中遮挡较严重，那么在其他视角中很可能得到该目标没有被遮挡时的图像信息。所以，本章提出使用多视角的信息来获取目标的三维空间位置，再对其进行跟踪，解决单视角中很难解决的遮挡问题。本章的研究内容对应的场景为多视角视频场景，每个场景有多个相机，包含室内、室外场景，只对人进行三维姿态估计和跟踪，相较于第二、三章中人体的分辨率更大，人群密度更低。

### 4.2 主要思路

#### 4.2.1 任务定义

本章主要研究的是如何利用多视角信息来解决单视角中的遮挡问题，任务名称是多视角三维人体姿态估计和跟踪，任务的输入是场景中多个视角相机的视频，输出是该场景下每个人的三维关节点坐标和身份。该任务如图 4-1 所示，输入是上侧和右侧的五个不同的视角下同一时刻的原始视频帧（RGB 图像）以及每个视角的相机参数，输出为左下角所示的三维空间中每个人的关节点坐标和身份，不同的颜色和数字代表不同的身份，点代表每帧中每个人骨盆关节点的运动轨迹，红色锥形为相机的位置。其中上侧和右侧中二维图像中的人体关节点坐标为输出的三维坐标通过相机参数投影回二维图像得到。可以看到右下角第五个视角中，身份为绿色“2”号的人遮挡非常严重，利用多视角信息能够将他的关节点坐标准确得到，并进行稳定跟踪。

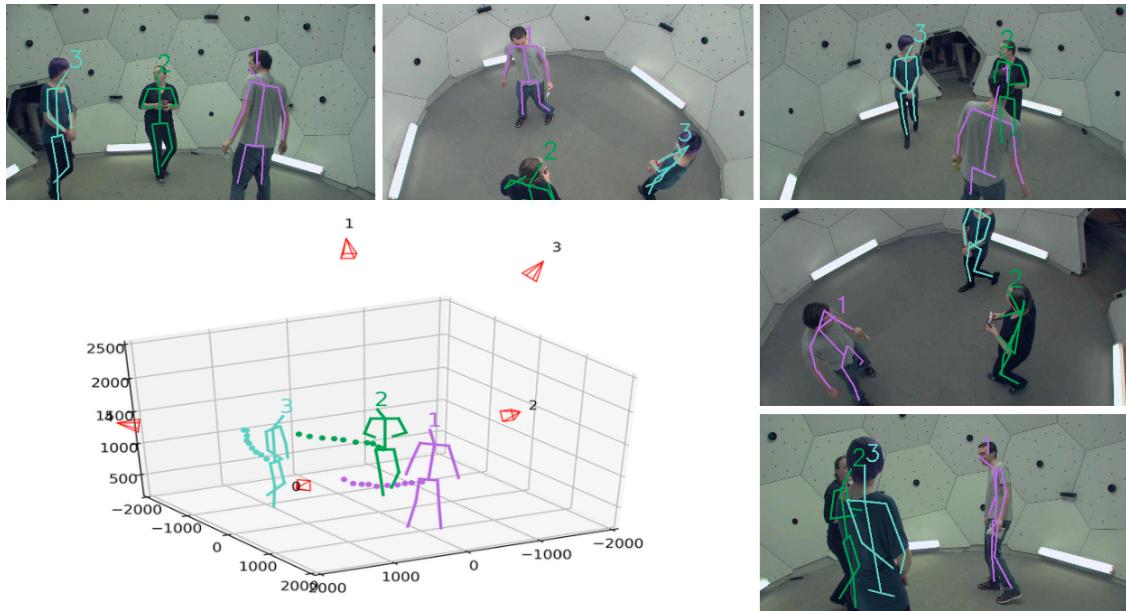


图 4-1 多视角三维人体姿态估计和跟踪示意图

### 4.2.2 整体框架

解决该任务的主要思路是将不同视角下的图像信息根据相机参数进行反投影，融合到三维空间中形成体素特征，之后通过三维网络学习得到每个人的三维姿态，并进行跟踪。该任务包括多个子模块，例如人体姿态估计、行人重识别、多目标跟踪等。本章提出一个基于体素（voxel）特征的多视角三维人体姿态估计和跟踪的简单框架 VoxelTrack，先估计出每个人的三维姿态，再使用三维坐标和重识别特征进行遮挡感知的跟踪。整个框架如图 4-2 所示，其中，三维姿态估计分为二维网络和三维网络，二维网络负责提取每个视角下的二维关节点热力图和每个人的重识别特征，三维网络负责得到每个人的三维关节点坐标。跟踪时需要用到三维关节点坐标，并对每个视角中的目标进行遮挡判断，再根据是否遮挡使用该视角下的二维重识别特征，最终同时利用关节点坐标和外观相似度对行人进行跟踪。该章中网络模型部分用到了第二章中基于点的重识别特征提取方法，数据关联部分用到了第二章中外观相似度计算和匹配的方法。

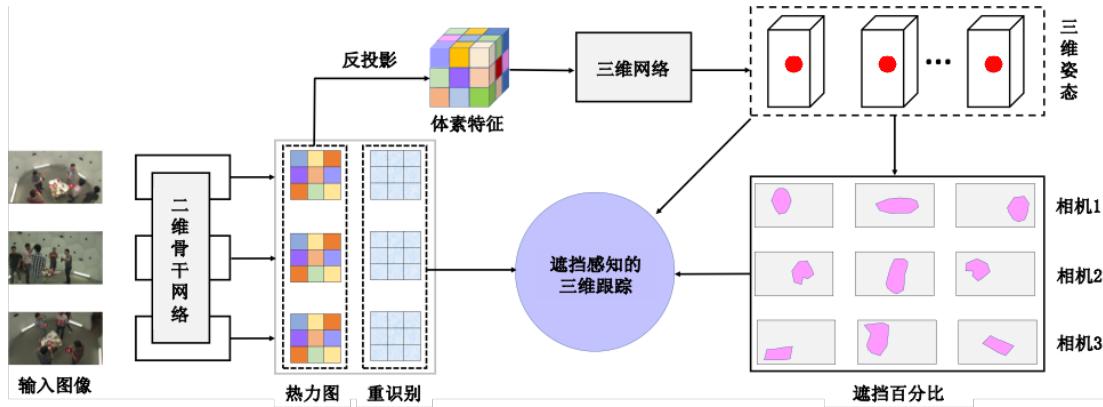


图 4-2 多视角三维人体姿态估计和跟踪方法 VoxelTrack 整体框架图

### 4.3 三维姿态估计

这部分主要介绍 VoxelTrack 的第一部分，估计场景中人的三维姿态（也叫关节点坐标），这部分包含二维姿态热力图估计，重识别特征提取和三维姿态提取。

#### 4.3.1 二维骨干网络

二维骨干网络使用 DLA-34<sup>[24]</sup>来提取每个视角下每帧图像的中间特征。DLA 网络<sup>[68]</sup>最先是被提出做图像分类的，本部分使用的是为了密集预测（Dense Prediction）任务而加入深层特征融合来增加输出特征图分辨率的改进版本。骨干网络的输入是单张图像  $I_v \in \mathbb{R}^{3 \times H \times W}$ ，来自第  $v$  个视角，输出是特征图  $F_v \in \mathbb{R}^{C \times \frac{H}{4} \times \frac{W}{4}}$ ，之后会被送进两个不同的分支，分别负责提取二维姿态热力图和重识别特征，如图 4-3 所示。

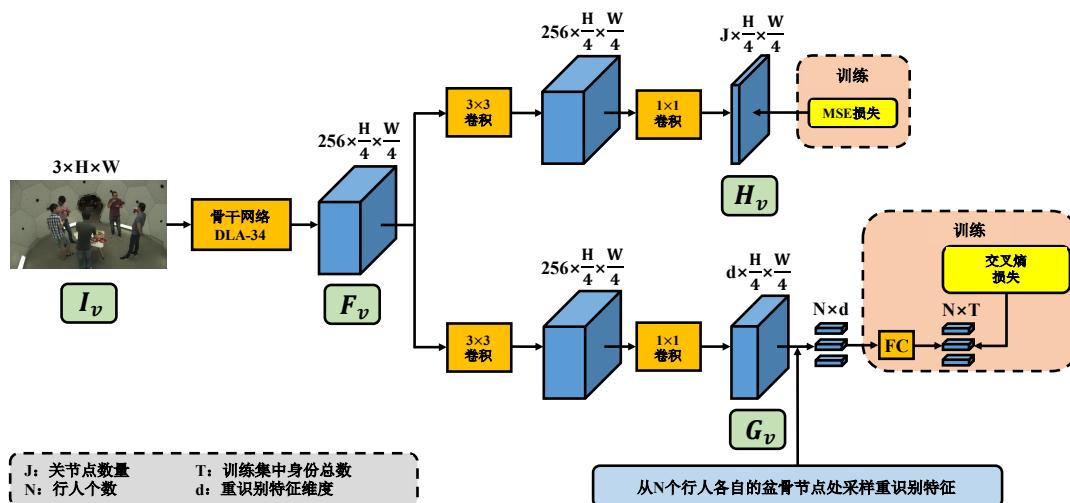


图 4-3 VoxelTrack 二维网络结构图

### 4.3.2 二维姿态热力图估计

在骨干网络特征 $\mathbf{F}_v$ 之后，一个简单的神经网络便可以提取二维姿态热力图 $\mathbf{H}_v \in \mathbb{R}^{J \times \frac{H}{4} \times \frac{W}{4}}$ ，其中 $J$ 是人体关节点的个数，比如鼻子、手腕、手肘、膝盖等。该网络包含两个卷积层，如图 4-3 所示。一个热力图表示逐像素的关节点可能性，这是在人体姿态估计中一种常用的表示。二维热力图的训练是通过最小化下式完成的：

$$L_{2D} = \|\mathbf{H}_v^* - \mathbf{H}_v\|_2 \quad (4-1)$$

其中 $\mathbf{H}_v^* \in \mathbb{R}^{J \times \frac{H}{4} \times \frac{W}{4}}$ 是二维姿态热力图的标签。和之前的三维人体姿态估计方法不同，本方法不在二维层面做比较难的决策，即不直接回归二维关节点坐标，因为从二维姿态热力图到二维关节点会引入误差，所以在 VoxelTrack 中直接将二维姿态热力图输入到三维网络中。

### 4.3.3 二维重识别特征提取

二维重识别特征提取同样使用一个简单的神经网络完成，由骨干网络特征 $\mathbf{F}_v$ 经过两个卷积层提取到重识别特征 $\mathbf{G}_v \in \mathbb{R}^{d \times \frac{H}{4} \times \frac{W}{4}}$ ，其中 $d$ 是重识别特征的维度，如图 4-3 所示。提取重识别特征的方法和第二章中 FairMOT 类似，对图像中的每个人，从特征图上这个人的骨盆节点处采样出一个维度为 $d$ 的重识别特征。在网络训练时，使用一个全连接（Fully Connected, FC）层和一个柔性最大值（softmax）操作将采样得到的重识别特征映射为一个 one-hot 向量 $\mathbf{P} = \{p(t), t \in [1, T]\}$ 代表人的身份。用 $L^i(p)$ 代表人的身份的标签的 one-hot 向量，使用一个分类任务的损失函数来训练重识别网络：

$$L_{ID} = -\sum_{i=1}^N \sum_{t=1}^T L^i(t) \log(p(t)) \quad (4-2)$$

其中 $N$ 是图中人的个数， $T$ 是训练集中总的人的身份数。在训练时，使用人的骨盆节点的标签坐标来提取重识别特征。在测试时，将估计出的三维骨盆节点的坐标通过相机参数投影回二维的图像上得到骨盆节点的二维坐标，以此来提取重识别特征，最后以全连接层之前的 $d$ 维向量作为每个人的重识别特征。

### 4.3.4 三维姿态热力图估计

首先要将整个三维场景划分成 $X \times Y \times Z$ 个体素格（voxels） $\{\mathbf{D}^{x,y,z}\}$ ，体素中每

# 华中科技大学硕士学位论文

---

一个格子代表三维关节点可能出现的一个位置。为了减少量化误差，通常将每个体素格的大小设置得尽可能小（本章中使用 62.5 毫米）。接着将每个体素格通过相机参数投影回每个二维视角的图像上，找到对应的二维姿态热力图中的值，将每个视角的值做平均，作为每个体素格的一个特征向量。用  $\mathbf{H}_v \in \mathbb{R}^{J \times \frac{H}{4} \times \frac{W}{4}}$  表示视角  $v$  的二维姿态热力图，其中  $J$  是人体关节点的总数。对每个体素格  $\mathbf{D}^{x,y,z}$ ，将它在视角  $v$  下的投影位置记为  $\mathbf{P}_v^{x,y,z}$ ，二维热力图在  $\mathbf{P}_v^{x,y,z}$  位置处的值记作  $\mathbf{H}_v^{x,y,z} \in \mathbb{R}^J$ 。每个体素格的特征向量通过下式计算：

$$\mathbf{V}^{x,y,z} = \frac{1}{V} \sum_{v=1}^V \mathbf{H}_v^{x,y,z} \quad (4-3)$$

其中  $V$  是相机视角总数。从公式 (4-3) 可以看出  $\mathbf{V}^{x,y,z}$  实际上表示三维关节点出现在体素格  $\mathbf{D}^{x,y,z}$  处的概率。值得注意的是特征量 (feature volume)  $\mathbf{V}$  通常带有很多的噪声，因为一些不含有关节点的体素格因为缺少深度信息可能也含有非 0 值。

本节提出关节点估计网络 (Joint Estimation Network, JEN) 从融合完每个视角的二维热力图的体素特征量  $\mathbf{V}$  中估计三维关节点热力图  $\mathbf{U} \in \mathbb{R}^{J,X,Y,Z}$ ，网络结构如图 4-4 所示。每个置信度得分  $\mathbf{U}^{j,x,y,z}$  表示有类型为  $j$  的关节点在体素格  $\mathbf{D}^{x,y,z}$  处的概率。所有关节点在所有体素格处的概率组成三维关节点热力图  $\mathbf{U} \in \mathbb{R}^{J,X,Y,Z}$ 。在训练时，三维关节点热力图中每个体素格的标签  $\mathbf{U}_*^{j,x,y,z}$  通过其和关节点位置的标签间的距离计算得到。具体来说，对每个关节点位置标签和体素格对，会根据它们之间的距离计算出一个高斯 (Gaussian) 得分，这个得分随着距离的下降呈指数下降。值得注意的是如果在场景中有几个人，一个体素格可能会有多个得分，对于这种情况只需要保留得分最高的。训练 JEN 网络的损失函数为：

$$L_{JEN} = \sum_{j=1}^J \sum_{x=1}^X \sum_{y=1}^Y \sum_{z=1}^Z \left\| \mathbf{U}_*^{j,x,y,z} - \mathbf{U}^{j,x,y,z} \right\|_2 \quad (4-4)$$

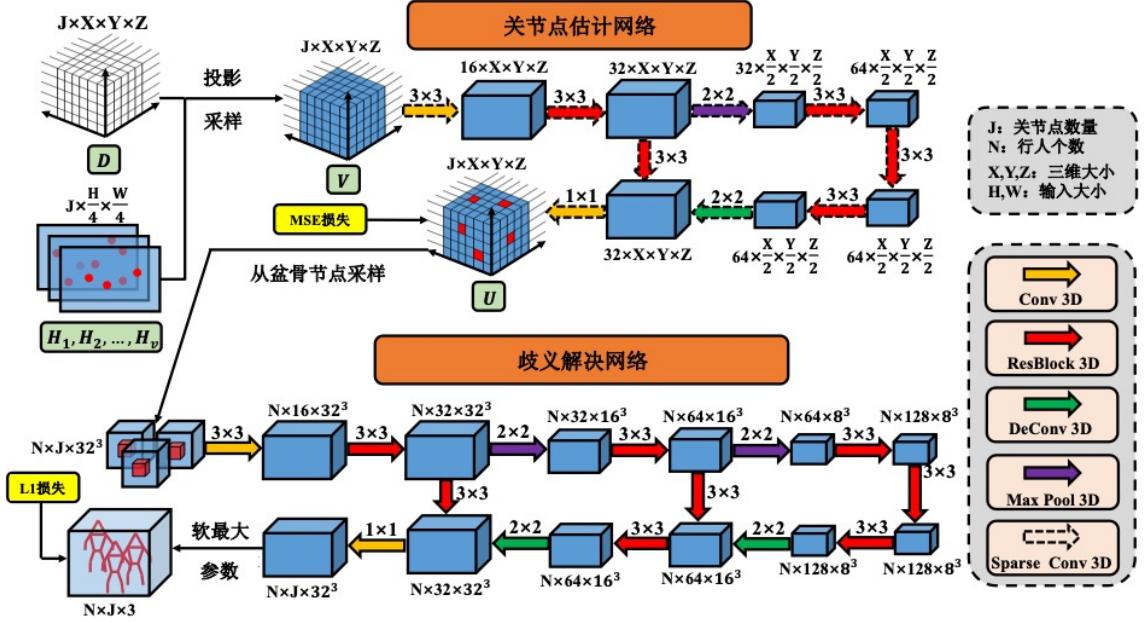


图 4-4 VoxelTrack 三维网络结构图

受到 voxel-to-voxel 网络<sup>[102]</sup>的启发，本节中使用三维卷积（3D convolutions）作为 JEN 的基本单元。因为输入的特征量  $V$  是非常稀疏的，而且具有比较明显的语义含义，本节提出一个比 voxel-to-voxel 网络更加简单的网络，如图 4-4 所示，其中黄色箭头代表普通的三维卷积（Conv 3D），红色箭头代表三维残差块（ResBlock 3D），绿色箭头代表三维反卷积（DeConv 3D），紫色箭头代表三维最大池化（Max Pool 3D），虚线箭头代表使用稀疏三维卷积（Sparse Conv 3D）。在一些比较大的场景中，比如足球场，会不可避免地带来非常高维的特征量，导致计算速度大幅降低。VoxelTrack 通过使用稀疏三维卷积<sup>[103]</sup>来解决这个问题，因为在特征量中只有少量非 0 的特征，尤其是在体素大小比较大的时候，该操作大幅提高推理速度。

#### 4.3.5 三维关节点分组

在得到三维姿态热力图之后，剩下的任务就是要将关节点分给对应的人，即给关节点分组。本章提出歧义解决网络（Ambiguity Resolution Network, ARN）来解决这个问题，如图 4-4 底部所示。首先，从三维姿态热力图中找到峰值，代表多个骨盆节点的位置，每个位置可能存在有一个人。接着使用基于热力图得分的非极大值抑制（Non-Maximum Suppression, NMS）筛选出局部峰值，即完成对人体的定位。然

后对每个人，以该人的骨盆节点为中心，从三维姿态热力图上抠出大小为固定  $X' \times Y' \times Z'$  的立方体，该立方体足够大，能够包围不同姿态的整个人体。在实验中，设置  $X' = Y' = Z' = 32$ ，相当于真实世界中的 2000 毫米。再之后将每个人的立方体特征输入到 ARN 中，对每个人的每个关节点  $k$  估计出一个三维姿态热力图  $\mathbf{A}_k \in \mathbb{R}^{X',Y',Z'}$ 。在这个 2000 毫米的立方体特征中，其他人的节点的响应会被 ARN 压缩至 0。最后，使用一个 soft argmax 的操作<sup>[104]</sup>生成关节点  $k$  的三维位置  $\mathbf{J}_k$ ，该操作通过计算  $\mathbf{A}_k$  的质心得到：

$$\mathbf{J}_k = \sum_{x=1}^{X'} \sum_{y=1}^{Y'} \sum_{z=1}^{Z'} (x, y, z) \cdot \mathbf{A}_k(x, y, z) \quad (4-4)$$

这里没有使用热力图  $\mathbf{A}_k$  中响应最大值的位置作为  $\mathbf{J}_k$ ，因为本节中使用的最小体素格大小 62.5 毫米的量化误差仍然很大，使用公式 (4-4) 中的计算期望的方法有效地减少误差。训练 ARN 时是通过最小化估计的三维姿态坐标和标签三维姿态坐标之间的 L1 距离完成的，如下式所示：

$$L_{ARN} = \sum_{k=1}^J \|\mathbf{J}_k^* - \mathbf{J}_k^k\|_1 \quad (4-5)$$

## 4.4 遮挡感知的人体跟踪

本节主要介绍 VoxelTrack 的第二部分，将估计出来的三维人体姿态坐标按照时间顺序关联起来，得到跟踪轨迹。这是一个单独的模块，不需要训练，核心是计算当前帧的三维姿态和过去帧的三维姿态跟踪轨迹之间的相似度矩阵。在得到相似度矩阵之后，关联就变成一个标准的线性二分图匹配的问题。

### 4.4.1 推理遮挡关系

因为二维重识别特征是从图像中提取的，所以会受到遮挡的影响。考虑到基准数据集中大多数遮挡都是人和人之间引起的，本节中假设人的姿态是被环境中的其他人的身体遮挡的，该想法同样能够用来解决人和物体之间的遮挡。

总的思路是假如某个视角中一个人被其他人严重遮挡时，就减少这个视角中该人的重识别特征的贡献，尽可能地使用该人在没有被遮挡的视角下的重识别特征。具体来说，对每个人的三维姿态坐标，使用相机参数投影回每个视角，得到每个视角下该人的深度，图 4-5 中展示了如何计算每个人被遮挡的百分比。具体来说，就是

使用每个关节点的平均深度作为一个人的深度。对于每个相机，一个将人体所有关节点紧密包围并和相机平面平行的二维包围框会被放置，框中间的数值代表该人在该相机下的深度，对于包围框中间的每个像素，能够通过和其他人的深度包围框的比较得到该像素是否被遮挡。具体计算过程是先计算出整张图的最小深度图，深度越小则证明该人被遮挡的可能性越小，接着将每个人的深度图和整图的最小深度图做比较，最后将每个人的包围框中没有被遮挡的像素的百分比作为该人的重识别特征的可靠性得分，假如一个人 100%都没有被遮挡，那么他的重识别特征的得分就为 1.0，相反，如果一个人 100%都被遮挡，那么重识别特征的得分就为 0.0。

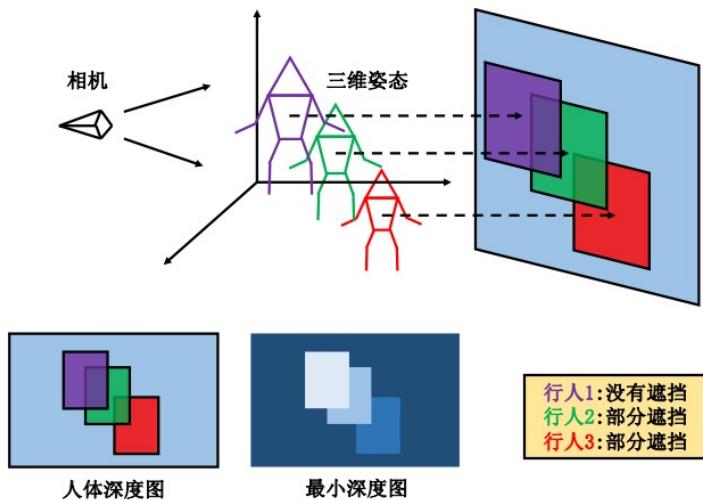


图 4-5 行人的遮挡百分比计算过程图

#### 4.4.2 计算相似度

两个三维人体姿态之间的相似度可以通过它们的外观特征和空间坐标计算得到。第一步是将每个人的三维骨盆坐标投影回每个视角的二维，根据二维骨盆坐标从每个视角的重识别特征图上提取每个人的重识别特征。将第  $i_{th}$  个姿态和第  $j_{th}$  个姿态在所有相机视角下的重识别特征记为  $\{\mathbf{G}_i^1, \mathbf{G}_i^2, \dots, \mathbf{G}_i^V\}$  和  $\{\mathbf{G}_j^1, \mathbf{G}_j^2, \dots, \mathbf{G}_j^V\}$ 。最终的重识别特征由每个相机视角下的重识别特征的加权求和得到，如下式所示：

$$\mathbf{G}_i = \sum_{v=1}^V \omega_i^v \mathbf{G}_i^v \quad (4-6)$$

其中， $\omega_i^v$  代表每个重识别特征的可靠程度的得分，就是 4.3.1 节中计算的未被遮挡的百分比。具体来说，如果一个人的姿态有 70% 都没遮挡，那么  $\omega_i^v$  将被设置成 0，因

# 华中科技大学硕士学位论文

---

为严重遮挡的人的重识别特征往往不可靠，所以完全不会被采用。最终的外观相似度通过融合视角过后的重识别特征之间的余弦距离计算得到。空间相似度是通过两个姿态的三维坐标之间的欧式距离计算得到。对于每个三维姿态跟踪轨迹，会将它和所有当前帧新检测到的三维姿态之间的欧式距离做归一化，使其和余弦外观距离都处在 0-1 之间，最后总的相似度为外观相似度和空间相似度的平均值。

## 4.4.3 整体跟踪框架

整个跟踪框架采取在线的跟踪模式，即只使用当前帧和过去帧的信息。在第一帧时，将所有估计出的三维姿态都新建跟踪轨迹。在接下来的帧中，先按照 4.4.2 节中的方法计算当前帧新估计（检测）出的三维姿态和跟踪轨迹之间的相似度，在使用匈牙利算法来完成身份的分配，如果两者之间的空间距离大于一定的阈值，那么直接拒绝此次匹配。如果当前帧新估计的三维姿态没有和任何跟踪轨迹匹配上，就新建一个跟踪轨迹。没有和新估计的姿态匹配上的跟踪轨迹，会被标记为丢失并保留 30 帧，以防在之后的帧中再次出现，如果丢失的帧数大于 30 帧，则会被从跟踪轨迹中永久删除。跟踪轨迹的重识别特征会使用 2.5 节中的移动平均的方式来更新以得到更加平滑的重识别特征。最终每帧都会输出当前帧的三维人体姿态坐标和对应的身份，当一个视频结束时，完成跟踪。

## 4.5 实验结果

### 4.5.1 评价指标

**三维姿态估计指标** 本章使用 PCP3D 指标<sup>[54]</sup>来评估三维姿态估计的准确性，具体来说，对每个三维姿态的标签，会找到离它最近的估计的三维姿态，并计算估计正确的部分的百分比。可以看出这个指标不会惩罚假阳性（False Positive, FP）的姿态，为了解决这个限制，本节将目标检测中的平均精度指标（Average Precision, AP<sub>K</sub>）<sup>[73]</sup>扩展到评估多人三维姿态估计的质量上。具体计算细节如下：如果所有节点的平均位置误差（Mean Per Joint Position Error, MPJPE）小于  $K$  毫米，便认为这个姿态是一个正确的估计。AP<sub>K</sub>计算召回率的值的平均精度，在 0 到 1 之间。

# 华中科技大学硕士学位论文

---

**三维姿态跟踪指标** 该指标是由标准的二维包围框多目标跟踪指标 CLEAR<sup>[79]</sup>和二维姿态跟踪指标<sup>[105]</sup>修改得到。对于每个三维人体关节点，都会单独地计算一个多个目标跟踪精度（Multi-Object Tracking Accuracy, MOTA）得分，类似二维姿态跟踪。预测的关节点和真实的关节点之间的距离如果小于半个头的距离（150 毫米）的话，就认为预测正确。MOTA 得分同时考虑姿态估计和姿态跟踪的准确性。对每个关节点也需要计算身份跳变的次数（identity switches, IDs）。还有一个指标 IDF1<sup>[80]</sup>得分也会用来评估轨迹的连贯性，其中 MOTA 和 IDF1 的计算方式在公式(1-1)和(1-2)中已经给出。

## 4.5.2 实验环境和实现

本章方法使用的主要软件环境为 Python 3.7 和 Pytorch 1.4.0，其他相关的 pip 安装包包括 json\_tricks、opencv-python、scipy、prettytable、tqdm、matplotlib、motmetrics、easydict、PyYAML、lap 等。硬件环境为一块 Nvidia 2080Ti 显卡，测试过程在单块显卡上五个视角的视频同时输入，整个系统的速度可以达到 15 FPS。

训练图像和测试图像在输入网络时大小会被重置为  $800 \times 608$ ，二维网络输出的二维热力图和重识别特征图的大小都是  $200 \times 152$ ，通道数根据各自所需的维度决定。二维骨干网络使用在 ImageNet<sup>[119]</sup>上经过分类任务预训练的 DLA-34。人体关节点的数量  $J$  为 15，和 COCO<sup>[73]</sup>数据集相一致。重识别特征的维度设置为 64，和第二章中的 FairMOT 保持一致。

数据集选用第 1.2.8 节中介绍的三维人体姿态估计和跟踪数据集 CMU Panoptic, Campus 和 Shelf 数据集，对于这三个数据集，场景大小都被设置成  $10 \text{ 米} \times 10 \text{ 米} \times 4 \text{ 米}$ ，对应  $160 \times 160 \times 64$  个体素格，每个体素格的大小大约是  $62.5 \text{ 毫米} \times 62.5 \text{ 毫米} \times 62.5 \text{ 毫米}$ 。值得注意的是因为最终的三维关节点位置是通过三维热力图的期望计算得到的，所以实际的误差会比 62.5 毫米小很多。在估计三维热力图时，本章使用的是三维稀疏卷积，只对三维特征量中非 0 的地方进行卷积计算。为保证特征量中的稀疏性，将初始值小于 0.3 位置处的值都置 0。对 ARN，以每个人骨盆节点为中心，从三维热力图中抠出  $2000 \text{ 毫米} \times 2000 \text{ 毫米} \times 2000 \text{ 毫米}$  的特征，相当于  $32 \times 32 \times 32$  大小的体素格。

# 华中科技大学硕士学位论文

---

在 Panoptic 数据集上训练 VoxelTrack 时，需要经过三个阶段，所有阶段都使用 Adam 优化器<sup>[81]</sup>。在第一个阶段中，只训练二维网络，包括二维姿态热力图网络和行人重识别特征提取网络。总共训练 20 轮迭代，初始学习率是  $10^{-4}$ ，在第 15 轮迭代之后降为  $10^{-5}$ 。在第二个阶段中，将二维网络的参数固定，只训练三维节点估计网络（JEN），总共训练 10 轮迭代，学习率为  $10^{-4}$ 。最后一个阶段只训练歧义解决网络（ARN），其他网络参数都固定，总共训练 10 轮迭代，学习率也保持  $10^{-4}$ ，整个三个阶段训练过程在单块 2080Ti 显卡上花费 30 个小时。如果有足够多的图像-三维姿态标注对的话，这三部分也可以一起训练，最终选择分三阶段分别训练的原因如下：(1) 当将模型应用到一个新的场景下时，可能没办法得到大量的图像-三维姿态对。所以，模型的二维部分能够在公开数据集上训练，三维部分通过根据新环境的相机参数生成二维姿态热力图和三维姿态对来训练。(2) 分三阶段训练能够使用更大的批大小 (batch size)，对训练的稳定性有帮助。

在 Campus 和 Shelf 两个数据集上训练时，为避免在这样的小数据集上过拟合，不使用数据集提供的图像和三维姿态，只使用相机参数来训练。二维骨干网络使用在 COCO 关键点数据集<sup>[73]</sup>上预训练过的 HigherHRNet<sup>[106]</sup>，有一点值得注意的是 COCO 数据集不提供身份的标注，所以不能直接在上面直接训练重识别分支。本章提出一个弱监督的训练方法来在 COCO 数据集上训练重识别特征，具体做法是对每个二维姿态赋予一个唯一的身份，将训练集中的每个物体实例视为一个单独的类别。这样的做法需要对全图做多种数据增强，使得每个实例能够具有不同的外观，数据增强包括随机翻转，旋转，缩放和仿射变换等。

在 Campus 和 Shelf 两个数据集上训练三维部分的时候，会使用相机参数生成许多合成的二维热力图。首先将一定数量的三维姿态（比如从 Panoptic 数据集中采样）随机地摆放到当前的场景上，使用相机参数将这些三维姿态投影到各个视角上，得到合成的二维热力图。这种训练方法具有非常高的实用价值，可以非常方便地迁移到底新的场景下，比如一个已知相机参数的零售商店中，如图 4-6 所示，该零售商店中有 12 个摄像头，中间有 4 个人在挑选商品，其中包含许多人和货架之间的遮挡，VoxelTrack 可以充分利用这 12 个摄像头的信息，解决单个摄像头中的遮挡问题，得

到稳定的跟踪结果，图 4-6 的上半部分是三维人体姿态跟踪的可视化结果，下半部分是将三维结果投影回每个视角下的二维姿态跟踪结果。

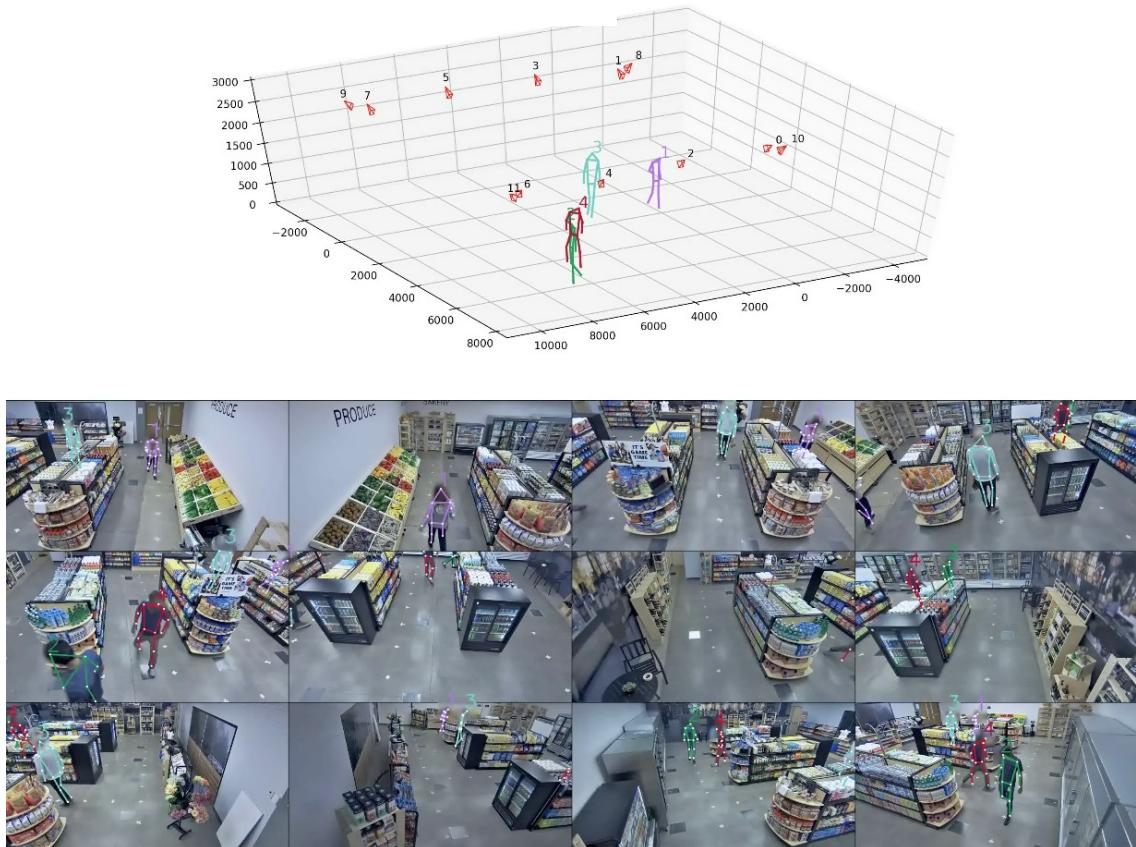


图 4-6 VoxelTrack 在零售商店场景中的应用

### 4.5.3 和当前最先进方法的比较

**三维人体姿态估计** 表 4-1 中展示了当前最先进（SOTA）方法在 Panoptic 数据集上的三维姿态估计结果。本章提出的 VoxelTrack 在所有方法中取得了具有竞争力的结果，和 VoxelPose<sup>[107]</sup>相比，本方法将平均节点误差（MPJPE）降低 0.7 毫米，这是由于使用了一个性能更强的二维骨干网络 HigherHRNet，并且加入额外的所有关节点的三维姿态热力图监督。与 VoxelPose 相比，本方法将运行时间从 347 毫秒缩短到 191 毫秒，这是一个非常大的进步，主要耗时缩短来自于加入稀疏卷积，尤其是在场景较大（或体素格划分的较精细时）减少大量的计算量。还有一点是在进行三维特征量的投影操作时，本方法只投影一次，而 VoxelPose 投影两次。比本方法更晚公布

# 华中科技大学硕士学位论文

---

的两个方法 MvP<sup>[108]</sup>和 TesseTrack<sup>[109]</sup>的三维姿态估计结果超过了本方法,这两个方法都是在 VoxelPose 的基础上进行改进的。MvP 使用一个更强的 transformer 模型<sup>[41]</sup>,巧妙地避免计算量昂贵的体积表示,因此降低 MPJPE 和耗时。TesseTrack 使用一个四维的卷积神经网络来得到每个人在时间上增强的表征,和现有方法相比,它大幅降低 MPJPE。但是, TesseTrack 的计算开销非常大,需要两块显存为 32 GB 的 V100 显卡才能放的下 batch size 为 1 的整个网络。相比之下, VoxelTrack 只需要 1 GB 的显存,在计算开销上有巨大的优势。

表 4-1 CMU Panoptic 数据集上 SOTA 方法的性能比较

方法	MPJPE	MOTA	IDF1	ID Switch	耗时
VoxelPose <sup>[107]</sup>	17.68 mm	-	-	-	347 ms
MvP <sup>[108]</sup>	15.80 mm	-	-	-	<b>170 ms</b>
TesseTrack <sup>[109]</sup>	<b>7.30 mm</b>	94.1	-	-	-
<b>VoxelTrack (本文方法)</b>	16.97 mm	<b>98.5</b>	<b>98.7</b>	<b>0</b>	191 ms

表 4-2 展示了 SOTA 方法在 Campus 和 Shelf 两个数据集上的表现,上半部分为 Campus 的结果,下半部分为 Shelf 的结果。从表中可以看出 VoxelTrack 同样取得具有竞争力的结果,比如在 Campus 数据集上将 PCP3D 指标从非常近期的一个工作 MvP 的 96.6% 提升到 96.7%,考虑到精度已经相当高,这是一个比较有力度的提升。在 4.5.1 节中已经讨论过,PCP3D 这个指标不会惩罚假阳性,但是由于 Campus 和 Shelf 这两个数据集中的三维人体姿态标签不完整(没有将所有的人都标出),所以 AP 指标也没有意义。图 4-7 展示了一些 Shelf 数据集上的可视化结果,其中包含第 0 帧,第 50 帧和第 100 帧的结果,每帧中包含五个视角的图像,不同颜色的人体姿态代表不同的身份,最底部是整个场景的三维人体姿态跟踪结果。可以看出本方法只要在至少两个视角可见下通常就能够得到一些精确的估计结果。

# 华中科技大学硕士学位论文

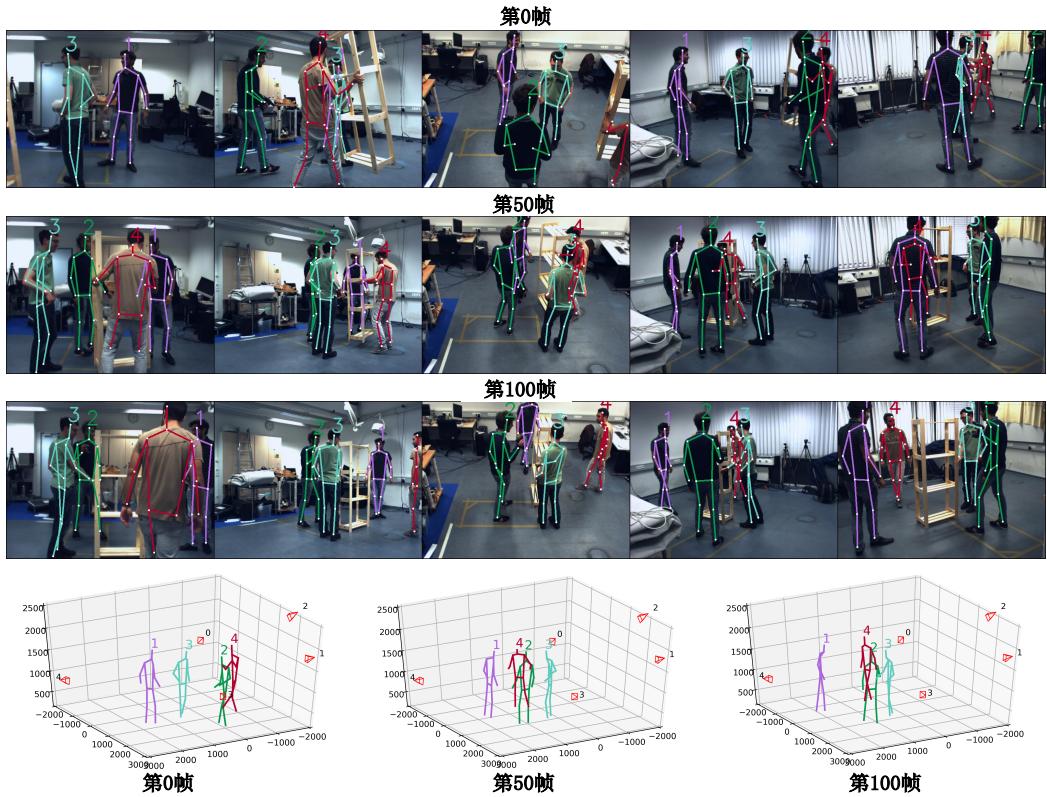


图 4-6 VoxelTrack 在 Shelf 数据集上的可视化结果图

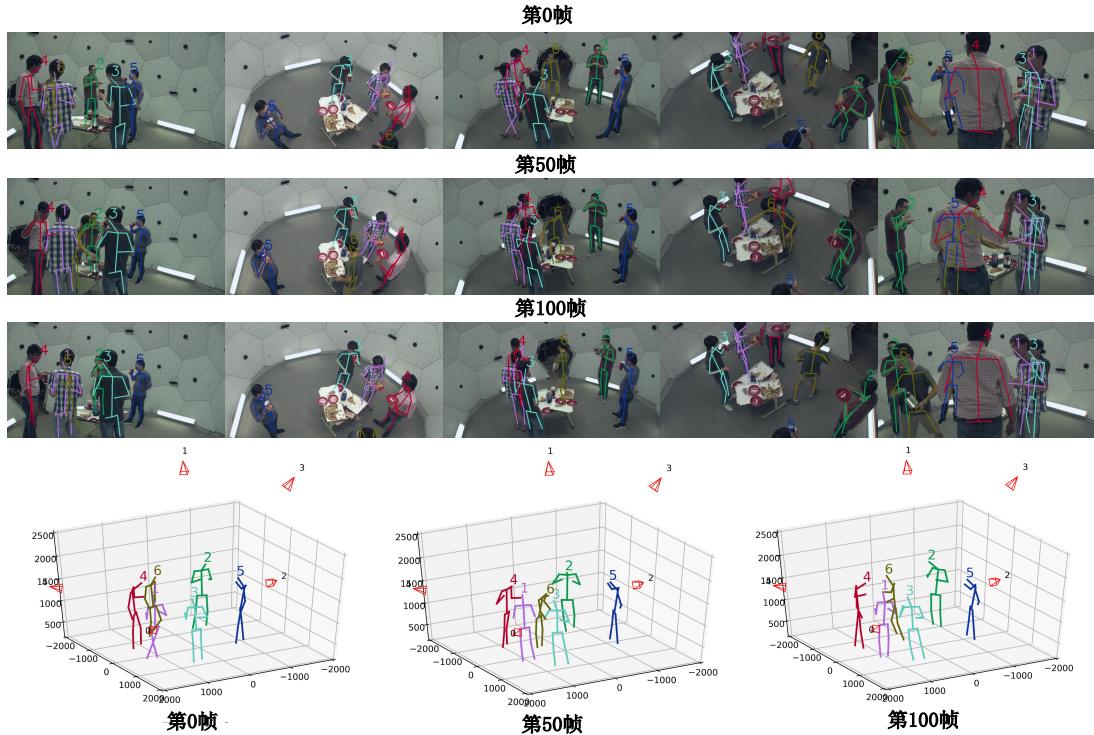


图 4-7 VoxelTrack 在 Panoptic 数据集上的可视化结果图

# 华中科技大学硕士学位论文

---

表 4-2 Campus 和 Shelf 数据集上 SOTA 方法的性能比较

Campus	演员 1	演员 2	演员 3	PCP3D	MOTA	ID	IDF1
Switch							
Belagiannis <i>et al.</i> <sup>[56]</sup>	82.0	72.4	73.7	75.8	-	-	-
Belagiannis <i>et al.</i> <sup>[57]</sup>	83.0	73.0	78.0	78.0	-	-	-
Belagiannis <i>et al.</i> <sup>[62]</sup>	93.5	75.7	84.4	84.5	-	-	-
Ershadi <i>et al.</i> <sup>[110]</sup>	94.2	92.9	84.6	90.6	-	-	-
Dong <i>et al.</i> <sup>[54]</sup>	97.6	93.3	98.0	96.3	-	-	-
Chen <i>et al.</i> <sup>[55]</sup>	97.1	94.1	98.6	96.6	-	-	-
VoxelPose <sup>[107]</sup>	97.6	93.8	98.8	96.7	-	-	-
MvP <sup>[108]</sup>	<b>98.2</b>	94.1	97.4	96.6	-	-	-
TesseTrack <sup>[109]</sup>	97.9	<b>95.2</b>	<b>99.1</b>	<b>97.4</b>	-	-	-
<b>VoxelTrack(本文方法)</b>	98.1	93.7	98.3	96.7	<b>89.3</b>	<b>0</b>	<b>94.6</b>
Shelf	演员 1	演员 2	演员 3	PCP3D	MOTA	ID	IDF1
Switch							
Belagiannis <i>et al.</i> <sup>[56]</sup>	66.1	65.0	83.2	71.4	-	-	-
Belagiannis <i>et al.</i> <sup>[57]</sup>	75.0	67.0	86.0	76.0	-	-	-
Belagiannis <i>et al.</i> <sup>[62]</sup>	75.3	69.7	87.6	77.5	-	-	-
Ershadi <i>et al.</i> <sup>[110]</sup>	93.3	75.9	94.8	88.0	-	-	-
Dong <i>et al.</i> <sup>[54]</sup>	98.8	94.1	97.8	96.9	-	-	-
Chen <i>et al.</i> <sup>[55]</sup>	<b>99.6</b>	93.2	97.5	96.8	-	-	-
VoxelPose <sup>[107]</sup>	99.3	94.1	97.6	97.0	-	-	-
MvP <sup>[108]</sup>	99.3	95.1	97.8	97.4	-	-	-
TesseTrack <sup>[109]</sup>	99.1	<b>96.3</b>	<b>98.3</b>	<b>98.2</b>	-	-	-
<b>VoxelTrack(本文方法)</b>	98.6	94.9	97.7	97.1	<b>94.4</b>	<b>0</b>	<b>97.2</b>

**三维人体姿态跟踪** 表 4-1 和表 4-2 中也分别展示了 VoxelTrack 在 Panoptic 和 Campus 和 Shelf 数据集上的三维人体姿态跟踪的结果。MOTA 指标是通过 FP, FN 和 ID Switch 三个指标共同计算得到的，整体考虑了行人检测，姿态估计和姿态跟踪的性能。IDF1 和 ID Switch 这两个指标能够更好地反映跟踪的性能。VoxelTrack 取得 0 的 ID Switch 和非常高的 IDF1 得分(Panoptic 上 98.7, Campus 上 94.6, Shelf 上 97.2)，这些数据集的单个视角的图像中的目标都具有非常严重的遮挡，结果证明本章提出

的多视角三维跟踪方法能够得到非常精准的跟踪结果。VoxelTrack 在 Panoptic 数据集上取得比 Campus 和 Shelf 数据集上更高的 MOTA (98.5 vs 94.4, 98.5 vs 89.3)，这是因为 Panoptic 数据集的三维姿态估计的结果更加地精准，因此提高了 MOTA。一个值得注意的点是在表 4-1 中，VoxelTrack 取得比 TesseTrack 更高的 MOTA，并且是在三维姿态估计结果 MPJPE 比 TesseTrack 差的情况下，这可能是由以下两个原因带来的：一是 MPJPE 指标不惩罚假阴性 (False Negative, FN)，而 MOTA 指标对三维姿态估计的限制很松，例如只要差异在 150 毫米内就算跟踪正确。二是 VoxelTrack 借助三维关节点坐标和从图像中学习到的遮挡感知的重识别特征来做跟踪，比 TesseTrack 中的四维体积特征在跟踪方面更加精准。

**定性分析** 图 4-6 中展示了 Shelf 数据集上 VoxelTrack 得到的一些三维姿态跟踪的可视化结果，可以看出所有单个视角的图像中目标都受到严重遮挡，但是，将多个相机视角的二维姿态热力图融合到一起之后，本方法获得了更加鲁棒的特征，能够非常简单地估计出三维姿态。值得注意的是本方法不需要在不同的视角下使用有噪声的二维姿态和一些复杂的技术做关联，找到相同的人在不同的视角下的姿态，这一点非常重要地提高了本方法的鲁棒性。从图 4-6 中可以看出身份为 2、3 和 4 的三个人正在围着一个书架走动，中间包含非常多的遮挡。总共 4 个人的身份在整个视频上都保持不变，验证了本章提出的多视角三维跟踪方法能够在遮挡严重的场景中取得非常精确的跟踪结果。

图 4-7 展示了 Panoptic 数据集上的可视化结果，同样每个视角的图像都包含许多人和人之间的遮挡。身份为 6 的人运动地非常剧烈，他来到桌子边然后再离开，在大多数的视角中都是被遮挡的，所以需要使用遮挡感知的重识别特征对其进行跟踪，只使用他没有被遮挡的视角下的重识别特征。本方法能够将这 6 个人的身份从视频开始保持到结束，始终有着稳定的跟踪结果。

#### 4.5.4 影响三维姿态估计精度的因素分析

本部分对 VoxelTrack 中影响姿态估计精度的一些因素进行消融实验，所有实验结果都是在 Panoptic 数据集上进行的，结果在表 4-3 和 4-4 中，同时评估三维姿态估计的精度，三维姿态跟踪的精度和运行时间。表 4-3 中展示了三维因素，4-4 中展示

# 华中科技大学硕士学位论文

---

了二维因素。

**体素格大小** 本部分采用三种不同大小的体素格： $160 \times 160 \times 64$ 、 $120 \times 120 \times 48$  和  $80 \times 80 \times 32$ ，整个场景的大小被设置成 $10\text{米} \times 10\text{米} \times 4\text{米}$ ，MPJPE 指标的单位是毫米。从表 4-3 的前三行可以看出，将体素格大小从  $80 \times 80 \times 32$  提高到  $120 \times 120 \times 48$  之后，三维姿态估计的性能得到了极大的提升，比如将 AP<sub>25</sub> 指标从 39.74 提升到 71.00，将 MPJPE 指标从 26.16 毫米降低到 19.83 毫米。当继续把体素格大小从  $120 \times 120 \times 48$  提升到  $160 \times 160 \times 64$  时，提升没有那么明显，这也是合理的，因为当体素格变得更加细粒度的时候精度将更难被提升。通过比较跟踪相关的指标如 MOTA，IDF1 和 IDs 可以看出，体素格大小对跟踪的精度影响不大。整个三维部分的运行时间随着体素格大小的增长而增长，为了达到一个比较好的精度和速度的平衡，在剩下的实验中都是用大小为  $160 \times 160 \times 64$  的体素格。

表 4-3 VoxelTrack 中的三维因素对姿态估计和跟踪性能的影响

体素格	JEN 类型	AP <sub>25</sub>	AP <sub>50</sub>	AP <sub>100</sub>	MPJPE	MOTA	IDF1	IDs	三维耗时
$160 \times 160 \times 64$	稀疏卷积	<b>79.34</b>	96.83	<b>99.58</b>	<b>18.49</b>	<b>98.45</b>	<b>98.67</b>	<b>0</b>	62 毫秒
$120 \times 120 \times 48$	稀疏卷积	71.00	97.04	99.45	19.83	98.27	98.52	<b>0</b>	38 毫秒
$80 \times 80 \times 32$	稀疏卷积	39.74	94.27	99.10	26.16	97.62	95.13	<b>0</b>	28 毫秒
$160 \times 160 \times 64$	卷积	74.09	96.87	99.55	19.05	98.32	98.39	<b>0</b>	146 毫秒
$120 \times 120 \times 48$	卷积	68.89	<b>97.06</b>	99.51	20.28	98.16	98.21	<b>0</b>	64 毫秒
$80 \times 80 \times 32$	卷积	38.66	94.40	99.17	25.93	98.27	98.56	<b>0</b>	<b>25 毫秒</b>

**稀疏卷积** 使用稀疏卷积代替普通卷积的原因是三维特征图中只有一小部分非 0 的元素。比较表 4-3 中稀疏卷积和普通卷积的结果可以看出，在体素格较大时（ $160 \times 160 \times 64$  和  $120 \times 120 \times 48$ ），稀疏卷积的运行时间比普通卷积要快很多，在体素格较小时（ $80 \times 80 \times 32$ ），稀疏卷积的运行时间比普通卷积要稍慢，这是由于稀疏卷积需要找到非 0 元素的位置，这需要花去一部分的时间。所以在 ARN 中，本章不采用稀疏卷积，因为其输入的单人三维热力图空间大小很小（ $32 \times 32 \times 32$ ）。稀疏卷积的三维姿态估计精度比普通卷积略高，这是由于在使用稀疏卷积时将三维特征图中小于 0.3 的值都置 0，三维特征图的稀疏性去除了一些歧义性，从而提高三维姿态估计的

# 华中科技大学硕士学位论文

---

精度。

**相机数量及位置** 本部分评估相机数量和位置对三维姿态估计精度的影响，从数据集中手动选择了 5 个视角交界最大的相机，与随机选择的 5 个相机做比较，并减少相机数量至 4、3 个。从表 4-4 的前三行中可以看出，根据相机位置和视角交界手动选择的 5 个相机比随机选择的 5 个相机具有更低的三维姿态估计误差，证明了位置好、视角好的相机对三维姿态估计帮助大。减少相机数量之后将平稳地提高三维姿态估计的误差，这是因为融合各个视角信息的三维特征图中的信息随着相机视角数的减小而减小。跟踪的精度受相机个数的影响不严重，可以看到身份跳变次数总是 0 次，使用 3 个相机视角已经能够精确地跟踪三维姿态。运行时间随着相机视角的下降而下降，主要是因为二维网络的输入变少带来的运行时间下降。

**二维骨干网络** 本部分评估三种不同的二维骨干网络，包括 DLA-34<sup>[24]</sup>，MobileNet-V2<sup>[111]</sup> 和 Higher-HRNet-W32<sup>[106]</sup>。对 MobileNet-V2，采用 Simple Baseline<sup>[112]</sup> 中的方法，将几个反卷积被加入到骨干网络之后得到分辨率为输入图像四分之一的特征图。结果在表 4-4 中，“D”代表 DLA-34，“M”代表 MobileNet-V2，“H”代表 Higher-HRNet-W32，其中 Higher-HRNet-W32 取得最高的 AP 指标和最低的 MPJPE 指标，MobileNet-V2 取得最快的运行速度，DLA-34 取得一个较好的精度和速度的平衡。

**输入图像大小** 本部分评估三种不同的输入图像大小，包括  $960 \times 512$ ,  $800 \times 448$  和  $640 \times 384$ ，结果在表 4-4 中，二维骨干网络都使用 DLA-34。降低图像大小稳定地提高三维姿态估计的误差，主要原因是更大的输入图像包含了更细节的信息。跟踪的精度几乎不受图像大小的影响。本部分说明得到精准的二维热力图对三维姿态估计的精度非常重要。

# 华中科技大学硕士学位论文

表 4-4 VoxelTrack 中的二维因素对姿态估计和跟踪性能的影响

相机数	骨干	图像大小	AP <sub>25</sub>	AP <sub>50</sub>	AP <sub>100</sub>	MPJPE	MOTA	IDF1	IDs	二维耗时
5	D	960×512	79.34	96.83	99.58	18.49	98.45	98.67	0	86 毫秒
5(随机)	D	960×512	72.34	96.54	99.51	19.23	98.39	98.54	0	86 毫秒
4	D	960×512	66.20	96.34	99.47	20.35	98.37	98.46	0	67 毫秒
3	D	960×512	49.09	92.44	97.62	24.93	95.77	93.08	0	55 毫秒
5	D	800×448	70.66	97.26	<b>99.70</b>	19.99	<b>98.61</b>	<b>98.99</b>	0	65 毫秒
5	D	640×384	55.96	96.78	99.65	21.67	98.37	98.45	0	45 毫秒
5	M	960×512	42.42	94.09	99.33	24.38	97.61	97.82	0	<b>28 毫秒</b>
5	H	960×512	<b>85.88</b>	<b>98.31</b>	99.54	<b>16.97</b>	98.51	98.73	0	129 毫秒

## 4.5.5 影响三维姿态跟踪精度的因素分析

在跟踪过程中，有三个主要组成部分：(1) 三维姿态，(2) 重识别特征，(3) 遮挡百分比。这部分将评估每个组成部分对最终跟踪结果的影响，跟踪结果在表 4-5 中。MOTA 和 IDF1 得分计算的是所有关节点的平均值，IDs 指标是所有关节点的身份跳变次数，是一个人关节点个数的倍数（例如 15）。

**三维姿态** 为验证三维姿态对跟踪的影响，本部分只使用三维姿态之间的归一化欧式距离来做跟踪，结果在表 4-5 的第一行。只使用三维姿态间的欧式距离做跟踪只能取得 93.82 的 IDF1 得分和 90 次 IDs，当姿态估计结果中出现假阳性时，往往伴随着身份跳变。总的来说，三维姿态还是较为可靠的，因为在三维坐标中几乎不存在遮挡。

**重识别特征** 为验证重识别特征对跟踪的影响，本部分只使用重识别特征之间的余弦距离做跟踪。对每个三维的人，将他在每个视角下的重识别特征直接融合，不做遮挡相关的处理。结果在表 4-5 的第二行，IDF1 得分（94.38 vs 93.82）和 IDs 次数（15 vs 90）都比只使用三维姿态要好。只使用重识别特征时，仍然存在一些身份跳变，这是因为一些视角中受到严重遮挡的人提供了一些非常不可靠的重识别特征，导致歧义性。

**遮挡百分比** 为验证遮挡百分比对跟踪的影响，本部分采用 4.3.1 节中基于深度

计算的遮挡百分比来融合各个视角的重识别特征。如果某个人在某个视角下被严重遮挡，那么就不使用这个视角下的重识别特征。结果在表 4-5 的第三行，可以看到使用遮挡百分比融合各个视角下的重识别特征之后能够取得最高的 IDF1 得分 98.67 并且不会产生 IDs。当使用三维姿态和加入遮挡百分比的重识别特征之后，跟踪结果保持不变。本部分说明遮挡感知的多视角融合过后的重识别特征具有非常强的区分能力。

表 4-5 VoxelTrack 中各个组件对跟踪性能的影响

重识别特征	遮挡百分比	三维姿态	MOTA	IDF1	IDs	耗时
	√		98.42	93.82	90	1 毫秒
√			98.44	94.38	15	1 毫秒
√	√		<b>98.45</b>	<b>98.67</b>	<b>0</b>	2 毫秒
√	√	√	<b>98.45</b>	<b>98.67</b>	<b>0</b>	2 毫秒

## 4.5.6 整个系统的运行时间

本方法的整个系统分成二维模块，三维模块和跟踪模块，本部分统计每个部分的耗时。三维模块的耗时如表 4-3 所示，其中三维模块的耗时主要来自于 JEN，ARN 因为输入较小，所以耗时相对少很多。二维模块和跟踪模块的耗时分别在表 4-4 和 4-5 中。从表 4-3 中可以看到稀疏卷积降低了很大部分的耗时，体素格大小对耗时的影响也很大。从表 4-4 中可以看到相机视角的数量，二维骨干网络和输入图像大小都会影响二维部分的耗时。从表 4-5 中可以看出跟踪模块的耗时基本可以忽略(2 毫秒)，这也表明本方法中跟踪模块的简单有效。本方法的轻量版本使用 MobileNet-V2 作为二维骨干网络，整个场景的体素格大小为  $120 \times 120 \times 48$ ，JEN 中使用稀疏卷积，五个视角的视频同时输入，整个系统的速度可以达到 15 FPS，大幅增加了本方法的实用价值。

## 4.6 本章小结

针对单视角多目标跟踪中的目标完全遮挡问题，本章提出了一个新颖的多视角

# 华中科技大学硕士学位论文

---

三维人体姿态估计和跟踪框架，利用多视角信息解决单视角中的遮挡问题。本方法使用一个多分支的网络来同时得到环境中所有人的三维姿态和重识别特征，使用多视角信息来对行人进行遮挡感知的跟踪。本方法的一个优势只在三维层面做比较困难的决策，不在遮挡严重的二维图像中做任何困难决策，避免了在二维层面做具有挑战性的行人关联。具体来说，每个相机视角中有噪声的，不完整的信息会被整合到三维空间，形成一个完整的三维特征量来进行三维姿态估计和跟踪。本章同样提出了一个遮挡感知的数据关联方法，能够利用三维信息推断每个人在每个二维视角中被遮挡的百分比，在三维层面对行人进行跟踪。在基准数据集上的实验表明本方法对遮挡情况具有很好的鲁棒性，可以达到 98.61 的多目标跟踪精度（MOTA）和 0 的身份跳变（IDs）。还有一点很重要的是本方法能够使用合成数据训练三维部分的网络，不需要三维数据的标注，带来很大的实用价值。

本章研究内容相关论文“VoxelTrack: Multi-Person 3D Human Pose Estimation and Tracking in the Wild”于 2022 年被国际计算机视觉顶级期刊 IEEE Transactions on Pattern Analysis and Machine Intelligence 接收，并已经在商超领域中得到应用。

## 5 总结与展望

### 5.1 全文总结

本文主要研究的是复杂场景下的高效率多目标跟踪技术，如何解决复杂场景中的人群密集，遮挡严重，运动模糊等问题，以及如何保证多目标跟踪的精度和速度是本文的研究重点。本文的创新如下：

1) 提出了一种基于目标中心点特征的单阶段多目标跟踪网络，按照每个目标中心点处的位置来提取目标检测和行人重识别所需要的特征，在同一个网络中公平地对待这两个任务，在密集场景中能够借助区分力足够强的基于点的重识别特征有效将多个目标分开，速度和精度都大幅领先当前单阶段和双阶段多目标跟踪方法，为多目标跟踪领域带来了一个非常简单有效的基线方法。

2) 提出了一种高低分检测框层次关联的多目标跟踪方法，与之前的数据关联方法不同，本方法加入包含大量背景的低分检测框，并使用先高分后低分的层级关联方法，从低分检测框中找出遮挡和模糊严重的物体，大幅减少漏跟踪和轨迹中断的情况。该方法具有非常强的泛化能力，适用于当前大部分的多目标跟踪网络，且十分简单有效。

3) 提出了一种基于体素特征的多视角三维人体姿态估计和跟踪的方法，融合多视角信息，在三维层面对物体进行跟踪，显式地解决了单个视角下的遮挡问题，能够对行人进行非常精准的跟踪，在跟踪的过程中目标不会发生身份跳变。

### 5.2 展望

本文提出的复杂场景下的几个高性能多目标跟踪方法已经在智慧城市、安防监控和智能商超领域得到了广泛的应用，如客流统计、车流统计、无人超市中顾客行为分析等。本研究在未来对自动驾驶的落地也有很大的帮助，如结合三维点云信息、十字路口处摄像头多视角信息感知路上行人和车辆的行进轨迹，并对之后时刻的车辆和行人的行为进行预测，再对自身车辆的行为进行决策。

# 华 中 科 技 大 学 硕 士 学 位 论 文

---

本文的研究中仍然存在一些不足，在相机运动过大的场景，例如车辆转弯时，车内相机会大幅运动，这时便会导致运动模型失效，带来许多跟踪错误的情况。后续将会针对相机运动剧烈这个问题，加入相机运动补偿，设计一些可学习的更鲁棒的运动模型，使得多目标跟踪能够在更加广泛的场景中得到应用。

# 华中科技大学硕士学位论文

---

## 致谢

七年前，我从家乡无锡来到武汉华中科技大学，在这里度过了丰富多彩的本科和研究生生活。首先要特别感谢我的导师刘文予老师和王兴刚老师，作为我科研路上的引路人，两位老师一起为我确定了科研方向，培养了我在科研上的兴趣，并教会了我如何发现问题和解决问题的思路，教会了我如何把握科研未来发展趋势和如何抓住做研究的机会。也非常感谢两位老师能够送我去微软亚洲研究院实习，开拓了我的视野，让我受益良多。

在微软亚洲研究院智能多媒体组实习期间，最要感谢的是我实习期间的导师王春雨老师，在他认真负责的指导下，我的科研基本功得到了很大的提升。研究院的大牛曾文军老师和王井东老师也给予了我很宝贵的指导，曾文军老师经常给我们说要做有影响力的工作，王井东老师教会了我如何细致地阅读科研论文。微软亚洲研究院的科研氛围非常浓厚，在这里我同样要感谢陪我一起科研一起生活的小伙伴们，他们有付靖文、文昱晴、张晓艺、谢荣昌、涂涵越、张哲、王光庭、张直政、徐良、胡姚姒、石恩升、何馨毅，感谢大家的陪伴，让我的“北漂”生活充实无比。

在字节跳动的应用视觉组实习期间，非常感谢我的导师喻冬东和江毅，他们花了大量的时间和我一起讨论论文的想法和创新点以及需要完成哪些实验，每周一次的论文分享也让我学习到了许多计算机视觉其他领域的知识。字节跳动的使命是“激发创造，丰富生活”，这里自由的科研氛围让我对科研的兴趣进一步加深，我也要感谢和我一起实习的同学们，向他们学到了许多，他们有孙培泽、严彬、韩楚楚、李新阳、张天舒、林闯、吴剑南、袁小丁、刁其帅。

此外，我也非常感谢学校里一起学习和生活的同学，他们有我的室友陈诚、江子文和顾昕，七年同学来之不易。还有实验室同学肖劲轩、周星升、翁付成、朱多旺、黄小虎、石以昂、王小康、郝振阳、李健锐、胡斌、程天恒、方杰民、廖本成、齐继扬、方羽新、冯嘉佩、刘艺璇、陈少宇、朱良辉，组成了一个温暖的大家庭。

最后，我要特别感谢我的家人，他们永远是我最坚强的后盾。

## 参考文献

- [1] 常亮, 邓小明, 周明全, 等. 图像理解中的卷积神经网络. 自动化学报, 2016, 42(9): 1300-1312
- [2] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016. Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016: 770-778
- [3] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述. 计算机学报, 2017, 40(6): 1229-1251
- [4] 蒋恋华, 甘朝晖, 蒋旻. 多目标跟踪综述. 计算机系统应用, 2010 (12): 271-275
- [5] 王宝树, 李芳社. 基于数据融合技术的多目标跟踪算法研究. 西安电子科技大学学报, 1998, 25(3): 269-272
- [6] 李天成, 范红旗, 孙树栋. 粒子滤波理论, 方法及其在多目标跟踪中的应用. 自动化学报, 2015, 41(12): 1981-2002
- [7] 刘钢, 刘明, 匡海鹏, 等. 多目标跟踪方法综述. 电光与控制, 2004, 11(3): 26-29
- [8] 黄常青, 郑链, 宋承天. 红外多目标跟踪算法研究. 红外与激光工程, 2005, 34(2): 188-191
- [9] N. Wojke, A. Bewley, D. Paulus. Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing, ICIP 2017. Beijing, China, September 17-20, 2017, IEEE, 2017: 3645-3649
- [10] Kalman R E. A new approach to linear filtering and prediction problems. 1960
- [11] Kuhn H W. The Hungarian method for the assignment problem. Naval research logistics quarterly, 1955, 2(1 - 2): 83-97
- [12] 尹宏鹏, 陈波, 柴毅, 等. 基于视觉的目标检测与跟踪综述. 自动化学报, 2016, 42(10): 1466-1489
- [13] 代科学, 李国辉, 涂丹, 等. 监控视频运动目标检测减背景技术的研究现状和展望. 中国图象图形学报, 2006, 11(7): 919-927
- [14] 罗会兰, 陈鸿坤. 基于深度学习的目标检测研究综述. 电子学报, 2020, 48(6):

# 华中科技大学硕士学位论文

---

1230-1239

- [15] A. Milan, L. Leal-Taix'e, I. D. Reid, S. Roth, K. Schindler. Mot16: a benchmark for multi-object tracking. CoRR. 2016, abs/1603.00831
- [16] P. F. Felzenszwalb, D. A. McAllester, D. Ramanan. A discriminatively trained, multiscale, deformable part model. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008. Anchorage, Alaska, USA, IEEE Computer Society, 2008
- [17] S. Ren, K. He, R. B. Girshick, J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017, 39(6): 1137-1149
- [18] F. Yang, W. Choi, Y. Lin. Exploit all the layers: fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016. Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016: 2129-2137
- [19] J. Xu, Y. Cao, Z. Zhang, H. Hu. Spatial-temporal relation networks for multi-object tracking. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019. Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019: 3987-3997
- [20] P. Chu, H. Ling. Famnet: joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019. Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019: 6171-6180
- [21] P. Bergmann, T. Meinhardt, L. Leal-Taix'e. Tracking without bells and whistles. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019. Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019: 941-951
- [22] G. Bras'o, L. Leal-Taix'e. Learning a neural solver for multiple object tracking. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020. Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020: 6246-6256

# 华中科技大学硕士学位论文

---

- [23] T. Lin, P. Goyal, R. B. Girshick, K. He, P. Doll'ar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2020, 42(2): 318-327
  - [24] X. Zhou, D. Wang, P. Kr"ahenb"uhl. Objects as points. *CoRR*. 2019, abs/1904.07850
  - [25] J. Redmon, A. Farhadi. Yolov3: an incremental improvement. *CoRR*. 2018, abs/1804.02767
  - [26] A. Bochkovskiy, C. Wang, H. M. Liao. Yolov4: optimal speed and accuracy of object detection. *CoRR*. 2020, abs/2004.10934
  - [27] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun. Yolox: exceeding yolo series in 2021. *CoRR*. 2021, abs/2107.08430
  - [28] A. Bewley, Z. Ge, L. Ott, F. T. Ramos, B. Upcroft. Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing, ICIP 2016. Phoenix, AZ, USA, September 25-28, 2016, IEEE, 2016: 3464-3468
  - [29] E. Bochinski, V. Eiselein, T. Sikora. High-speed tracking-by-detection without using image information. In: 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017. Lecce, Italy, August 29 - September 1, 2017, IEEE Computer Society, 2017: 1-6
  - [30] X. Zhou, V. Koltun, P. Kr"ahenb"uhl. Tracking objects as points. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow. UK, August 23-28, 2020, Proceedings, Part IV, Springer, 2020: 474-490
  - [31] P. Sun, Y. Jiang, R. Zhang, E. Xie, J. Cao, X. Hu, et al. Transtrack: multiple-object tracking with transformer. *CoRR*. 2020, abs/2012.15460
  - [32] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang, J. Yuan. Track to detect and segment: an online multi-object tracker. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021: 12352-12361
  - [33] S. H. Bae, K. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014. Columbus, OH, USA, June 23-28, 2014, IEEE Computer Society, 2014: 1218-1225
-

# 华中科技大学硕士学位论文

---

- [34] Z. Chen, V. Badrinarayanan, C. Lee, A. Rabinovich. Gradnorm: gradient normalization for adaptive loss balancing in deep multitask networks. In: Proceedings of the 35th International Conference on Machine Learning, ICML 2018. Stockholmsm"assan, Stockholm, Sweden, July 10-15, 2018, PMLR, 2018: 793-802
- [35] A. Kendall, Y. Gal, R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018. Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018: 7482-7491
- [36] O. Sener, V. Koltun. Multi-task learning as multi-objective optimization. In: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018. December 3-8, 2018, Montr'eal, Canada, : 525-536
- [37] Z. Wang, L. Zheng, Y. Liu, Y. Li, S. Wang. Towards real-time multi-object tracking. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow. UK, August 23-28, 2020, Proceedings, Part XI, Springer, 2020: 107-122
- [38] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, et al. Mots: multi-object tracking and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019. Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019: 7942-7951
- [39] K. He, G. Gkioxari, P. Doll'ar, R. B. Girshick. Mask r-cnn. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020, 42(2): 386-397
- [40] L. Chen, H. Ai, Z. Zhuang, C. Shang. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In: 2018 IEEE International Conference on Multimedia and Expo, ICME 2018. San Diego, CA, USA, July 23-27, 2018, IEEE Computer Society, 2018: 1-6
- [41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al. Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9. 2017, Long Beach, CA, USA, : 5998-6008

# 华中科技大学硕士学位论文

---

- [42] T. Meinhardt, A. Kirillov, L. Leal-Taix'e, C. Feichtenhofer. Trackformer: multi-object tracking with transformers. CoRR. 2021, abs/2101.02702
  - [43] F. Zeng, B. Dong, T. Wang, C. Chen, X. Zhang, Y. Wei. Motr: end-to-end multiple-object tracking with transformer. CoRR. 2021, abs/2105.03247
  - [44] L. Zhang, Y. Li, R. Nevatia. Global data association for multi-object tracking using network flows. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008. Anchorage, Alaska, USA, IEEE Computer Society, 2008
  - [45] J. Berclaz, F. Fleuret, E. T"uretken, P. Fua. Multiple object tracking using k-shortest paths optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2011, 33(9): 1806-1819
  - [46] A. R. Zamir, A. Dehghan, M. Shah. Gmcp-tracker: global multi-object tracking using generalized minimum clique graphs. In: Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II, Springer, 2012: 343-356
  - [47] A. Horn'akov'a, R. Henschel, B. Rosenhahn, P. Swoboda. Lifted disjoint paths with application in multiple object tracking. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020. 13-18 July 2020, Virtual Event, PMLR, 2020: 4364-4375
  - [48] C. Feichtenhofer, A. Pinz, A. Zisserman. Detect to track and track to detect. In: IEEE International Conference on Computer Vision, ICCV 2017. Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017: 3057-3065
  - [49] H. Luo, W. Xie, X. Wang, W. Zeng. Detect or track: towards cost-effective video object detection/tracking. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019. The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019: 8803-8810
  - [50] P. Tang, C. Wang, X. Wang, W. Liu, W. Zeng, J. Wang. Object detection in videos by high quality object linking. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020, 42(5): 1272-1278
-

# 华中科技大学硕士学位论文

---

- [51] W. Han, P. Khorrami, T. L. Paine, P. Ramachandran, M. Babaieizadeh, H. Shi, et al. Seq-nms for video object detection. CoRR. 2016, abs/1602.08465
- [52] B. Pang, Y. Li, Y. Zhang, M. Li, C. Lu. Tubetk: adopting tubes to track multi-object in a one-step training model. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020. Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020: 6307-6317
- [53] L. Bridgeman, M. Volino, J. Guillemaut, A. Hilton. Multi-person 3d pose estimation and tracking in sports. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019. Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019: 2487-2496
- [54] J. Dong, W. Jiang, Q. Huang, H. Bao, X. Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019. Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019: 7792-7801
- [55] L. Chen, H. Ai, R. Chen, Z. Zhuang, S. Liu. Cross-view tracking for multi-human 3d pose estimation at over 100 fps. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020. Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020: 3276-3285
- [56] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, S. Ilic. 3d pictorial structures revisited: multiple human pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016, 38(10): 1929-1942
- [57] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, S. Ilic. 3d pictorial structures for multiple human pose estimation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014. Columbus, OH, USA, June 23-28, 2014, IEEE Computer Society, 2014: 1669-1676
- [58] L. Leal-Taix'e, A. Milan, I. D. Reid, S. Roth, K. Schindler. Motchallenge 2015: towards a benchmark for multi-target tracking. CoRR. 2015, abs/1504.01942
- [59] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. D. Reid, et al. Mot20: a benchmark for multi object tracking in crowded scenes. CoRR. 2020, abs/2003.09003

# 华中科技大学硕士学位论文

---

- [60] W. Lin, H. Liu, S. Liu, Y. Li, G. Qi, R. Qian, et al. Human in events: a large-scale benchmark for human-centric video analysis in complex events. CoRR. 2020, abs/2005.04490
- [61] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, et al. Bdd100k: a diverse driving dataset for heterogeneous multitask learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020. Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020: 2633-2642
- [62] V. Belagiannis, X. Wang, B. Schiele, P. Fua, S. Ilic, N. Navab. Multiple human pose estimation with temporally consistent 3d pictorial structures. In: Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland. September 6-7 and 12, 2014, Proceedings, Part I, Springer, 2014: 742-754
- [63] H. Joo, H. Liu, L. Tan, L. Gui, B. C. Nabbe, I. A. Matthews, et al. Panoptic studio: a massively multiview system for social motion capture. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015. Santiago, Chile, December 7-13, 2015, IEEE Computer Society, 2015: 3334-3342
- [64] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, J. Yan. Poi: multiple object tracking with high performance detection and appearance feature. In: Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands. October 8-10 and 15-16, 2016, Proceedings, Part II, : 36-42
- [65] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian. Person re-identification in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017: 3346-3355
- [66] L. Wei, S. Zhang, W. Gao, Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018. Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018: 79-88
- [67] T. Khurana, A. Dave, D. Ramanan. Detecting invisible people. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021. Montreal, QC, Canada, October 10-17, 2021, IEEE, 2021: 3154-3164
- [68] F. Yu, D. Wang, T. Darrell. Deep layer aggregation. CoRR. 2017, abs/1707.06484

# 华中科技大学硕士学位论文

---

- [69] T. Lin, P. Doll'ar, R. B. Girshick, K. He, B. Hariharan, S. J. Belongie. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017: 936-944
- [70] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, et al. Deformable convolutional networks. In: IEEE International Conference on Computer Vision, ICCV 2017. Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017: 764-773
- [71] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2021, 43(10): 3349-3364
- [72] P. Chao, C. Kao, Y. Ruan, C. Huang, Y. Lin. Hardnet: a low memory traffic network. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019. Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019: 3551-3560
- [73] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, et al. Microsoft coco: common objects in context. In: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, Springer, 2014: 740-755
- [74] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, et al. Crowdhuman: a benchmark for detecting human in a crowd. CoRR. 2018, abs/1805.00123
- [75] A. Ess, B. Leibe, K. Schindler, L. V. Gool. A mobile vision system for robust multi-person tracking. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008. Anchorage, Alaska, USA, IEEE Computer Society, 2008
- [76] S. Zhang, R. Benenson, B. Schiele. Citypersons: a diverse dataset for pedestrian detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017: 4457-4465
- [77] P. Doll'ar, C. Wojek, B. Schiele, P. Perona. Pedestrian detection: a benchmark. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern

# 华中科技大学硕士学位论文

---

- Recognition (CVPR 2009), 20-25 June 2009. Miami, Florida, USA, IEEE Computer Society, 2009: 304-311
- [78] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang. Joint detection and identification feature learning for person search. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017: 3376-3385
- [79] K. Bernardin, R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing. 2008, 2008
- [80] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In: Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands. October 8-10 and 15-16, 2016, Proceedings, Part II, : 17-35
- [81] D. P. Kingma, J. Ba. Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015. San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings,
- [82] Y. Xiang, A. Alahi, S. Savarese. Learning to track: online multi-object tracking by decision making. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015. Santiago, Chile, December 7-13, 2015, IEEE Computer Society, 2015: 4705-4713
- [83] S. H. Bae, K. Yoon. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018, 40(3): 595-610
- [84] R. Sanchez-Matilla, F. Poiesi, A. Cavallaro. Online multi-target tracking with strong and weak detections. In: Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands. October 8-10 and 15-16, 2016, Proceedings, Part II, : 84-99
- [85] L. Chen, H. Ai, C. Shang, Z. Zhuang, B. Bai. Online multi-object tracking with convolutional neural networks. In: 2017 IEEE International Conference on Image Processing, ICIP 2017. Beijing, China, September 17-20, 2017, IEEE, 2017: 645-649

# 华中科技大学硕士学位论文

---

- [86] K. Fang, Y. Xiang, X. Li, S. Savarese. Recurrent autoregressive networks for online multi-object tracking. In: 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018. Lake Tahoe, NV, USA, March 12-15, 2018, IEEE Computer Society, 2018: 466-475
- [87] N. Mahmoudi, S. M. Ahadi, M. Rahmati. Multi-target tracking using cnn-based features: cnnttt. Multim. Tools Appl.. 2019, 78(6): 7077-7096
- [88] J. Peng, C. Wang, F. Wan, Y. Wu, Y. Wang, Y. Tai, et al. Chained-tracker: chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow. UK, August 23-28, 2020, Proceedings, Part IV, Springer, 2020: 145-161
- [89] S. Sun, N. Akhtar, H. Song, A. Mian, M. Shah. Deep affinity network for multiple object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021, 43(1): 104-119
- [90] J. Luiten, A. Osep, P. Dendorfer, P. H. S. Torr, A. Geiger, L. Leal-Taix'e, et al. Hota: a higher order metric for evaluating multi-object tracking. International Journal of Computer Vision. 2021, 129(2): 548-578
- [91] F. Yang, X. Chang, S. Sakti, Y. Wu, S. Nakamura. Remot: a model-agnostic refinement for multiple object tracking. Image and Vision Computing. 2021, 106: 104091
- [92] C. Liang, Z. Zhang, Y. Lu, X. Zhou, B. Li, X. Ye, et al. Rethinking the competition between detection and reid in multi-object tracking. CoRR. 2020, abs/2010.12138
- [93] C. Liang, Z. Zhang, X. Zhou, B. Li, Y. Lu, W. Hu. One more check: making "fake background" be tracked again. CoRR. 2021, abs/2104.09441
- [94] Q. Wang, Y. Zheng, P. Pan, Y. Xu. Multiple object tracking with correlation learning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021: 3876-3886
- [95] J. Pang, L. Qiu, X. Li, H. Chen, Q. Li, T. Darrell, et al. Quasi-dense similarity learning for multiple object tracking. In: IEEE Conference on Computer Vision and

# 华中科技大学硕士学位论文

---

- Pattern Recognition, CVPR 2021. virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021: 164-173
- [96] L. Zheng, M. Tang, Y. Chen, G. Zhu, J. Wang, H. Lu. Improving multiple object tracking with single object tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021: 2453-2462
- [97] Y. Xu, Y. Ban, G. Delorme, C. Gan, D. Rus, X. Alameda-Pineda. Transcenter: transformers with dense queries for multiple-object tracking. CoRR. 2021, abs/2103.15145
- [98] Y. Wang, K. Kitani, X. Weng. Joint object detection and multi-object tracking with graph neural networks. In: IEEE International Conference on Robotics and Automation, ICRA 2021. Xi'an, China, May 30 - June 5, 2021, IEEE, 2021: 13708-13715
- [99] E. Yu, Z. Li, S. Han, H. Wang. Relationtrack: relation-aware multiple object tracking with decoupled representation. CoRR. 2021, abs/2105.04322
- [100] P. Tokmakov, J. Li, W. Burgard, A. Gaidon. Learning to track with object permanence. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021. Montreal, QC, Canada, October 10-17, 2021, IEEE, 2021: 10840-10849
- [101] Y. Zhang, H. Sheng, Y. Wu, S. Wang, W. Ke, Z. Xiong. Multiplex labeling graph for near-online tracking in crowded scenes. IEEE Internet of Things Journal. 2020, 7(9): 7892-7902
- [102] G. Moon, J. Y. Chang, K. M. Lee. V2v-posenet: voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018. Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018: 5079-5088
- [103] Y. Yan, Y. Mao, B. Li. Second: sparsely embedded convolutional detection. Sensors. 2018, 18(10): 3337
- [104] K. Iskakov, E. Burkov, V. S. Lempitsky, Y. Malkov. Learnable triangulation of human pose. In: 2019 IEEE/CVF International Conference on Computer Vision,

# 华中科技大学硕士学位论文

---

ICCV 2019. Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019:  
7717-7726

- [105] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, et al. Posetrack: a benchmark for human pose estimation and tracking. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018. Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018: 5167-5176
- [106] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, L. Zhang. Higherhrnet: scale-aware representation learning for bottom-up human pose estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020. Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020: 5385-5394
- [107] H. Tu, C. Wang, W. Zeng. Voxelpose: towards multi-camera 3d human pose estimation in wild environment. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow. UK, August 23-28, 2020, Proceedings, Part I, Springer, 2020: 197-212
- [108] T. Wang, J. Zhang, Y. Cai, S. Yan, J. Feng. Direct multi-view multi-person 3d pose estimation. CoRR. 2021, abs/2111.04076
- [109] N. D. Reddy, L. Guigues, L. Pishchulin, J. Eledath, S. G. Narasimhan. Tessetrack: end-to-end learnable multi-person articulated 3d pose tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021: 15190-15200
- [110] S. Ershadi-Nasab, E. Noury, S. Kasaei, E. Sanaei. Multiple human 3d pose estimation from multiview images. Multim. Tools Appl.. 2018, 77(12): 15573-15601
- [111] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, L. Chen. Mobilenetv2: inverted residuals and linear bottlenecks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018. Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018: 4510-4520

# 华中科技大学硕士学位论文

---

- [112] B. Xiao, H. Wu, Y. Wei. Simple baselines for human pose estimation and tracking. In: Computer Vision - ECCV 2018 - 15th European Conference, Munich. Germany, September 8-14, 2018, Proceedings, Part VI, Springer, 2018: 472-487
- [113] H. Zhang, M. Ciss'e, Y. N. Dauphin, D. Lopez-Paz. mixup: beyond empirical risk minimization. In: 6th International Conference on Learning Representations, ICLR 2018. Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018
- [114] Z. Ge, S. Liu, Z. Li, O. Yoshie, J. Sun. Ota: optimal transport assignment for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021: 303-312
- [115] C. Wang, H. M. Liao, Y. Wu, P. Chen, J. Hsieh, I. Yeh. CspNet: a new backbone that can enhance learning capability of cnn. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020. Seattle, WA, USA, June 14-19, 2020, Computer Vision Foundation / IEEE, 2020: 1571-1580
- [116] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia. Path aggregation network for instance segmentation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018. Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018: 8759-8768
- [117] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. D. Reid, S. Savarese. Generalized intersection over union: a metric and a loss for bounding box regression. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019. Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019: 658-666
- [118] Z. Wang, H. Zhao, Y. Li, S. Wang, P. H. S. Torr, L. Bertinetto. Do different tracking tasks require different appearance models?. CoRR. 2021, abs/2107.02156
- [119] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009. Miami, Florida, USA, IEEE Computer Society, 2009: 248-255

# 华 中 科 技 大 学 硕 士 学 位 论 文

---

---

## 附录 1 攻读硕士学位期间取得的研究成果

### 发表与接收论文

- [1] **Yifu Zhang\***, Chunyu Wang\*, Xinggang Wang, Wenjun Zeng, Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking[J]. International Journal of Computer Vision, 2021, 129(11): 3069-3087. 华中科技大学为第一单位
- [2] **Yifu Zhang**, Chunyu Wang, Xinggang Wang, Wenyu Liu, Wenjun Zeng. Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022. 华中科技大学为第一单位