分类号：TP391.4

学校代码：10406

学号：1816085212017

南昌航空大学

硕 士 学 位 论 文

（专业学位研究生）

引入姿态信息的跨摄像头多目标跟踪方法

硕士研究生： 吴沛沛

导 师： 储珺教授

申请学位级别：硕 士

学科、专业： 计算机技术

所在单位： 软件工程学院

答辩日期： 2021-06

授予学位单位：南昌航空大学

Multi-Camera Multiple Target Tracking with Pose Information

A Thesis

Submitted for the Degree of Master

On Computer Technology

by **Peipei Wu**

Under the Supervision of

Prof. Jun Chu

School of Software Engineering

Nanchang Hangkong University, Nanchang, China

June, 2021

摘要

基金资助：国家自然科学基金(61663031)，江西省科技支撑计划项目(No.20161BBE50085，20192BBE50073)

目标跟踪广泛应用于各类工业应用，多传感器下的多目标跟踪任务更具实际应用意义。基于卷积神经网络的跨摄像头多目标跟踪技术的关键在于能否充分利用单传感器视野内的各个目标在时空域上的特征信息，以及跨传感器间的目标关联性。本文围绕这两点展开研究，主要工作包括以下部分：

(1) 提出一种引入姿态信息的联合检测多目标跟踪（JDE, Jointly learns the Detector and Embedding model）算法。传统的多目标跟踪算法是检测与数据融合分离式的算法(SDE, separate detection and embedding learning)，该类型的算法计算复杂，且多段式分离，处理耗时较长，无法达到实时性的需求。使用 JDE 跟踪算法可以合并检测和 embedding(数据融合)两个阶段，达到减少耗时，接近实时性的任务需求。然而当前 JDE 类型的算法检测部分精度较差导致 ID switch 数量上升，跟踪不稳定。本文提出了一种引入目标姿态信息的 JDE 多目标跟踪算法并替换了原有的 JDE 检测部分结构，提取多尺度的目标，提高了目标特征信息的丰富性，加强了目标表达模型的鲁棒性。在对算法实时性影响较小的条件下，提高了 JDE 跟踪算法的抗 ID switch(目标标识转换)能力，提高了跟踪可靠性。

(2) 提出一种集中式的跨摄像头多目标跟踪框架并探讨了部署该框架到物联网(IoT, Internet of Things 设备上搭建系统的可能性。将分布式传感器网络中的各个摄像头节点收集的图像信息，分别使用引入姿态信息的 JDE 框架得到相应的嵌入了 embedding 信息和姿态描述信息的检测结果。对检测结果通过多贝叶斯估计法进行数据融合得到精确的目标观测状态表达，提高了目标全局观测精度。将提高了观测精度的检测结果送入运动模型得到全局的跟踪结果。在尽可能减少耗时的情况下得到了一个精度较高，鲁棒性较强的跨摄像头多目标跟踪框架。

(3) 本文算法在 MOT，JTA 等数据集上完成了验证，与当前已有的性能较好的算法对比。实验证明了算法速度到达了近实时(Towards to real time)，算法精度和可靠性也有充分保证。

关键词： 行人跟踪，多目标跟踪，卷积神经网络，数据融合，分布式传感器管理

Abstract

With development of technology and complexity of application scenarios, multi-camera multiple targets tracking has more practical application significance.

The key to the multi-camera multiple targets tracking technology based on convolutional neural network is whether it can make full use of the features information of each target from each single-sensor's view in the temporal and spatial domain, and the target relevance between different sensors. The main contribution includes the following parts:

A joint learning detector and embedding model tracking (JDE) algorithm with posture information is proposed. Traditional multiple targets tracking methods are usually detection and data fusion separately, which called separate detection and embedding learning (SDE), those methods are computation complex and multiple stages separated. Thus, processing time-consuming is enormous that cannot reach the real-time requirements. Using the JDE tracking algorithm can combine the two stages, detection and embedding (data fusion), to reduce the time-consuming and improve towards real-time task requirements. In current JDE algorithms, the poor accuracy of the detection part leads to ID switches increasing and tracking unstable. This paper proposes a multiple targets tracking algorithm based on JDE with posture information introduced and the original structure of the JDE detection part is replaced. Extracting targets from different scales can improve the richness of target feature information, and it strengthens the robustness of the target expression model. Although a little bit speed of algorithm decreasing, anti ID switch ability of the JDE tracking algorithm and the tracking reliability is improved.

Propose a centralized multi-camera multiple targets tracking framework and discuss feasibility of deploying algorithms on IoT devices to build a target tracking system. The frame information collected by each camera from the distributed sensor network is used to obtain the corresponding detection result with embedding information and pose description via new JDE framework given in this paper. To fuse the detection results from each agent by multi-Bayesian estimation method can obtain a global accurate expression of the target observation state, which will improve the target tracking accuracy. Although a little bit speed downing, a multi-camera multiple

targets tracking framework with higher accuracy and robustness is obtained.

The algorithm in this paper has been tested on different datasets, including MOT and JTA. Also, it has been compared with the state-of-the-art algorithms. In the simulation experiment, the algorithm proved that the algorithm speed has reached near real-time (towards real-time), and the accuracy and reliability of the algorithm are also fully guaranteed.

Keywords: Pedestrian tracking, multiple targets tracking, convolutional neural network, data fusion, distributed sensor network management

目录

第 1 章 绪论

1.1 研究背景及意义

计算机科学技术的进步，促使了人工智能等计算机科学子学科快速发展，相关技术衍生出的产品在日常生活中被广泛使用。其中依托于机器感知技术的产品是当下最热门且在工业实际应用中有良好表现的一类。绝大多数的机器感知技术可依据不同的感知种类进行划分，包括视觉、听觉、触觉等不同的机器感知。其中，机器视觉和机器听觉是目前发展最好的两类机器感知技术，机器视觉技术衍生出的产品涵盖了自动驾驶，安防智能摄像头等，而基于机器听觉发明了智能翻译机、智能语音助手等。本文专注于机器视觉技术的研究。机器视觉的三大基础任务分别是目标检测，语义分割和目标跟踪。相较于前两项任务，目标跟踪任务具有条件复杂，难以实现的特点。同样，该任务又可以按照目标数量的不同分为单目标跟踪和多目标跟踪。在后续的研究发展过程中，考虑到传感器的数量不再局限单独一个，因而出现了单传感器跟踪和多传感器跟踪两大分支，后者多传感器多目标跟踪任务最为贴近实际应用。特别是在公安安全防范工作中，研究多传感器多目标跟踪的应用具有实际意义。当下，我国已经部署的监控摄像头数量迅速增长，公安系统在执行视频侦查任务时，常用的侦察手段仍然是使用人工进行处理，不仅大量警力被消耗，而且在侦察过程中存储了海量冗余视频片段，容易造成跟踪目标的遗漏或错误识别。为了解决实际应用中存在的困境，开发一个智能化的跨摄像头多目标跟踪系统是十分必要的。

完整的智能化跨摄像头多目标跟踪系统涉及多个不同技术，通常可以按照服务层级进行划分，通常情况下涵盖了图像处理，图像分析，智能决策等主要技术，如图 1-1 所示智能摄像头跟踪系统的服务层级划分。最底层服务以图像处理技术为基础，主要完成了图像采集和图像数据处理任务，便于数据在结点间的传输和为后续图像分析做好预处理，这其中包括了图像增强、滤波、恢复、编码等操作。次级服务则是以图像分析为主，在这一层中完成对目标进行跟踪，主要操作包括了特征提取、目标检测、目标跟踪等。最高层服务则是图像理解，在这一层中对跟踪等目标进行重识别并做出最终决策。

除了安全防范工程任务之外，跨摄像头多目标跟踪技术在其他任务中也得到了大量应用。在医院、养老院中该技术可以对病人和需要受照顾的老人进行监护，判断目标人的肢体语言及其位置，对诸如摔跤、求救等可能的突发事件及时报警；在自动驾驶中，跨摄像头多目标跟踪技术可以用于监控周遭车辆，当其他车辆进入本车安全车距时自动减速保证自动驾驶时行车安全；同样，在疫情期间网络视

频会议也迅速发展，跨摄像头多目标跟踪技术能提供寻找网络会议室中活跃用户并进行自动对焦的功能，保证会议质量。当然，跨摄像头多目标跟踪技术还是更多应用在安全防范工程任务之中。例如,在重要交通要道的路口处,公安部门通常会部署多个监控摄像头，对路口处的车辆及行人进行安全监控和行为研判工作。

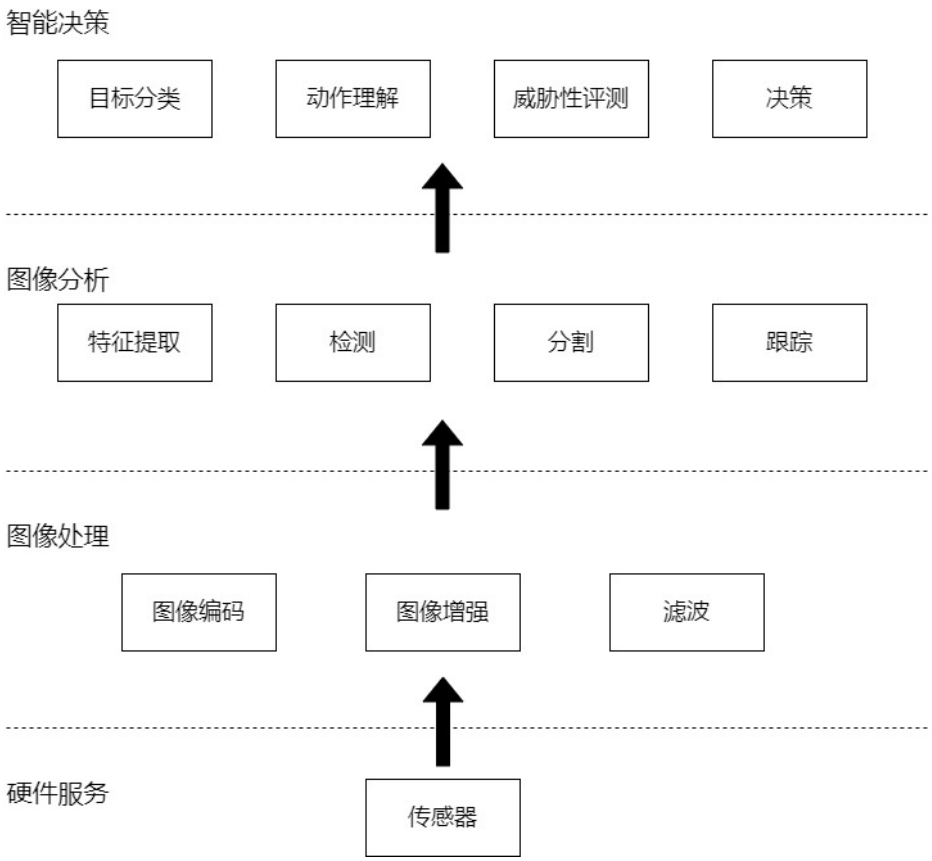

图 1-1 常见智能目标跟踪系统的服务层级划分

然而对多个目标实现跨摄像头跟踪是困难的。例如，在对嫌疑车辆或嫌疑人员进行监控跟踪时，存在目标会在某一个摄像头的监控视野中频繁出入的可能，或者在不同摄像头的视野中某个目标同时出现。与单摄像头目标跟踪任务不同，跨摄像头多目标跟踪任务还需要在不同传感器视野里对相同目标进行匹配，这无疑更增加了实现难度。因为上述难点的存在，导致了实现鲁棒且实时的跨摄像头多目标跟踪系统是一个巨大挑战。

本文考虑真实应用场景，重点关注于多目标跟踪和目标重识别关键技术并引入了姿态特征信息完成了实现鲁棒且接近实时的跨摄像头多目标跟踪系统的设计。该系统可以用于安全防范工程任务中，实现监控、预警的功能。

1.2 国内外研究现状

近些年来，多目标跟踪技术因其有重要的学术研究意义和巨大商业应用前景，是目前受学界工业界最为关注的研究方向，也是当前机器感知技术，尤其是机器视觉方向中最热门的研究话题。针对跨摄像头这一特殊的应用场景，当前也已经有了一定的研究基础，绝大多数工作都依赖于模板匹配的思想，缺少了其他种类信息的加入。本文引入了姿态特征信息用于跟踪任务，因此本小节将从目标检测、目标跟踪、跨摄像头目标重识别和姿态估计四个方向总结了当下国内外研究现状。

1.2.1 目标检测算法研究现状

目标检测被定义为在感知视野内对目标进行准确定位的任务。它是计算机视觉的基础任务之一，同时也是绝大多数目标跟踪算法的基础前提。目前成熟的目标检测方法被分为了两大类：运动目标检测、目标类别检测。

运动目标检测主要通过分析目标的运动信息完成检测任务，常用的方法有：背景去除法[1]、帧间差分法[2][3]和光流法[4][5]。背景去除法主要是通过对比统计前序帧的信息变化，找出不变的信息（背景）以获得背景模型。将新输入的图像帧与背景模型做差集获取运动前景。算法实现难度较低，但是鲁棒性较差，尤其是当运动目标运动较为缓慢，帧间差异不大时，算法检测效果较差。帧间差分法则通过比较连续数帧的灰度差异获取运动信息。算法优点是速度快，计算资源需求量小，但是无法较好的描述目标轮廓，只能进行粗略检测，精度较差。光流法无需和前序两类方法一样统计数帧信息，该方法只需要对比当前帧和前一帧的像素点差异或运动矢量信息的不同，图像中像素点出现差异的部分即为目标。即使背景和传感器位置发生变化时，该算法也能较好的进行检测。但是运算量大，鲁棒性较差也是该类算法的特点。

目标类别检测与运动目标检测相比，还需要对出现在图像帧中的每一个已知类别目标进行定位并且正确归类。在早期，传统方法主要聚焦于如何创建特征描述。较出名的方法包括了 SIFT 特征描述[错误!未找到引用源。]、色彩直方图描述[7]等。传统方法的主要特点是把整个算法框架分成两部分：特征提取和分类。特征提取部分不需要进行学习主要通过人为设计特征提取器实现；只有分类部分才需要进行学习。这类传统方法在场景不复杂的任务中有较好表现，但对不易人为设计特征描述的目标则无能为力。

卷积神经网络的提出使目标类别检测有了较大的发展。逐渐提出了将目标类别检测任务描述为对输入的图像帧中部分特定区域做分类操作的思想。1998 年，第一个用于分类任务的卷积神经网络 LeNet[8]由 Hinton 等人提出，整个网络包含

了 5 个卷积层和 3 个全连接层，同时对特征提取和分类进行学习。该方法极大程度的提高了目标检测的准确率。在此之后的近 15 年的时间里，受困于计算资源的限制，基于卷积神经网络的目标检测算法都没有较大的进展。2012 年，AlexNet[9]的提出再一次提高了算法的准确性并赢得了当年的目标检测竞赛，相较于 LeNet，网络结构的深度和参数数量都得到了加大。GoogleNet[10]提出了 Inception 结构，这一创新解决了过往卷积神经网络的输入必须是固定尺寸的困境。不同尺寸的输入都可以被送入到拥有 Inception 结构的网络中。VGG[11]提出了用多层小尺寸卷积核替代大尺寸卷积核的思想，这一思想使得卷积神经网络可以用更少的参数获取更多的特征信息，对硬件的计算量需求大幅度下降。ResNet[12]则进一步加深了卷积神经网络的深度，这使得精度进一步提升，同时解决了因为加深网络结构深度造成的梯度消失问题。卷积神经网络的深度相较之前得到了巨大的提高。在那之后，目标检测算法进一步发展形成了两大类分支：单阶段检测和多阶段检测。最著名的多阶段检测算法（或称为双阶段算法）是 R-CNN[13]，将之前的卷积神经网络成果与区域思想相结合，它先启发式地创建不同的感兴趣区域并把不同生成区域利用分类算法完成检测。很明显算法结构被分成了创建区域和对区域分类两阶段。基于这样的思想，形成了一系列检测算法，它们是 SPNet[14], Fast-RCNN[15]，Faster-RCNN[16]和 Mask-RCNN[17]，精度和速度同时得到了提高。但是多阶段的检测算法在实时性上还是有所缺陷，因此单阶段检测算法孕育而生，YoLo[18]和 SSD[19]是最著名的两个单阶段算法。绝大多数的单阶段检测算法都是基于搜索窗移动，通过在输入的图像上移动不同尺寸大小的视窗，每一个视窗都被用于分类。这一创新极大提高了检测速度，可以基本实现实时任务，尽管相比多阶段检测算法可能会有些许准确率下降。目标检测是大部分目标跟踪算法的一部分，可以为跟踪任务提供必要的先验条件。

1.2.2 多目标跟踪现状

目标跟踪是一个广泛的概念，可以利用不同感知方法对目标进行跟踪，在本文中该词特指视觉目标跟踪。目标跟踪可以被定义为在视频序列信息中，在各个时间帧内对感兴趣的目标进行定位，并对定位的目标进行正确归类和分配正确的身份标识。在跟踪的过程中，目标位置，目标尺寸，目标的运动轨迹等信息都会被获取或利用。如同前文所介绍过的，目标跟踪可按照不同跟踪的目标数量，分类成单目标跟踪（Single Object Tracking，SOT）和多目标跟踪（Multiple Object Tracking，MOT）。无论是单目标跟踪还是多目标跟踪，对于生物而言，该任务无疑是简单的，但对于机器则是一个艰巨的任务。在单目标跟踪任务里，往往会出现遮挡、目标在平面内外旋转、目标消失于视野等问题。对于多目标跟踪任务

而言，单目标跟踪任务所遇到的困难都同样出现在该任务里，同时还需要对跟踪到的目标和正确的身份标识进行匹配，克服目标之间的相互影响等，这使得任务难度进一步提升。

针对多目标跟踪任务，当前有两种不同的解决策略，分别是有检测跟踪（tracking-by-detection，TBD）和无检测跟踪（Detection-free-tracking，DFT）两种。有检测跟踪通常是基于在每一个时间帧上先进行目标检测，再依据检测的结果进行跟踪，因此检测器对于有检测跟踪来讲是必不可少的，跟踪精度较高。相反，无检测跟踪则是不需要检测器，该类跟踪算法是不需要基于检测结果的。同样的，跟踪算法还可以分成在线跟踪（online tracking）和离线跟踪（offline tracking）。这样的分类方法是依据算法如何进行数据关联来完成分类的。在线跟踪算法只需要当前帧和前一个时刻帧的信息完成跟踪任务。相反，离线跟踪算法可以任意使用前序帧进行跟踪，没有帧数限制。因此，离线跟踪算法可以更容易找到全局最优。除此之外，还有一种新型跟踪方法被分类为近在线跟踪，这类算法可以使用部分前序帧信息，这类算法也有很好的应用前景。接下来两段，给出当下效果显著的跟踪方法，并按照 DFT 和 TBD 两种不同策略进行介绍。

无检测跟踪算法 SAC[20]是基于注意力机制的跟踪算法，针对目标区域的权重高于背景区域权值，它引入了长短时记忆机制（long short time memory, LSTM），并对与当前帧处于不同时间间隔的视频帧进行了不同的处理。短时间间隔的帧间信息用于单个目标跟踪器，以提取该时刻的外观和位置信息。相应地，较长的时间间隔的帧间信息用于将相同的目标与其在不同帧中的标识进行匹配。 关联来自两个分支的信息，可以更新目标转换器的注意力，还可以使用匈牙利算法获得二分图，以匹配跟踪结果和检测结果。注意力机制是将目标与正确的识别进行匹配的好选择，但是利用不同时间帧的框架不能支持在线跟踪，这一限制无法较好的支持工业应用。相较于无检测跟踪算法，当前学界和工业界将更多精力放在了有检测跟踪算法之中。

有检测跟踪算法当下百花齐放，有不少算法已经具备了工业应用实战能力。有检测跟踪算法 DeepMOT[21]提出了一种端到端的多目标跟踪算法，其中提出了一种可微分的参数代理方法，以便可以直接学习准确性和准确性的误差。它使用多个单目标跟踪网络分支来达到多目标跟踪目的，每个目标都会分配到一个单目标跟踪器并获得距离矩阵。匈牙利算法用于距离矩阵，以获得优化的参数矩阵。综合损失函数是从优化矩阵获得的。尽管此解决方案可以提供较高的准确性，但也正是因为每个目标都拥有自己的专用单个目标跟踪器导致算法在目标数量过多时，算法对计算资源的需求量过高，只能适应应用场景较为简单的环境。FMA[22]提出了一种基于帧中运动信息和外观信息之间的相关性的方法。网络结构的输入

是一对视频帧，通过残差网络获得运动信息（两个帧之间的位置变化）和外观信息（特征提取），通过运动信息和外观信息得到预测结果，比较预测结果的重叠率，重叠率最大的预测结果将被作为最终的预测结果。但是，当提取的特征出现剧烈变化时，算法的准确性会受到影响。FAMNet[23]提出了一种结合不同维度信息关联的跟踪方法。在每个帧中执行目标检测，获取候选目标并将对其特征提取。已提取的特征被发送到对应子网络，包括运动信息，外观信息，跟踪器的预测信息，每一个分支都会分配不同的权重，该权重影响最终的跟踪结果。显然易见，这是一种计算资源需求高的跟踪方法。STRN[24]提出了一种基于时空相关的多目标跟踪方法，它可以分为两步。首先，输入的视频帧通过目标检测获得边界框，并给出相似度分数作为时空关联网络的输入。关联网络的输出是二分图，使用匈牙利算法给出最终的匹配结果。时空关联网络的先前跟踪结果和针对不同目标的当前检测结果被当作残差网络的输入，以获得时空相关性。该方法的时间复杂度高于其他方法，不适合实时任务。文献[25]基于实例的注意力机制，并与动态模型预测更新相结合完成多目标跟踪。网络首先执行目标检测，然后分析目标模型的预测结果。目标检测器确定的信息与先前的目标模型关联用于克服可能存在的目标遮挡。如果目标被阻挡了，则通过获取的身份验证信息来更新目标模型。更新的目标模型将删除消失的目标并添加新目标。该思路可以解决很目标数量较少的情况，当目标数目增加时，算法的效率会受到数量的影响。 MOTS [25]在完成多目标跟踪的同时完成了语义分割。该方法将特征提取应用于输入视频帧，并使用 3D 卷积内核执行 3D 卷积操作。因此，输出结果是具有增强的时间特征的特征图。将增强的特征图作为输入发送到 RCNN，并获得最终的跟踪结果和语义结果。文献[26]将不同级别的信息与高斯混合概率假设密度过滤器结合起来，以跟踪多个行人。首先，将目标检测应用于输入视频序列，该视频序列具有两个不同的焦点，分别是全身检测和身体部位检测。将检测结果发送到预测阶段，该阶段将增强检测到的目标的身份特性，然后使用增强的数据来驱动 GM-PHD 滤波器。最后，基于滤波处理后的结果完成跟踪确认和模型更新。该方法还收集了人体区域信息，因此可以解决重度遮挡的问题，这带来了更多的计算资源需求，而这些需求一直困扰着业界。FNSP-MOT [28]介绍了一种基于特征金字塔连体网络的多目标跟踪方法。首先将目标检测应用于已知目标，然后将其发送到骨架网络以进行特征提取。提取的特征将被发送到不同层级的孪生网络，以进一步提取特征细节。结合时空信息和运动信息之间的差异，将孪生网络的输出二值化以得到结果。

1.2.3 跨摄像头多目标跟踪研究现状

跨摄像头多目标跟踪又称多摄像机跟踪（MCT, multiple camera tracking），

是指在多个传感器（摄像机）的视场中跟踪目标[27]。多摄像机跟踪任务的要求可以定义为：特定摄像机视频中要跟踪的目标，并进行连续跟踪；当该目标出现在其他摄像机的视野中时，可以自动识别和跟踪目标。同样的，任务要求还可以被定义为是通过多个摄像机进行测量目标状态，融合了多个本地测量状态在全局层面上对目标进行跟踪。与单摄像机跟踪相比，多摄像机跟踪任务难以匹配检测到的目标特征。在不同的相机视图中匹配同一目标的功能以及匹配目标的运动信息（例如，轨迹段）是困难的。甚至对于人眼此项任务也是艰巨的挑战。因此，有必要从多个角度提高目标对象的特征相关性和目标匹配率。

在本文中，主要侧重点在于对来自不同摄像头对同一个目标测量状态进行匹配和数据融合。个人重识别（Re-Identification，Re-ID）是多摄像机跟踪中的一个特殊分支，专注于匹配不同摄像机视图中的同一目标以及同一传感器中的不同帧。它可以根据行人的衣服，身体形状和发型等信息来识别匹配行人。在大众潜意识中，人脸识别信息是一个能较好支持行人重识别任务的特征信息。由于实际应用场景中的数据非常复杂，跨摄像机跟踪的结果会受到很多因素的影响，因此在每帧中清晰捕捉人脸的可能性很小，也正是如此人脸识别信息并不是行人重识别任务力可利用特征信息中的最佳选择。经过分析，动作速率变化，衣服的差异，遮挡，亮度，相机分辨率低和背景变化等都是影响行人重识别效果的部分因素。当无法连续拍摄跨摄像机的行人时，并且无法清晰地拍摄到这些目标，则一种可行的解决方案是增强数据的时空连续性。综上所述，跨相机跟踪算法的基础仍然是提取目标的特征，将目标提取的特征与基础数据库中的数据进行比较，计算查询输入和模板之间的距离，并找到最佳的身份配对。常见的跨摄像头算法具有以下类型：表征学习，度量学习和局部特征学习。其中局部特征学习可以分为两类：基于局部区域增强的局部特征学习和基于姿态估计的局部特征学习，将在下一节中介绍。表示学习和度量学习则在本节中介绍。

表征学习是通过设计分类损失和对比度损失来实现网络结构的监督学习。文献[29]提出了一种名为 k-reciprocal 编码的方法，首先通过目标检测对输入视频帧进行处理以获得样本库，该样本库包含了提取出外观信息和 k 个特征。k 个特征是指来自样本的最接近要匹配的模板前 k 个特征。通过 Mahala Nobis 距离和 Jaccard 距离分别获得外观特征和 k 特征的相关性。对两个方程的结果进行回归以获得跟踪的样本距离。文献[30]首先进行目标检测，然后将提取的目标封城 4 个数据块；并分别对每个数据块进行特征提取，以提取相应块中最显著的特征。最后，将从每个块中提取的最显著特征进行组合以获得目标的特征表示。使用这些功能可以完成目标标识的匹配。文献[31]中提出了 group features 的方法，输入视频帧被发送到残差网络，样本库使用残差网络提取特征，这些特征将权重分配给

残差网络。通过两个残差网络提取的特征组进行相似度估计，并对估计结果进行匹配判断。

三重损失是一种常用的度量学习方法[32]。假设存在三个候选目标来源于两个不同的行人。其中两个候选目标同属一个行人，另一个候选目标则属于另一个行人。当尚未训练网络时，同一类别中两个候选目标之间的距离可能大于不同类别中两个候选目标之间的距离。训练后，相同类别的候选目标会更近，不同类型的候选目标会更远。文献[33]，首先在输入视频帧上进行目标检测，然后通过特征提取对那些检测到的简单样本进行处理。提取的特征将分为正样本和负样本，并通过三重态损失找到最佳的特征匹配。对于困难样本，需要分别在随机样本池和困难样本池中分别选择一个样本。将这两个选定的样本与查询样本结合在一起，通过三元组损失找到最佳匹配。文献[34]，将检测到的目标发送到残差网络以获取提取的特征，利用正负样本的特征计算三重态损失获得特征损失。引入中心损失并结合先前获得的特征损失，以通过归一化获得标签损失和标签平滑化的复合信息，从而提高了标签匹配的准确性。

1.2.4 姿态估计研究现状

姿态估计也称为人体姿态估计（Pose Estimation, PE），被定义为对指定图像或视频中的人体关节（也称为关键点，包括肘，手腕等）的位置进行检测并用该信息描述姿态。同样，也可将姿态估计定义为在描述关节姿势的姿态空间中搜索特定姿势。可以按照描述关键点的状态空间维度分为两类，即 2D 人体姿势估计和 3D 人体姿势估计。2D 人体姿势估计定义是根据 RGB 图像估计每个关节的 2D 姿势（x，y）坐标。类似地，将 3D 人体姿势估计描述为将人体姿势估计定义为基于 RGB 图像估计每个关节的 3D 姿势（x，y，z）坐标[35]。

经典的人体姿态估计方法的基础思想是通过组装关节点和关键点间躯体来描述人体姿态，将一系列可形变的组件组装起来构建不同的关节模型，并使用肢体结构模型连接关节模型组成全身姿态的姿态估计。而肢体结构模型是在输入的图像中将检测到的关节点匹配组装起来的外观模板。该方法可以在一定程度上恢复样品的原始外观，但是需要很严格的计算资源支持。当通过像素位置和梯度方向对目标姿态的部分区域进行参数化时，可以基于先前结构创建的结构建立关节运动的模型。然而，经典算法具有姿势模型的局限性，即它没有充分利用原始图像数据信息，通常不能成功在人体图像上分配姿势模型。而卷积神经网络的发明极大地改变了姿态估计领域。近年来的姿势估计算法通常将卷积层作为其主要组件，在很大程度上取代了人工设计特征提取器和图形模型。此策略大大改善了算法性能[36]。

DeepPose[36]是第一篇将深度学习应用于人体姿势估计的论文。它的性能击败了之前的算法，达到了当前最高性能。该方法将姿势估计描述为基于卷积神经网络的人体关节回归问题。并且该方法级连了基于卷积神经网络的回归器来优化姿态估计并获得更好的预测结果。最重要的是，该方法能够提供较好的鲁棒性，即使某些关节点被漏检，仍然可以按照全身结构进行姿态预测。

文献[37]通过并行运行多个分辨率库中的图像以同时捕获各种规模的特征来生成热图。输出是离散的热图，而不是连续型的回归数据。热图可预测在每个像素处出现关节的可能性。Pose Machine[38]的创新点在于将图像特征提取模块和预测模块合并在一起。Convolution Pose Machine[50]则完全不同，该方法的多层体系结构可以进行端到端训练。为学习丰富的隐式空间模型提供了序列预测框架，对于人体姿势非常有效。预测当前估计值的误差并执行迭代更正。因此，该方法不是立即直接预测输出，而是使用自校正模型通过反馈误差预测逐渐更改初始的预测结果。此过程称为迭代错误反馈(Iterative Error Feedback, IEF)。文献[40]介绍了一种新颖而直观的体系结构，金字塔网络。该网络由池化层和上采样层组成，它们看起来像金字塔，彼此堆叠。沙漏型的设计是源于提取特征的尺寸不同。尽管局部信息对于识别面部、手等特征至关重要，但最终的姿势估计是需要全局背景信息的。目标的肢体方向，四肢的排列布局以及相邻关节之间的关系是出色的信息线索，可以在不同比例图像下用于识别（较小的特征图会提供高阶特征以及全局上下文信息）。文献[40]中采用的网络结构与[38]中的类似。上采样用于提高特征图的分辨率，并将卷积参数放入其他块中，但是此方法以一种非常简单的构建方式将它们组合成解卷积。 HRNet[40]模型在关键点检测，多人姿势估计和 COCO 数据集中的姿势估计比所有现有方法表现更好。HRNet 遵循一个非常简单的想法。以前的大多数论文都是从高分辨率到低分辨率，然后再回到高分辨率的代表。HRNet 在整个过程中保持高分辨率，并且运行良好。

1.3 跨摄像头多目标跟踪技术难点分析

目前，已经问世了一部分性能较为出色的多目标跟踪系统，但是这类产品大多是使用了工业应用场景下的特殊条件，不具备一般性，推广使用难度较大。特别是在条件复杂的应用环境中，跟踪性能下降较为明显。本小节针对本文关注的跨摄像头多目标跟踪任务进行分析[41]，列举算法实现难点并加以阐述：

（1）场景复杂：

真实应用场景相较于模拟实验环境要更为复杂，背景光照亮度的变化，容易影响检测器的检测能力。尤其是当背景中存在目标类似的表征信息时，很容易对检测结果造成干扰，进一步影响跟踪结果。例如，室内背景的窗帘上存在人形图

案时，很容易影响行人检测器的检测结果。

（2）目标形变：

在真实场景中，运动的目标大概率会发生形变。目标的尺度大小、形状都可能发生改变。此外，运动的目标还可能存在旋转运动，该旋转分为平面内旋转和平面外旋转，二者都会造成目标外观特征改变，特别是平面外旋转，很容易造成目标外观特征的丢失。此现象很容易造成目标表征信息匹配的失败，甚至导致目标漏检、错检，跟踪效果大大降低。

（3）遮挡问题：

目标容易被背景遮挡，甚至目标之间互相也会产生影响，造成检测器无法检测到目标。尤其是遮挡现象严重时，可能造成目标特征无法被提取，进一步导致跟踪失败。

（4）跨摄像头目标匹配难点：

同一个目标在不同传感器视野中的表征信息很可能是不同的。例如对一个行人从其身前和身后两个角度进行观测，很可能因为前后衣服颜色的不一致导致提取出的特征信息差异较大，影响同一个目标在不同传感器中的匹配。

（5）鲁棒性和实时性的均衡：

为了追求算法的鲁棒性，需要尽可能的提取特征，并对提出的丰富特征信息进行感知理解。自然而然，高鲁棒性随之带来的是计算复杂度的上升，算法消耗的时间会变长，算法的实时性下降。如何均衡算法的鲁棒性和实时性也是跨摄像头多目标跟踪系统能否成功的关键。

1.4 本文结构安排

依据项目应用真实背景，参考现有的性能优越的检测算法，多目标跟踪算法，姿态估计算法以及匹配算法，本文设计了一种能够近实时（Towards to real time）处理跨摄像头多目标跟踪任务的鲁棒算法。首先是对 YoLo 检测算法进行了改进，在低时耗的前提下提升检测效果。再对 JDE 网络结构进行修改，并用改进后的 YoLo 算法作为跟踪框架中的检测部，降低了原有的时间复杂度。本文还加入了姿态估计信息，提高了算法的目标重识别能力。本文提出的算法达到了在近实时条件下的较高性能的跨摄像头多目标跟踪效果。全文结构安排如下：

第 1 章，介绍了本文研究背景及意义，针对当前的检测算法、多目标跟踪算法、跨摄像头跟踪算法及姿态估计算法阐述了其所在领域的国内外发展现状。列举并分析阐述了跨摄像头多目标跟踪系统实现难点，并给出了全文结构安排。

第 2 章，简述了本文相关技术理论基础。首先介绍了人工神经网络和卷积神经网络原理。其次对本文参考的目标检测算法——YoLo[4]算法和多目标跟踪算法骨架 JDE 进行了简要阐述。最后对二分图问题和匈牙利算法进行相关介绍。

第 3 章：提出并分析了引入姿态估计信息的单摄像头多目标跟踪算法。首先介绍了结构简化后的 darknet 特征提取结构，并修改 YoLo 目标检测算法的结构以满足目标检测需求。其次使用新得到的目标检测结构替换原有 JDE 框架中的检测结构。然后介绍了引入了姿态信息的基于 JDE 框架的多目标跟踪框架。最后在 JTA 和 MOT 数据集上对提出的算法进行验证并分析实验结果。

第 4 章：提出了一种集中式的跨摄像头多目标跟踪系统。将第 3 章提出的引入姿态估计信息的单摄像头多目标跟踪算法中的检测器部署到分布式传感器网络中的各个节点，各节点完成目标检测后，将检测结果（含有 embedding 信息和姿态描述信息的复合检测头）发送到数据融合中心节点，对各节点得到的检测头进行数据融合得到精确的全局目标观测状态描述模型，该模型用于运动模型中完成跟踪任务。算法在 JTA 数据集上进行模拟测试并分析了实验结果。

第 5 章：总结与展望。阐述了本文完成的工作内容，同时指出了第三章提出的算法和第四章提出的框架需改进的缺点，并对未来的可行的研究方向进行总结概括。

第 2 章 相关工作技术基础

2.1 引言

本章节主要讨论跨摄像头多目标跟踪任务的相关理论和技术基础。由于人工神经网络是人工智能的理论基础，也是本系统的核心技术，本章 2.2 节对人工神经网络、卷积神经网络和 YoLo 目标检测网络做了简要介绍。2.3 节介绍 JDE 多目标跟踪框架，该框架有别于传统的多目标跟踪算法，抛弃了多阶段跟踪模式，着重于目标的描述模型提出了单阶段跟踪框架，提供了一种解决多目标跟踪任务的新思路。2.4 节介绍了二分图匹配问题的定义，该问题能对目标重识别问题建模，解决该问题的经典算法——KM 算法也在此小节阐述。

2.2 基于人工神经网络的目标检测算法概述

2.2.1 人工神经网络

人工神经网络是深度学习的基础技术，是基于神经科学、心理学和计算机科学等学科的交叉产物，也是一种模拟生物感知系统的计算模型。1943 年，Warren McCulloh 和 Walter Pitts 模拟神经工作的方式，首次提出了神经网络模型 。1957 年康内尔大学教授 Frank Rosenblatt 提出了“感知器”（perceptron）模型，第一次用算法精确化定义了神经网络，是第一个具有自适应学习能力的数学模型。感知器模型出现的最大意义是使抽象的概念模型实例化。日后的众多神经网络模型的理论基础都源于感知器模型。神经网络结构通常如图 2-1 所示。

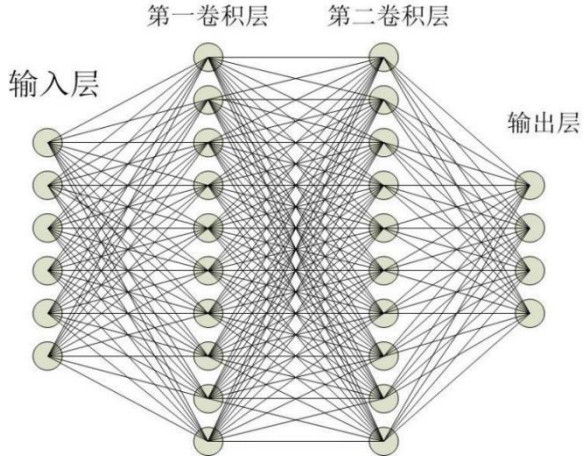

图 2-1 神经网络结构简化示意图

一般常见的神经网络都是多层网络，每层都由多个神经元组成，同层之间的神经元不存在连接，相互独立。不同层级间的神经元存在连接关系，每一条连接都有相对应权重，神经元间的权重和阈值在训练过程中进行调整。神经网络的每层输入都源于前一层或前几层的输出，通过将期望和输出之间的误差控制在规定范围内，使得神经网络本身能自适应调整权重和阈值以获取网络对某一种模式的表述，该过程又被称之为学习过程。传统神经网络有向前传播和反向传播两个步骤，前者获取输出，后者通过输出和期望间的误差反向修正网络模型中的参数。

2.2.2 卷积神经网络

卷积神经网络（Convolutional Neural Networ，CNN）是深度前馈网络模型的一种。通过引入卷积运算，减少了传统深层人工神经网络计算时占用的内存量。1998 年，LeCun 等人率先提出了第一个 CNN 网络结构 LeNet[8]，该网络结构模型如图 2-2 所示。与传统神经网络类似，LeNet 神经网络也包含数层结构，不同的是，LeNet 的每层不是简单的隐含层结构，而是多种网络结构组成，通常含有卷积层、池化层、全连接层等结构。

（1）输入层是将输入的数据以信号的形式传入网络的接口，与传统神经网络一致。

（2）卷积层中包含有卷积核，完成 CNN 的核心运算——卷积运算。其目的是通过卷积运算提取输 入信号中的特征信息。通常输入信号是离散信号（诸如图像信息），卷积运算被简化定义为层输入图像与卷积核间的点乘运算。

（3）池化层是通常连接在卷积层之后的结构层，其目的是减小特征图的尺寸大小，降低网络的运算成本也提高了网络的模型抽象能力。通常池化方法选择有最大池化、平均池化、中值池化等。

（4）全连接层，其目的是将特征图输出的多种特征映射到一个特征向量。在早期的 CNN 中被广泛使用，目前通常采用小尺寸的卷积核替代，如 1×1 大小。

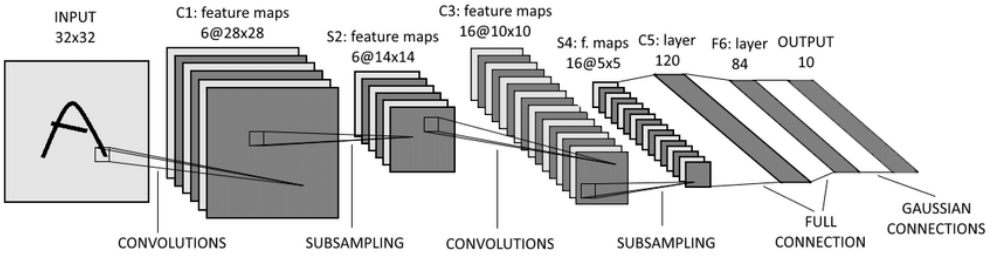

图 2-2 神经网络结构简化示意图

2.2.3 YoLo 目标检测网络

目标检测任务在 2012 年的 AlexNet[9]问世之后得到了极大的发展。RCNN 系列的目标检测网络证明了深度学习可以适应高精度的目标检测任务。在此基础之上，展开了对检测网络实时性的研究。基于传统的 two-stage 类型的目标检测网络，one-shot 类型的目标检测网络被提出，且在实时检测任务中得到了较好的应用。Yolo[18]目标检测网络是 one-shot 类型的目标检测网络中的佼佼者，它汲取了 VGG[11]，GoogLeNet[10]等优点，提出了 darknet 特征提取结构，结构如图 2-3 所示。该结构具有较强的特征提取能力，针对大型和中型目标，darknet 结构可以出色的完成特征提取任务，同时也能对小尺寸目标提供较好的特征提取支持。同时，通过滑动搜索窗（也称先验窗）替代了 two-stage 网络中的 ROI 生成过程，提高了算法效率。目前，Yolo 系列的目标检测网络已经发展到了第五代，其中前三代都由原作者提出[43][44][45]，后两代则是工业界修改后提出的。

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

图 2-3 darknet 特征提取结构

YoLo V3[43]论文中所讨论的预测边界框如图 2-4 所示。 预测边界框的格式是$[b_x, b_y, b_w, b_h]$，用质心坐标偏移量度量描述预测边界框的位置和尺寸。

$$b_x = \sigma(t_x) + C_x \tag{2-1}$$

$$b_y = \sigma(t_y) + C_y \tag{2-2}$$

t_x和t_y表示质心到质心所在的网格左上角的偏移量。σ 是激活函数，其目的

是实现比例转换。C_x和C_y代表到预测中心点所在的最小网格的左上顶点距全局图像左上顶点的偏移量。

$$b_w = P_w e^{t_w} \tag{2-3}$$

$$b_h = P_h e^{t_h} \tag{2-4}$$

P_w和P_h分别是搜索窗的宽度和高度，正因如此，检测框的尺寸比例仅与搜索窗的比例有关，与其坐标无关。

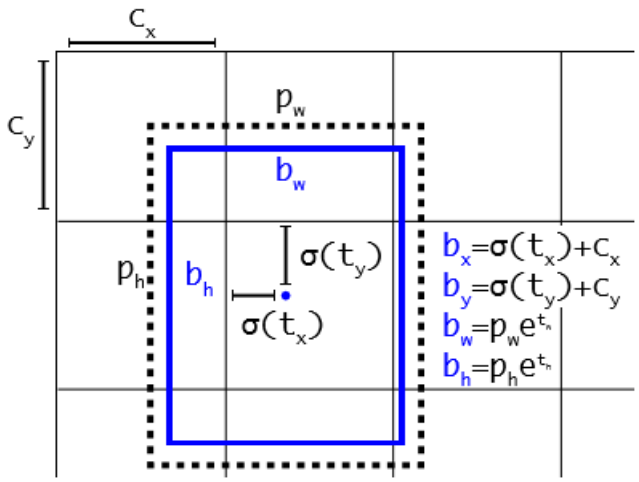

图 2-4 YoLo 算法的检测输出格式

由于检测过程中，搜索窗在整个输入图像中滑动，因此存在部分 bounding box 中不含有目标，于是检测产生的样本有三类：正样本（positive）、忽略样本（ignore）和负样本（negative）。从输入的图像的标注中读取真实框标注 Ground Truth，并使用所有的 bounding box 与其计算 IoU。IoU 值最大的 bounding box 则是正样本，正样本能且只能与一个 Ground Truth 匹配。得到的正样本会产生置信度损失、预测框损失和类别损失。类别标签置为 true，置信度标签置为 1（代表 bounding box 中存在目标）。

忽略样本则是与 Ground Turth 间的 IOU 大于阈值（本文中使用 0.5）的样本，但是其 IoU 值不是最大值（即不是正样本），此类不会产生任何损失。

负样本与 Ground Truth 的 IOU 小于阈值（0.5），负样本只有置信度会产生损失，并且置信度标签被设置为 0。

Yolo-V3 的结构示意图如图 2-5 所示。相比之前的检测算法，Yolov3 算法融合了不同层级的特征，模型得到了更强的图像理解能力。

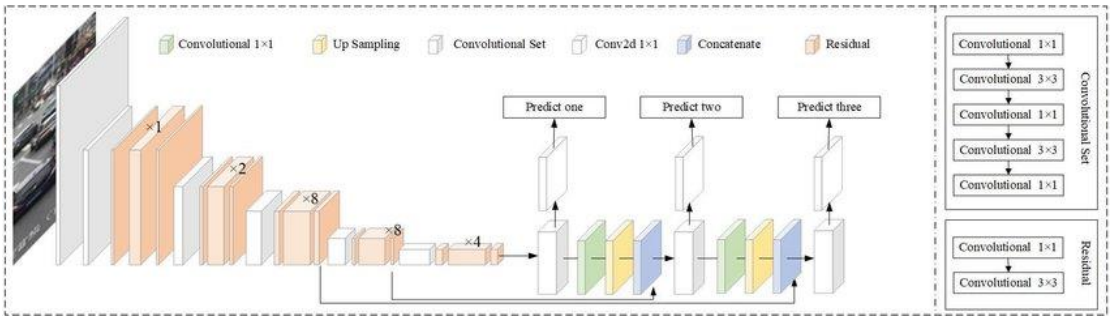

图 2-5 YoLo V3 结构简图

YoLo 检测算法使用了 predict head 作为输出并使用了多层级预测输出的思想，其训练策略也有所不同，Yolo V3 的损失函数定义如下：

$$
\begin{aligned}
loss_{N_1} &= \lambda_{box} \sum_{i=0}^{N_1 \times N_1} \sum_{j=0}^{3} 1_{ij}^{obj} [(t_w - t_w^{'})^2 + (t_h + t_h^{'})^2] \\
&\quad - \lambda_{obj} \sum_{i=0}^{N \times N} \sum_{j=0}^{3} 1_{ij}^{obj} \log(c_{ij}) \\
&\quad - \lambda_{noobj} \sum_{i=0}^{N_1 \times N_1} \sum_{j=0}^{3} 1_{ij}^{noobj} \log(1 - c_{ij}) \\
&\quad - \lambda_{class} \sum_{i=0}^{N_1 \times N1} \sum_{j=0}^{3} 1_{ij}^{obj} \sum_{c \in classes} [p_{ij}^{'}(c) \log(p_{ij}(c)) + (1 - p_{ij}^{'}(c) \log(1 - p_{ij}(c))]
\end{aligned}
\tag{2-5}
$$

其中，λ为权重参数，用于控制检测框损失，置信度损失，以及类别损失。

对于正样本而言，三种损失都存在；对于负样本而言，只有置信度产生损失；忽略样本则不产生损失；类别损失采用交叉熵作为损失函数。

2.3 JDE 多目标跟踪框架

JDE（Jointly Detector and Embedding Model Tracker）多目标跟踪框架是一种被用于多目标跟踪任务的新型跟踪器[46]。目前被广泛使用的多目标跟踪器往往都是多阶段模型。第一阶段是检测阶段，在此阶段中，视野里的目标将被检测出并定位；第二阶段则是数据融合阶段，检测到的目标将被分配正确的 ID 标记并与轨迹匹配。这意味着，基于多阶段模型的多目标跟踪算法都至少需要两个消耗大量计算资源的组件：检测器和目标匹配器（Re-ID）。此类型算法也通常被称为检测融合分离式（separate detection and embedding，SDE）算法。如前文所讨论的那样，传统的分段式的多目标跟踪算法不仅时间消耗很大，并且跟踪效果极度依赖于前一阶段的检测结果。一旦检测结果出现较大偏差，最终的跟踪结果也会收到极大影响。JDE 多目标跟踪框架也遵循了 SDE 多目标跟踪框架部分思想，

在此基础上去除部分原有结构（数据融合部），使用了合并检测和数据融合的单一阶段结构替代多段式结构，JDE 多目标跟踪框架的结构示意图如图 2-6 所示。

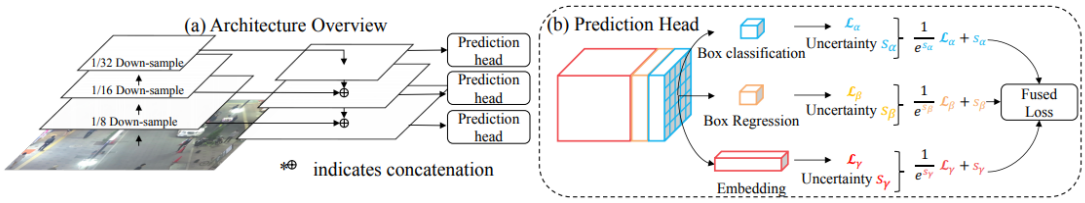

图 2-6 JDE 多目标跟踪框架结构简图

JDE 多目标跟踪框架的主要思想是让两个不同的任务（检测任务和数据融合任务）可以分享低级别特征信息，避免重复计算。JDE 多目标跟踪框架继承了 Yolo 目标检测算法的 embedding 思想，提出了输出结构（head）的优化方法，同时也采用了不同层级的特征融合的思想，提高了算法精度。输出结构包含了三个部分：目标类别（box classification）、目标位置（box regression）和 Embedding。每一个部分都有特殊目的，三个分支的输出被送入融合损失函数（fused loss）得到具体的损失函数。不同层级的 prediction head 对应着不同尺寸，不同等级的图像语义理解。JDE 多目标跟踪框架与 SDE 多目标跟踪框架最大的不同就是每一个的 prediction head 都包含有 embedding 信息，而 embedding 信息在本任务中恰恰就是 track ID 也就是目标的 ID 信息，因此在 prediction head 输出之后，Re-ID 操作不再是必须的。

JDE 多目标跟踪框架的训练方式也与 SDE 多目标跟踪框架不同。JDE 的检测部分继续延续了 faster-rcnn 的两个检测损失函数$\mathbf{L_a}$和$\mathbf{L_b}$，分别是目标/背景分类损失函数和 bounding box 回归损失函数，前者采用了交叉熵损失函数形式，后者则是 L1 损失函数形式。JDE 多目标跟踪框架的最终目标是为了训练得到 prediction head，尤其是得到含有 embedding 信息的 bounding box。因此，三重损失函数被应用于训练 JDE 框架以获取 embedding 信息。

$$L_{triplet} = \sum_i max(0, f^T f_i^- - f^T f^+) \tag{2-6}$$

f^T 是一个 mini-batch 中被选择出的实例，f^+ 是正样本，f^- 则是负样本。然而，训练三重损失函数时存在不稳定，不能收敛的情况，而且收敛速度会很慢。因此，为了稳定训练过程并提高收敛速度，使用了凸优化的思想，在一个光滑的上限上优化该损失函数：

$$L_{upper} = \log(1 + \sum_i \exp(f^T f_i^- - f^T f^+)) \tag{2-7}$$

$$L_{upper} = -\log \frac{\exp(f^T f^+)}{\exp(f^T f^+) + \sum_i \exp(f^T f_i^-)} \tag{2-8}$$

因此，三重损失函数可以被改写成交叉熵形式：

$$L_{CE} = -\log \frac{\exp(f^T g^+)}{\exp(f^T g^+) + \sum_i \exp(f^T g_i^-)} \tag{2-9}$$

综上所述，JDE 多目标跟踪框架的总损失函数是三个不同任务的损失函数之和，其形式为：

$$L_{total} = \sum_i^M \sum_{j=\alpha,\beta,\gamma} \omega_j^i L_j^i \tag{2-10}$$

2.4 KM 算法概述

Re-ID 任务可以被定义为一个二分图匹配问题：通过找出前一时刻与当前时刻检测到的目标进行最优匹配。KM 算法是用于求解二分图最优匹配任务的一个解决方案，该方法本质是一种带权值的匈牙利算法。本小节，将给出二分图匹配问题定义，其次对 KM 算法展开阐述。

2.4.1 二分图匹配问题定义

二分图，是一种特殊的图结构，也是图论中的一种特殊模型。假设图 G=<V，E>是一个无向图，顶点 V 可以被分成两个独立的子集 A，B，二者不相交，且图结构中的每条边（i，j）关联的两个顶点分别属于两个不同的顶点集 A，B，则无向图 G 是二分图。二分图示例如图 2-7 所示。

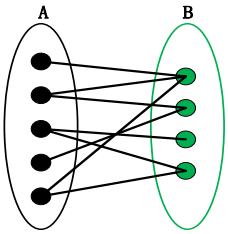

图 2-7 二分图示例

二分图的匹配问题可以被描述为：保证当前帧的某一个目标有且只有前一帧中存在的相同目标与之匹配（如图 2-8 红色边所示）。二分图的匹配可以看成是二分图的一个子图，该子图满足条件：在二分图 G 的子图 g 中，不存在任意的两条边依附于同一个顶点。

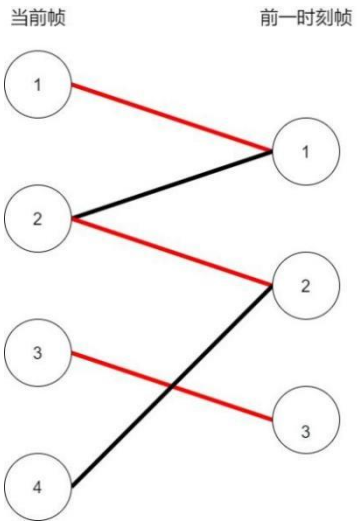

图 2-8 二分图匹配示例

常见二分图匹配方法有两种，第一个方法是最大匹配，即尽可能多地将 A 和 B 中的点进行配对；第二个方法是最佳匹配，该方法通常在在带权二分图中被使用，最佳匹配就是 A（B）中所有的点都与 B 中（A）的某一个顶点完成匹配，并且顶点间的连接（边）的权值之和最大。可证，最大匹配的子图可能不唯一，而最佳匹配的子图因为不同边可以存在相同权值，也可能不唯一。Re-ID 任务通常被定义为二分图最优匹配问题。

2.4.2 匈牙利算法

匈牙利算法[47]是寻找二分图最大匹配的一个好方法，也是 KM 算法的理论基础。匈牙利算法的基础思想是寻找当前二分图的增广路径。增广路径的定义如下：

设 M 为二分图 G 已完成匹配边的集合，若 P 是 G 中一条连通两个未匹配顶点的路径（P 的起点在顶点集 A 中，终点在顶点集 B 中，反之亦可），且已匹配和待匹配的边在 P 上交替出现，则称 P 为相对于 M 的一条增广路径。

1965 年匈牙利数学家 Edmonds 提出用增广路求最大匹配，即匈牙利算法：

（1）初始化以匹配边的集合 M 为空。

（2）找出一条增广路径 P，取反操作获得更大的匹配 M’代替 M。

重复步骤（2）直到找出所有增广路径。

寻找增广路径的算法通常使用的是 DFS（Deep First Search，深度优先搜索）：从顶点集 A 一个未匹配的顶点 i 开始，找一个未访问的邻接点 j（j 是顶点集 B 中的顶点）。顶点 i 存在两种可能的情况：

（1）如果顶点 j 未匹配，则已经找到一条增广路径。

（2） 如果顶点 j 已经匹配，则取出 j 的匹配顶点 w(w 是顶点集 A 中的顶点)，边(w, j)是当前已匹配的，进行取反操作将(w, j)更改为未匹配状态，(u, j) 设为匹配状态，则（ u， j ）就是一条以 u 为起点的增广路径。

2.4.3 KM 算法

KM 算法[48]是为了解决二分图的最佳匹配问题，前述的匈牙利算法是一种贪心算法，而 KM 算法是贪心算法的扩展。

首先定义顶点可匹配条件。对于原图 G 中的任意一个结点，给定一个函数 L(node) ，求出顶点的值。数组l $_A$(a)记录顶点集 A 中顶点 a 的值，数组l $_B$(b)记录顶点集 B 中顶点 b 的值。且对于原图 G 中任意一条边e d g e (x , y)都满足 $lx(x)+ly(y)>=weight(x,y)$ 。

同样，定义相等子图。相等子图是原图 G 的一个生成子图，包含所有顶点但不包含全部可能的边，生成子图满足 $lx(x)+ly(y)=weight(x,y)$ 的边为可行边。

因此，当条件满足时，匹配边的权重之和 K 达到最大时，该匹配称为最佳匹配：

$$\sum_{i=1}^{x_i\in X} lx(x_i)+\sum_{i=1}^{y_i\in Y} ly(y_i)=K>=\sum weight(x_i,y_i) \tag{2-11}$$

KM 算法流程可简化如下：

（1）初始化可行顶点值 (设定 lx,ly 的初始值)。

（2）用匈牙利算法寻找相等子图的完备匹配。

（3）若未找到增广路径则修改可行顶标的值。

重复步骤(2)(3)直到找到相等子图的最佳匹配为止。

2.5 本章小结

本章首先介绍了机器智能化的理论基础——人工神经网络，同时也介绍了卷积神经网络理论。其次，介绍了 YoLo 目标检测算法，该算法将作为本文研究内容的检测部提供基础框架；接下来介绍了 JDE 多目标跟踪算法框架，该框架是本文多目标跟踪算法的基础；最后介绍了 KM 算法，提供了 Re-ID 的解决方案。本章介绍的内容为第三章和第四章提出的算法与框架提供了理论和思想基础。

第 3 章 引入姿态预测信息的 JDE 多目标跟踪框架

3.1 引言

当下绝大多数多目标跟踪框架都采用了多阶段的跟踪方式，跟踪算法的效果和效率很大程度上取决于第一阶段——检测阶段的性能。本文采取了全新的建模方式，用更精确的检测算法替代原有检测方式，通过引入 embedding 信息避免了多阶段跟踪模式。在 3.2 节中首先介绍了改进的 Yolo 算法，作为 JDE 多目标跟踪框架的检测部，同时在检测部中还加入了使用人体姿态关键点描述的姿态信息。在 3.3 节中，介绍改进的 JDE 多目标跟踪框架。3.4 节介绍了框架开发环境和实验环境，以及训练方法，并与当前的部分多目标跟踪框架进行了比对。

3.2 改进的 Yolo 检测算法

目标检测是本文中设计的 MOT 和姿势描述的基础。本节主要介绍如何创建行人探测器以及人体关节关键点的探测。如前文中提到的，目标检测算法有两个的分支，单阶段检测和多阶段检测。由于单阶段检测的时间成本低于多阶段检测，因此本文选择了单阶段检测。分析被广泛使用的单阶段检测算法 YoLo，其主要步骤可以概括如下：首先，创建不同的比例搜索窗；其次，在特征图上滑动搜索窗并提取特征进行视觉感知；最后，回归类概率和边界框信息。传统的 Yolo 算法尽管速度较快，但依然有海量的参数需要参与运算，在计算资源受限的系统边缘设备上难以部署，依照 YoLo 的思想，本文改进了传统的 YoLo 算法用于行人检测，改进的思路是通过移除不同输出层级分支和简化特征提取结构中的特殊结构，减少了多尺度检测的尺度数量，同时减少了参与计算的参数量，并且可以和人体关节关键点检测复用网络前部的数层。下面详细讨论对 YoLo 算法的改进。

3.2.1 创建搜索窗

检测问题的关键的实质是对感兴趣区域（ROI）进行分类。对于单阶段检测算法，滑动视图的目的是为了查找 ROI 内的目标并提取特征。一个合适的比例搜索窗可以同时扩大 IoU 和提升分类正确概率。适当比例的搜索窗将更好匹配 Ground Truth 的边界，因此合适比例的搜索可以使模型提取特征时减少提取冗余特征的可能。因此要获得一个合格的行人探测器，首先创建合适比例的搜索窗是十分必要的。

与多标签检测任务相比，本文着重于行人检测，简化多目标检测任务为单类检测任务，因此创建搜索窗的难度有所下降，不用考虑不同类别的不同比例搜索窗，只需对人体身形的比例构建搜索窗的适当比例。考虑到人体身形的特点，搜索窗的长宽比不应为 1，搜索窗的形状应为高度大于宽度的矩形。要确定合适的搜素窗长宽比，需要针对已经标注的数据采用聚类算法提取。本文应用 K-Means 聚类算法针对男性、女性和未成年三个不同类别进行计算，获取最可能的三个长宽比值。获得的长宽比分别为 0.37（男性），0.35（女性）和 0.40（儿童）。长宽比值可能会因数据集的大小而有所差异，当数据集包含更多样本时，确定出长宽比值的准确性会更准确。再根据不同的长宽比值，针对不同尺寸的行人检测创建不同的搜索窗，并且可以支持小目标检测。

3.2.2 检测部骨架

与其他视觉任务一样，检测任务也基于特征提取。在 AlexNet[9]发明之后，有许多流行的特征提取框架被提出，例如 ResNet[12]，VGG[11]和 Inception[10]。作为最先进的骨干网之一，Darknet 被选为特征提取网络体系结构。在 Darknet 系列，最著名的模型是 Darknet-53，它结合了 ResNet[12]和 VGG[11]。

Darknet-53 的主要结构是卷积结构和残差结构。卷积结构用于特征提取，残差结构使骨干更深。在该结构中没有专门设置池化层或全连接层。因此，需要一种不同的方法来改变张量大小。原作者通过修改内核的步幅代替，实现张量大小转换。用 Darknet 进行特征提取时，网络结构不包括尾部的 3 层，这 3 层用于完成分类工作，无需进行特征提取。Darknet-53 的卷积结构层如图 3-1 所示，相较于 VGG[11]和 ResNet[12]中的传统卷积结构，池化层结构并没有被特意设置，取而代之的是添加了 Batch Normalization（BN）层和新的激活函数 Leaky ReLU。BN 层可以加快训练速度并允许使用较大的学习率值，并且还可以一定程度抵抗过度拟合。新的激活函数 Leaky ReLU 具有比传统激活函数 ReLU 更多的优势。ReLU 函数是一种取最大值的函数，公式为$\mathrm{ReLU} = \max(0，\mathrm{x})$。尽管此函数可以实现更高的速度，但是较为容易失效。例如，当较大的梯度值传递到 ReLU 神经元时，在更新参数之后，神经元将失去激活能力，且梯度始终保持为 0，也正是因此神经网络失去有效性，检测模型不能实现期望功能。要解决这一困境需要新的激活函数替代原有激活函数。Leaky ReLU 函数的发明可以解决 Dead ReLU 问题。ReLU 将所有负值设置为零。 而 Leaky ReLU 将所有负值指定非零斜率，可以将其表示为：

$$Leaky\ ReLU = \begin{cases} x_i, & x > 0 \\ \frac{x_i}{a_i}, & x \le 0 \end{cases}, a_i \in (1, \infty) \quad (3\text{-}1)$$

其中，ai 是一个固定值。

另外，卷积结构也有一个标志位。如果标志位表明是下采样类型，则卷积结构将输出比输入尺度小的结果。

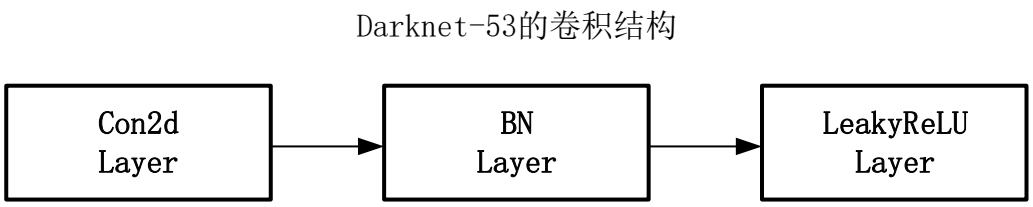

图 3-1 Darknet-53 的卷积结构

另一个关键结构，则是残差结构，其目的是使主干网络深度加深，该结构引入 Darknet 结构后可以使其加深至 53 层。残留结构的最大特点是使用 short cut 机制（类似于电路中的短路机制）来缓解由于神经网络深度增加而导致的梯度消失的问题，使神经网络更易于优化。它建立了一个直接的相关通道通过身份映射的方法输入和输出，使网络专注于学习输入和输出之间的残差。

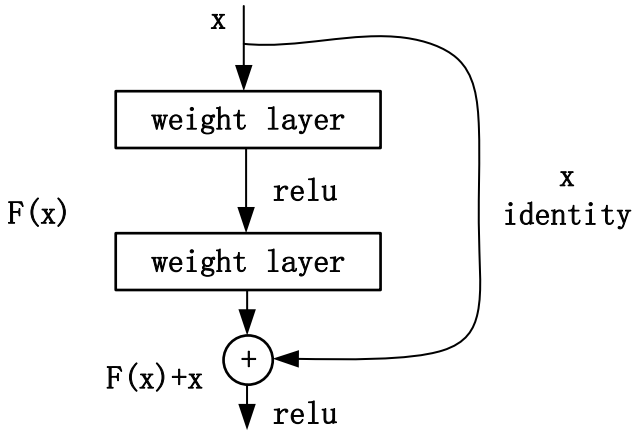

图 3-2 残差结构

然而 Darknet-53 中的参数过多，要训练该特征提取结构需要消耗大量的计算资源，这一因素可能会损害算法的效率，特别是在拥有较小 GPU 资源甚至没有 GPU 资源的机器条件下（例如边缘设备），这一因素会极大危害到算法效率。缩减检测模型规模，降低计算参数量势在必行。

考虑到本文的需求是单类对象检测任务，因此不需要在检测上浪费过多的计算资源，因为对前景和背景进行分类不需要太多的功能，冗余信息也可能影响以

后的匹配任务。显然在计算资源较少的条件下，Darknet-53 不是本文的最佳选择。但是，可以通过移除一些 darknet 结构中的模块，以达到减少骨干网中的参数的目的。本文尝试在 Darknet-19 和 Darknet-53 之间找到一个骨干。下面阐述本文所作的一些改进。

原有的 Yolo V3[43]算法是支持多尺度检测的检测算法，有三个不同尺度层级的检测输出，考虑到在本任务中，尺寸过小的目标不但容易引发误检，对目标的运动状态估计造成较大影响，另外小尺寸目标一般是距离摄像头传感器较远的目标，其特征细节很难提取，对基于目标匹配的数据融合也造成了重大影响。因此，放弃对小目标（远离摄像头传感器的目标）的跟踪是合情合理的。因而原算法中的第三层级输出分支被移除，该分支在原任务中主要负责对小目标和输入图像的细节特征提取，在本任务中该分支存在的意义不大。在此基础之上，网络输入的尺寸被扩大为 768×768。

放弃了第三层级尺度输出，不代表多尺度检测思想被否决。在本任务中，在传感器视野里存在多个行人目标，并且由于摄像机和目标之间的距离不同，因此在传感器视野里的行人目标的尺寸大小各不相同。为了解决该问题，利用金字塔特征图的思想，使用大尺寸的感受野来检测大尺寸的物体，而小尺寸的感受野则用于检测小尺寸的物体。

本文只使用两个不同尺度层级的检测分支用于检测稍小型目标和正常尺寸的行人。特征图的输出尺寸为$N \times N \times [2 \times （4 + 1 + 2）]$，N×N 是输出特征图的网格点数；如前所述，总共有 2 个 Anchor 框，每个框都有一个 4 维的预测框值$[tx，ty，tw，th]$，它分别表示目标质心坐标，宽度和高度；一维预测框置信度；二维的对象类别编号（行人或背景）。本文所改进网络结构见表 3-1。

表 3-1 网络结构改进表

残差块数目	种类	卷积核	尺寸	输入	输出
	Conv	32	5	768	256
	Conv	32	3	256	256
	Conv	64	3	128	128
	Conv	32	1		
	Conv	64	3		
1×	Res			128	128

	Pool			128	64
	Conv	64	1		
	Conv	128	3		
2×	Res			64	64
	Pool			64	32
	Conv	256	1		
	Conv	512	3		
8×	Res			32	16

3.3 引入目标姿态描述的 JDE 跟踪框架

MOT 的常规步骤如下：1）检测步骤，将单个视频帧中的目标定位。2）关联步骤，对检测到的目标区域进行签名并连接到现有轨道。这意味着该系统至少需要两个计算密集型组件：一个检测器和一个 Embedding 模型。传统的 MOT 算法通常是多阶段方法，不仅时间成本较高，而且性能基于前一阶段的结果，这意味着前一阶段的任何错误结果都会损害最终输出。下面详细论述本文改进的跟踪框架（参见图 3-3）。

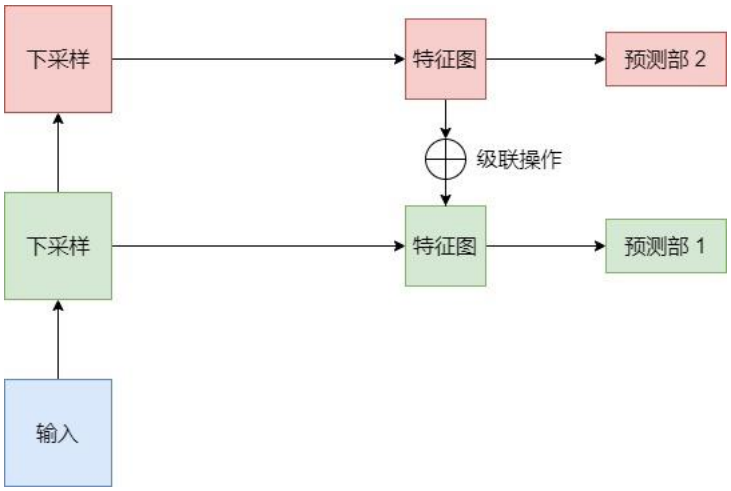

图 3-3 改进后的跟踪框架

参考 JDE 跟踪器，我们借鉴了其想法，将检测阶段和 Re-ID 组装到单一网络中。因此，这两个任务可以共享同一组低级功能并避免重新计算。但是，JDE

跟踪器的体系结构过于复杂，需要大量的计算资源支持。本文遵循 JDE 跟踪器的思路，与前一小节类似，放弃了细粒度检测分支，只保留了两个不同尺度的输出，保证了算法的效率，以实现低时延和低功耗。改进后的跟踪框架如图 3-3 所示，此框架可以回归目标的 ID 和位置，输出为预测部。

3.3.1 预测部(Prediction Head)

JDE 多目标跟踪框架也汲取了 Yolo 检测算法的思想，也提出了多尺度目标的跟踪输出。JDE 多目标跟踪框架的输出是名为预测部的结构。本文提出的跟踪框架继承了 JDE 跟踪器的思路，在输出的预测信息中包含位置信息（边界框）以及将外观信息嵌入其中。除此之外，本文将姿态描述信息引入跟踪框架，因而预测部结构亦发生了变化，新加入了姿态描述部。预测部如图 3-4 所示，包含 4 个标志位，分别为 Box 分类 、Box 回归、Embedding 和姿态描述。预测部是一个输出预测图，其大小为 $(6A+D+12P)\times H\times W$ 。其中，A是分配给比例的锚点模板的数量，2A用于框分类（前景和背景），而4A用于框回归（边界框信息）； D 确定 embedding 的尺寸（通常是 tracker ID）而 12P 是指姿态描述信息，其有关细节将在后文详细叙述。

图 3-4 预测部的标志位

3.3.2 人体关键点检测

传统的姿势估计方法易于可视化且十分直观，但是很难压缩成较小长度的编码，用于协助检测或者跟踪任务时包含了过多的冗余信息。相较于传统的人体姿态信息估计，本文提出了一种新的获取目标姿态信息的方法并命名为姿势描述。本文研究一种描述人体姿势信息的新方法，该方法可实现较小的存储成本，并易于量化描述姿态信息。

与其他姿势估计方法相同，本文采用的解决方案也基于关键点检测。OpenPose[49]关键点检测模块在该本文中用于完成人关键关节的检测。本文采用了 3.2.2 节讨论的检测部作为预训练模型，使 OpenPose 复用了检测部前面数层。

通常，关键点检测的输出是一个热度图 heat map（例如 OpenPose），它可以转换为包含此 heat map 的框。另外，OpenPose 作为一种自下而上的姿势估计方

法，如果可以检测到所有关键关节，则会在整个图形中输出所有关键关节。第一步是对检测到的关键关节进行归类分组。根据 bounding box 信息（人员检测器结果），如果关节位于 bounding box 中，则将该关节归类进对应的 bounding box。图 3-5 中对此进行了描述。如果返回了 4 个行人边界框，则将这些关键关节分组到同一框中。

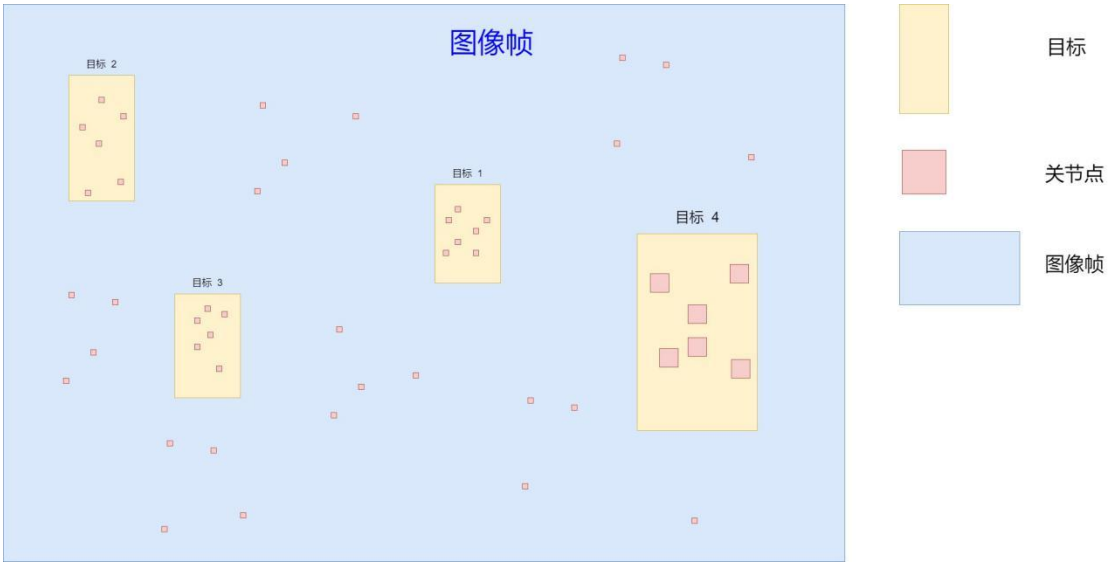

图 3-5 姿态描述信息示意简图

第二步是描述每个人的姿势，换句话说，是描述每个行人预测框中关键关节的布局。传统的方式是根据人的身体连接规则来连接每个关键关节，例如，手腕到肘部到肩膀来描述一只手臂。但是，这将花费太多时间和计算资源来完成对关键关节的贪婪搜索。本章方法放弃了由新的关节位置描述代替的传统思想。为便于叙述，此处假设只有 6 个关键关节用于姿势描述，而在实际应用中，将会有 12 个不同类别的关节点被用于姿态描述，如表 3-2 所示。

表 3-2 用于姿态描述的关节列表

右肩	右肘	右手腕	左肩	左肘	左手腕
右髋	右膝	右脚踝	左髋	左膝	左脚踝

计算每个关节框质心到人预测框质心的偏移量，如图 3-6 所示。红点是人边界框的质心；粉红色的盒子是人形盒子；黑线表示每个关节的偏移量。偏移量可以很容易地定义为$(\Delta x_i, \Delta y_i)$，i表示联合索引或类型。偏移距离可以通过 L2 距离来计算：

$$dist_i = \Delta x_i^{\,2} + \Delta y_i^{\,2} \qquad (3\text{-}2)$$

确定距离后，可通过以下公式将距离结果转换为比例：

$$d\_ratio_i = \frac{dist_i}{(\frac{w}{2})^2 + (\frac{h}{2})^2} \tag{3-3}$$

根据关键点检测返回的关节类型 id 排列这些关节距离偏移值，姿势描述组可以定义为：$\mathrm{pose_i} = (\mathrm{dist_1}, \mathrm{dist_2}, \mathrm{dist_3}, \mathrm{dist_4}, \mathrm{dist_5}, \mathrm{dist_6})$，i是人员预测框的 ID。但是，可能存在一种情况，即检测盒中有多个相同类型的关节。具有较小热图的关节将被较大的关节更新，即每一个类型的关节点在各 bounding box 中只出现一次（取 heat map 最大），每一个 bounding box 中存在的关节类型表如表 3-2 所示。如果发生更新，已确认的联接将从检测列表中删除；被其他关键点更新的那些关节将返回到检测列表。此操作可以通过非常短的长度代码将接头的布局投影到高维空间中。组中的关键关节越多，空间的尺寸将越大，这可以支持划分不同的姿势目标。同样，当盒子中的关节很少时，那些未检测到的关节的距离元素的位置为距离 0，当然，很少的关节会降低姿势描述的可信度。

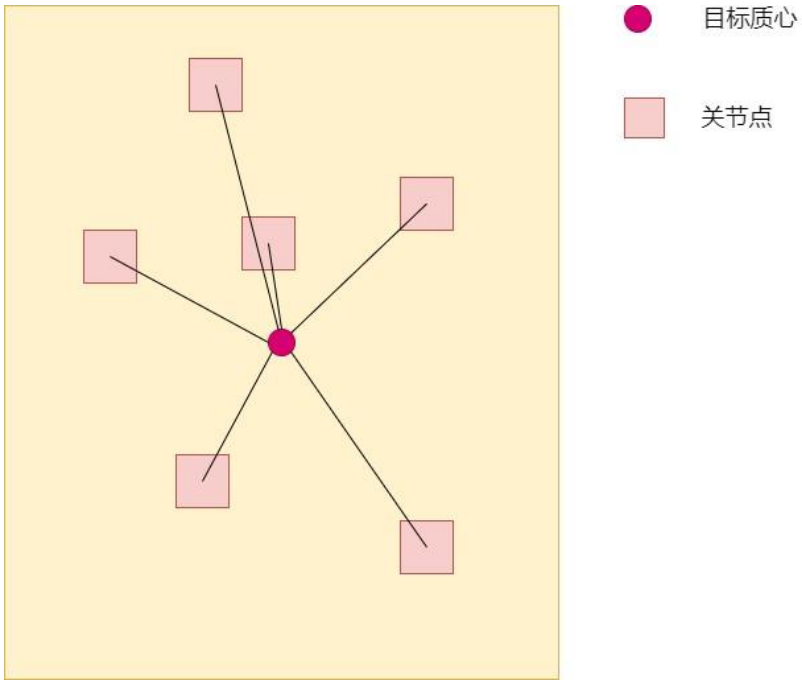

图 3-6 单目标的姿态描述信息计算示意简图

3.3.3 跟踪部

JDE 框架实现了检测和 Re-ID 两个不同任务的相结合，但是检测部并没有包含对下一个时刻目标运动状态的预测，还没有完全满足跟踪的需求。要完成跟踪的需求，尤其是满足 MOT 任务的跟踪需求，Kalman 滤波和二分图匹配操作需要

连接在 JDE 输出结果之后，对下一帧的目标状态进行预测。并且，原 JDE 框架采用了每步 Re-ID 操作保证目标匹配，并未完全脱离多阶段多目标跟踪模型的思想，同时每步 Re-ID 增加了过多耗时，因此引入了检查机制，尽可能减少 Re-ID 次数。跟踪部框架如图所示，首先算法会判断是否有新生目标出现，若无新目标出现则 tracker ID 是源于 JDE 输出结果的，则无需使用 KM 算法进行二分图匹配操作。若新出现了新生的目标，则需进行 Re-ID 操作完成 tracker ID 的新增与更新。跟踪框架流程图如图 3-7 所示。

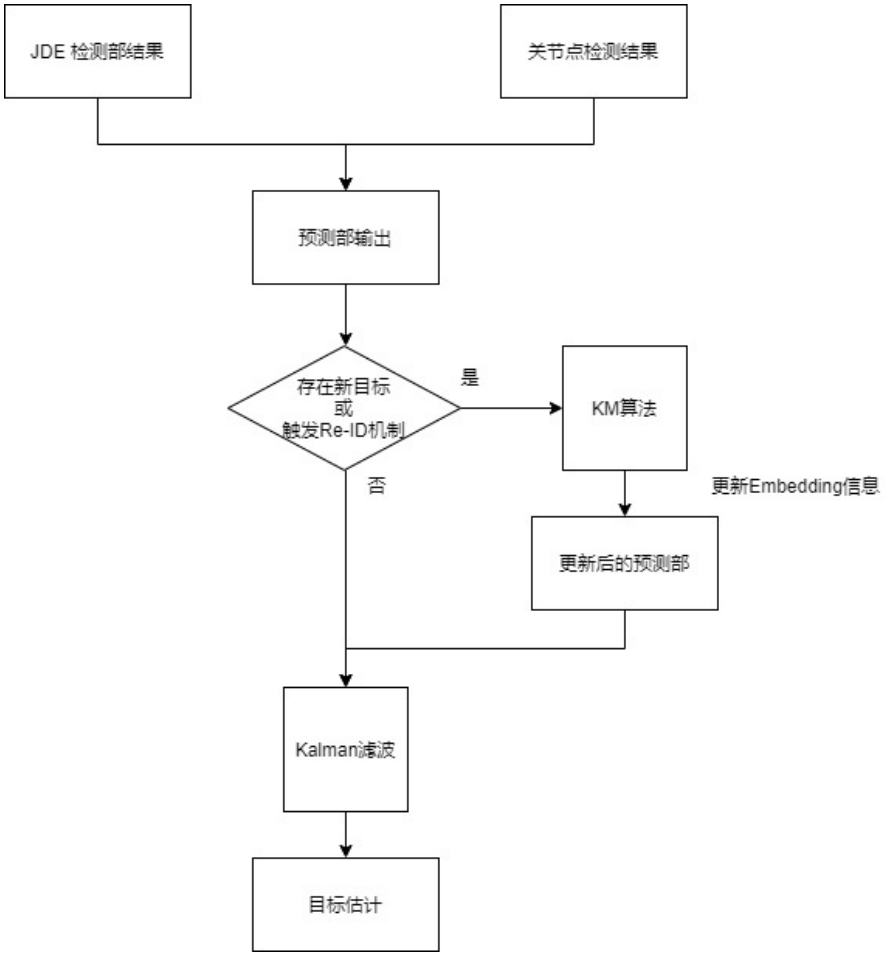

图 3-7 引入姿态信息的 JDE 多目标跟踪框架流程图

3.4 实验结果与分析

3.4.1 实验环境

本文使用 docker 环境分隔开发环境和实验环境。开发环境使用 pycharm IDE，python 版本为 3.7，深度学习框架选择的是 pytorch，开发的操作系统为 Windows 10。训练和评估环境使用的英国萨里大学 CVSSP 研究中心的 AI 服务器机组配

备有 condor 任务管理系统，操作系统选择的 Ubuntu 18.04，硬件配置为：Intel(R) Core(TM) i7-6700 8 核 16 线程，运行内存 32G, GPU 是 8 块 Nvidia Titan X。

3.4.2 实验数据集和数据集预处理

1）VOC 数据集及其预处理

VOC[50]数据集来自 PASCAL VOC 挑战（PASCAL 视觉对象类），这是世界级的计算机视觉挑战。PASCAL 的全名：模式分析，统计建模和计算学习，是由欧盟资助的网络组织。自 2005 年以来，PASCAL VOC 一直举办不同的挑战，每年的内容均不相同。挑战涵盖分类，检测，分割，人体布局和动作识别，数据集的容量和类型不断丰富。VOC 有 4 个主要类别和 20 个子类别，主要类别包括车辆，家庭，动物和人。常用 VOC 数据集有两个版本，VOC2007 和 VOC2012。前一个数据集包含 9963 张图片，并注释了 24000 个对象。相比之下，VOC2012 拥有的图片超过 15000 张，注释对象超过 33000 个。其中，人的标签包含最多的图片和注释，VOC 数据集中数据分布的一些详细信息，这意味着 VOC 数据集可用于训练人员检测器。与其他数据集相比，VOC 数据是视觉对象检测的初始数据集之一，被广泛用于检测算法的验证，具有较高的权威性。同时，数据的大小适宜，可在短时间内训练一种合格的模型。本文使用该数据集训练行人检测器，作为预训练模型供跟踪框架使用。

基于 YoLo 的行人检测器无法使用 VOC 数据集的原始格式和存储方法。有必要将原始格式和存储转换为合适的形式。

首先，编写一个脚本，以查找标签人中包含目标的每个图像，并删除文件夹中的所有其他文件，包括图像和注释。在此步骤中还将创建一个 txt 文件来存储所有有用的图像信息。

其次，再编写一个脚本从上一步创建的 txt 文件中读取所有图像信息，并在基于该信息的注释文件中搜索目标信息，这些被搜索的信息将被擦除并保存到以图像信息命名的 txt 文件中。转换后图像的注释的格式为：

[label_id，centroid_x，centroid_y，width height]

该 txt 文件中每一行仅包含一个目标，如果图像中有多个目标，则这些目标将存储在多行中。

最后，创建一个文件来存储在训练过程中使用的图像的绝对路径。

2）COCO 数据集及其预处理

COCO 数据集也称为 MS COCO[51]，起源于 2014 年，由 Microsoft 资助和维护。与 ImageNet 竞赛一样，是计算机视觉领域最受关注和权威的数据集之一。

COCO 数据集是一个大容量且支持检测，分割等不同任务的数据集。该数据集旨在使模型理解场景，数据主要从复杂的日常场景中截获，且图像中的目标标注十分精确。数据集中标注对象包括 91 个类别，328,000 张图像和 2,500,000 个标签。其中行人类别的标注数量超过 150 万。COCO 数据集还可以支持姿态关键点检测任务，所有注释都存储在 JSON 文件中，常规存储格式，注释格式的详细信息在下方给出。给出了目标（人）的边界框，并存储了关键点的坐标，关键点的可见状态也被添加其中。该数据集是用于检测和其他视觉任务的最大且最权威的数据

信息： 年份，版本，描述，贡献者，URL 链接，创建日期
许可： 许可 ID，许可名，URL 链接
图像： 图像 ID，宽，高，文件名，图像 URL，COCO 的 URL，捕获日期
标注： 关节点坐标，关节点数量，标注 ID，图像 ID，类别 ID，语义标志位，候选框位置，是否遮挡

集。由于数据集太大，该数据集上的训练模型需要高要求的设备以减少训练时间成本。本文拆分了 COCO 数据集，使用了其中的一半数据重新对 OpenPose 的关键点检测部分做了重训练。

3）JTA 数据集及其预处理

JTA 数据集[52]的全称是 Joint Track Auto，这是一个庞大的数据集，用于在城市场景（包括地铁站，街道和海滩）中进行行人姿势估计和跟踪。所有内容都是通过利用高度真实感的视频游戏（Rockstar North 开发的《侠盗猎车手 5》（Grand Theft Auto V））创建的。套装中有 512 个全高清视频（用于训练的 256 个用于测试的 256 个视频），每个都是 30 秒长和 30 fps。包含超过 500K 帧，并标记了超过 10 M 的身体姿势。而且，该数据集包括 3D 注释和遮挡注释。与其他 MOT 数据集（例如 MOT）相比，该数据集可以支持 MTMCT 任务，因为视频按组排列，每个组是从同一场景中的多个角度捕获的。

JTA 数据集预处理是预处理阶段中最复杂的操作。所有操作可以分为两部分：截取单帧和注释转换。为了适合模型训练，必须将视频片段截取成单帧，每帧都有唯一编号且按照出现顺序进行标注。使用开源软件 ffmpeg 将视频剪辑裁剪成帧，并按帧编号保存在文件夹中。在注释转换部分中，过程更加繁琐。需先创建 seqinfo.ini 文件来存储视频序列信息。其格式如下：

[*Sequence*]

$name = video\ clips\ name$

$imDir = image\ folder\ path$

$frameRate = frame\ rate$

$seqLength = total\ frame\ number$

$imWidth = width$

$imHeight = height$

$imExt = frame\ format$

然后创建检测注释文件。内容主要与目标检测信息没有区别，因此我们没有使用传统的 MOT 格式，而是结合了在人员检测器训练中使用的检测注释。格式描述如下：

$frame\_id\ \ target\_id\ \ centroid\_x\ \ centroid\_y\ \ width\ \ height\ \ confidence$

再创建 Ground Truth 文件。 基本事实的格式与检测注释文件非常相似。 也可以定义如下：

$frame\_id\ \ target\_id\ \ centroid\_x\ \ centroid\_y\ \ width\ \ height$
$ignore\_flag\ \ classes$

前 6 个元素与检测注释文件相同，第 7 个元素是用于评估期间是否忽视标志位，最后一个元素确定目标类型，包括 2 种：行人，遮挡目标。

3.4.3 实验评估标准

跟踪器的好坏需要用多个指标来评价。本小节讨论本文实验中用到的相关评估标准。为便于理解多目标跟踪的评价指标，先介绍几个基本评价指标。

1）TP：真正（True Positive，TP），经过计算被预测为正的正样本，可以用其数量描述判断为正的正确率，即 TP 数。

2）TN：真负（True Negative，TN），经过计算被预测为负的负样本，可以用其数量描述判断为负的正确率，即 TN 数。

3）FP：假正（False Positive，FP），经过计算被预测为正的负样本，可以用其数量描述误报率，即 FP 数。

4）FN：假负（False Negative，FN），经过计算被预测为负的正样本，可以用其数量描述漏报率，即 FN 数。

5）GT：真实基准（Ground Truth，GT），经过正确标注的数据或对象。

3.4.3.1 行人检测器的评估

行人检测器可用 GIoU、Objectness、Precision、Recall 等指标评估。下面讨论各个指标的含义。

1） GIoU[53]：新提出的一种称为广义交并比（Generalized Intersection over Union，GIOU）的指标，用来评估预测框（bounding box）与标注框（Ground Truth）之间的差异。其定义如下：

$$\mathrm{GIoU} = \mathrm{IoU} - \frac{\mathrm{A_C} - (\mathrm{A} \cap \mathrm{B})}{\mathrm{A_C}} \tag{3-4}$$

其中，IoU 是预测框和标注框的交并比，A 是预测框，B 是标注框，AC 是包含预测框和标注框的最小闭包区域。

GIoU 克服了 IOU=0 时不能反映预测框和标注框距离的大小，没有梯度回传，神经网络无法学习，以及无法精确的反映两者的重合度大小的缺点。GIoU 目前也被当作一种损失函数在目标检测领域中被使用。

2）Objectness：每个边界框的对象分（objectness score）。如果当前预测 bounding box 与 Ground Truth 的重合优于其他已存在的 bounding box，其分值是 1。如果当前预测的边界框不是最优，但与 Ground Truth 的重合率满足阈值要求，该 bounding box 被认为是忽略样本。由于本文任务着重于单类别检测任务，因此只有前景与背景存在，使用的阈值可以调低至 0.2。

3）Precision：精确度是指数据集中被分类器判定为正样本中真正的正样本所占的比重，其定义如下：

$$Precision = \frac{TP}{TP+FP} \tag{3-5}$$

其中，TP 是被正确判定的正样本数量，FP 是被错判为正样本的负样本数量。

4）Recall：召回率是指数据集中被分类器正确判定的正样本占总的正样本的比重，其定义如下：

$$Recall = \frac{TP}{TP+FN} \tag{3-6}$$

其中，TP 是被正确判定的正样本数量，FN 是被错判为负样本的正样本数量。

5）AP：平均精准度，对 PR 曲线上的 Precision 值求均值。

6）mAP：AP 是在单类别下的平均精度，mAP 是 AP 值在所有类别下的均值。本文中的 AP 和 mAP 无区别。

7）F1：对 Precision 和 Recall 进行整体评价，取值越大表明性能越好。

3.4.3.2 目标跟踪的评估

多目标跟踪算法的评价较为复杂，通常需要使用多项指标进行评价。本文使用了 MOTA，IDF1，MT，ML，IDs，FPSD，FPSA，FPS 等指标。下面讨论这些评价指标的含义。

1）MOTA：

多目标跟踪准确度（Multiple Object Tracking Accuracy），用于测量单个摄像机中多目标跟踪的准确性。其定义如下：

$$\mathrm{MOTA} = 1 - \frac{\mathrm{FN}+\mathrm{FP}+\phi}{\mathrm{T}} \tag{3-7}$$

其中，FN 是跟踪过程中所有帧的 FN-数总和，$\mathrm{FN} = \sum_t \mathrm{fn}_t$；FP 是跟踪过程中所有帧的 FP-数总和，$\mathrm{FP} = \sum_t \mathrm{fp}_t$；T 是所有帧中真正目标数的总和，$\mathrm{T} = \sum_t \mathrm{g}_t$；Φ是跟踪过程中跟踪轨迹的状态从“跟踪”到“未跟踪”的跳变数（Fragmentation），$\Phi = \sum_t \Phi_t$。公式中的 FN、FP 和Φ依次反映了漏报目标、误报目标和误配目标的情况。MOTA 越接近于 1 表示跟踪器性能越好。

2）IDF1：

识别 F 值（Identification F-Score），指每个行人框中行人 ID 识别的 F1 值，其定义如下：

$$\mathrm{IDF1} = \frac{2\mathrm{IDTP}}{2\mathrm{IDTP}+\mathrm{IDFP}+\mathrm{IDFN}} \tag{3-8}$$

其中，IDTP 是整个视频中正确检测（识别）的数量之和，IDFP 是整个视频误报的数量之和，IDFN 是整个视频漏报的数量之和。

对于多目标跟踪，除了 IDF1 指标外还有另外 2 个 ID 相关的指标，它们分别是 IDP 和 IDR。

IDP（Identification Precision）是指每个行人帧中行人 ID 的识别精度，其定义如下：

$$\mathrm{IDP} = \frac{\mathrm{IDTP}}{\mathrm{IDTP}+\mathrm{IDFP}} = \frac{\mathrm{IDTP}}{\mathrm{C}} \tag{3-9}$$

式中，C=IDTP+IDFP，是 ID 分配的总集。

IDR（Identification Recall）是指每个行人帧中行人 ID 识别的召回率，其定义如下：

$$\mathrm{IDR} = \frac{\mathrm{IDTP}}{\mathrm{IDTP}+\mathrm{IDFN}} = \frac{\mathrm{IDTP}}{\mathrm{T}} \tag{3-10}$$

式中，T=IDTP+IDFN，为正确的匹配 ID 数量。

IDF1、IDP 和 IDR 这三个指标可以根据任意其中两个推断出第三个的值。

其相互关系如下：

$$IDF1 = 2\frac{IDP \cdot IDR}{IDP+IDR} = \frac{IDTP}{\frac{T+C}{2}} \tag{3-11}$$

可以看出 IDF1 综合考虑了 IDP 和 IDR 指标。

3）IDs：

ID 切换数（ID switches），是指跟踪轨迹中行人 ID，由于跟踪算法的错误判断而导致的 ID 切换次数，能反映应跟踪的稳定性，数值越小越好。跟踪算法中的理想 ID switch 为 0。

4）MT：

多数跟踪数（Mostly Track），被正确跟踪目标轨迹大于 80％的跟踪轨道的数量。该值越大越好。

5）ML：

多数丢失数（Mostly Lost），丢失部分大于 80％的轨道数量。该值越小越好。

6）FPSD：

检测器每秒帧数（Frames Per Second of the Detector，FPSD）。

7）FPSA：

每秒数据融合操作次数（frame per second of the association step，FPSA）。

8）FPS：

每秒可处理帧数（frame per second of the overall system，FPS）。

3.4.4 生成预训练行人检测器

由于框架结构不同，因此当前已有的 YoLo 模型都不能当作预训练模型用于本文任务中。因此需要从初始状态训练行人检测器。整体训练被分为两个阶段，第一阶段检测模型在重组的 VOC 数据集上进行训练，重组的 VOC 数据集包含了 2007 和 2012 两个版本的数据；第二阶段在第一阶段得到的训练模型基础上冻结前八层（4 组 conv 和 max pooling 组合），在分割好的 COCO 数据集上再次训练进行调参。在测试不同的超参数组合后，选择超参数组合：epochs = 200； batch size= 16; momentum= 0.9; decay= 0.0005; learning rate= 0.001。训练和验证结果如图 3-8 所示。

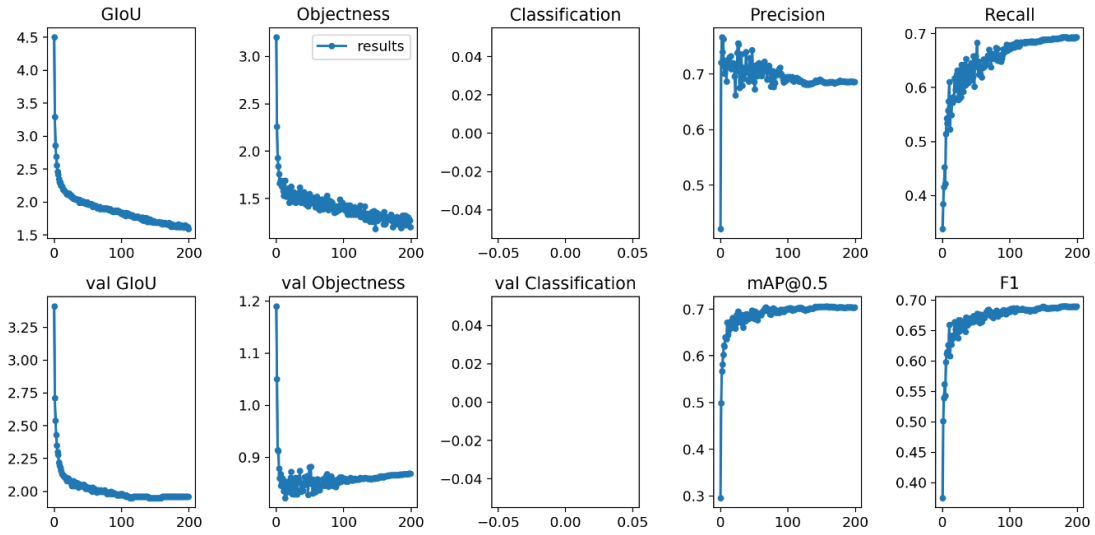

图 3-8 检测部的训练及验证结果

由于这是单一类检测任务，因此图中的第三列没有内容。在第一列中，训练和验证 GIoU 均从其初始值开始降低。如先前在方法论中提到的，GIoU 可以用作损失函数。GIoU 的下降趋势表明，bounding box 与 Ground Truth 之间的差异越来越小，逐渐逼近 Ground Truth。第二列“Objectness”显示，检测器经过训练后获取正样本的能力较强，错检或检测误差较大的现象较少。剩下的 4 个评估标准表明，本文训练的行人检测器的综合检测性能可以达到 70%，满足了支持跟踪算法的基础检测要求。图 3-9 现实了训练好的行人检测器对输入图像中的人体检测效果。从图中可以看出本文提出的行人检测器能有效检测输入图像中行人目标。

同时本文算法与 YoloV3 算法的三个不同版本在 COCO 数据集上进行了性能比对，性能比对如表 3-3 所示。从表中可知，本文提出的检测部 mAP 数值最高，但须指明的是，本文的检测是针对单类别进行的检测，因而占有优势。其余检测算法能提供多类别目标的检测功能，因而 mAP 会因为检测算法较好的泛化能力而受影响。但该数值仍然说明了本文提出的检测部是可以当作行人跟踪算法的检测部的。对比了单帧耗时，表明了本文除去了最小细粒度检测分支之后，算法的单帧耗时降低了不少。值得注意的是，本文提出的算法支持的图像输入大小为 768，原 YoloV3 算法推荐的输入是 256，从另一角度说明，在相同时间内本文算法能处理更多的图像细节。

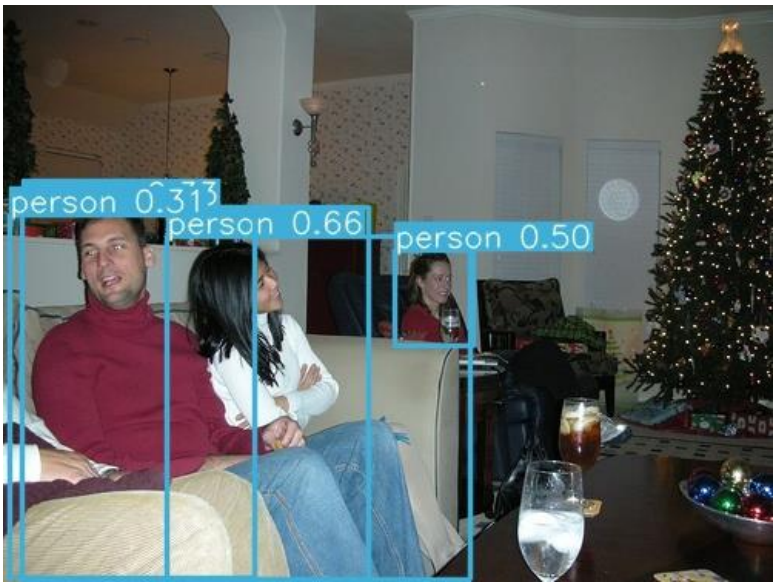

图 3-9 本文行人检测器对输入图像的人体检测结果

表 3-3 检测部性能比较

算法	mAP	单帧耗时(ms)
YoloV3-320[43]	51.5	**22**
YoloV3-410[43]	55.3	29
YoloV3-608[43]	57.9	51
Ours	**68.9**	32

3.4.5 训练引入姿态描述信息的改进 JDE 跟踪框架

先前训练的行人检测器当作预训练模型用于替换原有 JDE 模型中的 Darknet 检测部分，多尺度输出分支也如本文前述修改。网络模型使用标准的 SGD 训练了 25 个 epoch，learning rate 设置为 0.01，在第 12 个 epoch 和第 20 个 epoch，learning rate 进一步下调至原 learning rate 的十分之一。采用了不同的数据增强技术，包括随机旋转，随机缩放和颜色抖动等，以减少过拟合。训练集使用的是 JTA 数据集中的训练数据。为了与其他 MOT 算法比较性能，本文选择使用了 MOT16，与已有的 MOT benchmark 进行对比，本文对比了精度较高的几个 SDE 类型多目标跟踪算法和原 JDE 多目标跟踪算法，性能比较如表 3-4 所示（表上部为 SDE 类型多目标跟踪算法，下部为 JDE 类型多目标跟踪算法）。

表 3-4 中选取了不同的 SDE 类型多目标框架算法以及 JDE 算法的不同版本

和本文改进的两种算法进行比较。

（1）对比 MOTA 指标，本文改进的算法与原 JDE 算法基本相当，与 SDE 类型的多目标跟踪框架相比没有明显劣势。

（2）对比 IDF1 指标，SDE 类型算法要优于 JDE 类型算法，包括本文提出的框架。表明了 JDE 类型跟踪算法的稳定性仍有提高空间。

（3）对比 MT 指标和 ML 指标，本文算法与 SDE 类型算法和原 JDE 算法都没有明显差距，说明本文算法在跟踪完成度上是可靠的。

（4）在 IDs 指标上，SDE 类型算法要远优于 JDE 类型的算法和本文算法，这说明了 embedding 信息还不够精确，目标 ID 的归类出现了比较明显的性能下降，这是由于数据融合与目标检测被合并在一步导致的。但本文的算法在没有引入姿态信息的条件下，IDs 指标性能与原 JDE 算法相比，有微小的优势，则要归功于替换了检测部之后，算法对目标特征提取能力和检测能力都有所增强。在引入了姿态描述信息之后，IDs 数值有了明显下降，表明了引入姿态描述信息对提高跟踪算法抗 IDs 能力是可行的。

（5）同时，本文也对比了 FPSD，FPSA 和 FPS 指标。从表中可以看出，SDE 类型的算法在效率指标上远未达到实时甚至近实时的性能，每秒处理帧数都十分有限。与 JDE 类型的算法比较，本文提出的算法在没有引入姿态描述信息的条件下，每秒检测帧数与 JDE864 接近，但远好于 JDE1088，这是因为移除了小尺度目标跟踪输出分支。在引入了姿态描述信息后，检测帧数有所下降，这是由于关节点检测造成的资源消耗。对比了每秒数据融合帧，本文算法在没有引入姿态描述信息的情况下，要优于 JDE1088，在引入姿态描述信息的条件下，则无过多优势。对比了 FPS 指标，本文算法无论是否引入了姿态信息，都能在计算资源充足的条件下完成近实时的性能。虽本文在各单项指标上都没有达到最优排名，但跨摄像头多目标跟踪任务是一个综合任务，综合角度上本文算法没有明显缺陷，即使抗 ID Switch 能力相较于 SDE 并无优势，但也优于其余 JDE 模型。

表 3-4 多目标跟踪框架性能比较

Method	MOTA	IDF1	MT	ML	IDs	FPSD	FPSA	FPS
DeepSORT[54]	61.4	62.2	32.8	18.2	781	<15	17.4	<8.1
RAR16[55]	63.0	63.8	39.9	22.1	**482**	<15	1.6	<1.5
TAP[56]	64.8	**73.5**	**40.6**	22.0	794	<15	18.2	<8.2
CNNMTT[57]	65.2	62.2	32.4	21.3	946	<15	11.2	<6.4

POI[58]	**66.1**	65.1	34.0	21.3	805	<15	9.9	<6
JDE864[46]	62.1	56.9	34.4	**16.7**	1608	**34.3**	**259.8**	**30.3**
JDE1088[46]	64.4	55.8	35.4	20.0	1544	24.5	236.5	22.2
Ours（No Pose）	61.7	55.6	33.1	19.6	1444	33.8	247.7	29.3
Ours（With Pose）	62.2	55.9	34.2	18.5	1248	32.6	238.6	27.5

3.5 本章小结

本文提出了一种引入姿态信息的基于 JDE 框架的多目标跟踪算法。简化了 darknet 特征提取结构，创建一个新的检测算法替换了原 JDE 框架中的检测部。另外，本文使用 OpenPose 方法对姿势关键点进行检测，提供了一种姿态信息表达形式，并把姿态信息嵌入了检测部结果中。本文算法在近实时性上要优于传统 SDE 类多目标跟踪方法，但精度和可靠性相对不占优势；与原 JDE 跟踪算法相比，引入姿态信息之后，精度和可靠性得到了一定增强，且近实时性没有较大损害。在下一步工作中，将本章提出的算法用于跨摄像头多目标跟踪任务中。

第 4 章 集中式跨摄像头多目标跟踪算法

4.1 引言

随着近些年的硬件算力发展，并考虑到单传感器多目标跟踪算法的巨大提高与进步，跨摄像头多目标跟踪算法逐渐吸引了学界和工业界的兴趣，众多研究资源和工业项目设置趋向于该领域。当前，跨摄像头多目标跟踪算法有不同的发展方向，包括有集中式框架，分布式系统框架，涉及到节点数据的置信问题，数据融合问题等。本文结合了第三章得到的引入姿态信息的 JDE 多目标跟踪框架特点，选择了一种集中式的跨摄像头多目标跟踪框架。4.2 节中将详细介绍本文 MTMCT（Multi-Target Multi-Camera Tracking）算法流程。4.3 节将介绍 MTMCT 系统搭建细节。4.4 节将介绍系统性能并与其他 MTMCT 算法进行比较。

4.2 集中式跨摄像头 JDE 多目标跟踪框架

在前一章中阐述的引入姿态信息的多目标跟踪算法提供了一种近实时的多目标跟踪解决方案，其特点是在目标的观测结果中，加入了相应的 Tracker ID 信息，达到了减少 Re-ID 次数的目的，增加了算法处理速度。本章基于此算法提出一种集中式的跨摄像头多目标跟踪算法。

4.2.1 集中式跨摄像头目标跟踪问题定义

物体运动是一个连续过程，在实际应用中，跟踪系统一般采用离散非线性的运动系统模型对物体运动进行表达：

$$x_{k+1} = f(x_k) + w_k \tag{4-1}$$

其中，k 是时间指示，x_{k+1} 是物体在时刻 $k+1$ 时的真实状态，函数 $f(\cdot)$ 是转换函数又称过程函数，表示物体在两个时间点间的状态转移关系，w_k 是过程噪声。现有的检测方法，无论性能多么出色都存在一定的误差，因此还需要定义物体的观测模型：

$$y_k = h(x_k) + v_k \tag{4-2}$$

其中 y_k 是目标在时刻 k 的观测状态，函数 $h(\cdot)$ 也是转换函数但转换的是目标

真实状态到观测状态，v_k 则是传感器在时刻 k 时的观测噪声。得到了单一传感器的观测模型后，同样的，可以推导出全局中各观测节点的观测模型表达：

$$y_k^i = h^i(x_k) + v_k^i, i = 1,2,\ldots,N \tag{4-3}$$

其中 i 表示第 i 个传感节点，其取值是整数有限集，最大值 N 表示全局传感器网络中存在的传感节点最大数量。

而跟踪问题的本质是结合目标的观测状态信息预测下一时刻目标的真实状态，可表达为：

$$p_{k+1} = T(y_k) \tag{4-4}$$

其中， p_{k+1} 是预测的下一时刻的目标状态，函数 $T(\cdot)$ 就是跟踪模型又称为预测模型，该模型的性能评价由 p_{k+1} 和 x_{k+1} 间的误差决定，该误差越小则模型的跟踪能力越强。

因此想要构建一个性能较强的跟踪模型可以从两个方向着手，一个是提高 y_k 精度减少与 x_k 间的误差（提高观测模型的精度），另一个是提高 $T(\cdot)$ 的准确性。本文选择使用前者，通过构建集中式的系统，对各节点的观测模型结果进行数据融合以得到更接近真实状态的观测结果。

4.2.2 数据融合方法

数据融合方法有很多，包括了加权平均法、Kalman 滤波法、多贝叶斯估计法以及 D-S 证据推理方法等等。每一种数据融合的方法都有各自的优缺点，充分考虑到本算法的近实时性需求，选择了第三种方法——多贝叶斯估计法来完成各节点的检测数据融合。

部署算法框架的系统由两部分组成，一部分是分布式的传感器网络，另一部分是系统中心也是数据融合中心。首先考虑系统中的某一个单独传感节点 i，该节点会传输此节点检测结果至数据融合中心，所采用的检测算法是第 3 章论述的入姿态信息的 JDE 多目标跟踪算法。假设节点 i 处的检测结果是 $d_{i,k}^r$，其中 k 是指代时间标识，r 代表的是节点 i 处检测到的目标数量，同时也是检测结果的维度标识。检测结果 $d_{i,k}^r$ 与前一时刻的预测结果 y_k 求对应元素间的距离，得到距离

矩阵 $e_{i,k}^{r}$，若存在两个矩阵维度不匹配或元素缺失的情况，对缺失的元素其对应的距离元素为视野的对角线距离填充，保证矩阵 $e_{i,k}^{r}$ 与矩阵 y_k 中的元素一一对应。对得到的距离矩阵 $e_{i,k}^{r}$ 先进行倒数转换（更改了数据大小排列顺序，当距离为零时其倒数用同样视野的对角线距离替换）再求新矩阵的平均值 u 和方差 v，可得到正态分布

$$g_i(d_i \mid \theta) = N(u,v) = \frac{1}{\sqrt{2\pi v}} \exp(-\frac{(x-u)^2}{2v}), \tag{4-5}$$

该正态分布是传感器 i 的似然函数（数据确定但检测模型不确定），描述了预测模型吻合此传感器观测结果的程度。

考虑到部署算法框架的系统由两部分组成，一部分是分布式的传感器网络，另一部分是系统中心也是数据融合中心。因此，系统中的各个传感器节点被认为是相互独立的，各自检测结果互不干扰，可以得到的全局似然函数：

$$L_{global}(D;\theta) = f(d_1 \mid \theta) f(d_2 \mid \theta) f(d_3 \mid \theta) \cdots f(d_N \mid \theta) \tag{4-6}$$

对全局似然函数取对数可以简化得到：

$$\begin{aligned} LL_{global}(D;\theta) &= \log(f(d_1 \mid \theta) f(d_2 \mid \theta) f(d_3 \mid \theta) \cdots f(d_N \mid \theta)) \\ &= \log(f(d_1 \mid \theta)) + \log(f(d_2 \mid \theta)) + \cdots + \log(f(d_N \mid \theta)) \end{aligned} \tag{4-7}$$

对全局似然函数求解最大似然得到全局检测模型的参数。先取 LL_{global} 关于全局观测模型 θ 的一阶导数并等价于 0，再取 LL_{global} 关于全局观测模型 θ 的二阶导数使之取值为负值，求解方程得到全局观测模型的参数进而得到了数据融合后的检测结果。

4.2.3 集中式跨摄像头多目标跟踪框架的部署

前一节介绍了数据融合的方法，本节将介绍完整的集中式跨摄像头多目标跟踪框架。该框架部署在集中式系统中。相比去中心化系统和分布式系统，集中式系统拥有明确的数据融合中心，一般该中心是特殊节点拥有强大的算力资源，往往该节点是服务器。除了中心节点外，集中式系统中还有许多普通节点，这些节点构成了分布式传感器网络。

集中式跨摄像头多目标跟踪框架如图 4-1 所示。其中，图 4-1（a）为单节点处理流程，本文将其称为单节点操作，图 4-1（b）为数据融合中心的处理流程，

本文将其称为全局操作。单节点操作与前一章的算法相比，去除了 kalman 滤波操作，单独节点不再进行目标预测。单节点的输出为引入姿态信息的 JDE 检测结果。单节点输出将送往数据融合中心等待全局处理。下面分别论述单节点操作流程和全局操作流程。

4.2.4.1 单节点操作

单节点操作部署在分布式传感器网络的各个节点上。单节点操作的处理流程如图 4-1（a）所示，其原理基础是第三章中的多目标跟踪框架，但不包含有对目标下一时刻状态的预测。

第一步，分别获取 JDE 检测部结果和关节点检测结果。

第二步，基于前一步的结果生成预测部结果。

第三步，判断是否有新目标生成或达到了目标重匹配的时间间隔阈值要求；若满足条件，则触发 Re-ID 操作，更新预测部结果。

第四步，检测部结果被经过编码后发送到数据融合中心等待全局操作。

4.2.4.1 全局操作

全局操作部署在数据融合中心上，其处理流程如图 4-1（b）所示。

第一步，数据融合中心接受来自各单节点的输出数据，并立即进行数据转换处理。在数据处理阶段，依据预先定义好的场景参数（包含摄像传感器参数，传感器位置等）对检测结果中目标的位置信息进行转换，变成符合全局位置参数（目标的位置坐标被转换成全局俯瞰角度的位置坐标）。

第二步，数据融合中心判断框架是否需要进行初始化，若需要进行框架初始化，则对收集到的各单节点输出结果进行 KM 算法二分图匹配。对需要更改的的检测头，进行 Embedding 信息更新。

第三步，计算各单节点的距离矩阵并做维度判断，如果存在新生成目标或跟踪框架运行了十帧且没进行 Re-ID 操作，数据融合中心将进行 KM 二分图匹配并更新 embedding 信息。

第四步，进行数据融合，融合操作如 4.2.3 节所述，依据检测结果和前一时刻预测结果间的误差确认各节点的数据准确性，准确性越高的节点在数据融合时权重占比越高，数据融合的输出是全局俯瞰的目标位置坐标。

第五步，数据融合后的全局观测结果被送入 Kalman 滤波进行预测操作得到预测结果。如果数据融合中心进行了二分图匹配操作，更新后的观测信息将回传各节点以更新各节点中的观测信息。

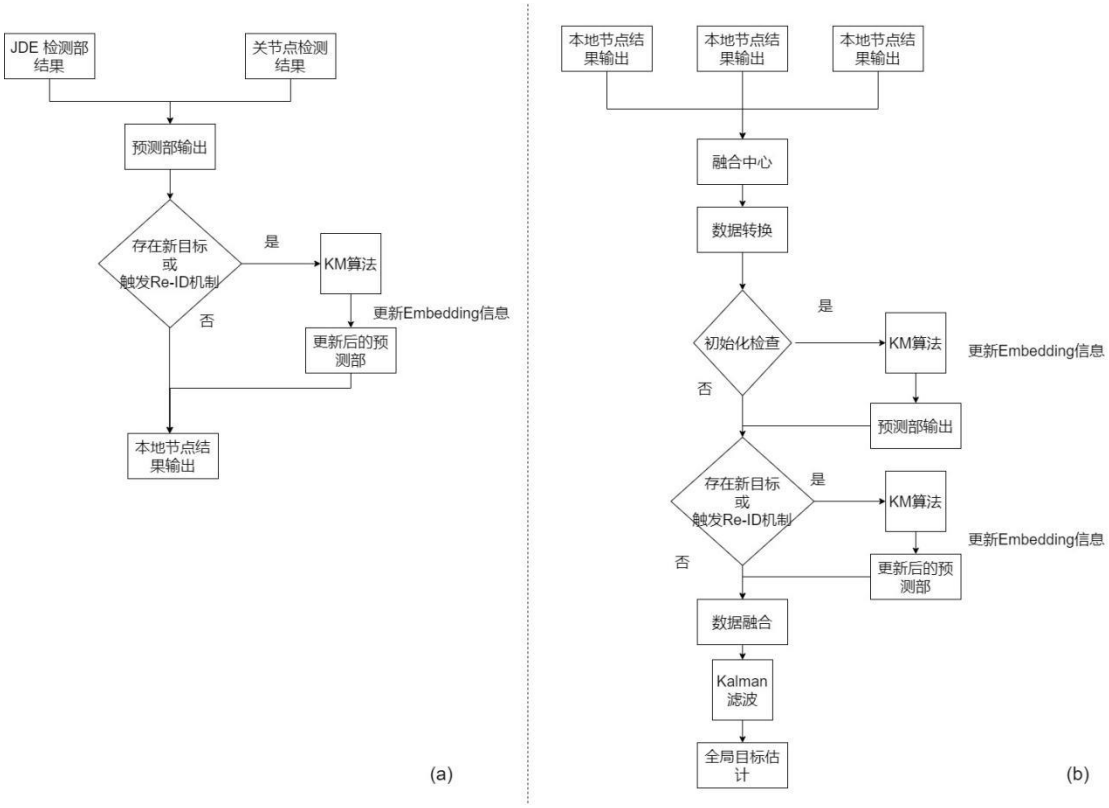

图 4-1 基于引入姿态信息的 JDE 多目标跟踪框架的 MTMCT 框架

(a)本地单节点操作；(b)全局操作

4.3 集中式跨摄像头多目标跟踪系统的组成

本节中介绍集中式跨摄像头多目标跟踪系统的组成。整个系统由智能感知节点、服务器、客户端组成，通过网络通讯连接。系统框架简图如图 4-2 所示。

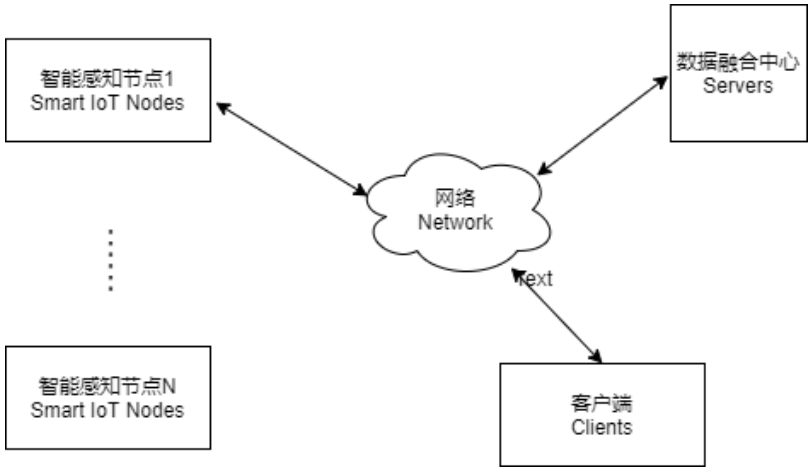

图 4-2 集中式跨摄像头多目标跟踪系统

4.3.1 智能感知节点的功能

智能感知节点由摄像机、数据处理模块和通信模块组成，其作用是捕获视频片段并检测视野中的行人目标。将检测到的数据进行编码，压缩到数据包中以便进行网络传输。由于要求具备一定的计算能力，在实际中可考虑采用支持 AI 计算的嵌入式平台实现智能感知节点，如英伟达的 JETSON XAVIER NX 平台，华为基于昇腾 310 AI 处理器的嵌入式平台等。

图 4-3 是智能感知节点的功能框图。其主要功能是帧捕获，行人检测，编码，数据打包、数据传输和系统设置控制等。

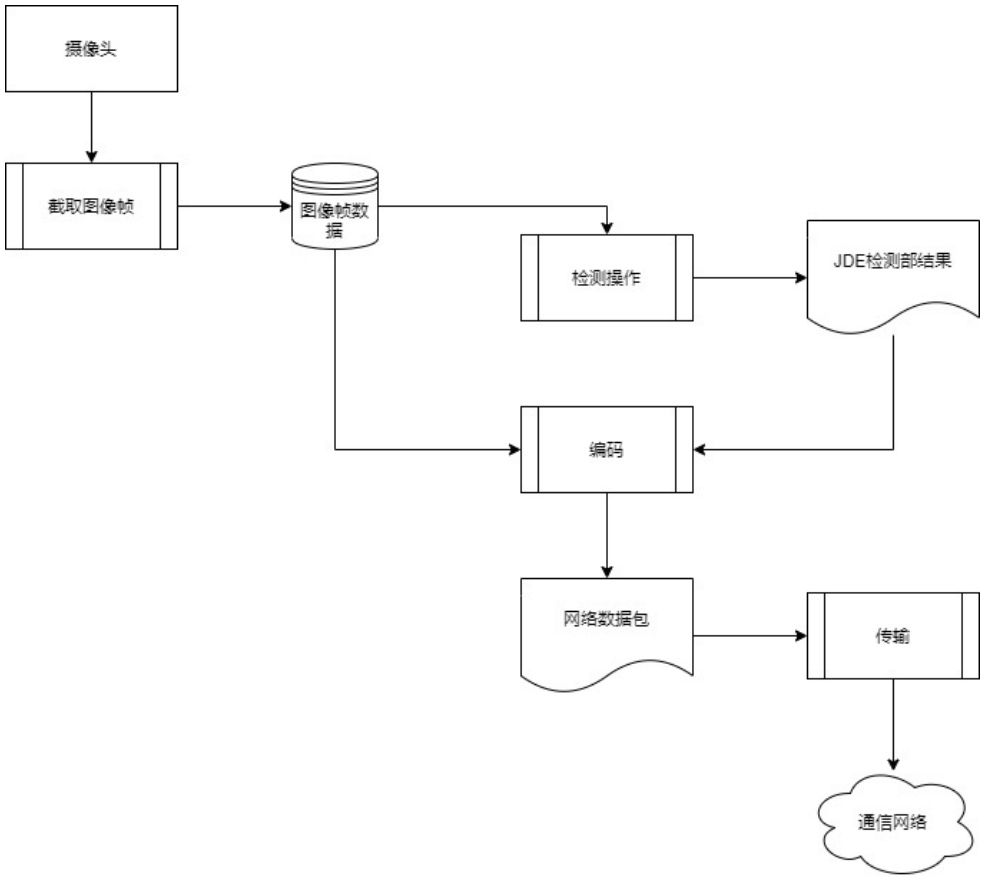

图 4-3 智能感知节点功能框图

1）帧捕获：读取从传感器收集的数据并将其转换为适合的数据类型以进行存储。

2）行人检测：运行单节点操作得到观测结果列表，并保存该数据等待编码。

3）编码：对检测到目标的帧进行压缩编码，并嵌入观测结果列表。

4）数据打包：将编码后的数据打包成网络传输数据包。为了满足数据流传输和不同网络协议的要求，可能会将数据切分成段。在传输过程中可能会丢失数据，从而破坏数据完整性。因此，选择将数据编码为二进制流。

5）数据传输：将打包好的数据发往数据融合中心。

6）系统设置控制是支持用户设置远程 IP 地址并监视智能 IoT 节点的工作状态的功能。

本文用树莓派 4B 开发板测试了算法部署到边缘设备上运行的可能性。由树莓派开发板、红外高清摄像头组成了一个智能感知节点，如图 4-4 所示。

图 4-4 智能感知节点硬件组成

由于树莓派开发板本身不带操作系统和算法需要的驱动软件，因此还需要对硬件设备安装操作系统和驱动程序以满足算法运行的需要。图 4-5 展示了成功安装操作系统和驱动程序，以及正确配置相关驱动程序的结果。

图 4-5 正确安装操作系统和驱动后的智能感知节点

4.3.2 数据融合中心的功能

数据融合中心是系统中最重要的节点，由服务器实现，是完成跨摄像头多目标跟踪任务的主要设备。在该设备中完成集中式系统中的全局操作。图 4-6 是数据融合中心的功能框图。其主要功能是数据流解码，数据处理、数据流编码和数据传输等。

1）数据流解码：接收来自智能感知节点的单节点操作得到的观测结果列表。

2）数据处理：实现全局操作，得到预测结果和观测信息。

3）数据流编码：

4）数据发送：将预测结果和更新后的观测信息发往各节点。

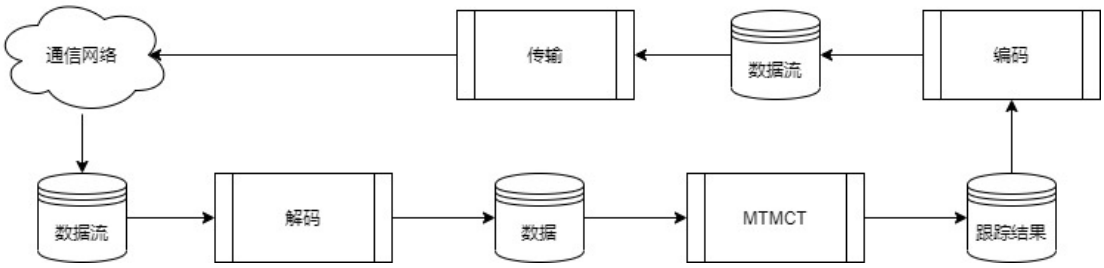

图 4-6 数据融合中心的功能框图

4.3.3 客户端的功能

客户端的主要目的是控制系统的运行状态，监视摄像机的视野，传输数据并显示跟踪结果。本文在一台 PC 机上实现客户端的功能。图 4-7 客户端 UI 界面运行测试图，由于系统中只接了一个智能感知节点，所以只有 Camera No.1 窗口显示了捕捉到的图像，其它几个窗口为空白，表示无智能感知节点接入。该程序用 PyQt 和 OpenCV 编程实现，包含以下功能：UI 界面，摄像头控制，网络通信控制，系统状态检查和控制。

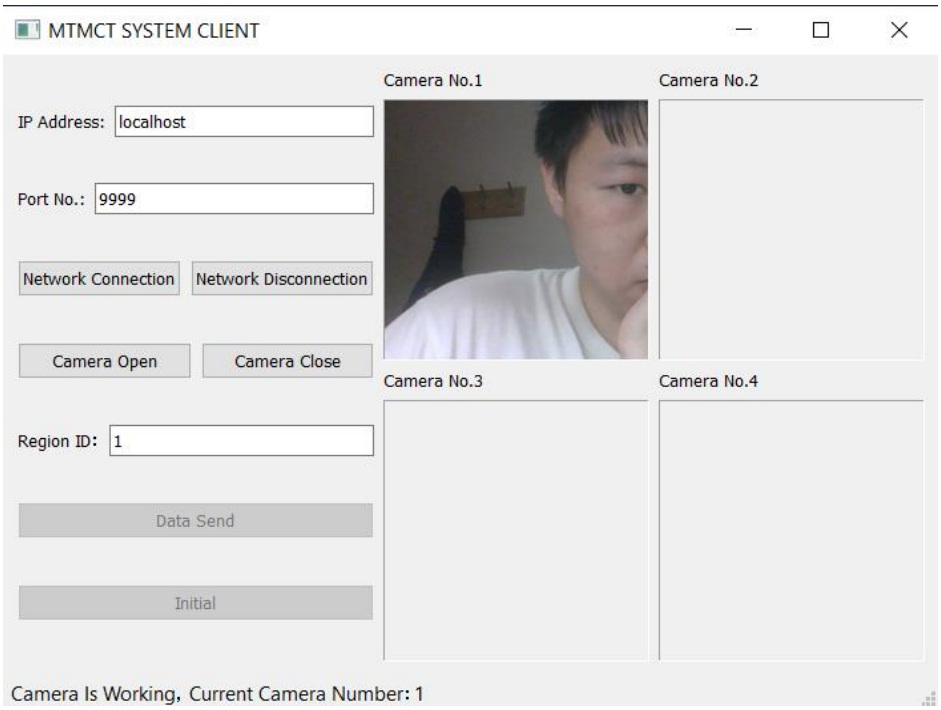

图 4-7 客户端应用程序 UI 界面

图 4-7 中，2 个输入框：IP Address 和 Port No，用来设置远程 IP 地址和端

口号，以便智能节点知道服务器在哪里；4 个按钮：Network Connection、Network Disconnection、Camera Open 和 Camera Close 实现了网络和摄像机控制；输入框 Region ID 用于以数字编号的方式定义智能节点的位置；按钮 Data Send 启动数据传输；按钮 Initial 启动系统检查节点状态并返回任何错误或警告。4 个窗口：Camera No.1、Camera No.2、Camera No.3 和 Camera No.4 用来显示智能节点捕获的图像和节点的状态。

4.4 跨摄像头多目标跟踪系统实验

4.4.1 实验环境的搭建

4.3 节讨论了集中式跨摄像头多目标跟踪系统的组成。整个系统由智能感知节点、服务器、客户端组成。集中式跨摄像头多目标跟踪算法包括了单节点操作和全局操作，单节点操作在智能感知节点上完成，全局操作在服务器上完成。需要智能感知节点具有一定的计算能力。因实验条件所限，特别是疫情期间的锁城状态，无法得到合适的嵌入式计算平台用于智能感知节点，所以将智能感知节点也搭建在服务器上，用来评估跨摄像头多目标跟踪算法性能。因此实验环境需要同时满足单节点操作和全局操作的要求。单节点操作主要实现引入姿态描述信息的改进 JDE 跟踪框架，是整个算法中运算量最大的部分。而全局操作没有新增深度学习运算（高算力），更多的是多贝叶斯估计和二分图匹配运算等，运算量远小于单节点操作的要求。综上，可以直接使用第 3 章所用的实验环境为本章实验所用。

4.4.2 实验数据集及其预处理

实验数据集使用的是含多摄像头数据的 JTA 数据集，该数据集已经在第 3 章详细介绍，此处不再赘述。数据集的预处理阶段，还需要全局信息标注，该标注格式有所区别：

frame_id	target_id	x_in_plain	y_in_plain

如上所示，全局标注格式如期所示。其中后两个坐标是依据相机模型和传感器位置信息等转换成的俯瞰坐标。

4.4.3 实验评估标准

尽管当下有越来越多的研究人员投入了跨摄像头多目标跟踪的研究之中，但

目前还没有普遍使用或公认的且完整有效的评估标准。除去第 3 章所述的多目标跟踪评价指标外，现阶段较受欢迎的评价标准还有跨摄像头目标跟踪精度(MCTA)[40]:

$$MCTA=\frac{2PR}{P+R}(1-\frac{M^{w}}{T^{w}})(1-\frac{M^{h}}{T^{h}}) \tag{4-6}$$

MCTA 一共由 3 项组成，第一项是 IDF1 结果；第二项则表示了单个节点中的目标匹配能力，M^{w}是单个摄像机中行人 ID 的错误匹配次数；T^{w}是单个摄像机中正确检测到的次数；第三项描述了跨多个摄像机后的目标匹配能力，M^{h}是摄像机之间对行人 ID 的错误匹配数，T^{w}是指跨摄像头后的正确检测的次数（例如，一个目标从一个摄像头视野中消失但出现在另一各传感视野中也能正确匹配目标 ID，或者分配了相同的 ID 的目标在多个摄像头视野中出现并正确匹配了他们的 ID）。

4.4.4 集中式跨摄像头多目标跟踪框架实验分析

基于前一章的实验操作后，将 JTA 数据集中同场景且摄像头固定的视频片段和对应标注取出，依照不同场景划分不同的实验组别并进行数据集预处理操作。为易于验证实验操作，每组实验中只使用两组不同传感器得到的数据。数据融合操作后的全局观测数据案例（同一目标在两个不同摄像头中的观测状态融合后得到全局观测状态）如表 4-1 所示：

表 4-1 数据融合示例

观测节点	目标中心点 x	目标中心点 y	宽度	长度
预测	338	425	20	31
传感器 1	335	422	23	33
传感器 2	352	418	28	38
融合	341	421	24	36

该示例来自 JTA 数据集中测试集场景 3 的第 69 帧目标 4，两个不同传感器对该目标的状态观测在转换成全局视角后有细微差别。基于前一刻的预测结果使用多贝叶斯估计法对该时刻的观测数据进行融合（传感器 1 的观测结果与预测结果相近，视传感器 1 的观测状态较传感器 2 的观测结果精准），得到结果如表所

示。集中式跨摄像头多目标跟踪实验结果如表 4-2 所示，验证数据融合后的跟踪效果是否提升。

表 4-2 集中式跨摄像头多目标跟与单摄像头多目标跟踪性能比对

Method	MOTA	IDF1	MT	ML	IDs	FPSD	FPSA	FPS
JDE864[46]	62.1	56.9	34.4	16.7	1608	34.3	259.8	30.3
JDE1088[46]	64.4	55.8	35.4	20.0	1544	24.5	236.5	22.2
Ours (No Pose)	61.7	55.6	33.1	19.6	1444	33.8	247.7	29.3
Ours (with Pose)	62.2	55.9	34.2	18.5	1248	32.6	238.6	27.5
Centralized-Our	71.3	69.9	36.8	16.7	1133	32.6	209.3	23.1

（1）对比第一列结果，集中式跨摄像头多目标跟踪算法大幅度提升了 MOTA 数值，这证明了对多源观测数据进行数据融合可以大幅度提升跟踪模型对目标真实状态的观测精度。

（2）IDF1 同样证明了数据融合后会提高模型的正确识别的检测，对模型的性能提升是有正向意义。

（3）MT 和 ML 指标也有相应改善，说明了集中式跨摄像头多目标跟踪算法可以提高跟踪模型的可靠性，ID switch 数量的下降也证明了这一点。

（3）FPSD 不受数据融合带来效率影响，但是 FPSA 有明显下降，这是因为数据融合的操作导致需要消耗一定时间，数据关联的效率被迫下降，同时算法的速度 FPS 值也下降较多。值得注意的是，当前系统中只有两个数据源，如果数据源数量继续上升，FPSA 数值将会继续下降，对跟踪框架的效率造成致命影响。

表 4-2 本文提出算法与其他 MTMCT 算法性能比对

Method	MOTA	IDF1	MCTA	FPS
State-Aware Re-ID[26]	82.8	77.1	-	~0
DeepCC[33]	73.0	75.6	-	~0
BIPCC[60]	75.3	83.2	79.7	~0
Centralized-Our	71.3	69.9	74.6	23.1

如前文所述，当前跨摄像头多目标跟踪研究没有广泛使用的评价标准，也没有被广泛使用的公开数据集，因此针对此类算法的横向比较很难有对应的

benchmark 矩阵。本文搜寻了当前较为出名的几类跨摄像头多目标跟踪算法，部分提供了 MCTA 数值，且三类算法都无法提供近实时的处理速度。比较 MOTA，本文算法与之有一定差距，尤其是与第一个算法相比，MOTA 数值差距较大。这是因为，前三种算法都采用的 SDE 跟踪思想，在跟踪过程中，光数据关联操作就进行了两次，算法的 ID switch 可能性较低，因而提供了较为不错的 MOTA 数值。通过 IDF1 数值比较可以得出，embedding 信息目前还不够稳定，造成了大量的 ID switch，因而正确识别的检测数量较少。对比 MCTA 本文与第三种算法相比也有一定差距，精度不够也是因为 IDF1 数值较低导致的。对比 FPS 数值，可以明显观察到只有本算法可以提供近实时的处理速度，这是 SDE 类型跟踪框架所不能提供的。

4.5 本章小结

本文提出了一种集中式跨摄像头多目标跟踪算法框架。利用多贝叶斯估计的数据融合方法，对多源的观测数据进行融合，得到了针对目标在运动系统中的真实状态较为精确的观测模型，为跟踪框架全局跟踪多目标提供了观测基础。本文提出的集中式算法相较于单源观测数据输入算法有了明显的性能提升，算法跟踪精度和抗 ID switch 的能力都有一定增强。与已有的基于 SDE 类型的跨摄像头多目标跟踪算法比较，在损失了一定跟踪精度和识别准确性的情况下，算法速度达到了近实时的需求。与已有的 JDE 算法相比，算法速度未有明显劣势。

第 5 章 总结与展望

5.1 工作总结

随着当前摄像头设备布设呈现出网络化趋势，视频数据海量激增。目标跟踪作为视频处理领域里十分重要且充满众多挑战的任务，通过对视频序列中存在的目标，智能化提取特征信息，获得传感视野里目标的观测状态信息（目标位置、速度、运动轨迹等），为后续其余计算机视觉任务提供基础。特别是在安防领域内，传统的人工跟踪目标的方式不再满足当下的需求，多目标跟踪，跨摄像头多目标跟踪等任务已经超出了原有人工处理能力。但是在应用场景中，受视野复杂背景、光强变化、目标存活状态的不确定、目标及障碍物间的互相影响等不同因素，容易造成跟踪目标的遗漏、错误跟踪、目标身份的确定失败等困境，更由于当前算法的复杂度较高，实时且鲁棒的跨摄像头多目标跟踪框架的搭建十分困难。近年来，基于软硬件性能的大幅度提高，跨摄像头多目标跟踪算法研究成为热点研究话题。本文基于应用场景，提出了一种引入姿态信息的基于 JDE 跟踪算法的跨摄像头多目标跟踪框架。本文完成的主要研究工作包含以下几点：

(1) 对原有的 JDE 多目标跟踪框架进行了改进。首先简化了 darknet 特征提取结构，创建一个新的检测算法替换了原 JDE 框架中的检测部。其次使用 OpenPose 方法对姿势关键点进行检测，并把姿态信息嵌入了检测部结果中。得到了一种引入姿态信息的基于 JDE 框架的多目标跟踪算法。该算法在精度和可靠性上比传统 SDE 类多目标跟踪方法略有下降，但比原 JDE 算法得到了一定增强；在算法时间效率上与原 JDE 算法相当，达到近实时水平，远高于 SDE 算法。

(2) 在得到的引入姿态信息的基于 JDE 多目标跟踪框架基础上，提出了一种集中式的跨摄像头多目标跟踪框架。该框架将系统中多个节点提供的目标观测状态结果融合成全局对目标观测状态的精确描述，提高了跟踪框架的跟踪精度。在 JTA 数据集上的实验结果证明了该框架性能的优越性和算法的有效性。相较于单源观测数据输入算法，本算法有了明显的性能提升，算法的跟踪精度和抗 ID switch 的能力都有一定增强。

5.2 研究展望

跨摄像头多目标跟踪技术是计算机视觉与系统控制领域里重要的研究课题之一，在实际运用中，兼顾算法的鲁棒性和实时性要求仍需大量研究。特别是当下的众多框架鲜有达到实时性需求，大多数是接近实时的处理速度。虽然本文框

架相较于已有的算法框架提高了算法速度达到接近实时的速度，同时算法也提供了较好的跟踪精度，但仍然有众多可以提高并进一步研究改进的地方：

(1) 由于 JDE 框架本身的特性，其在进行多目标跟踪任务时，ID switch 数量较高。本文提出的引入姿态描述的基于 JDE 多目标跟踪算法中，引入的姿态信息对提高跟踪框架抗 ID switch 能力有一定的效果，该方法减少了 JDE 框架的 ID switch 数量。但相较于多阶段式的跨摄像头多目标跟踪方法，其抗 ID switch 的能力仍显稚嫩，引入的姿态信息还有进一步研究与提升的空间。同时，JDE 框架本身只能提供含有 embedding 信息的检测结果，并没有将整体跟踪步骤嵌为一个框架，该思路仍有进一步研究的可能性。

(2) 本文在解决跨摄像头多目标跟踪问题时采用的是集中式系统的方法，融合了分布式摄像头网络中的多个节点检测信息，得到了较为精确的目标观测状态信息送入运动模型得到更精确的跟踪结果。该方法对已获得的近实时的性能损害较小，但在实际运用中该方法的鲁棒性较差，任意一个单独的节点发生故障，全局的检测数据结果的精度就存在受损的可能性。同时，中间节点的破坏会直接导致全局系统的故障。因此，将集中式的跨摄像头多目标跟踪框架改造成基于置信问题的分布式跨摄像头多目标跟踪框架不仅有必要跟有实际应用价值。

参考文献

[1] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]. 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[2] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015: 211-252.

[3] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Advances in neural information processing systems. 2012: 1097-1105.

[4] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[5] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[J]. ArXiv Preprint ArXiv, 2016:6517-6525.

[6] Lowe D G. Object recognition from local scale-invariant features[C]. Proceedings of the seventh IEEE international conference on computer vision. Ieee, 1999: 1150-1157.

[7] Novak C L, Shafer S A. Anatomy of a color histogram[C]. Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Champaign, IL, USA, 1992: 599-605.

[8] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, vol 86: 2278-2324.

[9] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012: 1097-1105.

[10] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[11] Simonyan, Karen, Andrew Zisserman. Very deep convolutional networks for large-scale image recognition[J/OL]. arXiv preprint arXiv:1409.1556 (2014).

[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[13] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[14] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015: 1904-1916.

[15] Girshick R. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[16] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. arXiv preprint arXiv:1506.01497.

[17] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[18] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[19] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]. European conference on computer vision. Springer, Cham, 2016: 21-37.

[20] Feng W., Hu Z., Wu W., et al. Multi-object tracking with multiple cues and switcher-aware classification[J/OL]. arXiv preprint arXiv:1901.06129(2019).

[21] Xu Y, Ban Y, Alameda-Pineda X, et al. Deepmot: a differentiable framework for training multiple object trackers[J/OL]. arXiv preprint arXiv:1906.06618 (2019).

[22] Zhang J, Zhou S, Wang J, et al. Frame-wise motion and appearance for real-time multiple object tracking[J/OL]. arXiv preprint arXiv:1905.02292 (2019).

[23] Chu P, Ling H. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[24] Xu J, Cao Y, Zhang Z, et al. Spatial-temporal relation networks for multi-object tracking[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[25] Voigtlaender P, Krause M, Osep A, et al. Mots: Multi-object tracking and segmentation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[26] Fu Z, Angelini F, Naqvi S M. Multi-level cooperative fusion of GM-PHD filters for online multiple human tracking[C]. IEEE Transactions on Multimedia, 21: 2277-2291.

[27] Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]. European conference on computer vision. Springer, Cham, 2016: 17-35.

[28] Lee S, Kim E Multiple object tracking via feature pyramid siamese networks[J]. IEEE access, 7: 8181-8194.

[29] Zhong Z, Zheng L, Cao D, et al. Re-ranking person re-identification with k-reciprocal encoding[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1318-1327.

[30] Liu W, Camps O, Sznaier M. Multi-camera multi-object tracking[J/OL]. arXiv preprint arXiv:1709.07065 (2017).

[31] Shen Y, Li H, Wang X. Deep group-shuffling random walk for person re-identification[C]. Proceedings of the IEEE conference on computer vision and pattern recognition 2018: 2265-2274.

[32] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.

[33] Ristani, E., Tomasi, C. Features for multi-target multi-camera tracking and re-identification[C]. Proceedings of the IEEE conference on computer vision and pattern recognition 2018: 6036-6046.

[34] Luo H, Gu Y, Liao X, et al. Bag of tricks and a strong baseline for deep person re-identification[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[35] Liu Z, Zhu J, Bu J, et al. A survey of human pose estimation: the body parts parsing based methods[J]. Journal of Visual Communication and Image Representation, 2015: 10-19.

[36] Toshev A., Szegedy C. Deeppose: Human pose estimation via deep neural networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition 2018: 1653-1660.

[37] Tompson J, Goroshin R, Jain A, et al. Efficient object localization using convolutional networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition 2015: 648-656.

[38] Carreira J, Agrawal P, Fragkiadaki K, et al. Human pose estimation with iterative error feedback[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[39] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]. European conference on computer vision. Springer, Cham, 2016.

[40] Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and tracking[C]. Proceedings of the European conference on computer vision (ECCV). 2018.

[41] Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]. European conference on computer vision. Springer, Cham, 2016: 17-35.

[42] Wang J, Sun K, Cheng T, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence (2020).

[43] Redmon J, Farhadi A. Yolov3: An incremental improvement[J/OL]. arXiv preprint arXiv:1804.02767(2018).

[44] Huang R, Pedoeem J, Chen C. YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers[C]. 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018: 2503-2510.

[45] Bochkovskiy A, Wang C, Liao H. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[46] Wang Z, Zheng L, Liu Y, et al. Towards real-time multi-object tracking[J/OL]. arXiv preprint arXiv:1909.12605, 2019, 2(3), 4.

[47] Kuhn H W. The Hungarian method for the assignment problem[J]. Naval research logistics quarterly 2.1-2 (1955): 83-97.

[48] Munkres J. Algorithms for the assignment and transportation problems[J]. Journal of the society for industrial and applied mathematics, 1957: 32-38.

[49] Cao Z, Hidalgo G, Simon T, et al. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields[J]. IEEE transactions on pattern analysis and machine intelligence, 2019: 172-186.

[50] Everingham M, Eslami S M A, Van Gool L, et al. The pascal visual object classes challenge: A retrospective[J]. International journal of computer vision, 2015: 98-136.

[51] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]. European conference on computer vision. Springer, Cham, 2014: 740-755.

[52] Fabbri M, Lanzi F, Calderara S, et al. Learning to detect and track visible and occluded body joints in a virtual world[C]. Proceedings of the European Conference on Computer Vision (ECCV). 2018: 430-446.

[53] Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 658-666.

[54] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]. Proceedings of the 2017 IEEE international conference on image processing (ICIP). IEEE, 2017: 3645-3649.

[55] Fang K, Xiang Y, Li X, et al. Recurrent autoregressive networks for online multi-object tracking[C]. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 466-475.

[56] Zhou Z, Xing J, Zhang M, et al. Online multi-target tracking with tensor-based high-order graph matching[C]. 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018: 1809-1814.

[57] Mahmoudi N, Ahadi S M, Rahmati M. Multi-target tracking using CNN-based features: CNNMTT[J]. Multimedia Tools and Applications, 2019: 7077-7096.

[58] Yu F, Li W, Li Q, et al. Poi: Multiple object tracking with high performance detection and appearance feature[C]. European Conference on Computer Vision. Springer, Cham, 2016: 36-42.

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

[60] Wang Z, Zheng L, Liu Y, et al. Towards real-time multi-object tracking[J]. arXiv preprint arXiv:1909.12605, 2019: 4.

致 谢

三年的研究生时光转眼即逝。三年前本科毕业来到昌航报道，如今要毕业离开校园，开始人生的新阶段。虽然未能走遍南昌的各个角落，这座城市依旧为我带来了无数美好的回忆。回首往事，在老师的关怀指导和同窗互助进步下，我要向大家致以真挚的感谢。

首先向我的导师储珺教授致以谢意，三年来储老师在科研和生活上的细致关怀，让我成长为合格的硕士研究生。储老师经常带领我参加课题相关领域的学术会议，从参加的每次会议中，我对研究课题相关领域知识储备得到丰富，同时找到了与优秀从业者的差距。也正是储老师的大力支持下，我才能得到出国交换的资格，开拓了我的视野。我在海外的时间，老师会定期与我沟通了解我的科研进展并关心我在海外的生活状况。特别是疫情发生后，老师特别关照了家在武汉的我，老师的关怀的让我倍感温暖。老师的认真负责才让我完善了本次论文。

其次，我要感谢课题组的老师和同窗同学，正是老师和同学的帮助，我才能远在海外依旧能够完成科研任务。同时我还需要感谢英国萨里大学的王文武教授，John 教授，Hinton 教授和前中科院声学研究员曹寅。王教授为我的研究方向提供了很多建设性思路并指导我在数据融合研究上的方向。John 教授和 Hinton 教授为我提供了授权，让我能够使用英国排名第一欧洲排名第三的的信号处理研究中心（CVSSP）的实验设备。曹寅研究员的指导让我熟悉并熟练操控研究中心的超算设备。

最后，感谢我的父母和家人，感谢二十五年来的照顾与关爱，是你们的支持，才让我能顺利完成硕士学业并远赴海外进修博士学位。

南昌航空大学硕士学位论文原创性声明

本人郑重声明：所呈交的硕士学位论文，是我个人在导师指导下，在南昌航空大学攻读硕士学位期间独立进行研究工作所取得的成果。尽我所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中作了明确地说明并表示了谢意。本声明的法律结果将完全由本人承担。

签名：吴沛沛 日期：2021 / 06 / 03

南昌航空大学硕士学位论文使用授权书

本论文的研究成果归南昌航空大学所有，本论文的研究内容不得以其它单位的名义发表。本人完全了解南昌航空大学关于保存、使用学位论文的规定，同意学校保留并向有关部门送交论文的复印件和电子版本，允许论文被查阅和借阅。本人授权南昌航空大学，可以采用影印、缩印或其他复制手段保存论文，可以公布论文的全部或部分内容。同时授权中国知网、中国科学技术信息研究所将本学位论文收录到《中国学位论文全文数据库》，并通过网络向社会公众提供信息服务。

（保密的学位论文在解密后适用本授权书）

签名：吴沛沛 导师签名：[illegible] 日期：2021 / 06 / 03