单位代码：10166

沈阳师范大学

硕士学位论文

基于卷积神经网络的车站人流量检测研究

论文作者：张依林

学科专业：计算机应用技术

指导教师：王学颖 教授

培养单位：科信软件学院

培养类别：全日制

完成时间：2022 年 3 月 18 日

沈阳师范大学学位评定委员会

编　号：

类别	全日制研究生	√
	教育硕士	
	同等学力	

沈阳师范大学

硕士学位论文

题　目：基于卷积神经网络的车站人流量检测研究

所　在　院　系：科信软件学院

专　业　名　称：计算机应用技术

指　导　教　师：王学颖 教授

研　　究　　生：张依林

完　成　时　间：2022 年 3 月

沈阳师范大学研究生处制

学位论文独创性声明

本人所呈交的学位论文是在导师的指导下取得的研究成果。据我所知，除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示了谢意。

作者签名：张依林 日期：2022年5月20日

学位论文使用授权声明

本人授权沈阳师范大学研究生处，将本人硕士学位论文的全部或部分内容编入有关数据库进行检索；有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版，允许论文被查阅和借阅；有权可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。保密的学位论文在解密后适用本规定。

作者签名：张依林 日期：2022年5月20日

基于卷积神经网络的车站人流量检测研究

中文摘要

高铁因为它的高安全性和高时效性，使得其成为几乎每个人首选的出行方式。随着高铁的普及率逐年上升，国内自主研发的新技术不断应用，每个车站的人流量也在逐年递增。本文进行研究与实验的目的是基于深度学习和目标跟踪算法在车站场景下进行人流量检测与统计。卷积神经网络(CNN)和视觉目标跟踪(Visual Object Tracking)也是近年来计算机视觉领域受到广泛关注和研究的方向。其中 YOLOX 目标检测算法是在 YOLO 系列的基础上吸收近年来目标检测的最新成果。但进行视觉跟踪的核心目的通常集中在以下两点，一是对物体在后继视频序列中的行动轨迹作出有效估计，二是明确物体在后继视频序列中的活动状态，并以此收集的信息数据为基础来剖析物体语义的深层内容。

立足于计算机视觉领域，以该领域中目标检测方向的最新研究成果，即 YOLOX 为导向，着重探索目标跟踪方向中的多目标跟踪板块，同时在研究过程中以简单的在线和实时深度关联度量跟踪(Deep SORT)为代表的多目标跟踪策略为基础，并将其嵌入到实际的跟踪任务中。

本文以传统目标检测算法为研究出发点，详细的介绍并分析了目前主流的二阶段和一阶段目标检测算法，而后又将传统目标跟踪算法以及多目标跟踪算法两大类别进行对比分析，为进行以 YOLOX 与 Deep SORT 为基础的人流量统计算法研究做好了充分的理论支持与铺垫。在阐述完 YOLOX 目标检测算法和 Deep SORT 目标跟踪算法的基本原则和流程后，对整个系统的最后一步即人流计数算法进行了介绍，最后将人流量统计算法在 MOT16 数据集上进行了测试，并在沈阳北站进行了视频采样，将人流量统计算法应用于实际的车站场景。

本文充分考虑目标检测算法所具备的特性，修改了 YOLOX 模型，提出了基于 CBAM 注意力机制的 YOLOX-AM 目标检测模型。率先在人流计数统计系统中混合应用了以下两种模式，一是 YOLOX-AM 目标检测算法，二是 Deep SORT 跟踪器，由于上述两种模式的嵌入，算法跟踪能够因此成功作用到行人多目标跟踪数据集 MOT16 上，计算了 12 种评价指标。同时在 MOT16 数据集上实现了人流计数功能，并应用于车站场景。经过实验验证与分析，基于 YOLOX-AM 与 Deep SORT 的人流量统计算法在车站环境的实际应用上取得了良好的效果，体现了本系统较好的鲁棒性，具有实际应用价值并可以推广到诸如机场、学校、商场等应用场景。

关键词：目标检测；YOLOX；Deep SORT；多目标跟踪；卷积神经网络；人流计数

Research on Station Passenger Flow Detection Based on Convolutional Neural Network

Abstract

Because of its high safety and high timeliness, high-speed rail has become the preferred travel mode for almost everyone. With the increasing popularity of high-speed rail and the continuous application of new technologies independently developed in China, the passenger flow of each station is also increasing year by year. The purpose of research and experiment in this paper is to detect and count the passenger flow in the station scene based on deep learning and target tracking algorithm. Convolutional neural network (CNN) and visual object tracking are also widely concerned and studied in the field of computer vision in recent years. YOLOX target detection algorithm absorbs the latest achievements of target detection in recent years on the basis of YOLO series. However, the core purpose of visual tracking usually focuses on the following two points: one is to effectively estimate the action trajectory of the object in the subsequent video sequence; the other is to clarify the activity state of the object in the subsequent video sequence, and analyze the deep content of object semantics based on the collected information data.

Based on the field of computer vision and guided by the latest research achievement in the direction of target detection, namely YOLOX, focuses on the multi-target tracking section in the direction of target tracking. At the same time, in the research process, it is based on the multi-target tracking strategy represented by simple online and real-time Deep SORT, And embed it into the actual tracking task.

This thesis makes a detailed analysis of the current two-stage target tracking algorithm, which is based on the research of the traditional target tracking algorithm and the statistical algorithm of Deep SORT and YOLOX, and then makes a detailed analysis of the current two-stage target tracking algorithm. After expounding the basic principles and processes of YOLOX target detection algorithm and Deep SORT target tracking algorithm, the last step of the whole system, namely the pedestrian flow counting algorithm, is introduced. Finally, the pedestrian flow statistical algorithm is tested on MOT16 data set, and the video sampling is carried out in Shenyang north station, and the pedestrian flow statistical algorithm is applied to the actual station scene.

In this thesis, the characteristics of target detection algorithm are fully considered, the YOLOX model is modified, and a YOLOX-AM target detection model based on CBAM attention mechanism is proposed. Firstly, the following two modes are mixed and applied in the pedestrian flow counting and statistics system. One is the YOLOX-AM target detection algorithm, and the other is the Deep SORT tracker, Due to the embedding of the above two modes, the tracking

algorithm can be successfully applied to the pedestrian multi-target tracking dataset MOT16, and 12 evaluation indexes are calculated. At the same time, the pedestrian flow counting function is realized on the MOT16 data set and applied to the station scene. Through experimental verification and analysis, the passenger flow statistical algorithm based on YOLOX-AM and Deep SORT has achieved good results in the practical application of station environment, which reflects the good robustness of the system, has practical application value, and can be extended to application scenarios such as airports, schools, shopping malls and etc.

Keywords: Target detection，YOLOX，Deep SORT，Multi target tracking，Convolutional Neural Network，Flow count

目　　录

第 1 章 绪　论

1.1 课题研究的背景及意义

1.1.1 研究背景

近年来，计算机视觉领域中的研究学者将关注点更多的集中于行人流量的统计上，这一统计在车站、学校或者商城等公共场所的应用价值尤为明显，核心原理在于以计算机视觉算法为桥梁，收集行人流量的相关数据，并以此为基础进行全面分析。同时，此类分析结果正是管理者进行决策或制定方案的客观基础，而实现行人流量统计的关键点是对以下三个方面的研究，一是目标检测，二是目标跟踪，三是目标计数。

在计算机视觉领域中，视觉跟踪属于数字图像处理方向的研究内容，其不仅能够为后期开展高级视觉任务提供数据基础，同时也是使得高级视觉任务能够顺利开展的前提。在分析交通道路状况、理解场景信息以及估计行人流量等传统应用方面，跟踪技术均已起到重要作用；而其在机器人导航等现代化高精尖技术行业中的应用价值也日益凸显。目标跟踪技术的应用是十分广泛的，包括但不限于智慧生活以及远程医疗等各个产业中。此外，跟踪在其他视觉技术的算法中能够起到辅助效果，应用价值较高，跟踪技术的引入能够在某种程度上促进其他任务的完成。其中，较为经典的是其能够直接作用于视频或语义连续的图像中，并在该类图像中精确捕捉目标所处的空间位置，降低了系统运行的复杂繁琐性。现有数据表明，如果能够实现对跟踪算法性能研究的进一步飞跃，将显著提高其在视频分析等领域中的应用价值。特别是在科学技术成熟度不断突破的今天，企业逐步强化了其在商业方面的实践能力，客观展现了跟踪技术在视觉应用环节将具备长效的发展前景。

在传统行人流量统计方式上，通常以人工计数或者传感检测为主，其与现代行人流量统计方式相较而言，缺点是显而易见的。在人工统计模式下，既需要庞大的劳动力群体，同时统计效率也难以得到保证，一旦行人流量过于集中，就会因存在人群粘连等问题导致统计误差的形成。传感检测尽管能够突破人工统计模式的部分弊端，但其实践应用过程会受到场所或者技术的约束，对行人流量的估计较为粗略，无法获取较为精确的统计结果，因此不能满足现代视频监控的运行标准。但以智能视频监控为基础的行人流量统计方式能够对行人流量进行较为精准的判定，包括人员流量以及行人运动状态等多个维度，并且该种方式不受运行时间的束缚，能够实现无间断且无需人力的持续性作业，已成为当前学术界研发课题中的首选方向之一。

1.1.2 研究意义

现如今，公共安全事件是频频影响和危害社会正常发展和进步的主要事件。一旦发生地震或海啸等重大危害公共安全事件，地区的公共设施等将出现相当大的混乱。人们通过怎样的路径从建筑物等逃离变得很重要。在人多的大型设施和群众聚集的场所，比如火车站和机场，为了避免混乱，有必要事先预测拥挤状况，并通过适当的诱导方法来缓解拥挤。

为了测量如此大规模的人流，作为测量基础的人物检测技术也是必要的。就针对区域内人流密度统计及控制的方式而言，目前应用价值最高的便是以机器视觉为基础的人流量统计，该种方式能够避免因单位时间内人流过大等问题造成的统计误差。本文的研究重点讨论在高铁车站测量及监测的数据。

近年来，如何正确应对并处理计算机视觉问题已成为各界的关注重点，而深度学习的出现成为破解该类问题的首选框架，同时提升了多目标跟踪(Multi-Object Tracking，MOT)的算法精度，这一点在深度神经网络的目标检测能力中得到了强有力的证实，几乎已经引领了多目标跟踪的性能趋势[1],这也可以从侧面反映了 MOT 目前存在的很多常见的问题，给 MOT 的未来发展指明了大方向。

目标检测器的性能在很大程度上决定了 MOT 方法的性能，较常见的 TBD(Tracking by Decetion)模式，该模式在提供检测目标的过程中是以每帧为基础进行的，这也是其推进跟踪进程的方式。在跟踪算法一致且所提供的检测集的目标检测器存在差异时，会使得由此所形成的跟踪结果存在较大差异性，判定检测结果优劣的标准主要是依据呈现出的趋势状况，跟踪结果应尽可能确保高精确度。如果 MOT 算法没有将检测模块产生的影响纳入考虑范畴，那么此时所得到的测试结果也是不具有可靠性的，往往与真实情况存在明显偏差。

由于卷积神经网络在特征提取方面的能力极为优越，因此当前绝大多数的多目标跟踪算法均将卷积神经网络界定为主流网络框架。但卷积神经网络的劣势也是较为突出的，其对于帧间的时间连续性信息存储不稳定，失帧现象时有发生，一旦目标出现目标遮挡或者形变等跟踪领域普遍场景，则会降低原有的鲁棒性。所以，多目标跟踪效果的提升仍需要构造性能更强大、复杂度更低的网络框架。尽管计算机视觉技术的层次是多样的，但当前多目标跟踪领域的核心发展方向是明确的，即对在层次上存在差异的计算机视觉任务展开优势互补，并在此基础上以反向逻辑方式作用到多目标跟踪问题的解决上。

1.2 国内外研究现状

1.2.1 人流量统计算法研究现状

开展行人流量统计的传统方式一般是通过人工劳动力进行的，而后随着科学技术的进步，以红外线感应为代表的传感设备逐渐被应用到计数领域。当前，视频监控在社会生活领域中应用程度已经得到广泛普及，以智能视频监控为基础的行人流量统计的优势也更加明显，其在压缩成本的同时能够实现高精确度的探测，并且整个过程的完成能够独立运行，无需借助其他设备的辅助。因此诸多学者对该课题进行了深入的研究，成功探索出具备行人计数性能的智能化监控设备。

就行人流量统计理论研究的起始点而言，国外要更具有前瞻性，包括麻省理工在内的众多高校以及企业或研究所均该领域进行了相关研究与实验，在部分重大国际会议上发表了众多成果[2]。1997 年，Carnegie Mellon 大学联合 MIT 公司正式推出 VSAM 系统，该系统的运行是以智能化视频监控以及分析技术为基础的，在军事以及民事中的应用最为常见，

这就大大压缩了对传统大规模人力成本的需要[3]。2005 年，日本 NEC 公司正式研发出 Smart Catch 智能视频监控处理系统解决方案，该方案在实践过程中发挥了关键作用，成功识别出旧金山国际机场所存在的安全隐患问题。英国雷丁大学在借鉴经验的同时就 VIEWS 项目进行了进一步探索，成功提升了识别跟踪技术在交通行人车辆的实践环节的价值，对维护交通安全及秩序起到了根本性的作用。

当前，受全球学者一致认可的核心课题之一是如何科学借助智能视频分析算法来有效统计行人流量。2005 年，Viola[4]从多角度对行人运动特征进行了集中性提取，依次作为开展人数统计工作的前提。2006 年，Antonin[5]将检测以及跟踪算法进行充分结合，将轨迹分析思路运用到了行人计数算法领域。2010 年，王强[6]等学者在研究行人计数方法时，融入了对颜色信息以及形状信息等因素的考虑，提高了单位时间内行人统计的效率。在当前计算机视觉技术不断突破的情况下，将机器学习算法嵌入到对行人流量统计的中已经成为现实。2015 年，Wang[7]等学者研究出以深度卷积神经网络回归模型为基础的统计方式，该方式能够计算图像中中的人数，而与此同时，Li[8]等学者研究出以人头检测以及跟踪为前提的统计方式，其能够避免由行人遮挡而产生的统计误差。周治平[9]等学者开创性地提出了一种新型人数统计算法，该算法的特点在于同时应用了检测以及特征回归的双重算法。

国内学者在行人流量统计方面的研究。其实点要明显晚于国外，但近年来随着科学技术的发展及国家政策的大力支持，对于智慧交通及智慧城市的追求也将成为当前社会发展的共识，这意味着对智能视频监控的应用将更为普遍，因此国内学者对该领域的研究热度也明显上升。以清华大学为代表的高校对行人流量统计进行了最初的研究，例如，清华大学设立了智能图文信息处理实验室，北京航空航天大学针对智能交通以及公共场所中的行人流量统计方式进行了深入改进[10-11]。

将国内与国外研究进程对比来看，国内的研究相对落后，以背景建模等传统算法为主，但是当前国家政策为行人流量统计发展注入了强劲的推动力，研究学者将更多的时间与精力投入到了此领域中，进一步带动了国内学者对行人流量统计技术的探究。

1.2.2 目标跟踪算法研究现状

通过开展视觉目标跟踪任务，即使是在一组相对复杂的图像序列中，也可以以较短的时间完成对既定目标的锁定。从任务内容的比较分析角度看，视觉目标跟踪和该领域中其他方式具有很大的相似性，均采取了相机摄像头来进行目标搜索，并对目标及其相关的数字图像信息进行持续跟踪与捕捉，最终借助人工智能来对上述搜寻结果进行解析及处理，使现代科技逐步取代了人工在目标跟踪中的作用。需要注意的是，跟踪目标是多样的，不仅能够对视频图像序列中目标运动信息的预测，同时能够获取时空信息等，从深层次的角度确保对目标运动状态的把控，为物体语义分析过程奠定可靠的数据前提。

跟踪的原始任务需求往往集中于单一运动目标，也就是所谓的单目标跟踪。在开展主流算法研究的过程中，大多是以滤波以及孪生网络为依据。最小输出误差平方和滤波器

(Minimum Output Sum of Squared Error Filter, MOSSE)的研发显著提高了跟踪的速度功能，其将信号处理中所运用的滤波技术以及跟踪领域中的相关问题纳入了同一考虑范畴[12]。不论是核函数逐点循环跟踪(Circulant Structure with Kernels, CSK)，还是核相关滤波器(Kemelized Correlation Filters, KCF)[13]，二者在进行算法口向的完善时均是以最小输出误差平方和滤波器为基础的。其中，前者能够在解决跟踪问题的过程中与核方法进行结合，而后者能够借助循环矩阵的嵌入提升跟踪速度与效率。与同期其他跟踪算法的不同在于，以滤波为基础的算法在鲁棒性方面表现得更为优越，兼具较高算法效率以及精度的特点，但其难以适应复杂场景的作业需求，仍存在发生漂移现象的可能性，因此这也使得有关滤波跟踪的性能受到了限制。虽然能够通过大边距循环跟踪(Large Margin with Circulant Feature, LMCF)等方法提高作业的精确度，但一旦采用该类方法，运行速度则无法得到保证，甚至会低于每秒 1 帧的频率，使有关滤波此前所具备的高速性能优势随之弱化。

对于回归深度学习框架来说，其在一定程度上促进了计算速度的提升，算法的实用性也随之进一步得到强化，这一特点在以孪生神经网络为构成原理的深度学习算法中，得到了诠释。全卷积暹罗网络 (Fully-Convolutional Siamese Networks, SiamFC)在衡量两个输入的相近程度时，主要是借助了权重一致的两个同框架网络进行的，通过深度卷积神经网络 AlexNeW 的 SiamFC[14]，把端到端形式的相关滤波和深度学习进行融合，并将其融到了跟踪领域的实践应用中。在优化以孪生思想为基础的算法精度时，暹罗网络(Siamese Region Proposal Network, SiamRPN)发挥了重要作用，其能够对目标的尺度以及位置进行先行预判，进而开展预测跟踪轨迹的任务。顾名思义，干扰感知暹罗网络(Distractor-aware Siamese Networks, DaSiamRPN)能够实现对同类目标的鉴别，因为其具有较为完备的数据支撑，特别是在视觉目标跟踪挑战(Visual Object Tracking, VOT)方面的性能更为优越。进化暹罗网络的视觉跟踪算法(Evolution of Siamese Visual Tracking, SiamRPN++)在 2015 在线目标跟踪评估基准(Online Object Tracking Benchmark 2015, OTB2015)上名列前茅，该种算法嵌入了多种网络结构，包括深度残差网络 (Residual Network, ResNet)以及区域候选网络(Region Proposal Networks, RPN)等类型。

1.2.3 目标检测算法研究现状

行人检测算法的目的是针对当前帧而言的，其需要在此范围内捕捉到所有行人的位置。基于特征在类型方面差异性，行人检测主要被划分为两大类别，一是以浅层机器学习为基础的检测算法，二是以深度学习为基础的检测算法。二者的区别在于，以浅层机器学习为基础的检测算法在对特征进行描述时，主要是依据行人静态以及动态等特征进行的，需要借助滑动窗口对其进行持续提取，并需要通过不同的分类器对上述特征展开分类；而以深度学习为基础的检测算法具备深层神经网络结构，行人检测过程的顺利开展，不仅需要依靠大量的真实样本，同时需要进行多次反复操作。

浅层机器学习检测方式被划分为以下两大类别，一是与全局特征为基础的，二是以部

件检测为基础的。前者在表达行人外观的信息时，不仅需要采用合适的特征，同时需要借助分类器的固有功能。2004 年，Viola 率先研究出 Haar 特征，并在融入积分图以及软级联的策略的同时，将此特征与 Adaboost 分类器的分类功能进行全面融合，使得原有检测效率实现了新的突破；2005 年，Dalal[15]在描述目标的过程中，借助了方向梯度直方图(Histogram of Oriented Gradient，HOG)的方式，并在目标分类环节采用了支持向量机(Support Vecor Machine，SVM)，此种算法模式的综合性较强，受到了学术界诸多学者的持续关注，为后期的算法改进提供了诸多可借鉴之处；2007 年，Sabzmeydani[16]等学者研究出 shapelet 特征，并将此特征与 Adaboost 分类器的分类功能进行全面融合，借助图像中所形成的曲线信息内容，能够开展人体边缘检测，极大的提升了算法的应用效率。

全局特征检测方法存在固有弊端，例如其无法避免因行人遮挡而产生的检测误差，但以部件为基础的检测方法的出现弥补了这一不足。2008 年，Felzenszwalb[17]等学者通过长期探索提出了一种新型的行人检测方法，并最终将其命名为 DPM，这一新型行人检测方法的检测逻辑是以 HOG 特征为基础的，结合了部件模板以及全局模板的双重特性，并在操作 SVM 分类器的过程中嵌入了难例样本挖掘算法。DPM 的准确程度不言而喻，其在 PASCAL VOC 行人检测比赛中曾夺得头名，加大了其在行人检测等应用领域的说服力。除此之外，由于 DPM 的精准度较高，因此在后来很长一段时间内出现的改进算法均是以 DPM 为基础的，该类算法凭借其独有的优势在行人检测比赛中连续多年拔得头筹。在以深度学习为基础的行人检测算法中，由于组成结构存在明显不同，一般会将其划分为多个类别，比较经典的有以自编码机(Auto encoder，AE)为基础的检测方法以及以限制玻尔兹曼机(Restricted Boltzmann machine，RBM)为基础的检测方法。此外，卷积神经网络结构能够满足图像处理的需要，因此绝大多数以深度学习为基础的行人检测往往是基于此特征来探究检测算法的形成的。

通过对以上两大类别的检测算法进行对比分析可知，以深度学习为特征的行人检测算法需要深层网络的支撑，在多次重复训练的前提下获取目标信息的抽象特征，进而提高检测精度，然而，由于此类算法受到检测速度的限制，往往难以在时间作业过程中得到广泛推广及应用。这就意味着，如何在提升算法运算效率的同时提高精度是当前学术界亟需破解的难题。

1.3 论文主要的研究内容

本文的研究内容是立足于多目标跟踪(Multi-object Tracking, MOT)领域的[18]。社会发展的更新迭代，人们对安全以及效率有了更为卓越的追求，计算机视觉任务也必须随之破除传统的模式，视觉追踪技术的有效应用使得多目标跟踪技术有了现实需求的必要性，较为常见的有视频监控以及自动驾驶等，在此类活动中所涵盖的活动轨迹通常是多个目标的，但在多数情况下借助单目标跟踪 (Single-object Tracking, SOT)尚不能达到既定任务的完成标准。这是由于目标跟踪过程并不是间断独立的，必须以双向的形式在图像前后目标之间

建立联系，若检测算法的应用仅局限于单个帧中，则此时能够完成识别任务，但彼此间的关联建立过程将中断；若借助单目标跟踪算法，则彼此间的关联建立过程可以持续进行，却无法起到识别作用。MOT 破解的核心瓶颈便是新旧目标的更换以及身份识别问题，此类跟踪算法的性能主要受到两个因素的影响，一是识别方法，二是关联策略。本文的研究内容集中于借助 MOT 来区分多个目标的方法以及 MOT 关联相同目标的策略上，具体应用领域以行人监控以及车辆驾驶两大方向为主，能够对安全防护产品中应用的 MOT 技术起到示范作用。

本文的研究细节可以划分以下三个层面，一是以 YOLOX 为基础的 DeepSORT 多目标跟踪算法，二是该算法在行人方面的跟踪效果，三是如何将该算法应用到实践过程。将深度学习方法和传统方法进行对比分析可知，前者在识别以及跟踪行人等层面不仅具有速度上的优势，同时精度也更为准确。在数据之间建立有效的关联是制约当前目标跟踪算法顺利执行的核心难点。检测与跟踪过程不同，前者需要针对图像中的物体进行内容识别。因此，当前绝大部分目标跟踪算法的建立均是以 TBD 为基础的，将检测结果视为基础信息，以此来克服多物体之间互相遮挡等问题的困扰，实现目标的一致性。Deep SORT 是多目标跟踪中应用频率较高的一类 TBD 方法，不仅能够在实现较高的精准度，同时能够在任意时点持续作业。就 You Only Look Once (YOLO)系列算法而言，其属于一阶段目标检测算法，以锚点为中介桥梁，实现了分类和目标定位的回归问题间的结合，进而构建出一款成熟度极佳的模型，即使在骨干网络缺失的情况下，也能通过其他框架的补位继续运作。

1.4 论文内容及组织安排

本论文设计了一套基于 YOLOX 和 Deep SORT 的人流量统计系统，达到在车站高铁站进行准确人流量统计的实际需求。全文涵盖下述几点内容：

第 1 章总结概述目标检测的研究背景，梳理全球有关行人流量统计系统的基本发展情况，明确本研究的具体内容以及行文结构。

第 2 章对本研究中运用到的以及出现的相关技术逐一进行介绍，比较了传统的目标跟踪算法和多目标跟踪算法的利弊，为下面的算法研究和实验做好铺垫。

第 3 章介绍了 YOLOX 目标检测算法和 Deep SORT 目标追踪算法，并对本研究所用到的人流计数算法进行了详细的论述。

第 4 章展示了实验结果以及对实验结果进行分析。介绍了 MOT16 数据集和 CrowdHuman 数据集，展示了 YOLOX 和 Deep SORT 在训练集上训练的数据，同时在 MOT16 数据集上实现了人流计数功能，并成功应用于实际高铁站场景。经过实验验证，基于 YOLOX 与 Deep SORT 的人流量统计算法在车站环境的实际应用上取得了理想的效果。

第 5 章对全文做了总结和展望，并提出了下一步对系统进行优化和改进的设想。

第 2 章 相关技术

2.1 目标检测算法

本章阐述了传统目标检测算法、深度学习目标检测算法以及传统目标追踪算法和多目标追踪算法。其中，深度学习目标检测算法涵盖二阶段及一阶段两类算法。

2.1.1 传统目标检测算法

传统的目标检测通常是借助滑动窗口框架进行的，具体步骤如下：

(1) 以滑动窗口尺寸的不同为划分依据，分别框住图中的某一位置，并将其界定为候选区域；

(2) 对候选区域中与视觉有关的特征进行收集提取。例如，在人脸检测环节所需的 Harr 特征等；

(3) 借助分类器的固有功能开展识别操作，例如，较为经典的 SVM 模型。

在传统目标检测环节中，多尺度形变部件模型 DPM（Deformable Part Model）[19]的优势尤为明显。DPM 将物体视为由多个部分共同构成的部件，通过对部件间的关系的充分分析来描述物体，这一特征与自然界中诸多物体的非刚体特征是具有较大相似性的。DPM 在某种程度上可以说是 HOG+SVM 的延伸，因为其借鉴了 HOG 以及 SVM 的优势，在进行人脸检测等任务上均具有良好的检测效果，然而 DPM 复杂度是偏高的，整体检测速度落后，因此改进方法的出现是大势所趋。在学术界致力于探索如何优化 DPM 性能时，以深度学习为基础的目标检测正式被提出，在受青睐程度上也一度超越了 DPM，大多数此前主攻传统目标检测算法的学者将逐渐转变研究方向，向深度学习领域发展。

2.1.2 二阶段目标检测算法

1. RCNN 算法

R-CNN 的全称是 Region-CNN，是首个将深度学习成功应用到目标检测上的算法。R-CNN 基于卷积神经网络(CNN)、支持向量机(SVM)[20]和线性回归等算法，实现目标检测技术。

RCNN 的过程分 4 个阶段：

(1) 候选区域提出阶段（Proposal）：采用 selective-search 方法，从一幅图像生成 1K~2K 个候选区域；

(2) 特征提取：针对单个候选区域依次采用 CNN 开展特征提取；

(3) 分类：分别将单个候选区域的特征依次放置于分类器 SVM 中，获取相应的分类结果；

(4) 回归：将收集的候选区域特征放置于回归器中，进而获取 bbox 的修正量。

2. Faster-RCNN

Fast R-CNN 算法在进行改进的过程中，融入了 R-CNN 算法以及 SPPNet 算法[21]的原有优势。其通过将整张图像作为输入的方式来达到对特征提取的目的，并且开创性地将感兴趣区域池化层(ROI) [22]这一概念引入到了 算法的执行中。ROI 层以尺寸具有差异的提议区域为目标，从中提取固定尺寸的特征，并将其视为全连接层的输入，从而展开分类以及回归操作。在本质上，Fast R-CNN 与 R-CNN 最大的区别在于，后者是针对单个提议区域分别执行卷积操作，但前者是采取直接的方式以整个图像为基础执行卷积操作，同时借助 ROI 池化以及映射关系得到尺寸相同特征，这一运行 原理避免了诸多重复性的计算环节。根据 PASCAL VOC 2007 数据集上的实验分析结果可知：R-CNN 的平均精确度(mAP)仅为 58.5%，而 Fast R-CNN 则直接提升到了 70.0%，在运行速度方面实现了质的飞跃。在多任务中，虽然 Fast R-CNN 可以满足端到端的操作需求，但缺点是在实践操作环节，作用于形成候选区域的选择性搜索算法的运行需要耗费大量时间，必须借助高效性更强的候选框提议方式进行替代。

就单一检测任务而言，其应完成对具体分类结果及其相关定位信息的输出。检测的输出信息记为［x,y,w,h,c］，各字母依次代表不同的含义，其中 x、y 代表目标的中心坐标偏移量，w、h 代表宽高的偏移量，c 则代表检测目标所属类别的可能性值。目前，较为流行的检测算法包括两类，一是二阶段检测算法(Two-stage detection)，二是一阶段检测算法(One-stage detection)，例如，YOLO[23-25]等。

在区别一阶段和二阶段检测算法时，需要对其初步定位步骤的特点进行判别。而且段检测算法的执行应当以定位的完成为前提，以此来实现对定位结构及分类结果的输出，此时所得到的候选框在形状上是具有差异性的，因此在一段时间内进行持续性分析时，必须借助检测网络模型所输出坐标偏移信息来开展，精准定位候选框的所处位置，最后针对分数最高的回归结果通过分类网络完成分类以及计算。

Faster-RCNN 属于二阶段检测算法，当前的应用频率较高，其在训练阶段以及测试阶段所需要的时间也是较短的。通过图 2-1 可知 Faster-RCNN 网络结构[26]的整体概况。网络结构的构成部分主要有以下四个板块，一是 CNN 特征提取网络、二是 RPN 网络、三是 ROI Pooling、四是学习模型。

(1) 卷积神经网络特征提取

Faster-RCNN 的输入环节会提供一幅不限定尺寸的 P×Q 输入图像，第一步需要将图像大小缩放至 M×N，即特征提取网络的输入。本文采取深度卷积网络(Visual Geometry Group-16, VGG16)[27-29]提取特征，借助 13 个以修正线性单元(Rectified Linear Unit,Relu)为激活函数的激活层，按照一定的顺序逐步深入到 4 个池化层(Pooling)，进而导出任务所需的特征图(Feature Map, FM)。这一过程中，输入图像的尺寸会发生改变，为 RQ 的 1/16。

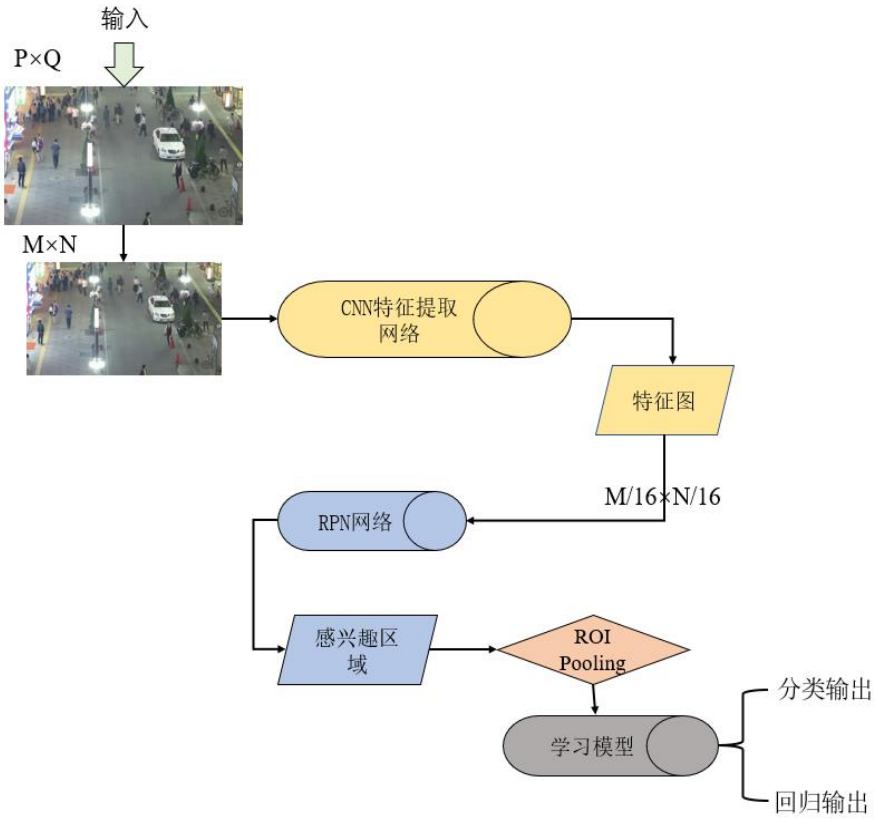


图 2-1 Faster-RCNN 网络结构图

(2) RPN

Faster-RCNN 提出了 RPN 网络结构提取候选区域，具体结构如图 2-2 所示。需要注意的是，RPN 的输入是其特征提取网络输出的特征图像。就特征图像的尺寸而言，是图像原来尺寸的 1/16，针对其进行 3×3 卷积处理，然后送入回归分支以及分类分支的运行中。

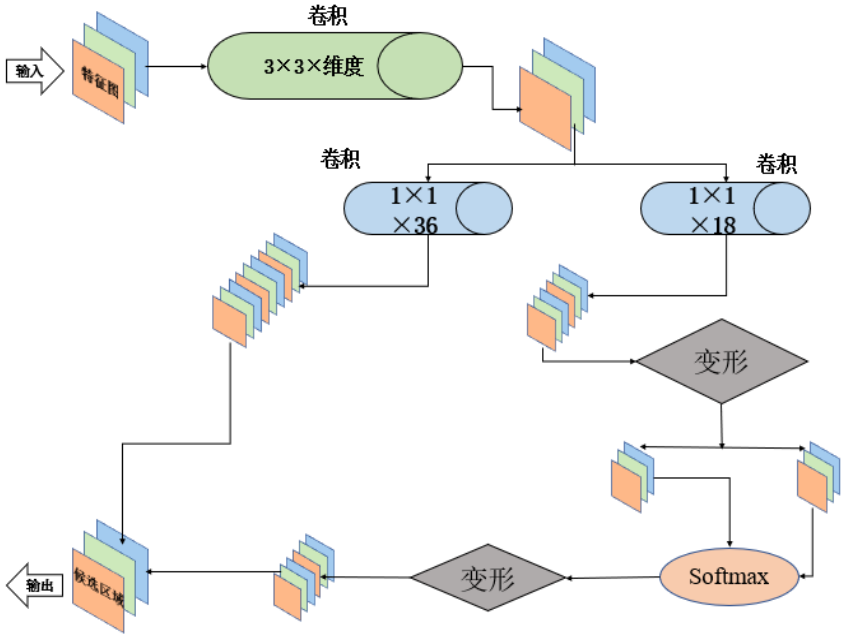


图 2-2 RPN 结构图

卷积在回归分支中会对通道数量进行更改，最终输出的通道数量是 36；而 1×1 卷积在分类分支中也会对通道数量进行更改，最终输出的通道数量是 18，并将这 18 个通道进行平均拆分为 2 个特征图，借助逻辑回归模型 SoftmaxOT[30-31]对特征图展开前景以及后景两个层面的类别划分，然后会将其融合为初始的 18 通道，同时输出特征图。此时，特征图上的信息主要涵盖以下三类，一是回归坐标偏移、二是前后景的类别划分概率、三是单个像素点在通过映射方式到原图所形成的 9 个锚点时所基于的特征信息网。

(3) ROI Pooling 尺寸归一化

Faster-RCNN 在本质上，是二阶段算法区域卷积神经网络(Region Convolutional Neural Networks, RCNN)系列[32]，因此必须形成所谓的感兴趣区域 (Region of Interest, ROI)。在原图完成 2.1.1 节(1)中所示的卷积网络操作后，能够得到所需的特征图，此时通过选择搜索或者 RPN 算法等桥梁的嵌入能够获得不同的目标候选框。本文借助 Faster-RCNN 来形成 ROI 候选框，其中 Faster-RCNN 是以 RPN 为基础的。因为候选区域样本尺寸会受到九种 Anchor 设置情况的影响，因此 Faster-RCNN 提出，在引入到训练模型时，必须借助 ROI Pooling 对候选区域样本尺寸展开标准化处理。就单个 ROI 而言，其相应的原始坐标以及大小都是独有的。如果尺寸为 10×14，那么必须借助 7×7 的尺寸将其嵌入到模型中，从而确保后续训练有序展开。第一步需要针对 RPN 所提取的特征图，按照比例划分成标准为 7×7 块的区域，若出现无法整除的块数，则将其将直接填充到最后的块中，就单个区域进行最大池化处理，从而为后续采样奠定基础。此外，最终所得的图像在尺寸以上应该具有一致性，为 7×7 的归一化图像。

2.1.3 一阶段目标检测算法

一阶段目标检测算法主要包涵了 SSD 以及 YOLO 系列，这一小节中介绍的是其中非常具有代表性的 YOLO 系列中的 YOLOv3 算法。

YOLO 的全称是“You Only Look Once”，取义为对精度及速度的双向追求，在某种程度上可以理解为对精度以及速度两方面的折中选择。YOLOv3 的运行原理吸收了 YOLOv1 以及 YOLOv2 的潜在优势，在确保不丧失 YOLO 家族速度的基础上优化检测精度，在目标物体的形态选择上以小型为主。YOLOv3 算法[33-35]是借助单一神经网络进行的，能够把图像分类为不同区域，进而实现对边界框以及不同区域概率的预判。

YOLOv3 的运行只借助了卷积层这一单一层面，属于全卷积网络(Fully Connected Network, FCN)[36]。YOLOv3 所借助的特征提取网络是新型的, 即图 2-3 中所体现的 Darknet-53。从命名表面含义可知，其内部具有 53 个卷积层，并且单个卷积层后面附着了批归一化(Batch Normalization)层和 leaky ReLU 激活层[37-38]。在没有池化层的情况下，则要借助步幅大小为 2 的卷积层进行替代，只有这样才具备对特征图进行采样的条件，进而规避因池化层缺失而弱化低层级特征的现象。

当在系统内输入$(m, 416, 416, 3)$时，随之输出的是具有识别类的边界框列表，并且单个边界框会通过$(p_c,\ b_x,\ b_y,\ b_h,\ b_w,\ c)$六个参数的形式进行表示。如需要表示出 80 个类别，则单个边界框会通过 85 个数字进行表示。在 YOLO 应用情境下，预测环节是借助一个1×1卷积进行的，因此输入的属于特征图。正是由于该过程借助了1×1卷积，所以随之产生的预测图与特征图尺寸一致，但需要注意的是，1×1卷积的作用仅仅是更改通道数值。在 YOLOv3 应用情境下，此时形成的预测图是单个 cell 预测固定数量的边界框。

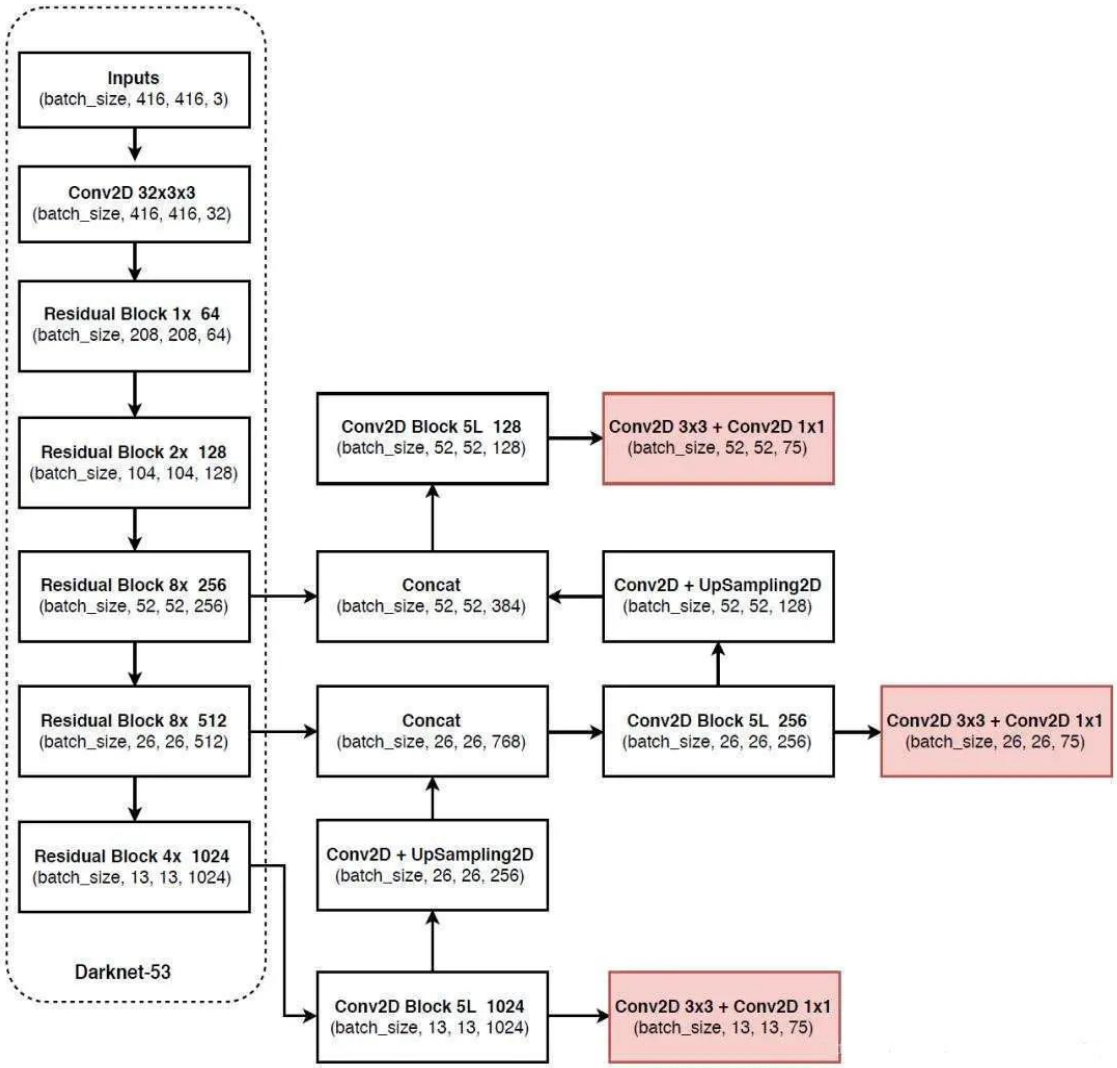


图 2-3 Darknet 53 网络结构图

以图 2-3 为例可知，此预测图的深度是 75，首先将预测图的深度用$B \times (5 + C)$来表示，其中B代表单个 cell 能够完成预测任务的边界框数量。上述B个边界框能够提前锁定检测到目标物体。就单个边界框而言，其均具有$5 + C$个特征，这些特征依次是对中心点坐标、物体宽高（四个）、物体分数（一个）以及C个类置信度的完整描述。此外，YOLOv3 所具有的单个 cell 能够同时实现对三个边界框的预测。

研究结果表明，若 GT 框中心位于 cell 感受野范围中，那么就预测图中的任何一个单元格而言，均能借助其中一个边界框完成对目标的预测。可以起到检测物体作用的边界框是单一的，必须明确该边界框的 cell 位置。原始图像在经过一系列预测识别及分析后，会

被分割为与预测图维度相同的网格。例如在图 2-4 中，最初的输入图像维度为416×416，步幅大小的值为 32，预测图维度为13×13，因此按照原始图像与预测图维度相同的原理，原始图像将被分割为13×13的网格。

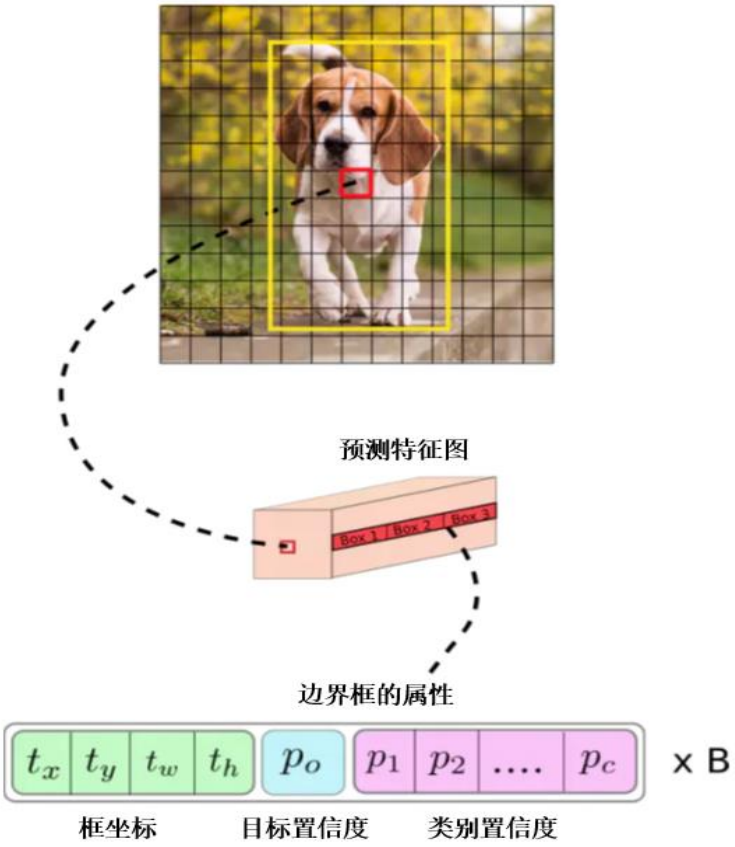


图 2-4 图像分割为网格示意图

边界框类是否有物体的得分被包含在 cell（输入图像中）中，相应的 cell 需要对预测物体负责。根据图 2-4 可知，狗的边界框，即黄色框的中心是被包含在红色框内的，这就说明对预测狗负有责任的应当是红色 cell。位于图中第 7 行第 7 列位置上的网格是红色框，而表现为粉黄色的长方体是预测特征图，能够对原图红色方格负有责任的是前者。换言之，预测特征图与原图中的 cell 在负责匹配上的关系是一一对应的。就单个 cell 而言，其能够对三个边界框完成预测，其中在计算边界框长宽时是依据 K-means 方法聚类[39-40]进行的，因此是否完成包含物体还有待商榷，需要对边界框的范围进一步修正。例如，5 是通过 4+1 得到的，其中边界框用 4 个代表，边界框类是否有物体的得分用 1 个代表。

如果对预测框的宽高进行直接预测，那么极有可能出现因训练不稳定而导致的梯度问题。所以，目前的大部分检测方法均对此进行了改进，引入了 log 空间转换或者偏移(offset)被称为锚框的预定义默认边界框。在进行预测的过程中，便将上述变换嵌入到锚框内。例如，YOLOv3 内部配置了三个锚框，能够对单个单元格中 三个边界框进行预测。

作为边界框的先验，锚框的计算是借助 k 均值聚类形式在 COCO 数据集[41-42]上进行的。通过对锚框的宽度值以及高度值进行逐一预测后，能够获得距聚类质心的偏移量。

以下公式描述了如何转换网络输出以获得边界框预测:

$$b_x = \sigma(t_x) + c_x,\ b_y = \sigma(t_y) + c_y \tag{2.1}$$

$$b_w = p_w e^{tw},\ b_h = p_h e^{th} \tag{2.2}$$

在上式中，通过预测获取的的中心坐标以及目标宽高值依次用b_x， b_y,，b_w， b_h进行表示；网络的输出通过t_x，t_y，t_w，t_h进行表示；网格从顶左部的坐标用c_x，c_y进行表示；锚框的维度则借助p_w，p_h进行表示。

在对中心坐标进行预测时，我们通常需要借助 sigmoid 函数[43]展开，此时数值的范围大小被假定在 0 和 1 之间。YOLO 仅能够对边界框中心的坐标进行一个偏移量的非精准预测，最后需要借助特征图 cell 对维度进行归一化处理。

以上述狗的图像为例可知，若将预测中心坐标界定为(0.4，0.7)，则说明中心位置为(6.4，6.7)，这是由于(6，6)是红色框左上角的坐标。然而在在所预测坐标高于 1 的情况下，则会出现以下问题，若预测坐标为(1.2，0.7)，则中心位置将位于(7.2，6.7)，此时的中心会在红色框右边形成，然而在对我们在对对象预测进行负责时仅可以采用红色框，因此必须嵌入 sidmoid 函数，将其大小直接限定在 0 和 1 之间。首先对输出应用进行对数空间形式的转换，其次将其和锚框相乘，最后便能够得到对边界框尺寸的预测。

单个边界框所涵盖的单个物体概率通过物体分数来表示，在红色框以及红色框附近的框表现为 1 的可能性居多，然后处于边角位置的框则大多表现为 0。在对物体分数概率值进行表示时，同样可以借助 sigmoid 函数进行。

如果所检测到的物体是具体类的概率值，则可以用类置信度进行表示，在传统的 YOLO 版本中是借助 softmax[44]来实现类分数向类概率的转换的。之所以采取 sigmoid 函数来进行替代处理，是因为在 softmax 假设类之间存在严重的互斥关系，常见的例子是，“Person”以及“Woman”不同同时被某一物体所包含，但在现实中这种现象是十分常见的。

为对大多数物体进行预测，特别是形态较小的物体，YOLOv3 采用了三个在尺度方面具有差异的步幅开展预测，即 32、16 以及 8。此时，输入416×416图像后，检测尺度也会呈现出不同的尺寸，依次表现为13×13、26×26以及52×52。

在 YOLOv3 中，单个类别的采样尺度下会预先设定 3 个先验框，能够完成 9 个尺寸存在差异的先验框的聚类。具体地，在 COCO 数据集上，上述 9 个尺寸存在差异的先验框表现形式如下。

(10×13)，(16×30)，(30×61)，(62×45)，(59×119)，(116×90)，(156×198)，(373×326)

其中，(116×90)，(156×198)和(373×326)是大范围感受野，(30×61)，(62×45)，(59×119)是中范围感受野，(10×13)，(16×30)，(30×61)是小范围感受野。

网络降采样输入图像直至首个检测层，步幅的大小为 32；其次，采用通道堆叠的形式将该层上采样 2 倍和上述形态一致的特征图进行处理，第二个检测层在形成过程中的步幅大小为 16；以此类推，执行一致的上采样环节，第三个检测层在形成过程中的步幅大小为 8。就单个尺度而言，单个 cell 均采用三个锚框进行预测，并且对边界框的预测也是一一对

应的，即共计为 9 个锚框。

YOLOv3 网络能够形成 10647 个锚框，但是在图像中仅存在一个狗，如果要把 10647 个锚框压缩为 1 个，则第一步需要借助物体分数对部分锚框进行过滤处理，第二步将低于阈值 0.5 的锚框进行逐一剔除，第三步借助非极大值抑制(NMS)[45]避免多个锚框检测单个物体的现象，这里 NMS 的作用便是对多个检测框进行舍弃。

在进行详细操作的过程中可以按照下述步骤进行：首先将分数较低的抛弃分数低的锚框进行剔除，因为分数较低的锚框往往代表检测类置信度低；其次在多个锚框高度重合且均对单个物体进行检测的情况下，仅能采用一个锚框。

如果要想实现非极大值抑制，必须对下述内容进行着重处理：对锚框的选择应当以分数最高为依据；然后对此锚框与其他锚框的重合程度进行测算，如果重合程度高于交并比(Intersection over Union，IoU)阈值，则需要对这些锚框进行剔除处理，最终重复步骤 1 迭代，直至不存在低于当前所选框时结束。

2.2 目标跟踪算法

2.2.1 传统目标跟踪算法

因为以相关滤波为基础的跟踪算法在速度方面的优越性极高，加之当前学术界对相关滤波的研究仍属于关键课题，因此本节采用了相关滤波领域中的两个代表算法来对 MOT 设计过程进行详细阐述，一是 MOSSE[46]，二是 KCF[47]。

相关性最初是在信号处理领域出现的，目的在对各个信号彼此间的关系进行解释，主要有自相关以及互相关两大类别。此后，计算机视觉领域逐渐引入互相关概念，以此来对因素间的关系进行分析，进而考究各个图像在同个区域范围内的匹配程度。相关滤波之所以能够起到定位的作用，是因为其采用了互相关原理，互相关原理能够使得物体在某一帧图像中的定位操作顺利开展。在相关滤波没有被提出，甚至深度学习也没有出现的时候，学术界缺乏对传统视觉跟踪方法的探索，整体跟踪性能的发展也较为滞后，但后期相关滤波跟踪方法的出现成功突破了算法研究受限的桎梏，率先推出了滤波跟踪器，奠定了算法研究的探索导向。自 MOSSE 被研究出来，相关滤波跟踪算法的研究进展也随之层层突破，其在跟踪层面的应用价值得到极大发掘。MOSSE 的运行原理为以相关滤波为基础对局部像素的相似程度进行衡量，其构建出一款新型滤波器模板，能够确保误差平方和最小。在初始化跟踪视频序列的第一帧后，MOSSE 能够自动形成具备稳定性的滤波器，该滤波器能够直接应用到后继视频图像中，通过模板来分析相似情况，并且在极限状态下的跟踪速度得到明显提高，为 669fps，此时原有的鲁棒性也不会受到威胁。MOSSE 在处理图像和滤波器的过程中，主要是借助了快速傅里叶变换(Fast Fourier Transfonn, FFT)[48-50]原理进行的，该原理的关键点在于把频域中图像以及滤波器的卷积转换为可运算的乘积。因为跟踪器仅关注对物体轨迹的预测，而忽略具体内容，因此无法主动辨别目标的差异，进而无法实现前后帧中的关联，导致跟踪结果中漂移或者丢帧现象层出不穷。

由于相关滤波借助原始像素提取特征的形式缺乏良好的科学依据，所以由此产生的算法精度也有待提高。为进一步提升算法精度，KCF 正式出现，其不仅能够保持相关滤波原有的高速性能，同时获得了更高的精度。

KCF 在 CSK 的基础上了进行了较大程度的改进，而 CSK 的建立则是已经吸收了 MOSSE 的优势。然而，在跟踪性能层面，KCF 的优势更为突出，其运行原理是依靠环矩阵的形式进行的，这就使得循环移位的运算变得较为简易，此外其借助了傅里叶变换，满足了对频域以及空域上进行实时切换额需求，顺利将矩阵乘法转换为点乘计算，运算难度及复杂度明显下降。此外，KCF 在进行跟踪操作时，还创新性地引入了检测器，能够对预测目标以及跟踪目标之间的相似程度作出衡量，就单个跟踪器预测输出的目标而言，必须以检测器为中介桥梁，对预测目标以及跟踪目标之间的相似程度进行先行衡量，并以此为基础判定跟踪轨迹的更新需求。

如果把 KCF 应用到多目标跟踪的任务中，其原有的运行速度会受到严重威胁，丢帧问题也随之产生，这就意味着 KCF 不管在速度上，还是精度上，均存在较大的改进空间。在单目标跟踪 SOT 领域中，“目标身份识别码不可用”的现象是关注度最集中的，一旦此现象出现，会扰乱多目标正常运行下的跟踪结果，严重的情况下甚至出现跟踪窗体漂移问题。多目标跟踪 MOT 的任务模型通常被应用到数据关联情景下，特别是在公共目标身份识别码被允许使用的过程中。传统跟踪器的功能较为单一，仅能完成定位操作，因此在嵌入检测技术后加大对多目标跟踪的研究是具有现实必要性的，本文课题的研究内容也正是在考虑了该背景后确定的。

2.2.2 多目标跟踪算法

多目标的跟踪策略被划分基于检测的跟踪(Tracking by Detecton, TBD)以及基于初始框的跟踪(Detection Free Tracking, DFT)[51]两大类别。DFT 和单目标跟踪在某种程度上是具备一致性的，均有必要在初始化目标的环节中通过人工方式对视频第一帧中的目标进行标记，并将检测与跟踪操作同步执行。因为通过人工方式进行初始化的操作不能标记第一帧中没有显现过的目标，同时，在多目标跟踪的情境下，新旧目标消失出现问题属于正常的波动范围，所以一旦在跟踪中形成了没有经过人工初始化的新目标，则无法继续执行跟踪操作。

人工标注的完整性存在不足，这也进而导致跟踪结果常常处于不稳定的状态，所以 TBD 在实际应用过程比 DFT 更受青睐，不论是在研究领域还是在工业领域也是十分常见的，本课题主要以 TBD 为基础展开后续研究。

TBD 是指以检测为基础展开的跟踪操作，以此为基础的 MOT 涵盖了三个过程，一是检测过程、二是检测结果、三是跟踪器轨迹的连接。不论是在确定 TBD 跟踪目标数量，还是判定其具体类型的过程中，均会受到检测算法结果以及检测成果这两大因素的影响，但由于现实中检测结果难以预测，所以检测成果的优劣对 TBD 跟踪目标数量以及类型是至关重要的。简单的在线和实时深度关联度量跟踪(Simple Online Realtime Tracking with Deep

Association Metric, DeepSORT)[52-53]属于以 TBD 策略为基础的 MOT 算法模式，其跟踪任务的实现需要检测结果以及跟踪预测结果的支撑。此外基于 TBD 的算法还有降低检测不稳定性影响的基于深度学习候选人选择与再识别的实时多跟踪(Multiple Tracking with Deeply Selection, MOTDT)[54]。

目前主流的多目标追踪算法大部分基于 SORT，SORT 算法介绍如下。

1. SORT 算法的基本原理

SORT (The Simple Online and Realtime Tracking)和它的改进版本 Deep SORT 算法是目前备受关注的 TBD 算法[55]。该算法主要应用于行人跟踪领域，以卷积神经网络为基础进行检测，属于 MOT 算法的范畴。在应用这一算法时，能够把 Faster R-CNN 运算的检测结果直接输出至跟踪算法，并通过卡尔曼滤波的辅助作用来预测物体当前及未来一段时间内的运动轨迹。需要注意的是，若要实现检测结果与跟踪 IoU 距离的关联操作，则应当在此引入匈牙利算法，因为匈牙利算法具备成本矩阵的计算功能，进而达到理想的跟踪效果。但是，SORT 算法的应用领域往往单一的集中在行人目标上，根据 Faster R-CNN 的检测结果，把其中行人目标可能性过半的目标输送至跟踪框架，同时针对单个目标分别建模，具体过程可参照(2.3)：

$$X = [u，\ v，\ s，\ r，\ \dot{u}，\ \dot{v}，\ \dot{s}]^T \tag{2.3}$$

上式中，目标边界框的中心坐标分别用 u 、 v 进行表示， 面积为 s， 长宽的比例值用 r 来表示。即使目标尺寸发生变动，但长宽的比例值是维持原有形态的。如果通过算法分析认为，当前帧中检测到的目标边界框会受到具体运用轨迹影响的情况下，轨迹会更新其运行状态，此时卡尔曼滤波开始发挥作用，自动对下一帧的运动轨迹进行预测。匹配的代价矩阵是一种 IoU 距离，一旦轨迹与检测边界框的 IoU 距离不足最低阈值IoU_{min}，则被列入“不匹配”的行列中。

2. 卡尔曼滤波

卡尔曼滤波本质上是一种最优估计算法，作为一种强有力的算法能够处理多种具备不确定性的信息。由于卡尔曼滤波算法仅仅对前向单个状态的信息进行存储，因此其几乎不需要占据过多的内存空间，且能够保持较高的实时性，这促进了该算法在现实作业过程中的进一步推广 [57]。

SORT 创造性地利用卡尔曼滤波器来预测目标的状态。卡尔曼滤波器作为状态估计器，能够完成加权平均的处理过程，该过程主要是依据对运动估计的结果以及实际测量所得到的运动状态展开的。通过式(2.4)、式(2.5)，能够清晰的认识到运动物体在位移、速度、加速度与时间之间存在的数量关系：

$$s(t) = s(t-1) + v(t-1)\Delta T + \frac{1}{2}\frac{F}{m}(\Delta T)^2 \tag{2.4}$$

$$v(t) = v(t-1) + \frac{F}{m}\Delta T \tag{2.5}$$

合并以上两式可得到式(2.6)：

$$\begin{pmatrix} s(t) \\ v(t) \end{pmatrix} = \begin{pmatrix} 1 & \Delta T \\ 0 & 1 \end{pmatrix} \begin{pmatrix} s(t-1) \\ v(t-1) \end{pmatrix} + \begin{pmatrix} \frac{(\Delta T)^2}{2} \\ \Delta T \end{pmatrix} \frac{F}{m} \tag{2.6}$$

状态空间模型为式(2.7):

$$x(t) = Ax(t-1) + Bu(t) + w(t) \tag{2.7}$$

令式(2.8):

$$x(t) = \begin{pmatrix} s(t) \\ v(t) \end{pmatrix},\ A = \begin{pmatrix} 1 & \Delta T \\ 0 & 1 \end{pmatrix},\ B = \begin{pmatrix} \frac{(\Delta T)^2}{2} \\ \Delta T \end{pmatrix} \tag{2.8}$$

又，在离散空间中，有式(2.9):

$$x[n] = Ax[n-1] + Bu[n] + w[n] \tag{2.9}$$

其中，位移等运动的具体状态通过向量 x[n]来表示，而加速度对于速度等直接对运动物体产生影响的通过向量 u[n]来表示。状态转移矩阵通过 A 进行表示，控制输入矩阵通过 B 进行表示，能够说明相邻时刻之间物体状态的相互影响。此外，过程噪声通过 w[n]进行表示，以 w[n]~N(0,Q)的高斯分布原则为基础。

此外，在离散空间的状态下，能够经式(2.10)测量分析得出状态模型：

$$z[n] = H[n]x[n] + v[n] \tag{2.10}$$

测量结果通过 z[n]进行表示，测量结果通过 H[n]进行表示，测量噪声通过 v[n]~N(0,R)进行表示，即测量误差。在大多数情境下，采用测量的方式便能获得物理量，无需任何其他辅助，卡尔曼滤波对测量值（Measurement）和预测值（Prediction）会展开修正处理。

状态矢量 x[n]属于 N 阶向量，观测矢量 z[n]是卡尔曼滤波器的输入，是一个 M 阶向量，预测值$\hat{x}[n|n-1]$，其以 n 时刻前的状态为估计基础，对 n 时刻状态结果进行分析，它由式(2.11)得到：

$$\hat{x}[n|n-1] = A\hat{x}[n-1|n-1] + Bu[n] \tag{2.11}$$

最小预测均方误差矩阵$P[n|n-1]$由式(2.12)得到：

$$P[n|n-1] = AP[n-1|n-1]A^T + Q \tag{2.12}$$

误差增益 K[n]预测值在总误差中的权重由式(2.13)得到：

$$K[n] = P[n|n-1]H^T[n]\{R[n] + H[n]P[n|n-1]H^T[n]\}^{-1} \tag{2.13}$$

预测的修正值$\hat{x}[n|n]$，采用了 0 至 n 时刻全部状态开展估计的结果，通过式(2.14)得到：

$$\hat{x}[n|n] = A\hat{x}[n|n-1] + K[n]\{z[n] - H[n]\hat{x}[n|n-1]\}^{-1} \tag{2.14}$$

修正后的最小均方误差矩阵为式(2.15)：

$$P[n|n] = \{I - K[n]H[n]\}P[n|n-1] \tag{2.15}$$

卡尔曼滤波器的运算流程主要是依据上述五个步骤完成的，式(2.11)以及式(2.12)能够起到预测作用，而式(2.13)、式(2.14)以及(2.15)能够起到更新作用。根据卡尔曼滤波器的具体推导进程可知，运动状态之间存在着明显的递推逻辑，这就意味着，对于某一时刻状态的认知是应该以前项时刻的状态为基础的。在实施 SORT 算法时，为确定某跟踪目标在 0 时刻的状态，必须把跟踪目标在-1 时刻的速度进行初始化处理。

3. 匈牙利算法

匈牙利算法（Hungarian Algorithm）[58-59]是一个寻找二分图最大匹配的算法。设 G 是具有二分类（X，Y）的二分图，$X=\{x_1，x_2，\ \ldots，x_n\}$，$Y=\{y_1，y_2，\ \ldots，y_n\}$，M 为 G 的任意一个匹配，如图 2-5 所示。

匈牙利算法涵盖下述三个步骤过程。以其中任一 M 为起始点：

步骤 1 若 M 饱和 X 的所有顶点，那么会出现终止现象。否则，设 u 是 X 中的 M 非饱和顶点。置$S=\{u\}$且$T=\emptyset$。

步骤 2 若$N(S)=T$，由于$|T|=|S|-1$，所以$|N(S)|<|S|$，则停止。否则，设$y\in N(S)\backslash T$。

步骤 3 若 y 是 M 饱和的，设$yz\in M$，用$S\cup\{z\}$来代替 S，$T\cup\{y\}$代替 T，并递进到步骤 2。否则，设 P 是 M 可扩$(u，\ y)$路，用$M'=M\Delta E(P)$代替 M，并转到步骤 1。

最终可以得到的匹配如图 2-6 所示。

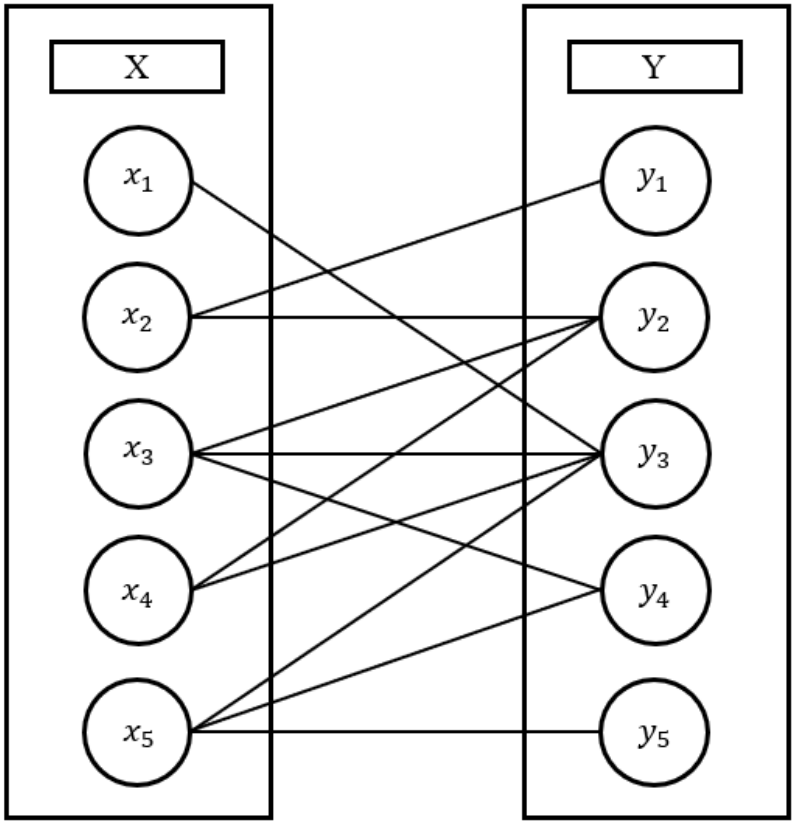


图 2-5 二分图 G 的任意一个匹配 M

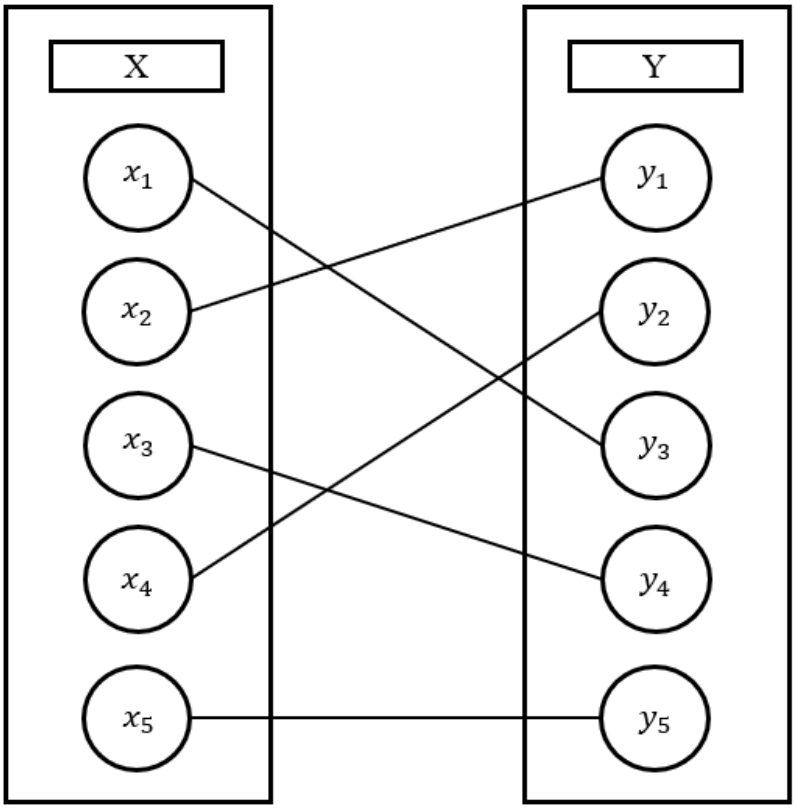


图 2-6 G 的最大匹配

2.3 本章小结

本章介绍了目标检测和跟踪过程中的相关技术。目标检测算法被划分为以下两大类别，一是传统目标检测算法，二是以深度学习为基础的目标检测算法，后者探究了二阶段目标检测算法（RCNN、Faster-RCNN）与一阶段目标检测算法（YOLO 系列）。目标跟踪算法分为传统的单目标跟踪，与基于目标检测的多目标跟踪。针对多目标跟踪领域，顺次分析了 SORT 算法、卡尔曼滤波以及匈牙利算法的运行原理。

第 3 章 基于 YOLOX 与 DeepSORT 的人流量统计算法研究

本章主要介绍了所提出的人流量统计系统中的目标检测算法 YOLOX、目标追踪算法 Deep SORT、以及实验部分所涉及到的人流量统计算法。

3.1 YOLOX 目标检测算法

YOLOX 目标检测算法是在 YOLO 系列的基础上吸收近年来目标检测学术界的最新成果，同时继承 YOLO 系列容易部署的特点[60]。同时需要在避免过拟合 COCO 的基础上，适度调参，以及在参数设置公平的条件下和 YOLO 系列做对比。YOLOX 的设计路线主要为：以 YOLOv3 作为模型的原始框架（YOLOv3 网络中使用的算子更加简单，应用范围更加广），然后设计 Decoupled Head、Data Aug、Anchor Free 以及 SimOTA 部件。

YOLOX 的设计，在大方向上主要遵循以下几个原则：

(1) 所有组件全平台可部署;

(2) 避免过拟合 COCO，在保持超参规整的前提下，适度调参;

(3) 不做或少做稳定涨 点但缺乏新意的工作(更大模型，更多的数据)。

所以可以看到，首发的 YOLOX 没有 deformable conv ，没有用额外数据做预训练，没有 momentum=0.937...

回到 YOLOX 设计的具体细节上，YOLOX 与之前 YOLO 最大的区别在于 Decoupled Head，Data Aug，Anchor Free 和样本匹配这几个地方。

1. Decoupled Head

原来的 YOLO 系列都采用了一个耦合在一起的检测头，同时进行分类、回归的检测任务。YOLOX 在结构上采用了 Decoupled Head，将特征平行分成两路卷积特征，同时为了降低参数量提前进行了降维处理，其好处在于：在检测的过程中分类需要的特征和回归所需要的特征不同，所以在 Decoupled Head 中进行解耦处理后学习的过程会变得更加简单。如图 3-1 所示，从 Decoupled Head 结构图中的左下角可以看到采用了 Decoupled Head 后，网络的收敛速度在训练早期要明显快于 YOLO head。

Decoupled Head 是学术领域一阶段网络一直以来的标准配置（RetinaNet，FCOS 等)。相比于它朴素的实现方法。起初未将对检测头的解耦纳入考虑范畴，直至后来在试图把 YOLOX 演进到“端到端”时发现，End2end 的 YOLOX 在调整损失权重以及控制梯度回传两个方面表现得较为乏力，与标准的 YOLOX 相较而言，平均要低 4~5 个点，这一现象显然与 DeFCN 中信息相矛盾。如果用 decoupled head 来取代传统的 YOLO Head ，那么上述差距会得到压缩，这意味着目前使用的 YOLO Head 在表达性能上仍存在提升空间。因此，在非 End2End YOLO 上嵌入 decoupled head 是可行的，能够实现峰值性能以及收敛速度的双重提升。与前面所进行的 End2end 实验结合来看，YOLO 系列所应用的检测头

存在配置不合理的可能性。

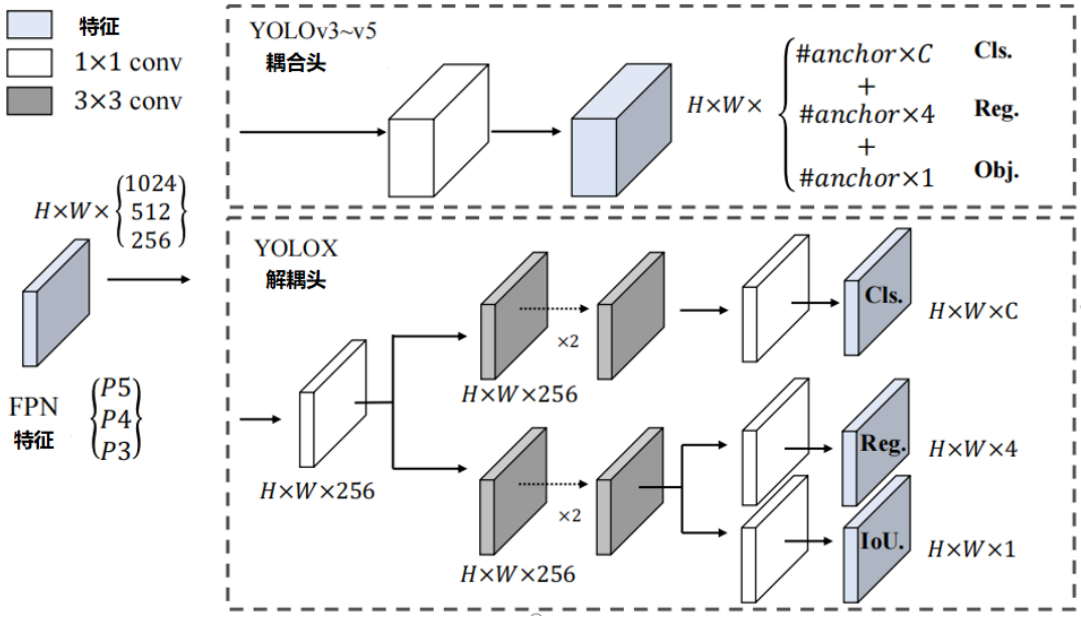


图 3-1 YOLOv3 Head 与 Decoupled Head 结构图

把检测头解耦无疑会增加运算的复杂度，但经过对速度以及性能上的得失进行全量考量后，决定首先通过 1 个 1x1 的卷积执行降维操作，同时在分类以及回归分支中分别采取 2 个 3x3 卷积，以此修正为对一点点参数的增加，YOLOX 在 s,m,l,x 模型速度上的轻微下降也全源自于此。表面上看，解耦检测头提升了 YOLOX 的性能和收敛速度，但更深层次的，它为 YOLO 与检测下游任务的一体化带来可能。例如：

(1) YOLOX + Yolact/CondInst/SOLO ，分割端侧实例;

(2) YOLOX + 34 层输出，检测处于端侧位置的人体的 17 个关键点;

YOLOX 设计了一个全新的检测头使得 YOLO 与检测的下游任务更深层次的结合，为检测与下游任务的端到端一体化带来一些变化。

2. 数据扩充

在数据扩充中，原来版本将先选择一张图然后再随机选择三张图再将其四张图拼接成一张，然后进行适量的放缩。在 YOLOX 中使用 Mosaic+MixUP 的方法[61]对图片进行加强比原版的 MixUP 效果更好。

Mosaic 在完成 YOLOv5 以及 v4 的检验后[62-63]，说明 Mosaic 在状态极强的基线上存在涨点。当模型容量足够大的时候，相对于先验知识（各种技巧，手工规则），更多的后验（数据/数据增强）才会产生本质影响。通过使用 COCO 提供的 ground-truth 掩膜标注，在 YOLOX 上试了试图像复制粘贴，在 48.6mAP 的 YOLOX-Large 模型[64]上，使用图像复制粘贴带来 0.8%的涨点。

但是图像复制粘贴的实现依赖于目标的掩膜标注，而掩膜标注在常规的检测业务上是稀缺的资源。而由于 MixUp 和图像复制粘贴有着类似的贴图的行为，还不需要掩膜标注，因此可以让 YOLOX 在没有掩膜标注的情况下吃到图像复制粘贴的涨点。但 YOLOX 实

现的 Mixup，没有原始 Mixup 里的伯努利分布和软标签 ，有的仅是 0.5 的常数透明度和图像复制粘贴里提到的尺度缩放。YOLOX 里的 Mixup 有如此明显的涨点，大概是因为它在实现和涨点原理上更接近图像复制粘贴，而不是原版 Mixup。数据扩充里面必须注意：Mosaic 和 Mixup 在训练结束前的 15 个 epoch 时就要关闭。因此不难看出，通过 Mosaic+Mixup 形成的训练图片是与自然图片的实际分布相互独立的，此外通过 Mosaic 的裁剪操作，会进一步弱化标注框的准确程度。

3. Anchor Free 与 Label Assignment

从原来 YOLO 系列的 Anchor Based 方法[65]切换到 Anchor Free 的操作是比较简单的，但 anchor free 的方法对于 label assign 的策略选取就有很大的空间。YOLOX 采用了 SimOTA 的标签分配策略, 最终在 OTA 方法的基础上进行了简化。

Anchor Free 的优点是全方位的。Anchor Based 检测器为了追求最优性能通常会需要对 anchor box 进行聚类分析，这无形间增加了算法工程师的时间成本； Anchor 的出现，会使得检测头的复杂程度和由此形成结果数量得到提高，导致原本处于 NPU 的检测结果转换到了 CPU 上，这点对于部分边缘设备来说是无法实现的；Anchor Free 所具有的解码及代码逻辑难度往往更低，可读性更高。至于为什么 Anchor Free 现在可以上 YOLO ，并且性能不降反升，这与样本匹配有密不可分的联系。

与 Anchor Free 比起来，样本匹配在业界似乎没有什么关注度。但是一个好的样本匹配算法可以天然缓解拥挤场景的检测问题(LLA、OTA 里使用动态样本匹配可以在 CrowdHuman[66] 上提升 FCOS 将近 10 个点)，缓解极端长宽比的物体的检测效果差的问题，以及极端大小目标正样本不均衡的问题。甚至可能可以缓解旋转物体检测效果不好的问题，这些问题本质上都是样本匹配的问题。

样本匹配有 4 个因素十分重要。loss/quality/prediction aware，基于网络自身的预测来计算 anchor box 或者 anchor point 与 gt 的匹配关系，充分考虑到了不同结构/复杂度的模型可能会有不同行为，是一种真正的 dynamic 样本匹配。而 loss aware 后续也被发现对于 DeTR 和 DeFCN 这类端到端检测器至关重要。与之相对的，基于 IoU 阈值 /in Grid(YOLOv1)/in Box or Center(FCOS) 都属于依赖人为定义的几何先验做样本匹配，目前来看都属于次优方案； center prior，考虑到感受野的问题，以及大部分场景下，目标的质心都与目标的几何中心有一定的联系，将正样本限定在目标中心的一定区域内做 loss/quality aware 样本匹配能很好地解决收敛不稳定的问题；不同目标设定不同的正样本数量(dynamic k)，YOLOX 不可能为同一场景下的西瓜和蚂蚁分配同样的正样本数。Dynamic k 的关键在于如何确定 k，有些方法通过其他方式间接实现了动态 k ，比如 ATSS、PAA ，甚至 RetinaNet ，同时，k 的估计依然可以是 prediction aware 的，具体的做法是首先计算每个目标最接近的 10 个预测，然后把这个 10 个预测与 gt 的 iou 加起来求得最终的 k，很简单有效，对 10 这个数字也不是很敏感，在 5~15 调整几乎没有影响；全局信息，有些 anchor box/point 处于正样本之间的交界处、或者正负样本之间的交界处，

这类 anchor box/point 的正负划分，甚至若为正，该是谁的正样本，都应充分考虑全局信息。

YOLOX 通过把样本匹配建模成最优传输问题，求得了全局信息下的最优样本匹配方案。但是 OTA 最大的问题是会增加约 20~25 %的额外训练时间，对于动辄 300epoch 的 COCO 训练来说是有些吃不消的，此外 Sinkhorn-Iter 也会占用大量的显存，所以在 YOLOX 上去掉了 OTA 里的最优方案求解过程，保留上面 4 点的前 3 点。由于相对 OTA 去掉了 Sinkhorn-Iter 求最优解的过程，将 YOLOX 采用的样本匹配方案称为 SimOTA （Simplified OTA）。在 Condinst 这类实例分割上用过 SimOTA，获得了 box 近 1 个点， seg 0.5 左右的涨点。同时在内部 11 个业务数据上也测试过 SimOTA，平均下来发现 SimOTA>FCOS>>ATSS，这些实验都满足我们不去过拟合 COCO 和 COCO style mAP 的初衷。没有复杂的数学公式和原理，不增加额外的计算时间，但是效果很好。

4. 端到端

端到端(无需 NMS)是个很诱人的特性，去年有不少相关的工作放出(DeFCN，PSS, DeTR)，但是在 CNN 上实现端到端通常需要增加 2 个卷积才能让特征变的足够稀疏并实现端到端，且要求检测头有足够的表达能力(Decoupled Head 部分已经详细描述了这个问题)，在 YOLOX 上实现 NMS Free 的代价是轻微的掉点和明显的掉 FPS 。所以 YOLOX 没有在最终版本中使用端到端特性。

3.2 YOLOX-AM

本文修改了 YOLOX 目标检测模型，为了提升 YOLOX 目标检测精度，在 YOLOX 中引入了注意力机制(Attention Mechanism,，AM)。本文将其命名为 YOLOX-AM，并将卷积模块注意力模型(Convolutional Block Attention Module，CBAM)作为注意力机制。

CBAM 作为注意力机制模块，实现了对空间(spatial)以及通道(channel)的双重结合。与 senet 进行对比分析可知，将注意力聚焦在通道上能够的实现更理想的效果。图 3-2 展示了添加 CBAM 模块之后的整体结构。由图中可见并分析，卷积层输出结果率先经过通道注意力模块，完成加权处理后会进入空间注意力模块，最终的加权结果则是出现在上述两个步骤之后。

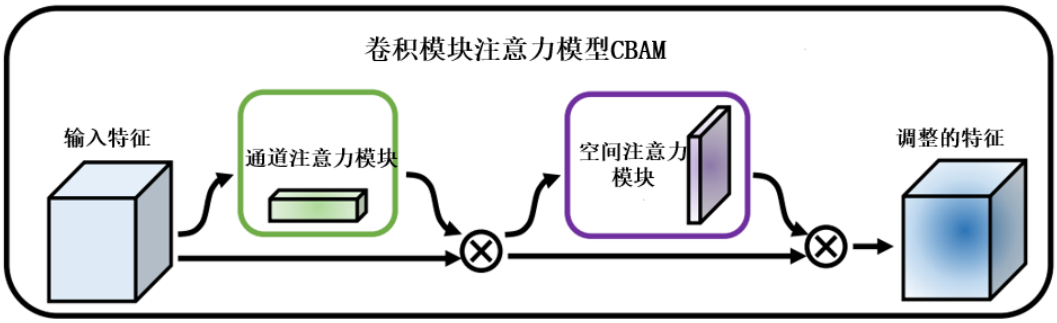


图 3-2 CBAM 结构图

通道注意力机制的步骤是首先将输入的特征图分别通过基于宽度和高度的全局最大

池化以及全局平均池化，其次按照一定的顺序经过多层感知机。以元素级别差异为基础，针对多层感知机输出的特征区域执行加和处理，然后通过 sigmoid 进行一一激活，此时便能形成通道注意力特征图。然后通过元素级别乘法处理，针对通道注意力特征图以及输入特征图展开处理，并形成在空间注意力模块中需要输入的特征。

以其他的另一个角度为切入点可知，通道注意力机制(Channel Attention Module)的作用范畴通常局限于空间维度，能够压缩空间中的特征图大小，使其变为一维矢量，并在基础上在展开后续操作。需要注意的是，若要对空间维度完成压缩处理，则必须同时把平均值池化(Average Pooling)以及最大值池化(Max Pooling)纳入考虑范畴，上述二者的存在能够作用到聚合特征映射的空间信息上，把其传输至共享网络，对输入特征图的空间维数进行持续压缩，先按照一定顺序对元素进行求和操作，完成求和后在进行合并处理，从而形成通道注意力图。对单张图片来说，最关键的是要明确通道注意力对图中哪部分内容是最为关注的。平均池化具有反馈性能，并且该反馈性能对于特征图上的所有像素点都是有效的，但最大池化则不同，其在展开梯度反向传播运算的过程中，梯度反馈仅出现在特征值中响应最明显的部分。

本模块输入特征图的来源是通道注意力模块输出的特征图。第一步需要进行全局性的最大尺化以及全局性的平均池化，这一过程是以通道为基础的；第二步是借助通道的作用，将上述两大结果进行连接处理；第三步是进行卷积处理，通过降低维度的操作，达到一个通道的目的。此后，借助 sigmoid 形成空间注意力特征。最终把这一特征与输入特征相乘，便出现了最终形成的特征。

与上述执行原理相同，空间注意力机制(Spatial Attention Module)是以通道为客体的，并在此基础上完成压缩处理，针对通道维度展开平均值池化以及最大值池化的双重操作。前者通过在通道上提取最大值，高乘以宽所得到的数值是具体需要提取的次数；后者通过通道上提取平均值，高乘以宽所得到的数值也是具体需要提取的次数；然后，把此前所获取的通道数量为 1 的特征图进行合并处理，形成 2 通道的特征图。

3.3 Deep SORT 目标追踪算法

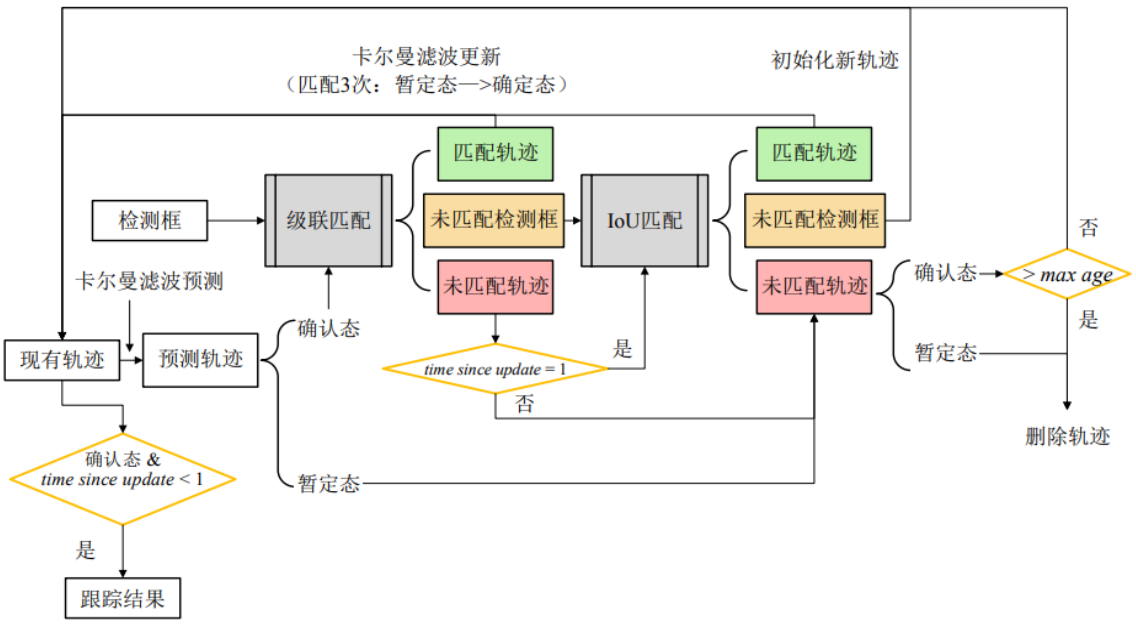


图 3-3Deep Sort 多目标跟踪算法流程图

Simple Online and Realtime Tracking（SORT）作为一种典型的多目标跟踪算法，不仅运行原理较为简便，而且兼具实用的性能。在该种多目标跟踪算法中，若要对卡尔曼滤波预测的 IOU 进行匹配则借助检测过程即可完成，便捷的执行过程提升了算法速度，不足之处在于其会继续出现身份跳变（ID switch）问题。Deep Sort 基于 SORT 引入级联匹配模式，该模式需要以目标的外观特征作为机会成本，对于单条轨迹的生命周期而言，能够在目标被遮挡的情况下进行找回，身份保持原有状态的可能性提升，多目标跟踪的鲁棒性也因此得到明显强化，具体流程可通过图 3-3 进行了解。

在判定人员流动状态时，身份跳变频率的降低对于人流量的统计来说是必不可少的。所以，本文将多目标跟踪算法的选择界定为 Deep Sort ，通过在算法可持续运行的前提下舍弃部分对速度的追求，以此换取多个头部跟踪稳定性的提升。

3.3.1 跟踪流程

1. 轨迹状态

将 Deep SORT 与 SORT 进行对比分析可知，后者的不同在于其具有轨迹状态，这一轨迹是随着目标初始化跟踪任务的完成而产生的，此时的检测环节是实时连续的，属于暂定态(Tentative)的范畴；但如果这一轨迹和其后面连续三帧中的检测目标存在明确的匹配关系，则属于确定态(Confirmed)的范畴。

2. 轨迹状态更新时间记录

若确定态轨迹在运行帧中不能实现对检测目标的准确匹配，那么确定态轨迹也能够继续保存。更新的时间会因为帧数的增加而逐渐递增，这种单次增加的幅度值会固定在 1 的大小上，若距离最近更新的时间是远远高于 max age 的，那么系统会启动删除程序，当更

新后的 max age 帧中能够和检测目标达到二次匹配状态时，卡尔曼滤波会对上述轨迹状态进行更新处理。该原理的存在，使得 Deep Sort 算法能够在跟踪目标被持续遮挡的情况下再次找回目标并继续执行跟踪。

3.跟踪流程

步骤一：借助目标检测器的功能，可以得到检测框(Detections)，在卡尔曼滤波的支持下，跟踪器可以以上一帧轨迹状态为基础，预测当前时刻下的帧轨迹信息(Predict tracks)，同时距离上次更新时间会随着检测流程而递增，单次增加幅度固定为 1。

步骤二：采取级联匹配的形式，对确定态的下述两部分内容进行结合操作，一是预测轨迹(confirmed predict tracks)、二是检测框(detections)，当更新后的 max age 帧中能够和检测目标达到二次匹配状态时，卡尔曼滤波会对上述轨迹状态进行更新处理。

步骤三：和级联匹配没有达到成功匹配标准的检测框展开 IoU 匹配，该匹配的前提条件包括以下两点，一是暂定态的预测轨迹(tentative predict tracks)以及级联匹配中没有达到成功匹配标准，二是距离上一次更新时间为 1，IoU 匹配。如果 IoU 匹配成功的轨迹属于确定态行列，则能够通过卡尔曼滤波对轨迹信息进行更新处理，而无需经过其他环节；但在暂定态的情况下，需要对轨迹信息进行更新处理，当且仅当成功匹配次数满足了三次的要求时，则能够试图转入到确定态中。

步骤四：若 IoU 没有达到成功匹配标准，那么与之相对应的检测框会回归到暂定态的新轨迹，等待进入后续预测环节。具体来说，若在归属于确定态的基础上距离前向更新时长低于 max age 的标准，则需要进行下一轮预测；若距离前向更新时长不低于 max age 的标准，那么会与暂定态未匹配成功轨迹同时被系统剔除。

步骤五：最终跟踪结果的输出需要满足以下两个标准，一是在轨迹中被归类为确定态，二是距离前向更新时长低于 n。需要注意的是，此处的 n 代表的是可选参数，如果其表现为 1，则仅对最近一次更新的轨迹进行输出操作，由此获得的跟踪精确也会更高，和观测值之间的误差更小。然而，若出现漏检现象，轨迹会中断。n 高于 1 时，一旦出现漏检现象，则会借助卡尔曼滤波的预测性能，就检测漏检目标展开弥补处理，能够使轨迹的连贯性始终维持在一个高水平的状态下，但由于卡尔曼滤波器预测值并不是一种真实框，其所进行的线性估计是以此前所获取的帧轨迹信息为基础的。综上所述，在具体的实践过程中，有必要从实际需要对 n 的值做出适当调整或修正。

3.3.2 卡尔曼滤波预测头部运动状态

在进行目标跟踪的过程中，卡尔曼滤波同样能够进行预测估计，它主要是基于观测方程以及状态转移矩阵两大部分来执行。在高铁站中，由于人流速度的水平是相对稳定的，因此在明确计算速度的基础上，借助卡尔曼滤波进行匀速运动建模是具有可行性的，能够实现对头部位置的良好估计。

在 Deep Sort 中进行目标运动建模时采取了匀速运动的形式，该方式的优点在于其能

够降低对运算的需要。(b，v)代表目标的运动状态，前者是目标位置，后者是目标位置参数的速度，通过式(3.1)能够对目标状态进行完整完整表达：

$$x = (x，\ y，\ a，\ h，\ v_x，\ v_y，\ v_a，\ v_h)^T \tag{3.1}$$

当卡尔曼滤波到达初始化阶段后，v 均以 0 的形式进行体现，该种情况下所出现的 b 是检测框的参数值；此后在卡尔曼滤波器的支持下，会进入第二阶段，也就是所谓的预测阶段。通过对匹配达标的检测框进行分析后，能够针对预测结果展开修正，在完成上述两个阶段后，意味着对目标状态的估计也随之结束，处于帧与帧之间的目标关联跟踪任务得以顺利完成。

3.3.3 关联匹配算法

Deep Sort 在开展跟踪任务时是以检测结果为基础的，通俗来说，目标检测结果需要和前一跟踪结果展开关联匹配。匈牙利算法在设置代价矩阵时，是通过检测结果以及跟踪结果的信息结合形成的，有助于实现当前及后续帧中的关联最大匹配。

Deep Sort 进行级联匹配时主要借助了以下两点，一是马氏距离，二是外观特征的加权值。但是，实践代码说明为提高找回目标的效率往往要依靠外观特征进行，根据图 3-4 能够明确级联匹配的具体流程。

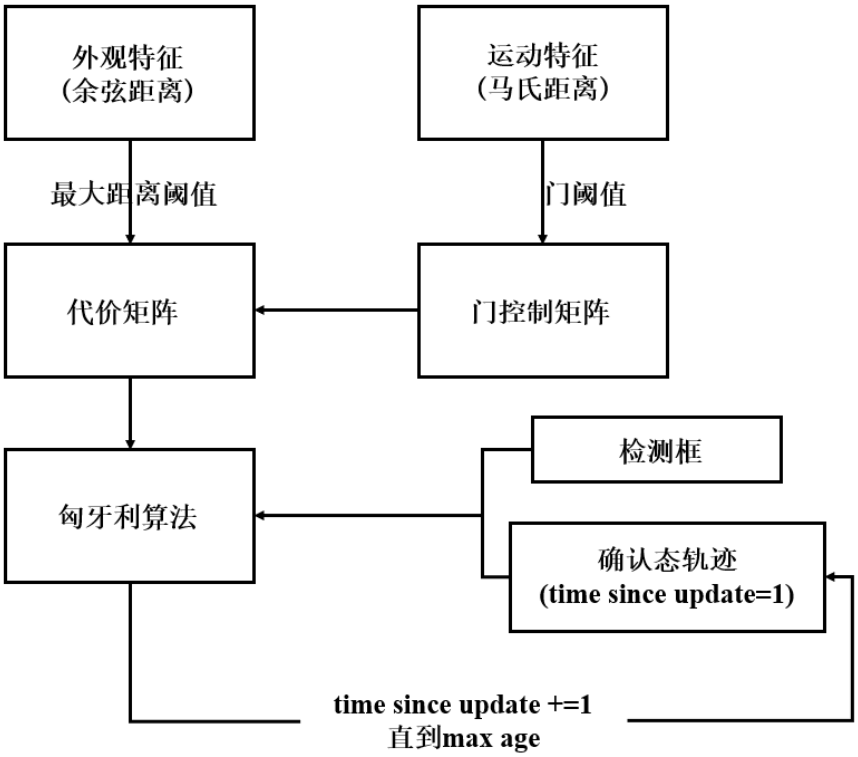


图 3-4 级联匹配流程图

步骤一：首先需要对外观特征的余弦相似度距离展开测算，这一过程主要是针对检测框以及确认态所对应的轨迹两个部分进行的，具体的参照标准以距离阈值的最大值为标准，对大于该距离的部分展开剔除操作，剩下的部分则能够组成代价矩阵的雏形。

步骤二：其次需要对马氏距离展开测算，这一过程主要是针对检测框以及确认态所对

应的轨迹两个部分进行的，依靠门域值的支撑构成门控制矩阵，进而提升其在外观方面的一致性，同时对于马氏距离较大的配对部分展开剔除操作，剩下的部分则能够组成代价矩阵的最终形态。

步骤三：以最终代价矩阵为基础，针对单个确定态跟踪轨迹引入匈牙利算法，进一步获取匹配检测框。在确定匹配顺序时，应当参照距离上次更新时间的长短来进行，秉持对丢失时长较短或者未丢失的轨迹进行优先匹配的原则，后续考虑对丢失时间较长的轨迹进行匹配。

IoU 匹配完成级联匹配后，会将没有满足匹配要求的轨迹和检测框进行 IoU 匹配，此时的机会成本是 IoU 距离，通过匈牙利算法为单个没有分配的轨迹配置检测框。IoU 距离可借助式（3.2）进行计算：

$$iou\_distance = 1 - IoU(detections，tracks) \tag{3.2}$$

3.4 人流计数算法

系统运行过程以计数为最后环节，在考虑了现实作业需求后，本研究在图像中标记出两条计数线，分别记做 A、B，以此来表示从此经过的具体人数。根据图 3-5 可知，统计区域为两条计数线之间的部分。

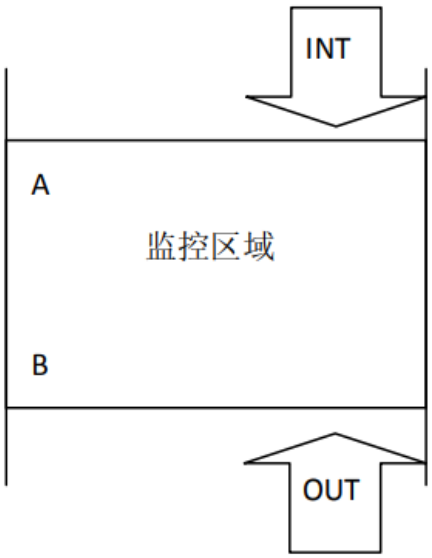


图 3-5 计数区域

1. 对人员进行聚类

在相邻两帧中需要对所检测人员的进入顺序进行区分，即判定该人员是否属在原记录中已经存在。为实现这一任务，可以采取聚类的形式对目标人员在连续两帧之间的距离进行计算，确定其头部的对应关系。以头部的位置为基础展开聚类分析，在检测第 i 帧时，能够获取其头部位置左下角的参考点坐标，其中横坐标用 tou_rect.x 进行表示，相应的纵坐标用 tou_rect.y 进行表示，同时将其存储至程序中。在对第 i 帧之前的各帧进行提取时，采用与上述一致的形式进行，记录全部人头的横坐标以及纵坐标，同时将其存储至程序中。

两帧中所有的人头参考点之间的距离用 L 表示，可以借助公式 3.1 对其进行计算：

$$L = \sqrt{[i.(tou\_rect.x) - (i-1)(tou\_rect.x)]^2 + [i.(tou\_rect.y) - (i1)(tou\_rect.y)]^2} (3.1)$$

在相邻两帧中，单个人员的位移量通常固定地处于 2-4 个像素之间，因此能够以阈值 L 为基础反对其进行聚类分析。具体来说，若 L 处于 2-4 之间，说明此时距离两侧为同一检测人员，此时人数保持不变；若阈值 L 比 4 更高，则意味着距离两侧所出现的人员不是同一人，此时人数以单位 1 的幅度顺次递增。根据图 3-6 可知，第 i 帧时所检测到的人头左下角用字母 A 进行代表，第（i-1）帧时所检测到的两个人头左下角能够借助字母 B 、C 进行代表，把前者依次与上一帧中的 B、C 两点做距离计算处理后可知：

(1)A 与 C 之间的距离若高于阈值 L，则说明不为同一人员，A 与 B 之间的距离若处于阈值允许范围之内，则说明为同一人员。

(2)A 与 B 之间的距离若高于阈值 L，则说明不为同一人员，A 与 C 之间的距离若处于阈值允许范围之内，则说明为同一人员。

(3)A 与 C 之间的距离若高于阈值 L，则说明不为同一人员，A 与 B 之间的距离若不处于阈值允许范围之内，则说明不为同一人员，此时人数以单位 1 的幅度顺次递增。

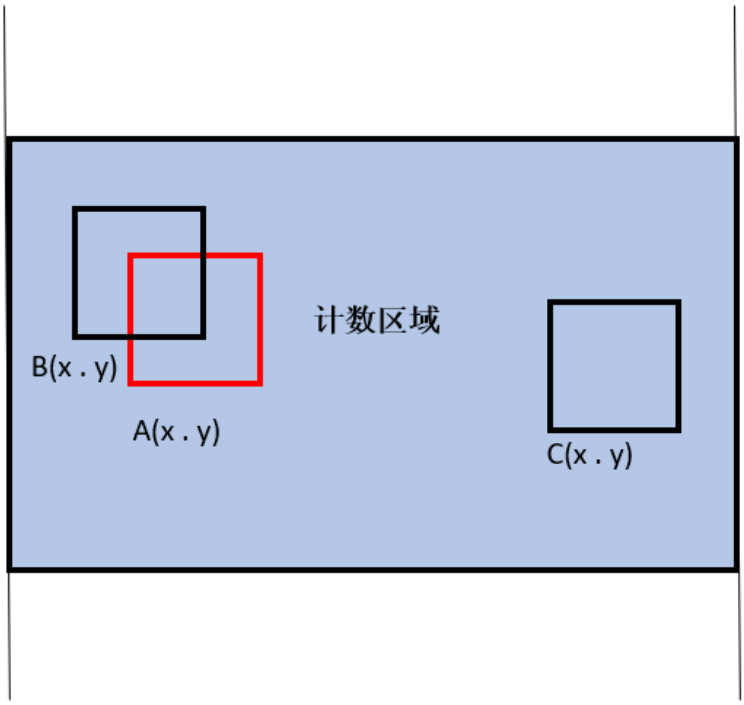


图 3-6 距离聚类示意图

2. 判断进出方向

以头部运动轨迹为基础，通过对进出口标志进行分析能够确定相应的进出方向。首先，通过一组线段的形式对进出口标志进行分别设置，而用户拥有自主设定位置的权限，在物体进出门的必经之处设置进出口标记，同时注意进出口标记位置其经过之处应当保持适当距离。在当物体碰到进出口标记后再次碰到出口标记，则被称为进门行为，反之，当物体先碰到出口标记随后再次碰到进口标记，为出门行为。此外，以进出口标志以及头部运动

轨迹为基础对进出行为进行相应判断。

3.5 本章小结

本章主要介绍了所提出的车站人流量系统所使用的目标检测算法 YOLOX、多目标追踪算法 Deep SORT 以及人流计数算法。YOLOX 主要基于端对端设计，由 Decoupled Head、数据扩充(Data Aug)、Anchor Free 与 Label Assignment 构成。Deep SORT 对每一帧的处理流程如下：检测器得到 bbox 然后生成 detections 接着进行卡尔曼滤波预测，最后借助匈牙利算法的运行原理，把已经完成预测的 tracks 匹配到帧的 detecions 中。人流计数算法首先是对人员进行聚类，然后再判断人员的进出方向。

第 4 章 实验结果与分析

4.1 实验数据

4.1.1 CrowdHuman 数据集

CrowdHuman 的训练集、验证集和测试集包括了 15000、4370 和 5000 幅图像。数据集进行了完全标注，其中包涵众多场景。在训练集和验证集中共计有 47 万个人体实例，每幅图中的平均行人数量为 22.6。图片中的人体实例还给出了三种标注，分别为人体可见区域边界框标注、头部区域边界框标注和人体整体边界框标注。

行人检测数据集的先驱工作有 INRIA，TudBrussels 和 Daimler，更大规模的数据集如 Caltech-USA 和 KITTI、大型多样化的行人检测数据集 CityPersons 等。这些数据集非常流行，但它们都存在着一个同样的问题即密度低，且拥挤人群的场景少。并且这些数据集的协议，由于完全标注人群区域非常困难，耗时太多，所以允许标注者忽略并抛弃大量人群聚集的区域。不同人体检测数据集的体积、密度和多样性如表 4.1 所示。

表 4.1 不同人体检测数据集的体积、密度和多样性

	Caltech	KITTI	CityPersons	COCOPersons	CrowdHuman
图像数量	42782	3712	2975	64115	15000
人体实例	13674	2322	19238	257252	339565
忽略区域	50363	45	6768	5206	99227
平均行人数量	0.32	0.63	6.47	4.01	22.64

1. CrowdHuman 数据集特点

(1) 数据集规模：CrowdHuman 训练子集共计有 15000 幅图像，标注了约 34 万个人体实例，约 9.9 万个忽略区域。与先驱工作的行人检测数据集如 CityPersons 相比，数量多了 10 倍，人体的总计数量也比其他数据集多了很多。

(2) 密度：在密度上，CrowdHuman 数据集中平均每幅图中有约 22.6 个人体实例。CrowdHuman 数据集与其他数据集相比，人体实例密度要大很多。Caltech 和 KITTI 的人体实例密度非常低，平均每幅图像只有不到 1 个人。CityPersons 的密度接近 7，增长很大，但行人仍然不够密集。对于 COCOPersons 来说，虽然其容量相对较大，但仍不能成为理想的人群场景的测试基准。多亏了我们数据集的预滤除和标注协议，CrowdHuman 数据集可以达到足够高的密度。

(3) 多样性：多样性是数据集的重要因素。COCOPersons 和 CrowdHuman 中的人体实例姿势多样，领域宽广，而 Caltech、KITTI 和 CityPersons 都是通过车辆在街道上录制的。

(4) 遮挡：为更好的分析遮挡程度的分布，我们将数据集分成“基本不遮挡”子集（遮挡小于 30%），“部分遮挡”子集（遮挡大于 30%小于 70%），和“严重遮挡”子集（遮挡大于 70%）。

2. CrowdHuman 数据集图像标注

人体实例对应的三种边界框：

(1) 对每个人体实例都详细标注了完全边界框。如果单个人体实例部分被遮挡，则要求标注者去补全不可见部分并画出一个完整的边界框；

(2)从图像中剪切出每个标注的例子，并将这些剪切出的区域给标注者用来画一条可见的边界框；

(3) 进一步将这些剪切出的区域标注出一个头部的边界框。所有标注都至少由另一位标注者进行二次检查以确保标注质量。

4.1.2 MOT16 数据集

2016 年，MOT16 数据集正式出现，其作用在于对多目标进行跟踪检测，应用领域集中于行人跟踪方面。

MOT16 的关注目标由以下两种，一是处于运动状态中的行人，二是处于运动状态中的车辆。其保留了 MOT15 的优势，并引入了更为详细的标注和 bounding box 数据集，画面丰盈度提升，在拍摄切入角和相机运动两个层面上表现出更多的差异性与创新性，涵盖各种天气状况的视频。研究人员在按照既定标注准则的前提进行人工标注，并借助双重检测方式确保标注信息的准确性。其中，2D 形式的运动轨迹来源于 MOT16。

1. MOT16 数据集的信息

在 MOT16 数据集中，视频序列的数量为 14 个，具有标注信息的训练集以及测试集各占据其中的二分之一。如下图 4-1 所示第一行为训练集，第二行为测试集。

图 4-1 MOT 16 数据集示意图

MOT16 数据集的构成，通常属于固定地文档组织格式。根据帧的不同，视频被归类为各图像，图像均是 peg 格式，并以 6 位数字的形式进行命名，采用 CSV 格式对目标以及轨迹信息文件进行标注，其行数与目标相关信息的个数保持一致。

2. MOT16 数据集的标注规则（Annotation Rules）

(1) Target Class-目标类别划分规则

MOT16 的标注对象集中在具有移动性的目标上，涵盖下述三大类别：

Target：行人（处于移动或站立状态）

Ambiguous：人或人造物，此时必须保持非直立的状态中（artificial representations）

Other：车辆和互相包含/遮挡的目标（vehicles and occluders）

首先，对行人（处于移动或站立状态）而言，是通过观察者进行主导标注的，其会可

观视野范围内任何处于移动或站立状态中的人进行标注，即使是处于滑板上的人也被包含在内，并且处于弯腰等类似形态的行人也包括。

其次，针对人或人造物，且此时必须保持非直立的状态而言，涵盖了所有的类似模特等 people-like 目标，属于模糊目标。如果行人佩戴了墨镜则属于 distractors。

最后，针对车辆和互相包含/遮挡的目标而言，能够标注两种目标，一是任何具有移动属性的车辆，二是具有潜在包含或遮挡联系的物体。此时获取的信息用途的单一的，往往仅供专业参赛者培训，被排除在评价目标检测方法准则之外。

(2) Bounding box alignment

Bounding box 的首要要求是应当涵盖全部的像素点，但必须确保像素点的排列是较为、紧凑的。当行人处于行走的持续移动状态时，其 bounding box 的长度宽度值也不会是固定地。一旦目标人物出现了部分被遮挡的现象，那么确定 box 的尺寸时就要借助影子等相关辅助信息进行考量。若人由于裁剪原因位于图像的边缘区域，则 box 能够通过使用高于该帧图像的尺寸完成对目标的确定。若物体存在一定的遮挡或者包含现象，例如错综复杂的树枝布局，则应当借助更多的 box 对物体进行粗略表示。需要注意的是，对于处于自行车上的人而言，Bounding box 会将目标锁定在人上，而不是自行车；但对于处于汽车内的人而言，Bounding box 是不会对其做任何标注的。

(3) Start and end of trajectories 起始与结束时间点

若标注者通过分析，认为目标物体被排除在了 ambiguous 范畴之外，那么：Start as early as possible, end as late as possible.

(4) Minimal size

尽管行人在图像中所占据的尺寸在很多情况下是较为微小的，但根据标注标准来看，只要标注者能够在人眼分辨能力允许的范围内进行标注操作即可。（In other words, all targets independent of their size on the image shall be annotated）

(5) Occlusions 遮挡

遮挡行为大多出现在跟踪标注的过程中，在确保物体识别过程是准确时，应当始终秉持容纳更大标记量的原则，如果物体在运动过程中出现了被完全遮挡的状况，那么当目标物体重新出现后，应当对其轨迹进行重新设定。

(6) Sanity check 检查

当视频完成操作标准后，借助精度较高的行人或者车辆检测方式进行最终的判别，避免其出现遗漏的现象，此过程的开展需要人工的协助。

3. MOT16 数据集对于各个检测识别算法的评价方法

数据集能够提供 ground truth 数据、评价算法指标以及训练的脚本内容，并且这三类内容均达到既定标准才能进行输出。通过上述做法，有助于采用数字的形式对目标检测跟踪算法的准确度进行表示，同时能够对多个检测识别方法所产生的错误信息进行及时有效的识别。

以下是评价方法（简要概述）：

(1) 对 bounding box 以及 ground truth 两部分的交集状况进行识别，确定距离值；

(2) 对语义以及标注 label 之间存在的相似程度作出判别；

(3) 其他轨迹跟踪的质量评价方法。

4.1.3 实验环境

实验环境如表 4.2 所示：

表 4.2 实验环境

硬件与软件	规格
操作系统	Windows 10
处理器	IntelCorei7-11700k(5.0GHZ)
CPU 物理内存	32GB
GPU 型号	RTX3080Ti
GPU 内存	12GB
编译语言	Python3.6
实验框架	Pytorch 深度学习框架
存储地址	本地磁盘

4.2 评价指标

1.评价依据

(1) 在短时间内能够完成对所出现目标的实时搜寻；

(2) 目标位置和与真实目标位置应当在最大程度上确保一致；

(3) 就单个目标而言，其均拥有相应的 ID，其在整个序列中能够始终维持原有状态。

2.评价过程

(1) 在目标以及假设最优间确定对应关系，该种对应关系简称为 correspondence；

(2) 针对各个 correspondence，计算其因位置偏移所形成的误差大小；

(3) 累积结构误差，对漏检数、虚警数以及跟踪目标发生跳变的次数分别进行计。

3.评价指标数学模型

(1) MOTA(Multiple Object Tracking Accuracy)

公式(4.1)为 MOTA 的具体计算公式及范围：

$$MOTA = 1 - \frac{\sum(FN+FP+IDSW)}{\sum GT} \in (-\infty，1] \tag{4.1}$$

FN 的全称是 False Negative, FP 的全称是 False Positve, IDSW 的全称是 ID Switch, GT 全称则是 Ground Truth，代表了物体的数量。MOTA 注重对 tracking 中全部对象匹配错误

进行处理，能以相对直接简单的方式对其在检测物体以及维持轨迹时的性能状况作出衡量，这独立于目标检测的精确度。

MOTA 的值是低于 100 的，然而一旦跟踪器形成的错误比场景中的物体还要多，那么 MOTA 会自动变为负数。MOTA&MOTP 是计算所有帧相关指标后再进行平均的，而并非计算每帧的 rate 并对其平均。

(2) MOTP(Multiple Object Tracking Precision)

公式(4.2)为 MOTP 的具体计算公式：

$$MOTP = \frac{\sum_{t,\ i} d_{t,\ i}}{\sum_t c_t} \tag{4.2}$$

在所有帧中，检测目标i以及为其提供的 ground truth 之间的平均度量距离用 d 进行表示，度量依据是借助了 bonding box 中的 overlap rate，需要注意的是，此时 MOTP 越大说明性能越为良好，然而在使用欧氏距离展开度量时，对 MOTP 大小的要求则相反；此外，在当前帧显示为匹配成功状态的数量用 c 进行表示。MOTP 的核心作用在于对检测器的定位精准度进行量化，和跟踪器作业性能的部分信息是互不干扰的。

(3) MT(Mostly Tracked)

除满足 Ground Truth 匹配不成功时间低于 20%的 track，在全部追踪目标中的相应比重。此时，MT 以及 ML 的状态不受当前 track 的 ID 的影响，唯一的条件是实现 Ground Truth 和目标的匹配。

(4) ML (Mostly Lost)

如果 Ground Truth 高于内匹配成功的 track 时长不足百分之二十，则在整体追踪目标中的确定该比重。

(5) ID Switch

ID 发生变动情况的次数，来源于 Ground Truth。

(6) FM (Fragmentation)

FM 对 Ground Truth 的 track 没有达到被匹配的次数要求的部分再次展开测算，如果轨迹状态由跟踪转换到了未跟踪情景，并且在此后一段时间内出现了与跟踪相同的轨迹，则进入 FM 自动计数阶段。在 FM 的计数过程中，要求 ground truth 的状态需要满足：tracked->untracked->tracked。需要注意的是，FM 与 ID 是否发生变化无关。

(7) FP (False Positive)

若帧预测的 track 不能和 detection 实现匹配，则被归类为预测错误的 track 点。此外，匹配成功与否会受到匹配时阈值因素的影响。

(8) FN (False Negative)

若帧预测的 track 不能和 detection 实现匹配，FN 是没有被匹配的 ground truth 点。

(9) ID scores

跟踪器出现错误的次数情况是 MOTA 需要关注的核心问题，但在航空等个别场景中，需要格外关注跟踪器所具备的跟踪时长能力。此问题可以借助二分图原理进行解决，针对 IDTP、IDFP、IDFN 进行计算。

(10) IDP

IDP 的公式如式(4.3)所示：

$$IDP = \frac{IDTP}{IDTP+IDFP} \tag{4.3}$$

(11) IDR

IDR 的公式如式(4.4)所示：

$$IDR = \frac{IDTP}{IDTP+IDFN} \tag{4.4}$$

(12) IDF1

正确识别的检测与真实数和计算检测的平均数之比。可以通过式(4.4)计算：

$$IDF1 = \frac{2IDTP}{2IDTP+IDFP+IDFN} \tag{4.5}$$

4.3 实验结果

4.3.1 目标检测实验结果

YOLOX 与 YOLOX-AM 在 CrowdHuman 训练集上进行了训练(头部标注)，并在测试集上进行了测试，结果如表 4.3 所示。本文提出的 YOLOX-AM 的目标检测精度明显高于原始的 YOLOX。

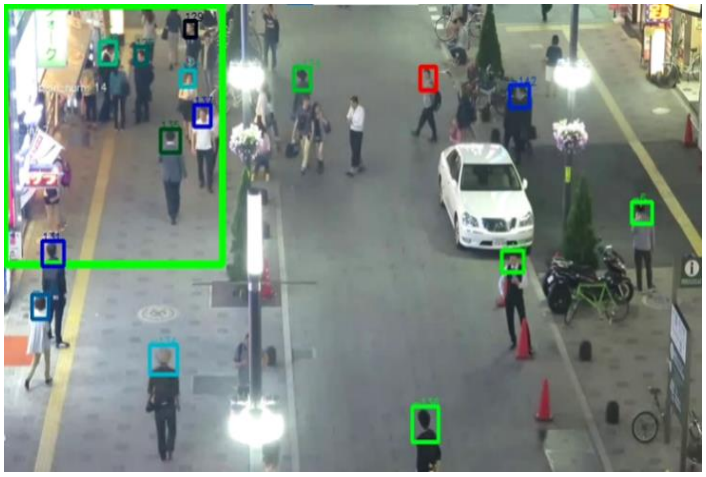
表 4.3 在 CrowdHuman 测试集上 YOLOX 和 YOLOX-AM 的结果

	Recall	AP	mMR
YOLOX	80.43	72.36	63.64
YOLOX-AM	83.10	79.95	65.06

4.3.2 目标追踪与人流计数实验结果

实验在 MOT16 数据集中选取了一段行人流动性较大的视频进行测试。人流量统计系统在 MOT 16 数据集上测试结果如图 4-2 和图 4-3 所示：

(a) 测试结果示意图



(b) 测试结果示意图

图 4-3 在 MOT 16 上的测试结果示意图

表 4.4-4.6 中体现了各评价指标在 MOT16 数据集测试中的结果：

表 4.4 在 MOT16 数据集上的测试结果

数据集	MOTA	MOTP	MT	ML
MOT16-02	33.972	78.934	12	30
MOT16-04	64.352	77.016	39	39
MOT16-05	58.699	78.933	39	77
MOT16-09	62.703	84.286	15	17
MOT16-10	54.113	77.571	19	29
MOT16-11	66.545	85.701	38	28
MOT16-13	40.867	75.68	23	54
各序列平均	55.622	78.609	161	244

表 4.5 在 MOT16 数据集上的测试结果

数据集	IDF1	IDR	IDP	FM
MOT16-02	39.515	27.506	4819	59
MOT16-04	66.42	59.387	28010	88
MOT16-05	69.387	61.779	4181	130
MOT16-09	57.922	54.181	2826	30
MOT16-10	59.097	47.797	5829	59
MOT16-11	62.509	59.035	5375	73
MOT16-13	53.111	39.609	4482	112
各序列平均	60.223	50.764	55498	521

表 4.6 在 MOT16 数据集上的测试结果

数据集	IDS	IDsw	FP	FN
MOT16-02	17838	88	404	11382
MOT16-04	47562	98	3477	13624
MOT16-05	6823	75	640	2146
MOT16-09	5262	47	635	1318
MOT16-10	12322	64	455	5206
MOT16-11	9179	30	1034	2062
MOT16-13	11455	54	454	6331
各序列平均	110412	430	7079	42051

通过对车站的实际应用情况的判断选取了两种场景分别是高铁站进站口附近以及高铁站检票口附近，所获得的部分实验结果如图 4-4 和图 4-5 所示。本次试验样本全部拍摄于辽宁省沈阳市沈阳北站北广场以及车站内部等地方。图 4-4 的视频样本拍摄于沈阳北站北侧进站连廊处，图 4-5 中的视频样本是拍摄于沈阳北站检票口附近处。

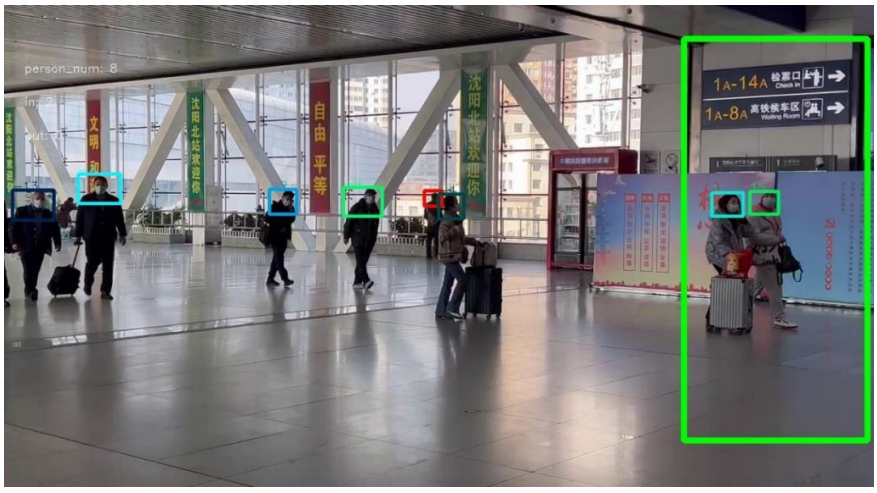


图 4-4 车站人流量系统在车站入口处的应用效果图

图 4-5 车站人流量统计系统在车站检票口处的应用效果图

4.4 分析

本文使用了部分 MOT16 数据集进行训练，剩余部分进行了测试，图 4-3 为 MOT16 测试集的一个可视化结果，并加入了人流计数功能，由图 4-3 可以明显看出，所设计的人流计数系统在MOT16的街景图像序列中可以对头部进行精确追踪，对于被短时遮挡的目标，依然可以做到连续追踪，体现了本文所设计系统的鲁棒性，实现了较好的人流计数功能。

除此之外本文对 MOT16 上的测试结果做了定量分析，计算了 12 种衡量指标，如表 4.2 所示。分别对 7 个序列进行了计算，并作汇总。本文所提出的基于 YOLOX + Deep SORT 的目标追踪系统在 7 个测试序列上均取得了比较理想的评价指标数值，体现了本系统的追踪高精度较高。

除了对开源的数据集做测试外，本系统还在真实的火车站进行了测试。图 4-4，图 4-5 为在沈阳北站所拍摄序列的测试结果图。除了精准计算了整个镜头内的人数外，还对特定区域的人流进出情况进行了精准统计。本系统在 MOT16 数据集上训练后迁移到了实际场景中。使本文所设计的系统不单单停留在研究中，还可以真正落地应用，再次体现了本系统的稳定性、可靠性以及鲁棒性，解决了深度学习落地难的问题。

4.5 本章小结

本章首先介绍了训练模型所使用的 MOT16 数据集，然后介绍了多目标追踪所使用的 12 个评价指标，之后将本文所设计的系统在 MOT16 上进行了测试，并计算了 12 个评价指标。最后将本系统在实际的车站场景下进行了测试，对测试结果做出了定性与定量分析。

第 5 章 总结与展望

5.1 总结

本文的研究内容集中在以 YOLOX 以及 Deep SORT 为基础的多目标跟踪算法上，并对人流计数算法进行了全面分析。以检测为基础的多目标跟踪策略，不论是在学术界，还是在工业实操应用中，均受到了极大的推崇与认可。本课题所提出的多目标跟踪算法，即 Deep SORT，是以一阶段检测算法 YOLOX-AM 为基础的，在 MOT16 数据集和真实的高铁站人流视频序列中都实现了良好的跟踪性能。

通过在一阶段检测算法和二阶段检测算法上进行对比，本文首次将 YOLOX-AM 和 Deep SORT 应用在车站人流量统计上。基于 YOLOX-AM 的多目标跟踪算法 Deep SORT 在跟踪效果上要优于同类算法，这既得益于跟踪策略的成熟性，同时离不开检测器所具备的突出的检测性能，并且 YOLOX-AM 在确定并分析检测结果的过程中提供了可靠支撑。在运用二阶段检测跟踪算法的过程中，性能能够保持在一个相对优越的状态中，但在工业行业的作业过程中要求具备较高的运算效率以及速度，因此在这一层面上，一阶段检测器的跟踪性能是更加契合市场需要的。在实际的车站场景中实现了高精度的人流计数功能，具有实际应用的落地价值。

5.2 展望

本文所提出的多目标跟踪算法不仅针对原有传统算法进行了改进，同时在实践应用方面也就更高的价值，能够深化多目标跟踪任务在实践作业中的意义。虽然当前的多目标跟踪算法是具有一定的跟踪性能的，但在模型体系方面仍然存在较大的改进空间，模型速度也有待提升。

同时，多目标跟踪研究成果会受到检测器检测效果等因素的影响，可以将检测算法的优化作为未来持续改进的任务之一。本文所应用的算法以一阶段以及二阶段检测算法为基础展开对比分析，若能在检测器的精度方面取得更大突破，则会形成更为优越的跟踪效果。

就多目标跟踪效果当前的发展程度而言，其精度以及速度等性能指标均尚处于滞后状态，必须强化对于检测器以及跟踪策略的研究。

参考文献

[1] Luo Wenhan,Xing Junliang,Milan Anton,Zhang Xiaoqin,Liu Wei,Kim Tae Kyun. Multiple object tracking: A literature review[J]. Artificial Intelligence,2020(prepublish).

[2] Haohua Du,Linlin Chen,Jianwei Qian,Jiahui Hou,Taeho Jung,Xiang Yang Li. PatronuS: A System for Privacy-Preserving Cloud Video Surveillance[J]. IEEE Journal on Selected Areas in Communications,2020,38(6).

[3] Detection and Tracking of Multiple Objects in Cluttered Backgrounds with Occlusion Handling[J]. Computer Science & Information Technology (CS & IT),2014,4(7).

[4] Paul Viola,Michael J. Jones,Daniel Snow. Detecting Pedestrians Using Patterns of Motion and Appearance.[J]. International Journal of Computer Vision,2005,63(2).

[5] Antonin Ponsich,Catherine Azzaro-Pantel,Serge Domenech and,Luc Pibouleau. Mixed-Integer Nonlinear Programming Optimization Strategies for Batch Plant Design Problems[J]. Ind. Eng. Chem. Res.,2006,46(3).

[6] 王强，冯燕. 基于颜色和形状信息的快速人数统计方法[J]. 计算机测量与控制，2010，18(09): 2157-2163

[7] Wang C, Zhang H, Yang L, et al. Deep people counting in extremely dense crowds[C]//ACM international conference. ACM, 2015: 1299-1302.

[8] Li B, Zhang J, Zhang Z, et al. A people counting method based on head detection and tracking[C]//2014 International Conference on Smart Computing. IEEE, 2014: 136-141.

[9] 周治平，许伶俐，李文慧. 特征回归与检测结合的人数统计方法[J]. 计算机辅助设计与 图形学学报，2015，27(03): 425-432.

[10] 肖杰，裴忠才，徐立新. 运动目标识别与跟踪系统的研究[J]. 微计算机信息，2007(34):1- 2

[11] 王铁军，张明廉. 一种二维耦合模型机动目标跟踪算法[J]. 航空学报，2006(03): 481-485.

[12] Kristan Matej,Matas Jiri,Leonardis Ales,Vojir Tomas,Pflugfelder Roman,Fernandez Gustavo,Nebehay Georg,Porikli Fatih,Cehovin Luka. A Novel Performance Evaluation Methodology for Single-Target Trackers.[J]. IEEE transactions on pattern analysis and machine intelligence,2016,38(11).

[13] 史超.基于 MTCNN 与改进的 KCF 人脸目标检测与跟踪算法研究[C]//.2019 第七届中国指挥控制大会论文集.[出版者不详],2019:322-326.

[14] Chenpu Li,Qianjian Xing,Zhenguo Ma. HKSiamFC: Visual-Tracking Framework Using Prior Information Provided by Staple and Kalman Filter[J]. Sensors,2020,20(7).

[15] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//international Conference on computer vision & Pattern Recognition. IEEE, 2005, 1: 886--893.

[16] Sabzmeydani P, Mori G. Detecting Pedestrians by Learning Shapelet Features[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2007: 1-8.

[17] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object Detection with Discriminatively Trained Part-Based Models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[18] 王彩霞,林寿英.基于深度学习的多目标跟踪算法综述[J].中阿科技论坛(中英文),2021(10):118-120.

[19] Jianfang Dou,Jianxun Li. Robust object detection based on deformable part model and improved scale invariant feature transform[J]. Optik - International Journal for Light and Electron Optics,2013,124(24).

[20] 郭明玮,赵宇宙,项俊平,张陈斌,陈宗海.基于支持向量机的目标检测算法综述[J].控制与决策,2014,29(02):193-200.

[21] 杨海舟,李丹.基于改进 SPPnet 的 YOLOv4 目标检测[J].电子制作,2021(22):52-54.

[22] 孙宇轩. 基于感兴趣区域池化的相关滤波目标跟踪[D].大连理工大学,2020.

[23] 李星辰,柳晓鸣,成晓男.融合 YOLO 检测的多目标跟踪算法[J].计算机工程与科学,2020,42(04):665-672.

[24] 刘彦清. 基于 YOLO 系列的目标检测改进算法[D].吉林大学,2021.

[25] 端辉. 基于 YOLO 的多尺度快速行人检测算法研究与应用[D].大连理工大学,2019.

[26] 李祥兵,陈炼.基于改进 Faster-RCNN 的自然场景人脸检测[J].计算机工程,2021,47(01):210-216.

[27] Liu B , Zhang X ,Gao Z , et al. Weld Defect Images Classification with VGG16-Based Neural Network[C]// International Forum on Digital TV and Wireless Multimedia Communications. Springer, Singapore, 2017.

[28] Tun N L , Gavrilov A , Tun N M , et al. Remote Sensing Data Classification Using A Hybrid Pre-Trained VGG16 CNN- SVM Classifier[C]// 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus). IEEE, 2021.

[29] 车大伟. 一种 VGG16 结合 FLANN 的人脸识别方法[J]. 移动信息, 2022(2).

[30] 陈宣,李怡昊,陈金立.基于知识图谱和 Softmax 回归的干扰信号识别方法[J].中国电子科学研究院学报,2021,16(09):856-861.

[31] Liu W , Wen Y , Yu Z , et al. Large-Margin Softmax Loss for Convolutional Neural Networks[J]. JMLR.org, 2016.

[32] Huo Z , Xia Y , Zhang B . Vehicle type classification and attribute prediction using multi-task RCNN[C]// International Congress on Image&Signal Processing. IEEE, 2016.

[33] Redmon J , Farhadi A . YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018.

[34] 张路达, 邓超. 多尺度融合的 YOLOv3 人群口罩佩戴检测方法[J]. 计算机工程与应用, 2021, 57(16):8.

[35] 沈震宇, 朱昌明, 王喆. 基于 MAML 算法的 YOLOv3 目标检测模型[J]. 华东理工大学学报（自然科学版）, 2021, 48(1):112.

[36] 章琳,袁非牛,张文睿,曾夏玲. 全卷积神经网络研究综述[J]. 计算机工程与应用, 2020(25-37).

[37] Mastromichalakis S . ALReLU: A different approach on Leaky ReLU activation function to improve Neural Networks Performance[J]. 2020.

[38] Li C L , Ravan Ba Khsh S , Poczos B . Annealing Gaussian into ReLU: a New Sampling Strategy for Leaky-

ReLU RBM[J]. 2016.

[39] 周爱武, 于亚飞. K-Means 聚类算法的研究[J]. 计算机技术与发展, 2011, 21(2):4.

[40] 王千, 王成, 冯振元,等. K-means 聚类算法研究综述[J]. 电子设计工程, 2012, 20(7):4.

[41] Silvia Rostianingsih,Alexander Setiawan,Christopher Imantaka Halim. COCO (Creating Common Object in Context) Dataset for Chemistry Apparatus[J]. Procedia Computer Science,2020,171(C).

[42] Simone Bonechi,Paolo Andreini,Monica Bianchini,Franco Scarselli. COCO_TS Dataset: Pixel-level Annotations Based on Weak Supervision for Scene Text Segmentation.[J]. CoRR,2019,abs/1904.00818.

[43] Van De Weijer J, Schmid C, Verbeek J, et al. Learning color names for Real-World Applications[J]. IEEE Transactions on Image Processing, 2009, 18(7): 1512-1523.

[44] 华逸伦,石英,杨明东,刘子伟.基于背景抑制和前景抗干扰的多尺度跟踪算法[J].红外技术,2018,40(11):1098-1105.

[45] Li Y, Zhu J. A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration[C]//European Conference on Computer Vision. Springer, Cham, 2014: 254-265.

[46] Danelljan M, Häger G, Khan F, et al. Accurate Scale Estimation for Robust Visual Tracking[C]// Proceedings of the British Machine Vision Conference. BMVA Press, 2014.

[47] Wang M, Liu Y, Huang Z. Large Margin Object Tracking with Circulant Feature Maps[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4021-4029.

[48] Kristan M, Leonardis A, Matas J, et al. The Visual Object Tracking VOT2016 Challenge Results[J]. 2016.

[49] Liu, Shu, et al. "Path aggregation network for instance segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[50] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." arXiv preprint arXiv:2004.10934 (2020).

[51] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European conference on computer vision (ECCV). 2018.

[52] Rezatofighi, Hamid, et al. "Generalized intersection over union: A metric and a loss for bounding box regression." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[53] Zheng, Zhaohui, et al. "Distance-IoU loss: Faster and better learning for bounding box regression." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 07. 2020.

[54] ROUAND O,HAVET M. Numerical investigation on the efficiency of transient contaminant removal from a food processing clean room using ventilation effectiveness concepts[J]. Journal of Food Engineering,2005,68:163-174.

[55] Zhong, Zhun, et al. "Random erasing data augmentation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 07. 2020.

[56] Singh K K, Hao Y, Sarmasi A, et al. Hide-and-Seek: A Data Augmentation Technique for Weakly-Supervised Localization and Beyond[J]. 2018.

[57] Chen, Pengguang, et al. "Gridmask data augmentation." arXiv preprint arXiv:2001.04086 (2020).

[58] Zhang, Hongyi, et al. "mixup: Beyond empirical risk minimization." arXiv preprint arXiv:1710.09412 (2017).

[59] DeVries, Terrance, and Graham W. Taylor. "Improved regularization of convolutional neural networks with cutout." arXiv preprint arXiv:1708.04552 (2017).

[60] Yun, Sangdoo, et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[61] Zhu, Pengfei, et al. "VisDrone-VID2019: The vision meets drone object detection in video challenge results." Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019.

[62] Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: the clear mot metrics[J]. EURASIP Journal on Image and Video Processing, 2008, 2008: 1-10.

[63] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Qinghua Hu, Haibin Ling. Vision Meets Drones: Past, Present and Future. arXiv preprint arXiv:2001.06303 (2020)

[64] Ioffe S , Szegedy C . Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. JMLR.org, 2015.

[65] 刘 博,王胜正,赵建森,等．基于 Darknet 网络和 Yolov3 算法的船舶跟踪识别[J].计算机应用，2019，39(06)：1663-1668.

[66] 覃勇杰. 基于机载系统的特定目标检测技术研究.

个人简历及在学期间的研究成果和发表的学术论文

一、个人简历

张依林（1996-），男，辽宁朝阳人。2014 年开始就读于沈阳师范大学软件学院的软件工程专业。2019 年考入沈阳师范大学软件学院计算机应用技术硕士研究生。

二、在研期间发表的学术论文

张依林，王学颖.基于 CNN-SVM 的车辆检测与类型分类[J].电子技术与软件工程.2022(07):182-185.

致　　谢

三年时光如同白驹过隙，我的硕士研究生生活即将接近尾声，而入学仿佛还是在昨天，初来乍到成为一名研究生的骄傲和场景犹历历在目。回忆起这三年的点点滴滴，喜怒哀乐，令我感慨不已，欣慰之余又深感庆幸无比。值得自己和家人感到欣慰的是在这三年的时光中学到了很多受益一生的东西，不仅仅是知识，还有对待生活的态度和精神，以及为人处世的道理。

论文即将付梓，不禁令我感叹到时光的荏苒。想要感谢的人太多，首先要感谢我的导师王学颖教授。转眼间我在王老师身边学习和生活已经过去了三年，她的厚爱，使我有了更加强大的信心，她的言传身教，使我学到了许多书本上没有的东西，这些也成为了我最宝贵的财富。

感谢李航老师、刘天华老师、范书国老师、杜庆东老师、周传生老师、赵永翼老师、吴鹏老师和朱宏峰老师在百忙之中多次抽出宝贵时间给我讲解毕业论文写作方法和技巧，使我受益匪浅。感谢高晓婴老师在三年的研究生学习中对我的关心和照顾。

感谢所有关心和帮助我的朋友们兄弟们及所有支持、帮助我的老师和同学。

感谢我的父母对我的大力支持使我得以完成学业。特别感谢一下我的父亲，您以您独到的见解和睿智，在我的学生时代从本科期间到硕士研究生期间给予了我无尽的鼓励和指点，为我以后的人生路打下了扎实的地基。您是我最爱的家人，更是我人生路上的明灯。

特别感谢百忙中抽出宝贵时间审稿、出席论文答辩的诸位专家、教授，向您们致以深深的敬意！